BAYESIAN ANALYSIS OF COSMOLOGICAL MODELS

Darell Moodley

University of KwaZulu-Natal

BAYESIAN ANALYSIS OF COSMOLOGICAL MODELS

Darell Moodley

Submitted in fulfillment of the academic requirements for the degree of Master of Science in the School of Mathematical Sciences, University of KwaZulu-Natal, Westville.

As the candidate's supervisor I have approved this dissertation for submission.

Signed: Name: Date:

ABSTRACT

In this thesis, we utilise the framework of Bayesian statistics to discriminate between models of the cosmological mass function. We first review the cosmological model and the formation and distribution of galaxy clusters before formulating a statistic within the Bayesian framework, namely the Bayesian razor, that allows model testing of probability distributions. The Bayesian razor is used to discriminate between three popular mass functions, namely the Press-Schechter, Sheth-Tormen and normalisable Tinker models. With a small number of particles in the simulation, we find that the simpler model is preferred due to the Occam's razor effect, but as the size of the simulation increases the more complex model, if taken to be the true model, is preferred. We establish criteria on the size of the simulation that is required to decisively favour a given model and investigate the dependence of the simulation size on the threshold mass for clusters, and prior probability distributions. Finally we outline how our method can be extended to consider more realistic N-body simulations or be applied to observational data.

PREFACE

The study described in this thesis was carried out in the School of Mathematical Sciences, University of KwaZulu-Natal during the period January 2007 to February 2010. This dissertation was completed under the supervision of Dr. K. Moodley, in collaboration with Dr. C. Sealfon at West Chester University of Pennsylvania.

This study represents original work by the author and has not been submitted in any form for any degree or diploma to another tertiary institution. Where use was made of the work of others it has been duly acknowledged in the text.

Signed: Name: Date:

DECLARATION 1 - PLAGIARISM

I, Darell Moodley, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    - Their words have been re-written but the general information attributed to them has been referenced.
    - Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: Name: Date:

This thesis is dedicated to my mother.

ACKNOWLEDGMENTS

Copious amounts of gratitude needs to be bestowed upon several people who assisted me with this project. A value cannot be placed upon the guidance and expertise that was rendered to me by Dr. Kavilan Moodley. His foresight and wisdom played a pivotal role in the polishing of this work. Dr. Carolyn Sealfon's regular insightful advice and tutelage proved to be a blessed catalyst with regards to completing my thesis.

I am forever grateful to my friends and family for supporting and believing in me. Members of the Astrophysics and Cosmology Research Unit (ACRU) and the school of Mathematical Sciences of the University of KwaZulu-Natal readily assisted with the smooth running of this project. Many thanks must be extended to God since it would not be possible for me to have adequately tackled an assignment of such large magnitude without his unconditional love and nurturing.

This study was supported by the South African Square Kilometer Array project and the National Research Foundation (NRF). I would also like to acknowledge support provided by West Chester University, Pennsylvania.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The $\Lambda-$Cold Dark Matter (ΛCDM) is referred to as the standard model of the big bang cosmology since it attempts to explain the existence and formation of large-scale structure and the cosmic microwave background (CMB). This model, along with many others, is based on the observational evidence that on large scales the universe is homogeneous and isotropic. This is known as the cosmological principle. Many cosmological observations support a flat ΛCDM model. The total energy density in this model is very close to the critical density and a universe that has precisely a density equal to the critical density is said to be flat or Euclidean. Geometry, however, is not the only feature of the universe, and there is evidence to suggest we currently live in an era in which our universe contains matter in two principal forms namely, baryonic and dark matter. Dark matter only interacts through gravitational attraction and is non-visible, therefore it is inferred from to its gravitational interaction with other baryonic matter. Clusters of galaxies are largely made of dark matter.

The observed dimming of type Ia supernovae is interpreted as the accelerated expansion of the universe, provided that the universe is described by the Friedmann-Robertson-Walker metric. The universe is expanding faster than it should be according to conventional knowledge of physics. This phenomenon has perplexed many cosmologists. The accelerated expansion is assumed to be caused by either of two phenomena, the first being due to modified gravity where general relativity is modified in a way that leads to the observed accelerated expansion [1; 2; 3]. The second is a result of a mass-energy component called dark energy. If the latter explanation is true it implies that dark energy constitutes most of the mass-energy ($\approx 71\%$) in the universe and there is no compelling theoretical explanation for the value of this component, therefore we are forced to rely on observational evidence to understand the nature of dark energy. Cosmological observations however, cannot distinguish between the two cases [4].

There are various observational techniques that are promising in determining the presence and nature of dark energy, respectively exploiting the phenomena of: Baryon Acoustic Oscillations (BAO); Galaxy Clusters (CL); Weak Lensing (WL); and Supernovae (SN). Dark energy is usually parameterised by an equation of state, and constraining the dark energy parameters means a deeper understanding of the true behaviour and nature of dark energy. The four major techniques that are proposed will begin to constrain these parameters and combining these techniques will further tighten these constraints [5]. These techniques incorporate surveys that detect the effect of dark energy on the relation between a measurable and observable quantity in the survey. The SN survey for example, detects the effect of dark energy on the luminosity distance-observable relation [6] while the CL survey detects the effect of dark energy on the mass-observable relation [7].

Galaxy clusters can be observed in various ways such as optically [8], via x-ray emission

[9; 10], the Sunyaev-Zel'dovich effect [11] or by weak gravitational lensing [12; 13; 14]. Optical detection relies on photon emission from the cluster's member galaxies while x-ray detection involves x-ray emission from the hot electrons contained within the gravitational potential well. These hot electrons also up-scatter the CMB photons leaving an apparent deficit of the low frequency CMB flux in their direction. This effect is called the Sunyaev-Zel'dovich effect. Regions of high concentrations of mass can cause the photon's path to curve around it resulting in a bending of light. This results in the image of the photon source being distorted and this phenomenon is known as gravitational lensing.

Poor forecasts of systematic error levels inhibit our ability to assess future capabilities of these techniques. In galaxy cluster surveys, these problems can be solved by combining lensing, Sunyaev-Zel'dovich and x-ray observations of large numbers of galaxy clusters to constrain the galaxy cluster mass-observable relationship. The direct measurement of cluster mass is difficult, therefore we require an accurate relation of mass to a proxy or observable quantity instead. For instance, the galaxy counts, the Sunyaev-Zel'dovich decrement or x-ray flux could be used as proxies depending on the type of survey. Observation of clusters have already made inference on cosmological parameters [15; 16; 17; 18; 19; 20]. Simulations of cluster formation provide a framework in which theoretical predictions can be compared to accurate observations in order to explain the evolution of galaxies and growth of cosmic structure. The Millennium N-body simulation [21] is an example of a large simulation that modelled a range of cosmic structures. These simulations were modified by [22] to analyse scaling relations for halo properties.

The mass function is an important ingredient in galaxy cluster counting since it measures the abundance of clusters of a given mass in the universe. The mass function is sensitive to dark energy due to its comoving volume element dependence, and cluster counts also depend on the

expansion history. The mass function is sensitive to the amplitude of density fluctuations since galaxy clusters serve as indicators of such variations in the early universe, and constraints on dark energy have been made using observed density fluctuations [23]. There are various models of mass functions: purely analytical; semi-analytical; and fitting functions to N-body simulations. The Press-Schechter model [24], is the pioneering mass function as it combines spherical top-hat collapse with the growth function for linear perturbations. The standard for precision determination of the mass function from simulations was initially set by producing a fitting function for the halo abundance accurate to $\sim 10 - 20\%$ [25; 26]. This mass function has thereafter been used to provide constraints on the normalisation of the density fluctuation power spectrum, σ_8 [7], the mean matter density, Ω_m, and the dark energy equation of state parameter, ω [16]. Improved models have been developed since then, namely, the Sheth-Tormen [27] model and Tinker [28] model. These models have proven to be the most popular in the cosmology community. With various models of mass functions available at present, and an expected increase in the future, one is faced with the task of optimally selecting the appropriate model that will describe the observations. In recent times, fitted-functions to N-body simulations [25; 27; 28] has become the preferred choice over analytic mass functions. The question however remains, as to how large a simulation is needed in order to discriminate one model over another. To answer this question we resort to *Bayesian* statistics.

The Bayesian statistics methodology has increased its popularity in the cosmology and astrophysics community [29]. This has been motivated by the increase in large and more complex data sets as well as the computational power available at present. The method was first introduced over 200 years ago [30], and eventually developed into the subject of Bayesian probability theory, used in a wide range of fields such as econometrics and biostatistics, to name a few. Cosmology is the most recent field to utilise the methodology, which proves to be quite suitable

given the large data sets currently available from experiments. There are many difficulties in cosmology, for example, the complexity of modelling theory to match observations is becoming more difficult. The accuracy of measurements must be well within a specified uncertainty since this could change the outcome of the result. Experiments are extremely costly in terms of resources, therefore we want to gain as much information from them as possible. Even very sophisticated experiments produce data that is of limited quality. Statistical analysis provides a means to overcome these problems and persistence in refining these methods has been recently motivated [31].

The statistic used in Bayesian model selection is called the *evidence*. This quantity is nearly always a multi-dimensional integral depending on the number of parameters within a model. It is for this reason the evidence becomes difficult to calculate and much work on numerical techniques has been devoted to efficiently evaluate this complex integral. The *Nested sampling algorithm*, [32] is widely used in cosmology [33; 34; 35; 36]. Recent computer codes based on the nested sampling algorithm, called *CosmoNest* [33; 34] and *MultiNest* [37] have been developed that simplifies this computation. MultiNest is used to explore multiple peaked functions, a problem often encountered when computing the evidence, and has proven quite efficient [37; 38].

Much work has been done on possible evidence for competing models to the ΛCDM model using various data such as WMAP3 [39; 40], Supernovae Type Ia [41; 42] as well as combinations of data sets such as WMAP1, Sloan digital sky survey (SDSS) and two-degree field galaxy redshift survey (2dFGRS) [43]. Models are sometimes nested within other models i.e. they are obtained from fixing certain parameters in the more complicated model to obtain the simpler one. This type of model selection can also be studied to determine whether addition of parameters is warranted by the data and is ideally suited for the Bayesian statistics approach rather than the

frequentist approach of null hypothesis testing [44]. This is because the latter approach does not take into account the information gained from the data and may provide contradictory results to the Bayesian statistics approach.

In this thesis we investigate how Bayesian statistics can be used to select between models, specifically we answer the question of how many particles are sufficient in a numerical simulation in order to discriminate between two competing models of the cosmological mass functions. We employ the Bayesian statistics approach and implement the *Bayesian razor* introduced by Balasubramanian [45]. The Bayesian razor assigns merit to a probability distribution, therefore comparing the razor of two distributions can provide a quantitative way to discriminate between distributions. We derive probability distributions for the mass functions considered in this thesis in order to apply the Bayesian razor to these models.

In chapter 2, we review the cosmological model. We discuss the physics of galaxy clusters, particularly the growth and properties of clusters. The statistics of large-scale structure is reviewed in this chapter since it forms the basis for the theory behind the mass function. We then introduce the various models of the mass function that we will compare in chapter 4.

In chapter 3, we look at the Bayesian statistics methodology to gain an understanding of how it can be used in model selection. We briefly review the theory of probability as well as properties of probability distribution functions and how these are relevant for model selection by referring to some examples. A more detailed example with mock data has been used to improve our understanding of the use of Bayesian evidence in assigning merit to a model. We then look at the Fisher information matrix and examine how it can be used to explore probability distributions. At the end of the chapter we review the Bayesian razor and its relevance in model selection.

Chapter 4 is devoted to applying the Bayesian razor to competing mass functions in order to determine how variations in simulation size affect the preference of one model over another. We obtain the relevant probability distribution functions, and the Bayesian razors for each of the three mass functions that we consider. We then analyse the razor ratio as a function of the number of particles in order to determine the size of the simulation needed. Our results consider the effect of varying different factors such as the minimum mass limit of a cluster, the fiducial model and the prior distribution, that all play a role in the computation of the razor.

In a concluding chapter we discuss how our method can be modified and extended to incorporate more realistic N-body simulations as well as how it can be applied to observational data.

CHAPTER 2

The cosmological mass function

2.1 The cosmological model

Cosmological models aim to explain the evolution and expansion of our universe. The big bang model is accepted by most cosmologists as the leading cosmological model, and is referred to as the standard cosmological model. The standard cosmological model has been supported by observational evidence such as the existence of the cosmic microwave background (CMB), Hubble's law that describes the expansion of the universe, and the abundance of light elements that were produced in an early hot and dense phase of the universe. Cosmological models in general provide the foundation for various theories and observations, and are dependent on several cosmological parameters. Among those parameters, is the Hubble parameter which is dependent on the scale factor $a(t)$, and describes the expansion of the universe. We have from Hubble's law that

$$v = H(t)d, \tag{2.1.1}$$

where d is the distance between two galaxies and v is the velocity at which these galaxies move apart. Hubble determined that the farther a galaxy is away from Earth, the faster it accelerates away. We generally use two methods to measure the distances to other galaxies, namely, Cepheid variables and Type Ia supernovae. Cepheid variables are examples of stars that vary in brightness but have a definite relation between distance and luminosity. We can determine the distance by comparing a star's absolute magnitude with the apparent magnitude of brightness. Cepheid variables are dimmer than Type Ia supernovae, therefore the supernovae allows us to measure the distances of galaxies further away. The luminosity of Type Ia supernovae are approximately the same therefore we can compute the distance to a galaxy by measuring the apparent brightness of Type Ia supernovae in that galaxy. The preferred unit of length in cosmology is the *parsec* which is approximately just under 31 trillion kilometers. Other units of length are also used such as *astronomical units* (au) and *light-years* (ly). If galaxies are close by then d is the usual Euclidean distance, but the greater the seperation then the concept of distance must be defined to avoid ambiguity. Luminosity distance is then used in this measurement.

The symbol H is known as the Hubble parameter and is related to the scale factor by

$$H(t) = \frac{\dot{a}(t)}{a(t)}, \tag{2.1.2}$$

where $\dot{a}$ refers to the derivative of a with respect to time t. At the present time t_0, $H(t_0) = H_0$ is known as the Hubble constant and is parameterised by h defined via

$$H_0 = 100h \,\text{km. s}^{-1}\text{Mpc}^{-1}$$

where $h = 0.71 \pm 0.07$ [46]. The evolution of the scale factor for a homogeneous and isotropic universe is described by

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\mathcal{K}c^2}{a^2} + \frac{\Lambda c^2}{3}.$$

This equation is known as the Friedmann equation. The symbol ρ represents the mass-energy density while p is the pressure resulting from that mass-energy. The symbol Λ represents the cosmological constant and accounts for the accelerated expansion of the universe. These quantities may be composite: The sum of several fluid species. The term $\mathcal{K}$ is the curvature of the universe and usually takes on discrete values for the three geometries, namely,

$$\mathcal{K} = \begin{cases} 1 & \text{Closed universe,} \\ 0 & \text{Flat universe,} \\ -1 & \text{Open universe.} \end{cases}$$

The symbol c represents the speed of light and can be scaled to 1 by using a geometric unit system. We can re-write the Friedmann equation as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\mathcal{K}}{a^2}, \tag{2.1.3}$$

where Λ has been set to zero. The local energy conservation equation is given by

$$\dot{\rho} + 3H(\rho + p) = 0. \tag{2.1.4}$$

The second term in equation (2.1.4) corresponds to the dilution of ρ due to the Hubble expansion and the third term represents the work done by the pressure of the fluid. So far we have three unknown quantities, i.e. ρ, p and a, therefore we require another equation relating these parameters. The third equation relates ρ and p by

$$p = \omega\rho, \tag{2.1.5}$$

where $\omega = \omega(\rho)$ depends only on the local energy density, and is known as the linear equation of state. The cosmological constant can be modelled by another fluid with $\omega = -1$. In this case, the Λ term is set to zero, as in (2.1.3). For a flat universe ($\mathcal{K} = 0$) dominated by only one fluid

with equation of state $\omega = constant$, we solve for ρ using (2.1.5) and (2.1.4) to obtain

$$\rho \propto a^{-3(1+\omega)}. \tag{2.1.6}$$

We use (2.1.6) to determine how ρ is related to the scale factor depending on which component dominates the energy density of the universe, namely, matter, radiation or the cosmological constant (Λ).

$$\text{Matter:} \qquad \omega = 0, \qquad \rho \propto a^{-3}, \tag{2.1.7}$$

$$\text{Radiation:} \qquad \omega = \tfrac{1}{3}, \qquad \rho \propto a^{-4}, \tag{2.1.8}$$

$$\Lambda\text{:} \qquad \omega = -1, \qquad \rho \propto constant. \tag{2.1.9}$$

We now derive an equation that describes the acceleration of the scale factor. We start by differentiating (2.1.3) with respect to time to obtain

$$2\frac{\dot{a}}{a}\left(\frac{\ddot{a}a - \dot{a}^2}{a^2}\right) = \frac{8\pi G}{3}\dot{\rho} + 2\frac{\mathcal{K}\dot{a}}{a^3}. \tag{2.1.10}$$

Substituting for $\dot{\rho}$ from (2.1.4) and using (2.1.3) again, equation (2.1.10) can be written as

$$\frac{\ddot{a}}{a} = \frac{-4\pi G}{3}(\rho + 3p), \tag{2.1.11}$$

which is known as the acceleration equation or Raychaudhuri equation [47].

We can relate the scale factor to time depending on the dominant component. It is convenient to normalise the scale factor so that it equals 1 at present time t_0 i.e. $a_0 = 1$. We know from

(2.1.7) that $\rho \propto 1/a^3$ for matter domination. Fixing the proportionality constant by the present density ρ_0 yields

$$\rho = \frac{\rho_0}{a^3}. \tag{2.1.12}$$

Substituting (2.1.12) into the Friedmann equation gives the following result

$$\dot{a}^2 = \frac{8\pi G}{3}\frac{\rho_0}{a}. \tag{2.1.13}$$

To solve this, we suppose a follows a power law, $a \propto t^q$; therefore $\dot{a} \propto t^{q-1}$. In the right hand side we have $a^{-1} \propto t^{-q}$. These powers must equal, therefore $q = 2/3$ which means

$$a(t) \propto \left(\frac{t}{t_0}\right)^{2/3}, \tag{2.1.14}$$

if matter is the dominant component. Similarly if radiation dominates,

$$a(t) \propto \left(\frac{t}{t_0}\right)^{1/2}, \tag{2.1.15}$$

so the universe expands faster if matter dominates instead of radiation. Similarly in a Λ dominated universe,

$$a(t) \propto e^t \tag{2.1.16}$$

and we have exponential growth.

The density parameters are generally shown as dimensionless quantities when we relate them to the critical density $\rho_{crit} = \frac{3H^2}{8\pi a^2 G}$. We use subscripts r, m and Λ to denote the radiation, matter and dark energy component respectively to simplify the notation. The density parameters expressed as ratios of ρ_{crit} depending on the dominant component are

$$\Omega_r = \frac{\rho_r}{\rho_{crit}}, \tag{2.1.17}$$

$$\Omega_m = \frac{\rho_m}{\rho_{crit}}, \tag{2.1.18}$$

and

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_{crit}} = \frac{\Lambda}{8\pi G\rho_{crit}}, \tag{2.1.19}$$

with curvature given by

$$\Omega_\mathcal{K} = \frac{\rho_\mathcal{K}}{\rho_{crit}} = \frac{-3\mathcal{K}}{8\pi G a^2\rho_{crit}}. \tag{2.1.20}$$

These quantities are constrained by

$$\Omega_r + \Omega_m + \Omega_\Lambda + \Omega_\mathcal{K} = 1, \tag{2.1.21}$$

which is a dimensionless form of the Friedmann equation. The current density ρ_0, excluding curvature, in terms of the current critical density is given by

$$\Omega_0 = \Omega_{r0} + \Omega_{m0} + \Omega_{\Lambda 0}, \tag{2.1.22}$$

where Ω_{r0}, Ω_{m0}, and $\Omega_{\Lambda 0}$ are the current density for radiation, matter and the dark energy component respectively. The Friedmann equation is an important equation in cosmology as it describes the expansion of the universe. The scale factor is related to the redshift of distant objects by

$$a = (1+z)^{-1}, \tag{2.1.23}$$

where we have set $a_0 = 1$ as mentioned previously. Using this relation, we can write the Friedmann equation in terms of redshift to yield

$$H^2(z) = H_0^2\left[\Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_\Lambda + \Omega_k(1+z)^2\right]. \tag{2.1.24}$$

This equation explains how the different components of the universe scales with redshift. Since redshift is analogous with time, equation (2.1.24) shows that radiation was the dominant component during the early universe and the cosmological constant will dominate eventually. The

determination of these densities are therefore vital in our understanding of the evolution of our universe. Matter is made of two non-relativistic components, namely, the visible baryonic and cold dark matter components such that $\Omega_m = \Omega_b + \Omega_{CDM}$ where Ω_b and Ω_{CDM} are the baryonic and cold dark matter densities relative to the critical density respectively. Cold dark matter has many candidates, namely: WIMPs, Axions, WIMPzillas and Primordial black holes [48]. In a similar way, radiation is made up of relativistic photons and neutrinos. The CMB photons are the dominant source of photons in the universe and we therefore neglect those generated from stars since their contribution to the density is very small. We denote the density of CMB photons and neutrinos as Ω_{CMB} and Ω_ν respectively with $\Omega_r = \Omega_{CMB} + \Omega_\nu$.

2.2 Growth of clusters

Most of the matter in the universe is dark matter, which comprises the majority of the content of galaxies. The process of galaxy formation is initiated by perturbations in the density distribution of cold dark matter. The density distribution in the initial perturbation will determine the rate of increase in the growth of matter. These perturbations result in matter accumulating through the gravity of dark matter. The gravitational pull slows down the expansion of matter resulting in a deviation from the Hubble flow. Eventually the structures that form undergo gravitational relaxation and stop expanding. This process forms a gravitational well, accreting more matter to eventually form a cluster at the center of these perturbations.

Previous work emphasised that anisotropic collapse characterises the evolution of structure in a universe filled with pressureless matter [11]. Self-similar collapse solutions with high density regions exhibiting spherical symmetry were explored to explain this process [49]. These solutions provide insight into results from numerical simulation.

For a perfectly spherically symmetric geometry, the behaviour of an individual mass shell in the presence of a homogeneous Ω_Λ field can be described by the equation of motion,

$$\ddot{r}_{sh} = -\frac{GM_{sh}}{r_{sh}^2} - \frac{1+3\omega}{2}\Omega_\Lambda H_0^2(1+z)^{3(1+\omega)}r_{sh}, \tag{2.2.1}$$

where r_{sh} is the radius of the shell, and M_{sh} is the mass enclosed within the shell. We obtain a parametric solution for r_{sh} by assuming M_{sh} is constant and Ω_Λ is negligible during the early evolution of the density perturbation. This parametric solution is given by,

$$r_{sh} = r_{ta}\left[(1 - cos\theta_M)/2\right], \tag{2.2.2}$$

with

$$t = t_c\left[(\theta_M - sin\theta_M)/2\pi\right], \tag{2.2.3}$$

and r_{ta} is the turnaround radius given by,

$$r_{ta} = \left[(2GM_{sh}t_c^2)/\pi^2\right]^{\frac{1}{3}}, \tag{2.2.4}$$

with t_c being the time it takes for a shell to collapse to the origin. After collapse, the mass and radii of the shell may vary due to shells on different trajectories crossing. As time progresses, the gravitational potential changes causing the velocities of in-falling particles to follow a velocity distribution where the temperature is proportional to the particle mass. This process is known as violent relaxation, and results in the following relation between total kinetic energy E_k and the total potential energy E_G,

$$E_G + 2E_k = 4\pi P_b r_b^3, \tag{2.2.5}$$

where P_b is the effective pressure due to in-falling particles at the boundary radius r_b. This state in (2.2.5) is called *virial equilibrium*. The virial theorem suggests that the bounding radius of the cluster, after it collapses and relaxes, should be approximately $r_{ta}/2$.

2.3 Clusters as cosmological probes

The evolution of cluster properties provides insight into the evolution of dark matter and dark energy. Certain cosmological models explain how structure formation is influenced by dark matter. By using cluster observations, we are able to measure these model's parameters, such as the matter density Ω_m, the density due to the cosmological constant Ω_Λ, the present Hubble parameter h, normalisation of density perturbations σ_8, and the equation of state that characterises dark energy ω. Surveys of galaxy clusters are a useful method for testing models of structure formation in the universe since galaxy clusters enable us to constrain the cosmological parameters. The constraints on these cosmological parameters are dependent on our ability to determine the mass of galaxy clusters, which is calculated from an observable physical effect that is related to mass. In order to accurately measure the parameters of the cosmological model we would therefore require the uncertainty in the observable-mass relation of clusters to be reduced as much as possible. These surveys promise to be competitive with Supernovae and CMB surveys [50].

Galaxy cluster surveys are carried out using various methods, among which the most widely used are the near-infrared (NIR) optical light from galaxy clusters, the cluster X-ray emission, the weak lensing of background galaxies, and the effect that hot electrons within clusters have on the CMB photons. Examples of galaxy cluster surveys include the optical NIR ESO Imaging Survey (EIS) [51], the X-ray XMM-Newton Large Scale survey (XMM-LSS) [52] and the Planck thermal Sunyaev-Zeldovich all sky cluster survey (tSZ) [53]. The Deep Lens Survey is an example of a weak gravitational lensing survey used to produce unbiased maps of the large-scale structure of the mass distribution at extremely high redshifts [54]. In what follows, we briefly discuss each survey method and their ability to determine the mass of clusters.

The optical observing power of clusters continue to grow due to technological advancements. Cluster catalogues by George Abell and collaborators [55] formed the foundation for much of our understanding of clusters today. Cluster identification techniques have been refined and modified since then to incorporate galaxy colours, which aid in identifying distant clusters since many galaxies appear redder than other galaxies at a similar redshift due to the amount of star formation within the cluster. Optical surveys of galaxy clusters are useful probes since the mass of the cluster can be related to its luminosity.

Cluster masses can also be measured through gravitational lensing of background galaxies [56]. The mass of the cluster deflects photons toward our line of sight through the cluster's center thereby distorting the image of the photon source. Since the unlensed galaxy is generally unknown, the shear distortion and redshift distribution of an entire field of background galaxies is used to determine the mass of the cluster. Weak lensing techniques are however hindered by the mass-sheet degeneracy transformation therefore techniques that combine strong and weak lensing can break this degeneracy [57].

X-ray surveys are useful in determining the properties of galaxy clusters as well since clusters are rich X-ray sources. The baryonic gas in the intergalactic medium is compressed by the deep potential wells of galaxy clusters causing the gas to increase to energies in the X-ray range. The temperature indicates the depth of a cluster's gravitational potential well, and the spectrum also indicates the pressure of the baryonic gas. The pressure is then related to its density, thereafter the mass of the cluster is determined from its relation to the temperature and density of the heated baryonic gas. This is based on the assumption of spherical symmetry and that the baryonic gas is isothermal. Isothermal equilibrium is essential when using the β-model to determine the temperature profile, however we are not restricted to this model as hydrostatic equilibrium is also used

to determine the matter-density profile [58].

The effect of the hot intergalactic gas on CMB photons creates a signature that enables the gas to be detected. The baryonic gas Compton scatters the CMB photons to higher energies thereby distorting the virtually perfect black body spectrum of the CMB. This effect on the CMB spectrum is called the Sunyaev-Zeldovich (SZ) effect [11]. The SZ effect is nearly independent of distance to the cluster and allows clusters to be detected at high redshifts. The distortion of the spectrum is quantified by a parameter that depends upon the probability that a photon will undergo Compton scattering while passing through the cluster and the amount of energy a scattered photon may gain. Integrating this parameter over the cluster's projected surface area provides a measure of the cluster mass.

2.4 Statistics of large-scale structure

Semi-analytic techniques have been developed recently that predict the properties and abundance of different galaxy types. The *Press Schechter* theory [24] forms the basis for this work. Galactic scales that are smaller than $10h\text{Mpc}^{-1}$ have already gone non-linear and this poses a problem for predicting the number density of galaxies, however there have been recent work on nonlinear growth of structure [59] including non-standard cosmology [60]. Galaxy clusters arise from an initial perturbation in a region that is overdense and clusters with mass up to $10^{15}\text{M}_{\odot}$ arise from perturbations on just the right scales. The perturbation can be quantified as the difference in the density in that region compared to the average background density, denoted by

$$\delta = \frac{\rho(\mathbf{x}) - \bar{\rho}(\mathbf{x})}{\bar{\rho}(\mathbf{x})}, \tag{2.4.1}$$

where $\rho(\mathbf{x})$ is the density of that region and $\bar{\rho}(\mathbf{x})$ is the average density of the background. The density perturbation simplifies in Fourier space and we use the convention

$$A(\mathbf{x}) = \int \frac{1}{(2\pi)^3} A(\mathbf{k}) \exp(i\mathbf{k}.\mathbf{x}) d^3\mathbf{k}, \tag{2.4.2}$$

to obtain

$$\delta(\mathbf{x}) = \int \frac{1}{(2\pi)^3} \delta(\mathbf{k}) \exp(i\mathbf{k}.\mathbf{x}) d^3\mathbf{k}, \tag{2.4.3}$$

where $\delta(\mathbf{k})$ is the Fourier component of $\delta(\mathbf{x})$. From isotropy, we have all the $\mathbf{x}_i$ as being identical and the perturbation distribution can be characterised by the power spectrum,

$$P(k) = \left\langle |\delta(k)|^2 \right\rangle, \tag{2.4.4}$$

where $\langle . \rangle$ is the expectation over many realisations. If δ follows a Gaussian distribution then the power spectrum provides a complete description of the perturbations since we know that the mean of $\delta(k)$ is zero, and the variance is given by the power spectrum. For a smoothing spherical window function $W(\mathbf{r})$, the mass perturbation smoothed over the window is,

$$\frac{\delta M}{M}(\mathbf{r}) = \int \delta(\mathbf{x}) W(|\mathbf{x} - \mathbf{r}|) d^3x. \tag{2.4.5}$$

The variance in the smoothed density field over the window function is given by

$$\Sigma^2 = \left\langle \left| \frac{\delta M}{M} \right|^2 \right\rangle. \tag{2.4.6}$$

We use the convolution theorem to relate the variance to the power spectrum and window function in Fourier space. From (2.4.3), the normalising constant of the Fourier transform of $\delta(\mathbf{x})$ is given by $\frac{1}{(2\pi)^3}$. Taking the expectation of $|\frac{\delta M}{M}|^2$ with respect to the window function and using (2.4.5) yields,

$$\begin{aligned} \Sigma^2 &= \left\langle \left| \frac{\delta M}{M} \right|^2 \right\rangle = \frac{1}{(2\pi)^3} \int \left\langle |\delta(k)|^2 \right\rangle |W(k)||W(k)| d^3k \\ &= \frac{1}{(2\pi)^3} \int P(k) |W(k)|^2 d^3k, \end{aligned} \tag{2.4.7}$$

where $W(k)$ is the Fourier transform of $W(r)$.

Using linear theory, we say a region is overdense if for some dimensionless δ_c,

$$\frac{\rho - \bar{\rho}}{\bar{\rho}} > \delta_c, \tag{2.4.8}$$

where δ_c is the critical overdensity for spherical collapse and is dependent on time, that is, $\delta_c = \delta_c(z)$. If we extrapolate it to the present time $\delta_c = 1.69$ for an Einstein-de Sitter cosmology.

We quantify the fraction of collapse by combining spherical top-hat collapse with the growth function for linear perturbations. If the average inhomogeneity in a particular volume is zero, then significant deviations from this average can cause gravitational collapse in those regions, which is why galaxies form in these regions of overdensity. Particles in this region are now trapped in the local gravitational field. The fraction of collapse is given by [24],

$$f(m(R), z) = \frac{2}{\sqrt{2\pi}\sigma(m, R, z)} \int_{\delta_c}^{\infty} d\delta \, e^{-\delta^2/2\sigma^2(m,R,z)}, \tag{2.4.9}$$

where $R = (3m/4\pi\bar{\rho})^{1/3}$ is the radius over which the density field has been smoothed, and $\sigma(R, z)$, given by

$$\begin{aligned} \sigma^2 = \sigma^2(R, z) &= D^2(z)\Sigma^2 \\ &= \frac{D^2(z)}{(2\pi)^3} \int P(k)|W(k, R)|^2 d^3k, \end{aligned} \tag{2.4.10}$$

is the variance in the initial density fluctuation field when smoothed with a top-hat filter of scale R, extrapolated to the present time using linear theory. The function, $W(k, R) = 3(\sin kR - kR\cos kR)/(kR)^3$, is the Fourier transform of a spherical top-hat window function that encloses

mass M. Here we assume that all density perturbations continue to grow according to the linear growth rate $D(z)$. The power spectrum is normalised so that $\sigma(M_8, 0) = \sigma_8$ for

$$M_8 \equiv (8h^{-1}\text{Mpc})^3 H_0^2 \Omega_M / 2G = 6 \times 10^{14} \Omega_M h^{-1} M_\odot.$$

The fraction of collapse (equation (2.4.9)) resembles a Gaussian distribution and since we are integrating from δ_c, we are in essence considering only the regions of overdensity, i.e. the regions of collapse. The possible issues with this formula are that it assumes the distribution of inhomogeneities are Gaussian and that we have ignored non-linear effects. Numerical simulations and recent work however have justified this formula [61; 62].

2.5 The cluster mass function

Cosmological time scales are long which makes it difficult to observe how individual clusters evolve but the demographics of the entire cluster population provide insight into the evolution since the properties change with redshift. This is why the cluster mass function, denoted $n(m, z, \delta)$, is so useful. The cluster mass function yields the number density (number per comoving volume element) of galaxy clusters with mass greater than m at redshift z. Previous work [24] has led to a widely used semi-analytical method for expressing the cluster mass function in terms of the cosmological parameters and has been refined and extended since then [63; 64; 65]. Here we assume that all clusters have the same density $\bar{\rho}$.

A functional form of $n(m, z, \delta)$ is given by

$$n(m, z, \delta)dm = \frac{\bar{\rho}}{m} f(m, z, \delta)dm, \tag{2.5.1}$$

where we have multiplied the fraction of collapse, (2.4.9), by a small interval dm and divided by

the total volume. This gives us the fraction of volume collapsed into objects with mass between m and $m + dm$. This implies that the shape of the mass function depends only on the mass variance σ. The mass function is often written as a differential form of the mass variance by

$$\frac{dn}{d\ln\sigma^{-1}} = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}\delta_c}{m\sigma}e^{-\delta_c^2/2\sigma^2}. \tag{2.5.2}$$

Models have recently been developed that explain the shape and evolution of the mass function of collapsed dark matter halos, as well as for the evolution of the spatial distribution of these halos [27; 25; 28]. Many of these functions are derived as fitting functions to $N-$body simulations. Among those mass functions that are commonly used [20; 66; 67] is the Tinker mass function [28] given by

$$\frac{dn}{d\ln\sigma^{-1}} = f(\sigma)\frac{\bar{\rho}}{m}, \tag{2.5.3}$$

where

$$f(\sigma) = B\left[\left(\frac{\sigma}{b}\right)^{-a} + 1\right]e^{-c/\sigma^2}. \tag{2.5.4}$$

The function $f(\sigma)$ may be an analytic, semi-analytic, or reasonably fitted function to cosmological $N-$body simulations. Recently mass functions have been developed in an effort to increase statistical precision, with the Tinker mass function reaching errors that are less than 5% (for a fixed cosmology at $z = 0$) [28]. In this work, however, we consider the *normalisable Tinker* (NT) mass function in appendix C of [28] given by

$$u(\sigma) = B\left[\left(\frac{\sigma}{e}\right)^{-d} + \sigma^{-h}\right]\exp\left[\frac{-g}{\sigma^2}\right], \tag{2.5.5}$$

since this function is normalisable. This is because the mass function is well behaved and we can compare like probability distributions with each other. This mass function contains four free parameters, namely, d, e, g and h with B given by the normalising constraint,

$$\int u(\sigma)\, d\ln\sigma^{-1} = 1. \tag{2.5.6}$$

In this thesis the mass function will be written as a function of the dimensionless variable

$$\nu = \left(\frac{\delta_c}{\sigma}\right)^2, \tag{2.5.7}$$

following [27]. This re-scaling allows the mass function to be written in a universal functional form that is independent of redshift and the power spectrum. The NT mass function can be transformed to ν as follows.

The mass function in general, with respect to ν is given by

$$\nu F(\nu) = \frac{m}{\bar{\rho}} \frac{dn}{d\ln m} \frac{d\ln m}{d\ln \nu}, \tag{2.5.8}$$

as in [68]. From (2.5.8), we have

$$\frac{m}{\bar{\rho}} = \frac{d\ln m}{dn} \frac{d\ln \nu}{d\ln m} \nu F(\nu). \tag{2.5.9}$$

Similarly from (2.5.3),

$$\frac{m}{\bar{\rho}} = f(\sigma) \frac{d\ln \sigma^{-1}}{dn}. \tag{2.5.10}$$

Equating (2.5.10) and (2.5.9) we obtain,

$$f(\sigma) \frac{d\ln \sigma^{-1}}{d\ln \nu} = \nu F(\nu), \tag{2.5.11}$$

and using

$$\frac{d\ln \sigma^{-1}}{d\ln \nu} = \frac{1}{2}, \tag{2.5.12}$$

equation (2.5.11) can then be expressed as

$$\frac{1}{2\nu} f(\sigma) = F(\nu). \tag{2.5.13}$$

Replacing $f(\sigma)$ with $u(\sigma)$ yields the NT mass function in terms of ν given by,

$$T(\nu) = \frac{B}{2\nu} \left[\left(\frac{\delta_c}{e\sqrt{\nu}}\right)^{-d} + \left(\frac{\delta_c}{\sqrt{\nu}}\right)^{-h} \right] \exp\left[-g\nu/\delta_c^2\right]. \tag{2.5.14}$$

Other mass functions that have been extensively utilised in the literature are the Press-Schechter [24] and Sheth-Tormen [27] mass functions. The Sheth-Tormen (ST) mass function [27] is given by

$$F_{\mathcal{ST}}(\nu) = \frac{A}{\nu}\sqrt{\frac{a\nu}{2\pi}}\left(1 + (a\nu)^{-p}\right)e^{-a\nu/2}, \tag{2.5.15}$$

where typically $a = 0.707$, and $p = 0.3$, measured from numerical simulations [27]. The parameter A is given by the normalising constraint,

$$\int_0^\infty F_{\mathcal{ST}}(\nu)d\nu = 1. \tag{2.5.16}$$

The Press-Schechter (*PS*) mass function [24] is given by

$$F_{\mathcal{PS}}(\nu) = \frac{1}{\nu}\sqrt{\frac{\nu}{2\pi}}e^{-\nu/2}, \tag{2.5.17}$$

which is free of parameters. The Sheth-Tormen model is a modification of the Press-Schechter model that overcomes some of its shortcomings. We will later look at how to derive a probability distribution function for any given model of the mass function.

CHAPTER 3

Bayesian statistics

3.1 Probability theory

In this thesis we adopt the Bayesian methodology for model testing. Probability theory forms the foundation for Bayesian inference. Richard Cox [69] proposed a way to express our beliefs in the truth of various propositions, for example, the probability that there will be a car crash on a certain freeway. This would depend on different factors, like the number of drivers on the freeway that are under the influence of alcohol as well as the number of drivers that drive excessively fast on that freeway. He suggested using real numbers in expressing probabilities so that a quantitative value would enable us to compare our degree of belief. This presents the problem that if there are several different ways of using the same information then it is possible that we may arrive at different conclusions. He solved this problem by asserting two rules of probability,

$$p(X|I) + p(\bar{X}|I) = 1, \tag{3.1.1}$$

and

$$p(X, Y|I) = p(X|Y, I) \times p(Y|I), \tag{3.1.2}$$

where X stands for a proposition which asserts that something is true and $\bar{X}$ for the same proposition being false. The symbol '|' means 'given' which implies that all symbols to the right of the sign are taken to be true. The letter I in the brackets refers to the relevant background information which is always present and affects our prior information of the proposition. There is always information that would affect our state of knowledge of a proposition and it should therefore be contained in I. The comma between the two propositions X and Y in (3.1.2) indicates that both X and Y are true, therefore it can be referred to as an 'AND' operator in *Boolean algebra*. The term $p(X, Y|I)$ is also known as the joint probability of X and Y. Equations (3.1.1) and (3.1.2) are known as the *sum* and *product* rule respectively. Equation (3.1.2) can also be written as

$$p(X, Y|I) = p(Y, X|I) = p(Y|X, I) \times p(X|I), \tag{3.1.3}$$

since (X, Y) are interchangeable which implies the occurrence of both propositions. Equating both right hand sides of equation (3.1.2) and (3.1.3) we obtain

$$p(X|Y, I) = \frac{p(Y|X, I)p(X|I)}{p(Y|I)}, \tag{3.1.4}$$

which is known as *Bayes' theorem*. This is the foundation of all Bayesian inference methods. The denominator in equation (3.1.4) is the proportionality constant. Substituting X and Y for hypothesis and data respectively, we obtain

$$p(hypothesis|data, I) = \frac{p(data|hypothesis, I)p(hypothesis|I)}{p(data|I)}. \tag{3.1.5}$$

We are interested in calculating the term on the left hand side but is generally difficult to evaluate. It is easier to evaluate the right hand side of equation (3.1.5), and thereby indirectly evaluate $p(hypothesis|data, I)$. The terms in equation (3.1.5) have specific names. The term

$p(hypothesis|data, I)$ is known as the *posterior* probability of the hypothesis and the terms $p(data|hypothesis, I)$ and $p(hypothesis|I)$ are known as the *likelihood* and *prior* probability respectively. The marginal likelihood encapsulates the probability of obtaining the measured data had our hypothesis been true. The prior probability represents our state of knowledge about the hypothesis before analysing the data. This can be based on previous experiments for example. The term in the denominator, $p(data|I)$, is the proportionality constant and is important in model selection. It is called the *evidence*. Suppose we have n number of hypotheses and let H_i represent the proposition asserting the truth about a particular hypothesis for $i = 1, 2, \cdots, n$. The set of all H_i denoted $\{H_i\}$ is called the hypothesis space. Suppose $\boldsymbol{D}$ is a set of N data points. The probability of obtaining $\boldsymbol{D}$, given H_i and I are true is denoted by $p(\boldsymbol{D}|H_i, I)$. This is the likelihood function which can also be written as $L(H_i)$. The prior probability for the respective hypothesis is denoted by $p(H_i|I)$ and the posterior probability of the hypothesis H_i is represented by $p(H_i|\boldsymbol{D}, I)$. We derive an expression for the evidence, $p(\boldsymbol{D}|I)$, from the condition

$$\sum_{i=1}^{n} p(H_i|\boldsymbol{D}, I) = 1, \tag{3.1.6}$$

which holds for the posterior probabilities. We evaluate $p(\boldsymbol{D}|I)$ by the following,

$$p(H_i|\boldsymbol{D}, I) = \frac{p(\boldsymbol{D}|H_i, I)p(H_i|I)}{p(\boldsymbol{D}|I)}, \tag{3.1.7}$$

which is taken from Bayes' theorem. Taking the sum, we obtain

$$\sum_{i=1}^{n} p(H_i|\boldsymbol{D}, I) = \sum_{i=1}^{n} \frac{p(\boldsymbol{D}|H_i, I)p(H_i|I)}{p(\boldsymbol{D}|I)} = 1,$$

therefore the evidence is given by

$$p(\boldsymbol{D}|I) = \sum_{i=1}^{n} p(\boldsymbol{D}|H_i, I)p(H_i|I). \tag{3.1.8}$$

The condition $\sum_{i=1}^{n} p(H_i|\boldsymbol{D}, I) = 1$ must hold since it is an extension of the sum rule where there are more than just two outcomes. In the case of a continuous hypothesis space, the summation then changes to an integral, i.e.,

$$p(\boldsymbol{D}|I) = \int_H p(\boldsymbol{D}|H, I)p(H|I),$$

where we integrate over the entire hypothesis space, H. The hypothesis space encompasses the volume in the parameter space that is occupied by all possible models.

We will often encounter problems where we are dealing with more than one parameter. Suppose we have a set of discrete parameters given by $\theta_1, \theta_2..., \theta_n$ such that the joint probability distribution function is denoted by $p(\theta_1, \theta_2, ..., \theta_n|\boldsymbol{D}, I) = p(\boldsymbol{\theta}|\boldsymbol{D}, I)$ where $\boldsymbol{D}$ and I are the data set and prior information respectively as before. If we are interested in the probability distribution of θ_1 only, denoted $p(\theta_1|\boldsymbol{D}, I)$, then we can derive it from the joint probability distribution as follows. In this case we are not concerned with $\theta_2, \cdots, \theta_n$.

In Boolean algebra, the logical operator 'OR' is denoted by '$+$', therefore the event of θ_2 or θ_3 or $\cdots$ or θ_n being true is equivalent to $(\theta_2 + \theta_3... + \theta_n)$ being true. By stating the compound proposition that θ_1 and θ_2 or θ_3 or $\cdots$ or θ_n is true, we use the notation $(\theta_1, [\theta_2 + \theta_3, + \cdots +, \theta_n])$. The probability distribution of this proposition can be expanded using the product rule in (3.1.2) to

$$\begin{aligned} p(\theta_1, [\theta_2 + \theta_3 + ... + \theta_n]|\boldsymbol{D}, I) &= p(\theta_2 + \theta_3 + ... + \theta_n|\boldsymbol{D}, I) \\ &\times \; p(\theta_1|[\theta_2 + \theta_3 + ... + \theta_n], \boldsymbol{D}, I) \end{aligned} \tag{3.1.9}$$

The first term on the right hand side given by $p(\theta_2 + \theta_3 + ... + \theta_n|\boldsymbol{D}, I)$ is equal to unity since

the event is true. The second term in equation (3.1.9) reduces to

$$p(\theta_1|[\theta_2 + \theta_3 + ... + \theta_n], \boldsymbol{D}, I) = p(\theta_1|\boldsymbol{D}, I) \tag{3.1.10}$$

since $[\theta_2 + \theta_3 + ... + \theta_n]$ is true and is contained within our prior information I. Substituting (3.1.10) into equation (3.1.9) and using the fact that

$$p(\theta_2 + \theta_3 + ... + \theta_n|\boldsymbol{D}, I) = 1,$$

we obtain

$$p(\theta_1, [\theta_2 + \theta_3 + ... + \theta_n]|\boldsymbol{D}, I) = p(\theta_1|\boldsymbol{D}, I).$$

The proposition $(\theta_1, [\theta_2 + \theta_3 + \cdots + \theta_n])$ can be expanded using the following rule from Boolean algebra:

$$\theta_1, [\theta_2 + \theta_3 + ... + \theta_n] = \theta_1, \theta_2 + \theta_1, \theta_3 + ... + \theta_1, \theta_n,$$

if we consider the 'AND' operator denoted by the comma as normal multiplication in arithmetic which takes preference over '+' represented by the 'OR' operator. Since these events are mutually exclusive, taking the probabilities of these events yields

$$\begin{aligned} p(\theta_1, [\theta_2 + \theta_3 + \cdots + \theta_1, \theta_n]\,|\boldsymbol{D}, I) &= p(\theta_1, \theta_2|\boldsymbol{D}, I) + p(\theta_1, \theta_3|\boldsymbol{D}, I) + \cdots \\ &+ p(\theta_1, \theta_n|\boldsymbol{D}, I) \\ &= p(\theta_1|\boldsymbol{D}, I). \end{aligned} \tag{3.1.11}$$

Rearranging equation (3.1.11) we obtain

$$p(\theta_1|\boldsymbol{D}, I) = \sum_{i=2}^{n} p(\theta_1, \theta_i|\boldsymbol{D}, I).$$

This can be extended to the continuous case where we have a set of continuous parameters $\theta_1, \theta_2, \cdots, \theta_n$ by the following operation,

$$p(\theta_1|\boldsymbol{D}, I) = \int \cdots \int p(\theta_1, \theta_2, \cdots, \theta_n|\boldsymbol{D}, I) d\theta_2 d\theta_3 \cdots d\theta_n.$$

This problem is encountered in many cases and the parameters that we are summing or integrating over are generally referred to as *nuisance parameters*. In some scientific applications, these nuisance parameters could be measurement errors for example. In high dimensional probability distribution functions (distribution functions with a large number of parameters), it is difficult to visualize the function. We therefore *marginalise* over the parameters that we are not interested, and project the distribution over the remaining parameter space to gain insight into how the distribution function behaves in a particular parameter space or to determine any degeneracy between the parameters.

Our main application in Bayesian statistics will be model selection, that aims to determine which model from a set of theoretical models is preferred.

3.2 Model selection

We first describe the Bayesian methodology and consider an example afterwards. In model selection, one wishes to know which model is preferred, regardless of the parameter values. With nested models this is in a sense the same as determining whether or not the data supports the introduction of a new parameter in our model. An example [44] is whether CMB data supports the inclusion of the spectral index of scalar perturbations.

Suppose the two models we want to compare are M_0 and M_1 and we have the data vector $\boldsymbol{D} = d_1, d_2, \cdots, d_N$ for N data points. We denote parameters in M_0 and M_1 by $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^1$ respectively. Using equation (3.1.4) and setting $X = M_0$ as well as $Y = \boldsymbol{D}$, we have

$$p(M_0|\boldsymbol{D}, I) = \frac{p(\boldsymbol{D}|M_0, I)p(M_0|I)}{p(\boldsymbol{D}|I)}, \tag{3.2.1}$$

where $p(M_0|\boldsymbol{D}, I)$ is the posterior probability of model M_0 given the data. Similarly for M_1,

using the same argument,

$$p(M_1|\boldsymbol{D}, I) = \frac{p(\boldsymbol{D}|M_1, I)p(M_1|I)}{p(\boldsymbol{D}, I)}. \tag{3.2.2}$$

In order to compare M_0 with M_1, we look at the posterior probability ratios,

$$\frac{p(M_0|\boldsymbol{D}, I)}{p(M_1|\boldsymbol{D}, I)} = \frac{p(\boldsymbol{D}|M_0, I)p(M_0|I)}{p(\boldsymbol{D}|M_1, I)p(M_1|I)}. \tag{3.2.3}$$

The left hand side of equation (3.2.3) is not easy to evaluate. If we assume equal weighting on the priors for the models i.e. $p(M_0|I) = p(M_1|I)$, then we are left with

$$\frac{p(M_0|\boldsymbol{D}, I)}{p(M_1|\boldsymbol{D}, I)} = \frac{p(\boldsymbol{D}|M_0, I)}{p(\boldsymbol{D}|M_1, I)}. \tag{3.2.4}$$

For any model M, containing a set of parameters $\boldsymbol{\theta}$, the evidence can be written as,

$$p(\boldsymbol{D}|M, I) = \int p(\boldsymbol{D}|\boldsymbol{\theta}, M, I)p(\boldsymbol{\theta}|M, I)d\boldsymbol{\theta}, \tag{3.2.5}$$

by marginalising over $\boldsymbol{\theta}$ and using the product rule. Hence (3.2.4) now becomes

$$\frac{p(M_0|\boldsymbol{D}, I)}{p(M_1|\boldsymbol{D}, I)} = \frac{\int p(\boldsymbol{D}|\boldsymbol{\theta}^0, M_0, I)p(\boldsymbol{\theta}^0|M_0, I)d\boldsymbol{\theta}^0}{\int p(\boldsymbol{D}|\boldsymbol{\theta}^1, M_1, I)p(\boldsymbol{\theta}^1|M_1, I)d\boldsymbol{\theta}^1}, \tag{3.2.6}$$

where $p(\boldsymbol{D}|\boldsymbol{\theta}^i, M_i, I)$ is the likelihood function with respect to the parameters while $p(\boldsymbol{\theta}^i|M_i, I)$ is the prior probability of the parameters given the model M_i for $i = 0, 1$. Equation (3.2.6) is denoted as

$$B_{01} = \frac{\int p(\boldsymbol{D}|\boldsymbol{\theta}^0, M_0, I)p(\boldsymbol{\theta}^0|M_0, I)d\boldsymbol{\theta}^0}{\int p(\boldsymbol{D}|\boldsymbol{\theta}^1, M_1, I)p(\boldsymbol{\theta}^1|M_1, I)d\boldsymbol{\theta}^1}, \tag{3.2.7}$$

and is given the special name of the *Bayes' factor*. We must keep in mind that the ratio of the posterior probabilities is equal to the ratio of the evidences provided that the priors for the models M_0 and M_1 have equal weighting.

The evidence for model M_0 and M_1, given by the numerator and denominator respectively in equation (3.2.7), is computationally intensive to evaluate. The product of the likelihood and the prior is integrated over the entire parameter space. It is difficult to determine the region in parameter space where the integrand is peaked and in some cases, it may have multiple peaks of equal amplitude. Calculation of the integrand at individual points in parameter space may be computationally intensive and increases exponentially with the number of parameters. Since $\ln B_{01}$ is used as the statistic of choice to select the preferred model by the data according to table 3.1, the uncertainty in the calculation of equation (3.2.7) needs to be well below 2.5 on the Jeffreys' scale to ensure a reliable result, as it may determine whether a model is preferred or not.

Algorithms are available to evaluate the evidence to reasonable accuracy and have been utilised in model selection of cosmological models. Among the algorithms that are used is *nested sampling* [32] that has become popular in cosmological applications [33; 34; 35; 36]. The algorithm explores the parameter space in an ascending likelihood order thereby ensuring that the region of high likelihood is sampled. This is done by transforming the integral to a 1-dimensional problem and using a modified version of the Monte Carlo sampling algorithm. Recently, code has been developed called *CosmoNest* [34; 33], that uses nested sampling to calculate the evidence accurately. A recently developed multimodal nested sampling algorithm called *MultiNest*, [37], is used to explore functions of multiple peaks and is much faster than other methods such as MCMC, when dealing with multimodal functions.

A recently proposed method using Markov Chain Monte Carlo algorithms, shows promising results in cosmology [38] and is suitable for high dimensional functions. Another method is *parallel tempering*, [70] that is designed to explore multiple peaked functions, for example, the MultiNest algorithm. *Thermodynamic integration* provides accurate results but is not compu-

tationally feasible; however the algorithm has been applied in cosmology [43] with reasonable results. Approximation methods such as the *Savage Dickey density ratio* have also provided reasonable results [44]. This method requires one model be nested within the other and that the maximum of the more complex model be situated near the parameter value of the embedded model. For the sake of simplicity, we analyse the evidence for very simple models with only a single parameter in this chapter.

3.2.1 Model selection using the odds ratio

The odds ratio for N possible models is given by

$$O_{ij} = \frac{p(M_i|\boldsymbol{D}, I)}{p(M_j|\boldsymbol{D}, I)}. \tag{3.2.8}$$

Using equation (3.2.2), the odds ratio for the case of two models is

$$\begin{aligned} O_{01} &= \frac{p(M_0|I)}{p(M_1|I)} \times \frac{p(\boldsymbol{D}|M_0, I)}{p(\boldsymbol{D}|M_1, I)} \\ &= \frac{p(M_0|I)}{p(M_1|I)} \times B_{01}. \end{aligned}$$

Note that the denominator of equation (3.2.2) for each model cancels out since it is the same for each model provided that we are using the same set of data, $\boldsymbol{D}$. In most cases the prior ratio of the two models is taken to be unity since we assume the same amount of knowledge of each model before looking at the data, therefore $O_{01} = B_{01}$. From the same condition in equation (3.1.6) a useful relation between the posterior probability of a model and the odds ratio can be derived as follows. Given the condition

$$\sum_{i=0}^{N-1} p(M_i|\boldsymbol{D}, I) = 1,$$

and taking $j = 1$ in equation (3.2.8) we obtain,

$$O_{i1} = \frac{p(M_i|\boldsymbol{D}, I)}{p(M_1|\boldsymbol{D}, I)}. \tag{3.2.9}$$

Taking the sum over all possible models and rearranging, gives

$$p(M_1|\boldsymbol{D}, I) = \frac{1}{\sum_{i=0}^{N-1} O_{i1}}, \tag{3.2.10}$$

which is obtained from using the condition in (3.1.6). Combining equation (3.2.9) and (3.2.10) we obtain the posterior probability of any model i as,

$$p(M_i|\boldsymbol{D}, I) = \frac{O_{i1}}{\sum_{i=0}^{N-1} O_{i1}}.$$

Since $O_{ii} = 1$, the posterior distribution for M_0 is

$$p(M_0|\boldsymbol{D}, I) = \frac{O_{01}}{1 + O_{01}}. \tag{3.2.11}$$

The above result is useful when we know the posterior probability of either model.

We now consider an example to illustrate the concepts that were discussed. Suppose we have two theories about the age of the universe, namely, H_0 and H_1 which are,
H_0: proposition that the age of the universe is 13 billion years.
H_1: proposition that the age of the universe is 10 billion years.
We have obtained some data that indicates $t = 13.5$ billion years. The hypothesis H_0 is equivalent to $H_0 \Rightarrow t = t_0 + e$, where $t_0 = 13$ billion years and e is some measurement uncertainty that follows a normal distribution, i.e.,

$$p(e|I) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{e}{1\text{Gyr}}\right)^2\right],$$

where 'Gyr' represents a gigayear and is equal to 10^9 years. H_1 is equivalent to, $H_1 \Rightarrow t = t_1 + e$ where $t_1 = 10$ billion years. From Bayes' theorem

$$p(H_0|D, I) = \frac{p(D|H_0, I)p(H_0|I)}{p(D|I)},$$

and

$$p(H_1|D,I) = \frac{p(D|H_1,I)p(H_1|I)}{p(D|I)}.$$

Since we wish to compare the proposition of each theory or model we compute the odds ratio which is simply given by the ratio of the two posteriors,

$$\begin{aligned} \text{odds ratio} &= \frac{p(H_0|D,I)}{p(H_1|D,I)} \\ &= \frac{p(D|H_0,I)}{p(D|H_1,I)} \times \frac{p(H_0|I)}{p(H_1|I)} \\ &= B_{01} \times \frac{p(H_0|I)}{p(H_1|I)}. \end{aligned} \tag{3.2.12}$$

We assume that there is no prior information for either model to lead us to prefer one model over the other, hence $p(H_0|I) = p(H_1|I)$. The only term that plays a role in the model selection is therefore B_{01}. For $O_{01} \gg 1$, H_0 is the preferred model and if $O_{01} \ll 1$, then H_1 is the preferred model according to the data. We generally use the *Jeffreys' scale* employed in [71] as a guideline. The Jeffreys' scale is illustrated in table 3.1.

Table 3.1: Jeffreys' scale

$\ln \lvert B_{01} \rvert \leq 1$	not worth more than a bare mention
$1 < \ln \lvert B_{01} \rvert \leq 2.5$	substantial
$2.5 < \ln \lvert B_{01} \rvert \leq 5$	strong to very strong
$\ln \lvert B_{01} \rvert > 5$	decisive

Note that the statistic used is $\ln B_{01}$ and not B_{01} with a positive value favouring H_0 and a negative value favouring H_1. We evaluate the likelihood for H_0 first. Model H_0 is equivalent to $t = t_0 + e$ and since $p(D|H_0,I)$ assumes H_0 to be true, the only way that the data, t, could be different from t_0 is because of the measurement error e. This is equivalent to $e = t - t_0$ and

hence,

$$\begin{aligned} p(D|H_0, I) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-t_0}{1\text{Gyr}}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(13.5-13)^2\right] \\ &= 0.35207. \end{aligned}$$

Similarly,

$$\begin{aligned} p(D|H_1, I) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(t-t_1)^2\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(13.5-10)^2\right] \\ &= 0.00087. \end{aligned}$$

Taking the natural log of (3.2.12) we obtain

$$\begin{aligned} \ln B_{01} &= \ln\left[\frac{p(D|H_0, I)}{p(D|H_1, I)}\right] \\ &= \ln\left[\frac{0.35207}{0.00087}\right] \\ &= 6.00309 > 5. \end{aligned}$$

According to table 3.1, model H_0 is the preferred theory according to the data.

3.3 Choosing a prior distribution

We have only considered cases where the prior distributions for the models we wish to compare are equal. This is not always the case. The prior distribution can affect the odds ratio significantly thereby causing the data to favour a different model as compared to excluding the prior contribution to the odds ratio. In chapter 4, we examine cases in which the choice of the prior

distribution significantly contributes to the model selection results. The prior distribution is dependent on the model parameters. There are two types of parameters, namely *scale* parameters and *location* parameters. Scale parameters can only take on positive values from 0 to ∞ while location parameters can take on both negative and positive values i.e. $-\infty$ to ∞. An example of a scale parameter would be the length L of a racetrack. If the units of length were later found to be incorrectly taken in kilometers instead of meters then this should not affect our prior probability function $p(L|I)$. If $p(L|I)$ was a *uniform* or *flat* distribution then clearly this change in units would affect the probability distribution function (pdf). This problem is overcome by using the *Jeffreys' prior*, which is given by

$$p(L|I) = \begin{cases} \frac{k}{L} & \text{for } L_{min} \leq L \leq L_{max}, \\ \\ 0 & \text{otherwise,} \end{cases}$$

for some constant k. The constant can be obtained from the constraint,

$$\int_{L_{min}}^{L_{max}} p(L|I)dL = 1,$$

therefore

$$k\left[\ln(L_{max}) - \ln(L_{min})\right] = 1,$$

hence,

$$k = \frac{1}{\ln\left(\frac{L_{max}}{L_{min}}\right)}. \tag{3.3.1}$$

The prior is therefore given by

$$p(L|I) = \begin{cases} \frac{1}{L \ln\left(\frac{L_{max}}{L_{min}}\right)} & \text{for } L_{min} \leq L \leq L_{max}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

The choice of prior is essential in Bayesian analysis and accounts for why our state of knowledge or degree of belief is regarded as subjective. Generally this probability is acquired from previous experiments or observations and is updated accordingly. When our prior probability distribution is unknown, it becomes necessary that two people with the same prior information, I, should assign the same pdf.

Consider the case where we have a location parameter X, which is the proposition that a lighthouse is located in some vicinity of the shore, say $x + dx$. Furthermore the origin x, has a pdf denoted by $f(x)$. If later, knowledge has come to our attention that the origin has shifted by an amount c, then we are interested in the probability distribution $p(X'|I)$ with X' being the event that $x' = x + c$. If no prior knowledge about the location was obtained, then the choice of prior must be invariant to a shift in location. This is the same as saying,

$$p(X|I)dX = p(X'|I)dX' = p(X'|I)d(X + c) = p(X'|I)dX, \tag{3.3.2}$$

therefore,

$$f(x) = f(x') = f(x + c), \tag{3.3.3}$$

which means that $f(x) = k = constant$. Hence we obtain a flat pdf with the upper and lower limit as the range of the parameters. To evaluate the constant k, we use the normalising condition

$$\int_{x_{min}}^{x_{max}} p(X|I)dX = \int_{x_{min}}^{x_{max}} kdX = 1, \tag{3.3.4}$$

therefore

$$k = \frac{1}{x_{max} - x_{min}}. \tag{3.3.5}$$

The parameter limits x_{max} and x_{min} are important in model selection problems since it contributes significantly to the Occam factor. If the limits of the parameter are unknown then $p(X|I)$ is referred to as an improper prior and can be used in parameter estimation problems but not model selection.

As a simple example we let Y be the proposition that the distance between Earth and Mars lies between $y + dy$ astronomical units. If the unit of the distance is now changed to, say light years, then we would need a prior pdf that would be invariant under different scales of the parameter. If the initial prior pdf is denoted by $p(Y|I)$ and the probability density of y is $g(y)$ then under some new information that comes to light, that suggests a new scale $y' = \beta y$, we would need to find a prior pdf that would be scale invariant. We want to find the prior pdf that satisfies

$$p(Y|I)dY = p(Y'|I)dY' = p(Y'|I)d(\beta Y) = \beta p(Y'|I)d(Y), \tag{3.3.6}$$

therefore

$$g(y) = \beta g(y') = \beta g(\beta y), \tag{3.3.7}$$

which yields the solution $g(y) = \frac{constant}{y}$, i.e.,

$$p(Y|I) = \frac{constant}{y}. \tag{3.3.8}$$

Hence, for an invariant prior pdf, it would be optimal to use the Jeffreys' prior. Finally, the constant can be evaluated as in equation (3.3.1) to yield

$$p(Y|I) = \begin{cases} \frac{1}{y \ln\left(\frac{y_{max}}{y_{min}}\right)} & \text{for } y_{min} \leq y \leq y_{max}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

3.4 Evaluating the evidence and Occam's razor

In order to calculate the odds ratio we first need the evidence of each model. This is a measure of how well the data fits the model and is simple in the case where the model has no free parameters. The problem arises when the model has unknown parameters. We suppose a vector $\boldsymbol{\theta}$ represents the set of n parameters that is, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_n\}$. We look at the case where model M_1 has one unknown parameter θ. By marginalising the joint probability distribution and using the product rule we obtain

$$p(\boldsymbol{D}|M_1, I) = \int_\theta p(\boldsymbol{D}, \theta|M_1, I)d\theta = \int_\theta p(\boldsymbol{D}|\theta, M_1, I)p(\theta|M_1, I)d\theta. \tag{3.4.1}$$

The second term in the integral, denoted by $p(\theta|M_1, I)$ is the prior probability for the parameter θ and if we have some idea of what the range of the possible parameter values are then we may assign a flat distribution to the prior pdf,

$$p(\theta|M_1, I) = \begin{cases} \frac{1}{\theta_{max}-\theta_{min}} & \text{for } \theta_{min} \leq \theta \leq \theta_{max}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Later we study other possible choices of the prior probability distributions but for the purpose of this example we use the simplest case which is the flat prior distribution. The first term in (3.4.1) is the likelihood function $L(\theta)$ and we will assume that there is some value of θ say, $\hat{\theta}$, that maximises $L(\theta)$, such that

$$L(\hat{\theta}) = L_{max}(\theta).$$

If we assume that the likelihood follows a Gaussian distribution with characteristic width $\delta\theta$, then we write $L(\theta)$ for model M_1 as

$$\begin{aligned} L(\theta, M_1) &= \int_{\theta_{min}}^{\theta_{max}} \frac{1}{\theta_{max} - \theta_{min}} p(D|\hat{\theta}, M_1, I) \exp\left[-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\delta\theta}\right)^2\right] d\theta \\ &= \frac{1}{\theta_{max} - \theta_{min}} \times p(D|\hat{\theta}, M_1, I) \int_{\theta_{min}}^{\theta_{max}} \exp\left[-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\delta\theta}\right)^2\right] d\theta \\ &= \frac{\delta\theta\sqrt{2\pi}}{\theta_{max} - \theta_{min}} \times p(D|\hat{\theta}, M_1, I), \end{aligned}$$

where the last step follows from

$$\int_{\theta_{min}}^{\theta_{max}} \exp\left[-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\delta\theta}\right)^2\right] d\theta = \delta\theta\sqrt{2\pi}, \tag{3.4.2}$$

which is the property of a Gaussian distribution provided the limits of θ are large enough to encompass the tails of the probability distribution. We compute the Bayes' factor using equation (3.2.4), which gives

$$\begin{aligned} B_{01} &= \frac{p(D|M_0, I)}{p(D|M_1, I)} \\ &= \frac{p(D|M_0, I)}{p(D|\hat{\theta}, M_1, I)} \times \frac{\theta_{max} - \theta_{min}}{\delta\theta\sqrt{2\pi}}. \end{aligned}$$

We assume equal prior probabilities of each model, therefore the odds ratio becomes

$$O_{01} = B_{01} = \frac{p(D|M_0, I)}{p(D|\hat{\theta}, M_1, I)} \times \frac{\theta_{max} - \theta_{min}}{\delta\theta\sqrt{2\pi}}. \tag{3.4.3}$$

The second fraction in equation (3.4.3) is the *Occam factor* and serves as a penalty function for model M_1 since it has a free parameter. The numerator, given by $\theta_{max} - \theta_{min}$ indicates that the greater the range (or uncertainty) of the parameter θ, the greater the penalty.

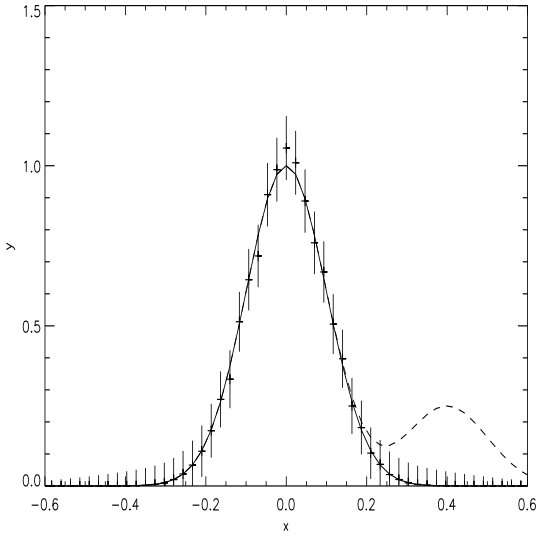

Figure 3.1: The mock data with associated measurement errors used to discriminate between model A (solid line) and model B (dashed line). In this plot the free parameter b in model B is set to 0.25.

We use a concrete example with mock data to illustrate these concepts. In this example we wish to discriminate between two competing models using the mock data that we have created using 60 data points with measurement errors of 10%. The fiducial model and data set $\{D_k\}$ is plotted in figure 3.1 along with its associated measurement errors $\{\sigma_k\}$. The fiducial model that we use is

$$y = \exp\left[-\frac{1}{2}\left(\frac{x}{0.1}\right)^2\right].$$

The fiducial model is depicted by the solid line in figure 3.1. The models that we wish to compare are

$$\text{Model } A\text{: } y = \exp\left[-\frac{1}{2}\left(\frac{x}{0.1}\right)^2\right],$$

which we take to be our fiducial model and,

$$\text{Model } B\text{: } y = \exp\left[-\frac{1}{2}\left(\frac{x}{0.1}\right)^2\right] + b\exp\left[-\frac{1}{2}\left(\frac{x-0.4}{0.1}\right)^2\right],$$

with an unknown parameter b. The posterior ratio in this case is denoted by

$$\text{posterior ratio} = \frac{p(A|\boldsymbol{D}, I)}{p(B|\boldsymbol{D}, I)}. \tag{3.4.4}$$

We use Bayes' theorem to extend equation (3.4.4) to

$$\frac{p(A|\boldsymbol{D},I)}{p(B|\boldsymbol{D},I)} = \frac{p(\boldsymbol{D}|A,I)}{p(\boldsymbol{D}|B,I)} \times \frac{p(A|I)}{p(B|I)}. \tag{3.4.5}$$

Our prior probabilities are $p(A|I)$ and $p(B|I)$ for models A and B respectively and indicates our knowledge of the models before analysis of the data. Since we are just as ignorant for our theory of model A as we are for model B we assume this ratio to equal unity. In evaluating $p(\boldsymbol{D}|A,I)$ and $p(\boldsymbol{D}|B,I)$, we compare the predictions of each theory to the data. We assume our data set is given by $\{D_k\}$, for $k = 1, 2, 3, \cdots, 60$ i.e. we use 60 data points.

We evaluate the chi squared statistic denoted by χ^2, which is given by

$$\chi^2 = \sum_{k=1}^{N} \left[\frac{F_k - D_k}{\sigma_k}\right]^2, \tag{3.4.6}$$

where σ_k is the respective measurement error for the k^{th} datum and F_k is the model prediction. Equation 3.4.6 is the sum of the square of the normalised residuals. The likelihood is approximated by

$$p(\boldsymbol{D}|A,I) \approx e^{-\frac{\chi^2}{2}}. \tag{3.4.7}$$

We find that $p(\boldsymbol{D}|A,I) = 1$ since model A is the fiducial model. The likelihood for Model B however, is not as straightforward to compute due to the parameter b as we cannot make predictions without knowing b. We therefore use the sum and product rule to relate the probability we require to other pdfs which might be easier to evaluate, which gives

$$\begin{aligned} p(\boldsymbol{D}|B,I) &= \int_{-\infty}^{\infty} p(\boldsymbol{D},b|B,I)db \\ &= \int_{-\infty}^{\infty} p(b|B,I) \times p(\boldsymbol{D}|b,B,I)db, \end{aligned} \tag{3.4.8}$$

where $p(b|B,I)$ is referred to as the prior probability for b under model B. A flat prior distribution yields

$$p(b|B,I) = \begin{cases} \frac{1}{b_{max}-b_{min}} & \text{for } b_{min} \leq b < b_{max}, \\ \\ 0 & \text{otherwise,} \end{cases}$$

where b_{min} and b_{max} are our minimum and maximum values of b respectively. We find a value $\hat{b} = 0$ that maximises the likelihood for model B. The maximum of the likelihood will therefore be denoted by $p(\boldsymbol{D}|\hat{b},B,I)$. A reasonable fit to the data can be represented by the Gaussian pdf

$$p(\boldsymbol{D}|b,B,I) = p(\boldsymbol{D}|\hat{b},B,I)\exp\left[-\frac{(b-\hat{b})^2}{2\sigma_b^2}\right]. \tag{3.4.9}$$

We now calculate σ_b from our fit. Substituting equation (3.4.9) in equation (3.4.8) together with the prior distribution for $p(b|B,I)$ yields

$$\begin{aligned} p(\boldsymbol{D}|B,I) &= \frac{1}{b_{max}-b_{min}} \times p(\boldsymbol{D}|\hat{b},B,I)\int_{b_{min}}^{b_{max}} \exp\left[-\frac{(b-\hat{b})^2}{2\sigma_b^2}\right] db \\ &= \frac{1}{b_{max}-b_{min}} \times p(\boldsymbol{D}|\hat{b},B,I) \times \sqrt{2\pi}\sigma_b, \end{aligned}$$

since

$$\int_{b_{min}}^{b_{max}} \exp\left[-\frac{(b-\hat{b})^2}{2\sigma_b^2}\right] db = \sqrt{2\pi}\sigma_b, \tag{3.4.10}$$

is the property of a Gaussian distribution provided b_{min} and b_{max} are chosen to cover a large enough range so as to incur a negligible error in the integral. We choose $b_{min} = -1.1$ and $b_{max} = 1.1$. Numerically we find

$$\int_{b_{min}}^{b_{max}} \exp\left[-\frac{(b-\hat{b})^2}{2\sigma_b^2}\right] db = 0.034, \tag{3.4.11}$$

hence we solve for σ_b using equation (3.4.10) and (3.4.11) to obtain $\sigma_b = 0.01$. The maximum of the likelihood for model B is given by $p(\boldsymbol{D}|\hat{b}, B, I) = 1$. The odds ratio is therefore evaluated by

$$\begin{aligned} \text{odds ratio} &= \frac{p(A|\boldsymbol{D}, I)}{p(B|\boldsymbol{D}, I)} \\ &= \frac{p(\boldsymbol{D}|A, I)}{p(\boldsymbol{D}|\hat{b}, B, I)} \times \frac{b_{max} - b_{min}}{\sigma_b\sqrt{2\pi}} \\ &= \frac{1}{1} \times \frac{2.2}{0.01\sqrt{2\pi}} \\ &\approx 9. \end{aligned}$$

The odds ratio is greater than 5, therefore we would be inclined to favour model A. This can also be shown by looking at each model's respective evidence individually. The evidence for model A is given by $p(\boldsymbol{D}|A, I)$ which we found to be 1. The evidence for model B however is more complicated to evaluate and is given by

$$p(\boldsymbol{D}|B, I) = p(\boldsymbol{D}|\hat{b}, B, I) \times \left(\frac{\sigma_b \times \sqrt{2\pi}}{b_{max} - b_{min}} \right). \tag{3.4.12}$$

Substituting the relevant values from above into (3.4.12) we obtain

$$p(\boldsymbol{D}|B, I) = 0.01.$$

We can see that the evidence for model A is greater than the evidence for model B. We should expect this as the maximum likelihood estimate, $\hat{b}$, was equal to 0. With $b = 0$, this reduces model B to model A which is the fiducial model we used to generate the data. The Occam factor here is given by $\left(\frac{\sigma_b\sqrt{2\pi}}{b_{max} - b_{min}} \right)$ and penalises model B for the extra parameter, b. Generally the model with more free parameters will have a greater probability of preference over the simpler model but the Occam factor serves to penalise the model whose prior is less informative over the likelihood space. In the next chapter we will study the behaviour when a model with more free parameters is taken to be the fiducial model.

3.5 Bayes' factor in model selection

In the previous example, Model B had a free parameter, denoted by b. We now look at how the log of the Bayes' factor varies with models that differ by the fiducial value for b. The models we wish to compare are

$$\begin{aligned} \text{Model } A \ : y &= \exp\left[-\frac{1}{2}\left(\frac{x}{0.1}\right)^2\right] \text{ and} \\ \text{Model } B_b \ : y &= \exp\left[-\frac{1}{2}\left(\frac{x}{0.1}\right)^2\right] + b \times \exp\left[-\frac{1}{2}\left(\frac{x-0.4}{0.1}\right)^2\right], \end{aligned}$$

where model B_b is specified by its fiducial value for b. We take a discrete sample of b values in the range $[-0.4, 0.4]$ to analyse the effect on $\ln B_{01}$. We use model A as the fiducial model. The Bayes' factor, or evidence ratio, is given by

$$B_{01} = \frac{p(\boldsymbol{D}|A, I)}{p(\boldsymbol{D}|B_b, I)},$$

where $\boldsymbol{D}$ is the given data and I is the relevant background information as before. The Bayes' factor indicates the preference of one model over another, according to the data. In this case we are comparing model A to B_b. For $B_{01} \gg 1$, the data is inclined to favour model A and if $B_{01} \ll 1$ then it would favour model B_b, however, if $B_{01} = 1$ then not much can be said about which model is a better fit to the data.

For illustrative purposes, we have plotted the inverse of the log Bayes' factor i.e. $(\ln B_{01})^{-1}$. We denote this quantity by

$$\beta = (\ln B_{01})^{-1}. \tag{3.5.1}$$

This transformation in conjunction with table 3.1 tells us that for $\beta \leq -1.2$, model A is favoured and if $\beta > -1.2$, it is inconclusive as to which model is preferred.

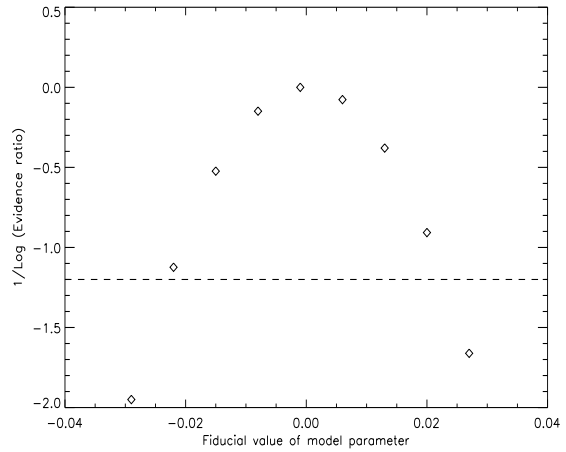

Figure 3.2: The inverse of $\ln B_{01}$. The preference for model A over the various alternative models follows a parabolic relationship. The inverse of the log of the Bayes' factor for model A against the various models is indicated by the diamond shapes. The dashed line indicates the critical value for the Bayes' factor that is used as a guideline to choose a model. Points below the dashed line indicate decisive preference for model A.

Referring to figure 3.2 we see that model A is preferred when $|b| \geq 0.02$. The Bayes' factor in the region $|b| < 0.02$ is not strong enough to indicate a model that should better fit the data, hence for the alternate model B_b with $-0.02 < b < 0.02$, we are unable to decisively choose one model over the other. When $b = 0$, model B_b reduces to model A. This is an example of model selection between nested models where in this case, model A is said to be nested in model B_b.

3.6 Fisher information matrix

We now describe the *Fisher information matrix* since it plays a role in our model selection techniques when considering different forms of the evidence, such as the distributions of section 3.7 and chapter 4. Suppose we have a model consisting of a parameter set $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_M)$ and we wish to know how accurately we can estimate these parameters from a given data set. An example of a data set is the counts-in-cells of a galaxy redshift survey. We denote the data set of N real numbers by $\boldsymbol{D} = (d_1, d_2, \cdots, d_N)$ and likelihood denoted by $L(\boldsymbol{D}|\boldsymbol{\theta})$ as before. If there exists an estimator for the arbitrary parameter θ_α, say $\hat{\theta}_\alpha$ such that

$$\left\langle \hat{\theta}_\alpha \right\rangle = \theta_\alpha, \tag{3.6.1}$$

where $\langle . \rangle$ denotes the expectation computed by averaging over many realisations, and the standard deviation

$$\Delta\theta_\alpha = \sqrt{\left\langle \hat{\theta}_\alpha^2 \right\rangle - \left\langle \hat{\theta}_\alpha \right\rangle^2},$$

is a minimum, then $\hat{\theta}_\alpha$ is referred to as a *Best Unbiased Estimator* (BUE), of θ_α. If condition (3.6.1) only is satisfied then $\hat{\theta}_\alpha$ is just an unbiased estimator of θ_α.

The *maximum likelihood estimator* (MLE) is the parameter, say θ_0, that maximises the likelihood, i.e. $L(\theta_0) = L_{max}$. Since the MLE maximises the likelihood, we have

$$\frac{\partial L}{\partial \theta}|_{\theta_0} = 0. \tag{3.6.2}$$

Expanding $-\ln L$ around the MLE estimates $\boldsymbol{\theta_0}$ in a Taylor series yields

$$-\ln L \approx -(\theta_\alpha - \theta_{0\alpha})(\theta_\beta - \theta_{0\beta})\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta}|_{\boldsymbol{\theta_0}} - \cdots , \tag{3.6.3}$$

with higher order terms in the Taylor expansion neglected. If L is sharply peaked around the MLE values, it would imply that the errors on the parameters are small enough for the third term to be neglected. The likelihood function is therefore approximately Gaussian near the MLE. The width of the distribution will depend on $\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta}$ and tells us how sharply peaked the likelihood is around its MLE values.

The Fisher information matrix is given by

$$F_{\alpha\beta} = \left\langle -\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle .$$

The Fisher information matrix is an important quantity because of the *Cramer Rao inequality* that states that $\Delta\theta_\alpha \geq \sqrt{F^{-1}_{\alpha\alpha}}$ (for a proof of this refer to Appendix A). The Cramer-Rao inequality places a lower bound on the uncertainty of the parameter estimates. Since the Fisher information

matrix does not incorporate the data, the uncertainty bounds hold no matter which method is used to determine the errors from the data. The MLE and BUE are linked to the Fisher information matrix by the following theorem:

If

$$\left\langle \hat{\theta}_\alpha \right\rangle = \theta_\alpha, \tag{3.6.4}$$

and

$$\Delta\theta_\alpha = \sqrt{F^{-1}_{\alpha\alpha}}, \tag{3.6.5}$$

then $\hat{\theta}_\alpha$ is the BUE of θ_α. The lower bound on the marginal errors of the parameters are obtained from the diagonal elements of the inverse Fisher information matrix by using the inequality

$$\Delta\theta_\alpha \geq \sqrt{F^{-1}_{\alpha\alpha}}. \tag{3.6.6}$$

3.7 The Bayesian razor and the log razor ratio

We now extend our notion of the Bayes' factor between two parametric models to that of the *Bayesian razor* between two probability distributions. Suppose we have a set of N data generated from a *true* distribution, that we are trying to model. The problem is that we have two models, say $\mathcal{M}_0$ and $\mathcal{M}_1$, and we want to pick the one that best describes the true distribution. To solve this problem, we use a statistic in the context of Bayesian inference that assigns a merit to a model. This statistic is called the Bayesian *razor* and is an index of the accuracy and the simplicity of a model family as a description of a given true distribution [45]. The razor is similar to the evidence since both are integrated over the product of the likelihood and prior distribution parameter space. The razor is the expectation of the evidence. In the case of the razor, we assume that the data follows a true distribution denoted by $p_{true}(\vec{\nu})$, for some vector $\vec{\nu}$ that represents the physical quantity measured by the data. Suppose we have a distribution given by $p(\vec{\nu}|\boldsymbol{\theta}, \mathcal{M})$,

where $\boldsymbol{\theta}$ are the parameters contained within model $\mathcal{M}$. We expect this distribution to be in good agreement with the true distribution. We have the following relationship for the likelihood:

$$L \propto \exp\left[-D(p_{true}(\vec{\nu})\|p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M}))\right], \tag{3.7.1}$$

where

$$D\left(a\|b\right) = \int d\nu\, a \ln \frac{a}{b},$$

is the Kullback-Liebler distance between two probability distributions $a = a(\vec{\nu})$ and $b = b(\vec{\nu})$. The proportionality constant is not significant since it is the same for both models, therefore it cancels out when calculating the ratio of the razors for both models. This form of the likelihood is similar to our previous approximation where we use the χ^2 statistic instead of D to measure the fit to the data. The Kullback-Liebler distance is a non-symmetric measure of the distance between two probability distributions and in our case, the distributions are $p_{true}(\vec{\nu})$ and $p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})$. The last component of the razor that is required is the prior distribution. We require the prior to be scale invariant, hence we use the Jeffreys' prior which is given by

$$\text{Jeffreys' prior} = \frac{\sqrt{\det\left(J_{ij}(\boldsymbol{\theta})\right)}}{\int d^n(\boldsymbol{\theta})\sqrt{\det(J_{ij}(\boldsymbol{\theta}))}}, \tag{3.7.2}$$

where $J_{ij}(\boldsymbol{\theta})$ is the Fisher information matrix, i.e.,

$$J_{ij}(\boldsymbol{\theta}) = \left\langle -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln L(\boldsymbol{\theta}) \right\rangle.$$

Following Balasubramanian [45] we define the razor as,

$$\mathcal{R}(\mathcal{M}) = \int d^n(\theta) \frac{\sqrt{\det(J_{ij}(\boldsymbol{\theta}))}e^{-D(p_{true}(\vec{\nu})\|p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M}))}}{\int d^n(\boldsymbol{\theta})\sqrt{\det(J_{ij}(\boldsymbol{\theta}))}}.$$

The razor can also be extended to use a flat prior instead, which we examine in chapter 4. The log razor ratio is a form of the log evidence ratio, applicable to probability distribution functions,

and can be used in the same way to discriminate between models. The form of the log evidence ratio we use is the log razor ratio given by

$$\mathcal{R}_{01} = \ln \frac{\mathcal{R}(\mathcal{M}_0)}{\mathcal{R}(\mathcal{M}_1)}$$

and we use the table 3.1 to discriminate between model $\mathcal{M}_0$ and model $\mathcal{M}_1$, treating $\mathcal{R}_{01}$ as the Bayes' factor B_{01}. More details of the razor are given in the next chapter, when we calculate the razor ratio for different cosmological mass functions.

CHAPTER 4

Discriminating mass functions using the Bayesian razor

In cosmology, $N-$body simulations are used to study the process of structure formation, such as the process of forming galaxy halos from dark and baryonic matter. These simulations consist of a number of particles with a particular minimum mass. The goal of this chapter is to determine how many particles in a simulation would be required for the choice of the cluster mass function to become relevant. In section 4.2 we use the log of the Bayesian razor, $\ln \mathcal{R}_{01}$, reviewed in chapter 3.7, as a statistic to discriminate between the mass functions described in chapter 2.5. In section 4.1, we begin by establishing a probability distribution for the cosmological mass function since this distribution is incorporated into the razor. We then use this methodology to quantitatively discriminate between two mass functions, while also determining the number of particles that would be required in order for the choice of the mass function to matter. Constructing the prior and likelihood distributions may require intensive computation due to the number of parameters in the mass function. This integration is therefore done numerically over a grid of points in the parameter space, after which we use interpolation methods to obtain the functional

form of the distribution. These distributions were computed using Mathematica software.

4.1 Probability distribution for a cosmological mass function

We wish to obtain a probability distribution for the mass function using the theory we have laid out in chapter 3. The parameter set $\boldsymbol{\theta}$ will refer to the parameters in the mass function, for example, the parameters in the *ST* mass function will be $\boldsymbol{\theta} = (a, p)$. The observations $\boldsymbol{D}$ are based on the masses of clusters. In general, the mass function is dependent on the model parameters i.e. $n(m, z|\mathcal{M}) = n(m, z|\boldsymbol{\theta}, \mathcal{M})$. To simplify our analysis, we ignore the dependence on redshift z (reserving it for a future study) and will write $n(m|\boldsymbol{\theta}, \mathcal{M})$ instead. The problem is complicated since not all particles of mass are found in clusters. Those that are not in clusters are said to be in *dust*. The question we want to ask is: if we pick a bit of mass, δm, at random in the universe, what is the probability that this bit of mass is in a cluster of mass m or in dust? Our treatment follows the notation used in [68].

To answer this question, we begin with the simplifying assumption that all mass are in clusters so that there is no dust. The probability that a tiny bit of mass, δm, is in a cluster with mass between m and $m + dm$ is given by

$$p(m|\boldsymbol{\theta}, \mathcal{M})dm = \frac{m}{\bar{\rho}} n(m|\boldsymbol{\theta}, \mathcal{M})dm = F(\nu)d\nu. \tag{4.1.1}$$

This can also be thought of as the fraction of mass contained in clusters that have mass in the range m to $m + dm$. We have

$$\int_0^\infty m\, n(m|\boldsymbol{\theta}, \mathcal{M})dm = \bar{\rho}, \tag{4.1.2}$$

therefore

$$\int_0^\infty p(m|\boldsymbol{\theta}, \mathcal{M})dm = 1, \tag{4.1.3}$$

which satisfies the normalisation condition. We suppose all the mass in a particular comoving volume of the universe is divided into N particles that we label $(P_1, P_2, \cdots, P_N)$, each of mass δm. The probability that particle P_i is found in a cluster with mass between m_i and $m_i + dm_i$ is thus $p(m_i|\boldsymbol{\theta}, \mathcal{M})dm_i$. Since the particles are independent, the joint pdf can be obtained by multiplying out the individual pdfs for each m_i,

$$p(m_1, m_2, \cdots, m_N|\boldsymbol{\theta}, \mathcal{M})d^N m = \prod_{i=1}^{N} p(m_i|\boldsymbol{\theta}, \mathcal{M})dm_i. \tag{4.1.4}$$

We now introduce dust into the model. Suppose N_c is the total number of particles in clusters above some minimum mass M_{min} from a simulation of N particles. The number of particles in dust is $N - N_c = N_d$. The likelihood is then given by,

$$L(\boldsymbol{\theta}) = p(m \leq M_{min})^{N_d} \prod_{j=1}^{N_c} p(m_j|\boldsymbol{\theta}, \mathcal{M}). \tag{4.1.5}$$

We now change variables from mass, m, to a dimensionless version of mass, ν defined in (2.5.7). In terms of ν, the mass function also gives the probability that a tiny particle with mass $\delta m \ll m$ is in a cluster with a dimensionless mass parameter between ν and $\nu + d\nu$, i.e.,

$$p(m_j|\boldsymbol{\theta}, \mathcal{M})dm_j = p(\nu_j|\boldsymbol{\theta}, \mathcal{M})d\nu_j, \tag{4.1.6}$$

hence (4.1.5) can be written as,

$$L(\boldsymbol{\theta}) = f_d(\boldsymbol{\theta}, \mathcal{M})^{N_d} \prod_{j=1}^{N_c} p(\nu_j|\boldsymbol{\theta}, \mathcal{M}), \tag{4.1.7}$$

where

$$\begin{aligned} f_d(\boldsymbol{\theta}, \mathcal{M}) &= p(m \leq M_{min}|\boldsymbol{\theta}, \mathcal{M}) \\ &= p(\nu \leq \nu_d|\boldsymbol{\theta}, \mathcal{M}) \\ &= \int_0^{\nu_d} F(\nu|\boldsymbol{\theta}, \mathcal{M}) d\nu, \end{aligned} \tag{4.1.8}$$

for some parameter ν_d where one can no longer distinguish among particles with $\nu < \nu_d$ [68]. Taking the natural log of the likelihood,

$$\ln L(\boldsymbol{\theta}) = N_d \ln f_d(\boldsymbol{\theta}, \mathcal{M}) + \sum_{j=1}^{N_c} \ln\, p(\nu_j|\boldsymbol{\theta}, \mathcal{M}). \tag{4.1.9}$$

4.2 Applying Bayesian statistics to cosmological mass functions

We apply the Bayesian razor to the cosmological mass function in order to assign merit to the models we consider. From chapter 3.7, we established that the razor is the integral over the parameter space of the product of the likelihood and prior distribution. The likelihood is approximated using the Kullback-Liebler distance measure between two distributions and the prior is calculated using the Fisher information matrix. We use the Jeffreys' prior given in chapter 3.7. We therefore establish the Fisher information matrix for the cosmological mass function which we use to derive a functional form of the Jeffreys' prior. The final step is to establish the Kullback-Liebler distance measure after which we derive the functional form of the likelihood. For the mass functions in general, the Fisher information matrix is

$$J_{ij}(\boldsymbol{\theta}, \mathcal{M}) = N \left[f_d \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f_d(\boldsymbol{\theta}, \mathcal{M}) + \int_{\nu_d}^{\infty} d\nu\, p(\nu|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\, p(\nu|\boldsymbol{\theta}) \right]. \tag{4.2.1}$$

Refer to Appendix B for a derivation. In the trivial case of no dust this simply reduces to,

$$J_{ij(\text{no dust})}(\boldsymbol{\theta}, \mathcal{M}) = N \left[\int_0^\infty d\nu \, p(\nu|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln \, p(\nu|\boldsymbol{\theta}) \right]$$

The probability distribution is just the cluster mass function, i.e.,

$$p(\nu|\boldsymbol{\theta}, \mathcal{M}) = F(\nu|\boldsymbol{\theta}, \mathcal{M}), \tag{4.2.2}$$

where $F(\nu|\boldsymbol{\theta}, \mathcal{M})$ is given in chapter 2.5 for a given model of the mass function. The first model that we will look at is the Sheth-Tormen model ($\mathcal{M}_{\mathcal{ST}}$), whose mass function is given by (2.5.15) where $\boldsymbol{\theta} = (a, p)$, so the Fisher information matrix including dust is now dependent on two parameters and is a 2×2 matrix denoted by

$$J_{ij}(\boldsymbol{\theta}, \mathcal{M}) = \begin{bmatrix} J_{aa} & J_{ap} \\ J_{pa} & J_{pp} \end{bmatrix}.$$

Since this matrix is symmetric about the diagonal, i.e. $J_{ap} = J_{pa}$, we need only evaluate $\frac{n(n+1)}{2} = 3$ terms, for $n = 2$ parameters. The Jeffreys' prior is given by

$$p(\boldsymbol{\theta}|\mathcal{M}) = \frac{\sqrt{\det(J_{ij}(\boldsymbol{\theta}, \mathcal{M}))}}{\int d^n\theta \sqrt{\det(J_{ij}(\boldsymbol{\theta}, \mathcal{M}))}}. \tag{4.2.3}$$

The prior becomes more computationally intensive as the number of parameters increases. The denominator of (4.2.3) is a normalising constant which ensures that

$$\int d^n\theta \, p(\boldsymbol{\theta}|\mathcal{M}) = 1.$$

The ranges for the parameters are $a \in (0, 1.5]$ and $p \in [0, 0.5)$. Figure 4.1 is a plot of the Jeffreys' prior including dust. Note that we have chosen an arbitrary limit of $\nu_d = 0.1$ as in most cases in this work. This limit has not been supported by observation or simulations and requires calibration which will form part of work in the near future.

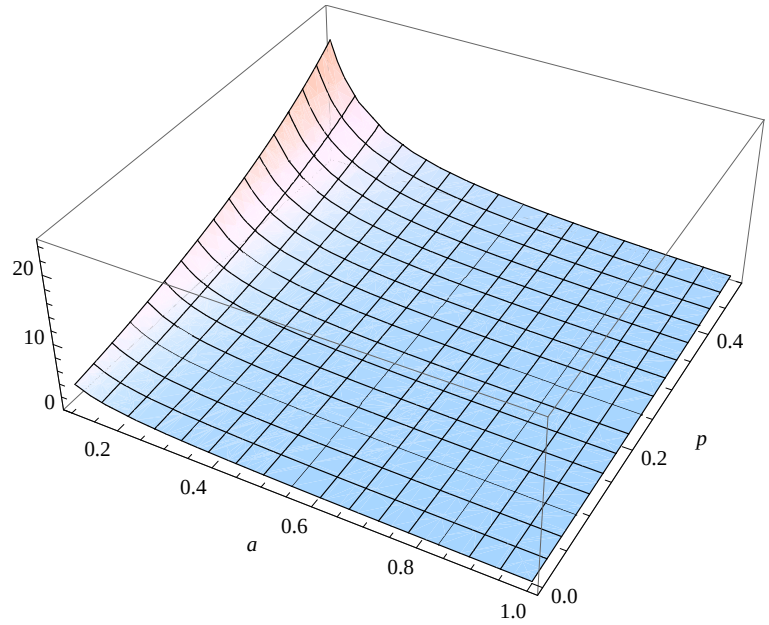

Figure 4.1: The Jeffreys' prior for the Sheth-Tormen model, ($\mathcal{M}_{\mathcal{ST}}$), given by (4.2.3). The prior is a function of two parameters within the Sheth-Tormen model, namely, a and p and essentially depends on the Fisher information matrix.

The Jeffreys' prior increases significantly as a and p approach their region of singularity of $a \to 0$ and $p \to \frac{1}{2}$.

Now that we have the prior distribution, the next step is to obtain the exponent of the Kullback-Liebler distance since

$$L(\boldsymbol{\theta}|\mathcal{M}) \propto e^{-D}, \tag{4.2.4}$$

where for any model in general, the Kullback-Liebler distance, described in chapter 3.7, is given by

$$D = D\left(p(\vec{\nu}|\boldsymbol{\theta}_{true})||p(\vec{\nu}|\boldsymbol{\theta})\right) = N\left[f_d(\boldsymbol{\theta}_{true}) \ln \frac{f_d(\boldsymbol{\theta}_{true})}{f_d(\boldsymbol{\theta})} + \int_{\nu_d}^{\infty} p(\nu|\boldsymbol{\theta}_{true}) \ln \frac{p(\nu|\boldsymbol{\theta}_{true})}{p(\nu|\boldsymbol{\theta})} d\nu\right]. \tag{4.2.5}$$

Refer to Appendix C for a derivation. The computation of (4.2.1) and (4.2.5) is intensive due to the number of parameters in a model. The computation becomes more intensive as the number of parameters increase. An analytic evaluation proves too impractical since it requires an excessively large amount of time. To compute (4.2.1), we consider a sample of points in the parameter

space within the specified range and evaluate the Fisher information matrix at each point in the sample, creating a table of points. We then interpolate the table of points to obtain a function that is representative of the Fisher information matrix. The Kullback-Liebler distance measure is evaluated using the same method as for the Fisher information matrix. A large enough sample of points is chosen such that it captures the behaviour of the distribution at high density regions. This can be done by visually examining the distribution, however for mass functions that have more than 2 parameters, it is best to examine the behaviour in each individual plane of the parameter while holding the other parameters constant, preferably at the fiducial values.

We denote the Kullback-Liebler distance for $\mathcal{M}_{\mathcal{ST}}$ as $D_{\mathcal{ST}}$. We now compare two models given specific fiducial values for $\boldsymbol{\theta}$. The model we will compare with $\mathcal{M}_{\mathcal{ST}}$ is the Press-Schechter mass function, $\mathcal{M}_{\mathcal{PS}}$, given by (2.5.17). Notice that $\mathcal{M}_{\mathcal{PS}}$ closely resembles that of Sheth-Tormen. This is because $\mathcal{M}_{\mathcal{PS}}$ is nested within $\mathcal{M}_{\mathcal{ST}}$, i.e., we can obtain $\mathcal{M}_{\mathcal{PS}}$ by setting $a = 1$ and $p = 0$ in $\mathcal{M}_{\mathcal{ST}}$. We consider the case of $\mathcal{M}_{\mathcal{PS}}$ being the fiducial model. Hence $\boldsymbol{\theta}_{true} = (a_{true}, p_{true})$ where $a_{true} = 1$ and $p_{true} = 0$.

The Kullback-Liebler distance, for $\mathcal{M}_{\mathcal{ST}}$ is a function of a and p. We now have the likelihood, and we are in a position to evaluate the razor for $\mathcal{M}_{\mathcal{ST}}$, which is

$$Razor_{\mathcal{ST}} = \frac{1}{\int_a \int_p dp\, da \sqrt{\det(J_{ij}(\boldsymbol{\theta}, \mathcal{M}))}} \int_a \int_p dp\, da \sqrt{\det(J_{ij}(\boldsymbol{\theta}, \mathcal{M}))} e^{-D_{\mathcal{ST}}}, \qquad (4.2.6)$$

where $D_{\mathcal{ST}}$ is given in (4.2.5) with $\boldsymbol{\theta} = (a_{true}, p_{true})$. The integral in (4.2.6) is a function of the number of particles, N, and is calculated numerically for different values of N.

We turn our attention to the razor for the Press-Schechter model. This is much simpler to evaluate since $\mathcal{M}_{\mathcal{PS}}$ does not contain any free parameters, therefore there is no prior, or we can

think of the prior as just being equal to unity. Therefore we need only focus on the Kullback-Liebler distance. This is given by $D_{\mathcal{PS}} = 0$. This is because we assumed $\mathcal{M}_{\mathcal{PS}}$ to be the fiducial distribution, hence, we are calculating the distance between the same distributions. The razor for the Press-Schechter mass function is then simply

$$Razor_{\mathcal{PS}} = e^{-D_{\mathcal{PS}}} = 1. \tag{4.2.7}$$

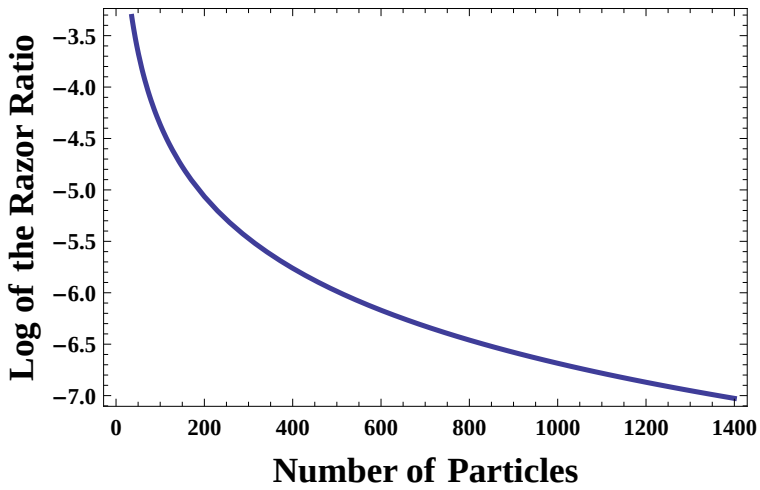

Figure 4.2: Plot of the log razor ratio for the Sheth-Tormen model versus the Press-Schechter model ($\ln \frac{Razor_{\mathcal{ST}}}{Razor_{\mathcal{PS}}}$) against the number of particles, N. Negative values for $\ln \frac{Razor_{\mathcal{ST}}}{Razor_{\mathcal{PS}}}$ indicate preference for the Press-Schechter model. The fiducial model used is the Press-Schechter model.

Figure 4.2 is a plot of the log razor ratio for $\mathcal{M}_{\mathcal{ST}}$ versus $\mathcal{M}_{\mathcal{PS}}$ against the number of particles N. The log razor ratio for the Sheth-Tormen versus Press-Schechter model is given by

$$\ln \mathcal{R}_{01} = \ln Razor_{\mathcal{ST}} - \ln Razor_{\mathcal{PS}}. \tag{4.2.8}$$

If there is more merit assigned to the Press-Schechter model, then

$$\ln Razor_{\mathcal{PS}} > \ln Razor_{\mathcal{ST}}, \tag{4.2.9}$$

which is equivalent to $\ln \mathcal{R}_{01} < 0$. Using the Jeffreys' scale (table 3.1) to discriminate the models, the log razor ratio tends to -5 as N approaches 200 particles. This tells us that we would need more than 200 particles in order to very strongly favour $\mathcal{M}_{\mathcal{PS}}$ over $\mathcal{M}_{\mathcal{ST}}$. The Press-Schechter model is favoured as expected since we assumed it to be the fiducial model.

We now look at the outcome when considering $\mathcal{M}_{\mathcal{ST}}$ to be the fiducial model with typical values of $a_{true} = 0.7$ and $p_{true} = 0.3$. The fiducial values will only affect $D_{\mathcal{ST}}$ and not the prior, i.e., it will only be accounted for in the likelihood. We compute the razor as in (4.2.6) with $\boldsymbol{\theta}_{true} = (0.7, 0.3)$. The Kullback-Liebler distance measure for $\mathcal{M}_{\mathcal{PS}}$ is

$$D_{\mathcal{PS}} = N \left[f_d(\boldsymbol{\theta}_{true}) \ln \frac{f_d(\boldsymbol{\theta}_{true})}{f_{d\mathcal{PS}}} + \int_{\nu_d}^{\infty} p(\nu|\boldsymbol{\theta}_{true}) \ln \frac{p(\nu|\boldsymbol{\theta}_{true})}{p(\nu|\boldsymbol{\theta}, \mathcal{M}_{\mathcal{PS}})} d\nu \right], \tag{4.2.10}$$

where

$$p(\nu|\boldsymbol{\theta}, \mathcal{M}_{\mathcal{PS}}) = F_{\mathcal{PS}}(\nu), \tag{4.2.11}$$

and $f_{d\mathcal{PS}}$ is a constant given by

$$f_{d\mathcal{PS}} = \int_0^{\nu_d} d\nu \, F_{\mathcal{PS}}(\nu). \tag{4.2.12}$$

With a dust limit of $\nu_d = 0.1$, we find $f_{d\mathcal{PS}} = 0.248$. The fraction of dust for the fiducial model is also a constant since

$$f_d(\boldsymbol{\theta}_{true}) = \int_0^{\nu_d} d\nu \, \frac{A}{\nu} \sqrt{\frac{a\nu}{2\pi}} \left(1 + (a\nu)^{-p}\right) e^{-a\nu/2},$$

and setting $a = 0.7$ and $p = 0.3$ we find $f_d(\boldsymbol{\theta}_{true}) = 0.44262$. The model we use will influence the fraction of particles that are in dust, in this case we have more particles in dust when the fiducial model is Sheth-Tormen. The resulting Kullback-Liebler distance for $\mathcal{M}_{\mathcal{PS}}$ is

$$D_{\mathcal{PS}} = 0.09682.$$

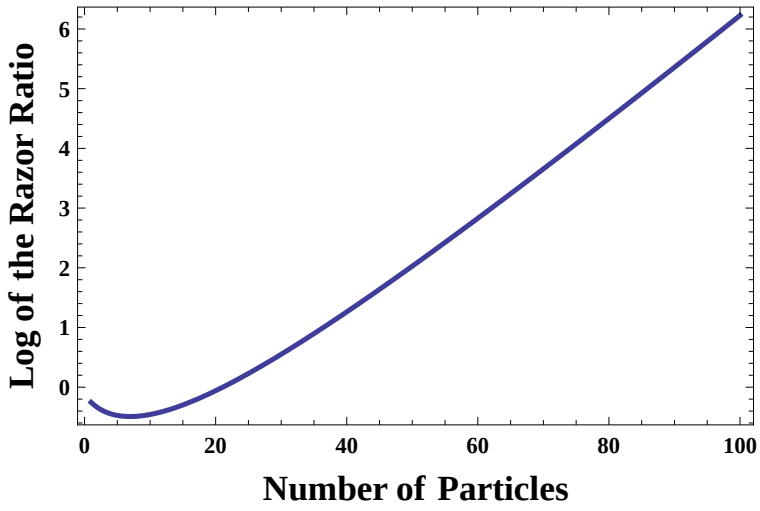

Figure 4.3: Plot of the log razor ratio ($\ln \frac{Razor_{\mathcal{ST}}}{Razor_{\mathcal{PS}}}$) for the Sheth-Tormen ($\mathcal{M}_{\mathcal{ST}}$) versus the Press-Schechter ($\mathcal{M}_{\mathcal{PS}}$) model against the number of particles observed. The fiducial model used is $\mathcal{M}_{\mathcal{ST}}$. Positive values indicate a preference for the Sheth-Tormen model over the Press-Schechter model with $\ln \frac{Razor_{\mathcal{ST}}}{Razor_{\mathcal{PS}}} \geq 5$ being a 'decisive' preference.

Figure 4.3 is a plot of the log razor ratio using $\mathcal{M}_{\mathcal{ST}}$ as the fiducial model. It is evident that the log razor ratio exceeds 5 when we have approximately more than 90 particles. The ratio is initially negative but is still not significant (refer to table 3.1) before turning over to positive values when $N > 20$ particles. The initial preference for $\mathcal{M}_{\mathcal{PS}}$ is due to the Occam's razor effect. In the presence of little data, the simpler model is preferred. The Press-Schechter model is situated in the same plane of the parameter space as the Sheth-Tormen model, since it is nested within $\mathcal{M}_{\mathcal{ST}}$. At low N, the likelihood is spread out over the parameter space for both models. The razor is proportional to the likelihood divided by the volume of the prior. The prior distribution volume for the simpler model, $\mathcal{M}_{\mathcal{PS}}$, is less than the prior volume for the more complex model, $\mathcal{M}_{\mathcal{ST}}$, since $\mathcal{M}_{\mathcal{ST}}$ has more free parameters. At low N, the razor will therefore favour the simpler model. At larger N however, the likelihood becomes more sharply peaked for both models, and the likelihood contribution from $\mathcal{M}_{\mathcal{PS}}$ becomes negligible as compared to the likelihood for $\mathcal{M}_{\mathcal{ST}}$. The increased likelihood space results in a higher razor for $\mathcal{M}_{\mathcal{ST}}$ than $\mathcal{M}_{\mathcal{PS}}$ at large N, hence favouring $\mathcal{M}_{\mathcal{ST}}$. It would be worth investigating the outcome if ν_d is varied, which we study now.

4.2.1 Changing the minimum mass limit

In this section we vary the dust limit, ν_d, and investigate the effect on the log razor ratio. We vary the dust limit relative to $\nu_{max} = 25$ and plot the log razor ratio for two cases, namely, when $\mathcal{M}_{\mathcal{PS}}$ and $\mathcal{M}_{\mathcal{ST}}$ are the fiducial models. We refer to the log razor ratio plot for different dust limits when $\mathcal{M}_{\mathcal{PS}}$ is the fiducial model, illustrated in figure 4.4. When the dust limit is 0.1% of ν_{max} the log razor ratio, $\ln \frac{Razor_{\mathcal{ST}}}{Razor_{\mathcal{PS}}}$, reaches a 'decisive' value of -5 in favour of $\mathcal{M}_{\mathcal{PS}}$ when the number of particles, N, is approximately 180. When the dust limit is increased to 1% of ν_{max}, we require 250 particles to reach a log razor ratio of -5. Finally, when ν_d is 10% of ν_{max} the number of particles required is 600.

Figure 4.5 illustrates the log razor ratio for different dust limits when $\mathcal{M}_{\mathcal{ST}}$ is the fiducial model. When ν_d is 0.1% and 1% of ν_{max}, we require 58 and 130 particles respectively, to reach a 'decisive' log razor ratio of 5 in favour of $\mathcal{M}_{\mathcal{ST}}$. When the dust limit is increased to 10% of ν_{max}, the log razor ratio is initially negative, indicating a slight preference for $\mathcal{M}_{\mathcal{PS}}$. The preference is not very strong since the log razor ratio reaches a minimum of -2.3 at $N = 200$, thereafter increasing to obtain a log razor ratio of 5 at $N \approx 2500$. We come to the conclusion that when the dust limit is increased, more particles are required in order to strongly discriminate between the two models. The reason this result occurs is because increasing ν_d implies that there are more particles in dust and hence fewer clusters. With less clusters in the simulation, discriminating the mass functions becomes inefficient.

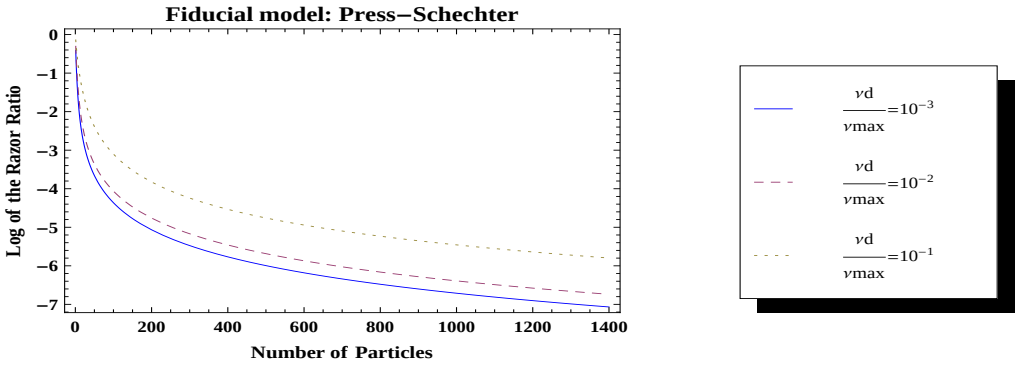

Figure 4.4: Variation of the log razor ratio with dust limit. The fiducial model is Press-Schechter. The dust limits considered are 0.1%, 1% and 10% of the maximum dust limit, $\nu_{max} = 25$.

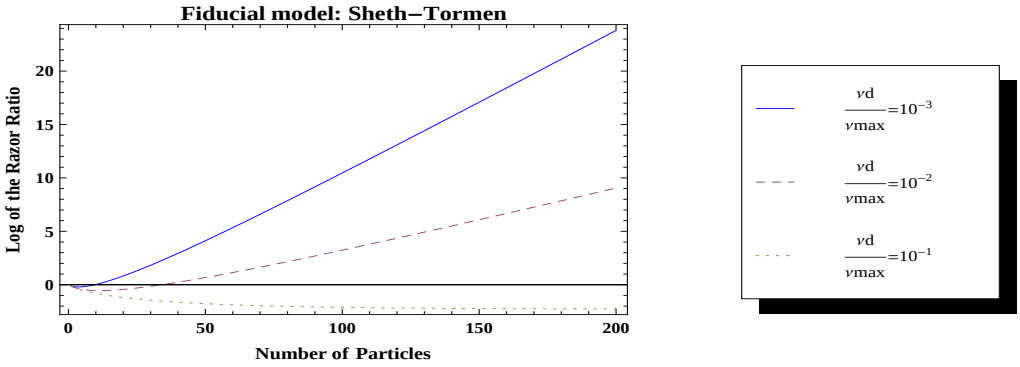

Figure 4.5: Variation of the log razor ratio with dust limit. The fiducial model is Sheth-Tormen. A log razor ratio of 5 indicates a decisive preference for the Sheth-Tormen model. When the dust limit is 10% of ν_{max}, the log razor ratio is initially negative, thereafter increasing to a value of 5 at approximately 2500 particles.

4.2.2 Flat prior vs Jeffreys' prior

In the previous sections we used the Jeffreys' prior for the Sheth-Tormen mass function. The Jeffreys' prior is generally computationally intensive to analyse due to the large parameter space. In the case of the Sheth-Tormen model, the Jeffreys' prior is dependent on two parameters but there are other models that are dependent on more than two parameters which we examine in 4.3. Computing the Jeffreys' prior for these high dimensional models can therefore take a long time depending on the computational resources available. Hence it is good to compare the results obtained thus far using the Jeffreys' prior with those of other prior distributions. We now investigate what the outcome would be if we use a flat prior instead. The flat prior is much easier and

less intensive to compute than the Jeffreys' prior since the flat prior occupies a constant volume that is only dependent on the range of each parameter. The razor for a model using a flat prior in general is

$$\begin{aligned} Razor(\mathcal{M}) &= \int d^n\theta p(\boldsymbol{\theta}|\mathcal{M})e^{-D} \\ &= \int d^n\theta \frac{1}{\Delta\theta_1\Delta\theta_2\cdots\Delta\theta_n}e^{-D}, \end{aligned} \tag{4.2.13}$$

so for the Sheth-Tormen mass function, the prior is given by

$$p(\boldsymbol{\theta}|\mathcal{M}) = \begin{cases} \frac{1}{(a_{max}-a_{min})(p_{max}-p_{min})} & \text{for } a_{min} < a \leq a_{max} \text{ and } p_{min} \leq p < p_{max}, \\ 0 & \text{otherwise.} \end{cases}$$

The log razor ratio plots are similar to those in the Jeffreys' prior case so we plot the case of both priors on the same set of axes for comparison. The log razor ratio plot we obtained when considering $\mathcal{M}_{\mathcal{PS}}$ as the fiducial model for both prior distributions is shown in figure 4.6. We consider the fiducial values $a = 1, p = 0$ when taking $\mathcal{M}_{\mathcal{PS}}$ to be the fiducial model. The Jeffreys' prior case is represented by the dashed line and the flat prior case is represented by the solid line. Using a different prior changes the number of particles required to decisively favour $\mathcal{M}_{\mathcal{PS}}$ over $\mathcal{M}_{\mathcal{ST}}$. The number of particles required is approximately 190 and 825 for the Jeffreys' prior and flat prior respectively, hence, using a flat prior distribution requires more particles to distinguish between the models when $\mathcal{M}_{\mathcal{PS}}$ is the fiducial model. The model that depends on the prior distribution is the Sheth-Tormen model since the Press-Schechter model is independent of free parameters. Changing the prior distribution will therefore only affect the razor for $\mathcal{M}_{\mathcal{ST}}$.

Since the log razor ratio requires more particles to favour $\mathcal{M}_{\mathcal{PS}}$ for the flat prior case, the razor for $\mathcal{M}_{\mathcal{ST}}$ using a flat prior distribution must be greater than the razor when using a Jeffreys' prior. We project the relation of the prior and likelihood onto the $p = 0$ plane to get an

understanding of why the razor, using a Jeffreys' prior, for $\mathcal{M}_{\mathcal{ST}}$ is lower than the razor using a flat prior. Figure (4.7) is the projection of the product of the prior and likelihood, which we call the integrand since it is integrated to yield the razor, onto the $p = 0$ plane. The solid curve is the integrand for the flat prior case and the dashed curve is the Jeffreys' prior case. From this plot, it is evident that the area for the flat prior case is greater than the area for the Jeffreys' prior case. This explains why the razor ratio is lower when using a Jeffreys' prior as compared to a flat prior.

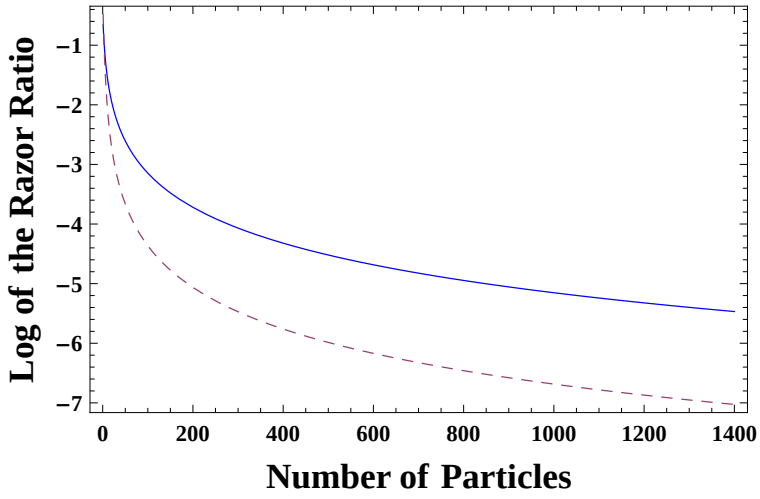

Figure 4.6: The log razor ratio between Press-Schechter and Sheth-Tormen as a function of the number of particles in a simulation. The log razor ratio for different priors assuming Press-Schechter as the fiducial model is illustrated in the plot. The solid line represents the flat prior and the dashed line represents the Jeffreys' prior. The dust limit used is $\nu_d = 0.1$.

The razor favours a given model more when the model likelihood uses as much of the prior volume of the model. The more informative the prior, the more volume of the prior is used by the likelihood and the greater the razor. We plot the likelihood against each prior individually, projected onto the $p = 0$ plane, in figure (4.8). The plot in the left panel is the Jeffreys' prior and the right panel is the flat prior. The prior distributions are shown by the dashed lines. By looking at the shaded region we notice that the likelihood uses more of the flat prior space than

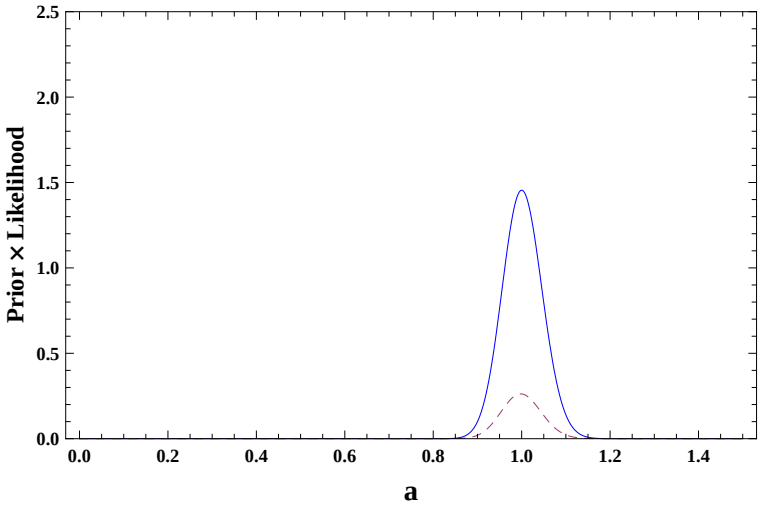

Figure 4.7: The product of the likelihood and prior for the Sheth-Tormen model, $\mathcal{M}_{\mathcal{ST}}$. We call this quantity the integrand of $\mathcal{M}_{\mathcal{ST}}$, since it is integrated to yield the razor for $\mathcal{M}_{\mathcal{ST}}$, hence this quantity is proportional to the razor for $\mathcal{M}_{\mathcal{ST}}$. This model depends on two parameters, namely a and p. The Jeffreys' prior (dashed line) and flat prior (solid line) are projected onto the $p = 0$ plane. The plot depicts the behaviour of the integrand over the parameter space for a. For illustrative purposes, we considered 1000 particles.

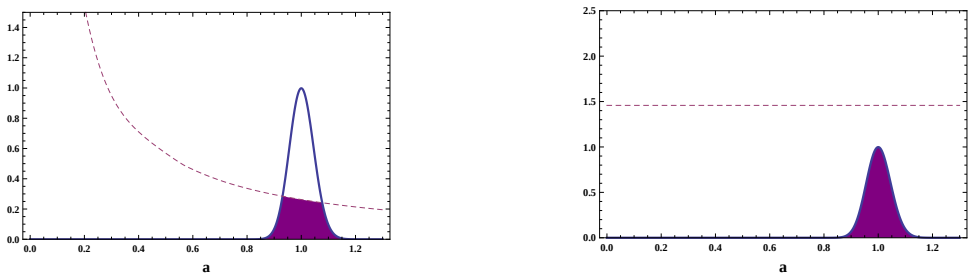

Figure 4.8: The prior distributions (dashed line) against the likelihood (solid line) projected onto the $p = 0$ plane, for a fixed number of particles. The Jeffreys' prior is illustrated in the left panel with the flat prior in the right. The shaded region indicates the area of the prior that is used by the likelihood.

the Jeffreys' prior space. This results in a higher razor for the flat prior case.

In figure 4.9 we illustrate the the outcome of the log razor ratio when $\mathcal{M}_{\mathcal{ST}}$ is the fiducial model. When taking $\mathcal{M}_{\mathcal{ST}}$ as the fiducial model, the razor ratio for the Jeffreys' prior is less than the razor ratio for the flat prior case for the same reason as just discussed. The difference is that the razor for $\mathcal{M}_{\mathcal{ST}}$ is greater than the razor for $\mathcal{M}_{\mathcal{PS}}$, therefore $\ln \frac{Razor_{\mathcal{M}_{\mathcal{ST}}}}{Razor_{\mathcal{M}_{\mathcal{PS}}}} > 0$. In the flat prior case, we would require slightly fewer particles to distinguish between models as compared to the

Jeffreys' prior. The behaviour at low N of a negative log razor ratio is due to the Occam's razor effect as discussed in chapter 4.2.

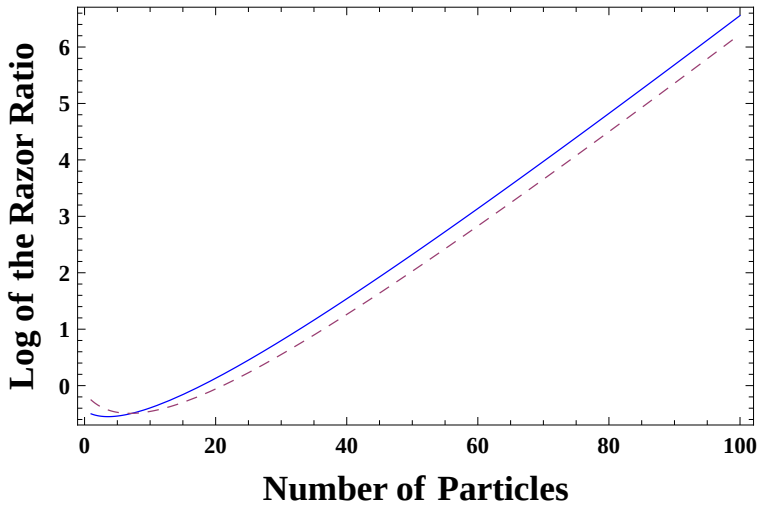

Figure 4.9: Log razor ratio using a flat prior (solid line) and Jeffreys' prior (dashed line) as a function of the number of particles, N. The fiducial model is Sheth-Tormen. We would require slightly more particles to favour $\mathcal{M}_{ST}$ when using the Jeffreys' prior as compared to the flat prior distribution. The Occam's razor effect is evident at low N for both prior cases, where the simpler model, $\mathcal{M}_{PS}$ is slightly favoured.

4.3 Normalisable Tinker mass function against the Sheth-Tormen mass function

We now examine the normalisable Tinker (NT) mass function, as discussed in chapter 2.5, compared to the Sheth-Tormen mass function. The Sheth-Tormen model is nested within the NT model by the following result that we establish now.

Consider the generalised Sheth Tormen mass function proposed by R. Sheth (private communication, 2009),

$$F(\nu) = \frac{A}{\nu}\sqrt{\frac{a\nu}{2\pi}}\left[c(a\nu)^{-t} + (a\nu)^{-b}\right]e^{-a\nu/2} \tag{4.3.1}$$

that is dependent on four parameters a, b, c and t. The Sheth-Tormen mass function [27] is nested within this model since it can be obtained by setting $c = 1$ and $b = 0$. The parameter A is derived from the constraint

$$\int_0^\infty F(\nu) = 1. \tag{4.3.2}$$

The model that we wish to compare to $F(\nu)$ is the NT mass function given by (2.5.14) in chapter 2, which we rewrite here for easier reference.

$$T(\nu) = \frac{B}{2\nu}\left[\left(\frac{\delta_c}{e\sqrt{\nu}}\right)^{-d} + \left(\frac{\delta_c}{\sqrt{\nu}}\right)^{-h}\right] e^{-g\nu/\delta_c^2}. \tag{4.3.3}$$

From the above form of $T(\nu)$, it is not clear as to whether it can be obtained from a transformation of $F(\nu)$ so we write it in a form that is more clear:

$$\begin{aligned} T(\nu) &= \frac{\sqrt{\nu}}{\nu}\frac{B}{2}\left[\left(\frac{\delta_c}{e\sqrt{\nu}}\right)^{-d}\nu^{-1/2} + \left(\frac{\delta_c}{\sqrt{v}}\right)^{-h}\nu^{-1/2}\right] e^{-g\nu/\delta_c^2} \\ &= \frac{\sqrt{\nu}}{\nu}\left[\frac{B}{2}\left(\frac{\delta_c}{e}\right)^{-d}\nu^{-(1-d)/2} + \frac{B}{2}\delta_c^{-h}\nu^{-(1-h)/2}\right] e^{-g\nu/\delta_c^2}. \end{aligned} \tag{4.3.4}$$

We also write $F(\nu)$ in an alternate form that will enable us to compare the parameters of the two models:

$$\begin{aligned} F(\nu) &= \frac{\sqrt{\nu}}{\nu}\left[A\sqrt{\frac{a}{2\pi}}c(a\nu)^{-t} + A\sqrt{\frac{a}{2\pi}}(a\nu)^{-b}\right] e^{-a\nu/2} \\ &= \frac{\sqrt{\nu}}{\nu}\left[A\frac{ca^{1/2-t}}{\sqrt{2\pi}}\nu^{-t} + A\frac{a^{1/2-b}}{\sqrt{2\pi}}\nu^{-b}\right] e^{-a\nu/2}. \end{aligned} \tag{4.3.5}$$

By comparing the coefficients of ν we obtain the relation between the parameters of the generalised Sheth-Tormen and NT mass functions as follows,

$$a = \frac{2g}{\delta_c^2}, \tag{4.3.6}$$

$$t = (1 - d)/2, \tag{4.3.7}$$

$$b = (1 - h)/2. \tag{4.3.8}$$

The transformation of c, however, is not clear and we show the derivation below for reference. Comparing (4.3.4) and (4.3.5) we obtain

$$\frac{B}{2}\left(\frac{\delta_c}{e}\right)^{-d} = Ac\frac{a^{1/2-t}}{\sqrt{2\pi}}, \tag{4.3.9}$$

$$\frac{B}{2}\delta_c^{-h} = A\frac{a^{1/2-b}}{\sqrt{2\pi}}. \tag{4.3.10}$$

Dividing equation (4.3.9) by equation (4.3.10) and solving for c yields,

$$\frac{e^d}{\delta_c^{d-h}} = ca^{b-t}, \tag{4.3.11}$$

$$c = a^{t-b}\left(\frac{e^d}{\delta_c^{d-h}}\right). \tag{4.3.12}$$

We want c in terms of the parameters in the NT mass function so using (4.3.7) and (4.3.8) we have

$$t - b = \frac{1}{2}(h - d), \tag{4.3.13}$$

and substituting (4.3.6) and (4.3.13) into (4.3.12), we finally obtain

$$c = \left(\frac{2g}{\delta_c^2}\right)^{(h-d)/2}\left(\frac{e^d}{\delta_c^{d-h}}\right). \tag{4.3.14}$$

Hence the NT mass function is equivalent to the generalised Sheth-Tormen (ST) mass function. The fiducial values given in appendix C of [28] are listed in table 4.1 for $\Delta = 200$ (where Δ is defined such that the mass, M_Δ, is the amount of matter contained within a radius r_Δ in which the density of that enclosed region is $\Delta \times \rho_{crit}$). Using (4.3.6), (4.3.7), (4.3.8), (4.3.14) and table 4.1, we can now obtain the fiducial parameter values of the NT mass function in terms of the

Table 4.1: Fiducial values

Δ	NT mass function				generalised ST mass function			
	d	e	g	h	a	b	c	t
200	1.97	1	1.228	0.51	0.86	0.245	0.519	-0.485

parameters of the generalised Sheth-Tormen model as

$$a = 2(1.228)/1.69^2 = 0.86 > 0 \tag{4.3.15}$$

$$t = (1 - 1.97)/2 = -0.485 < 1/2 \tag{4.3.16}$$

$$b = (1 - 0.51)/2 = 0.245 < 1/2 \tag{4.3.17}$$

$$c = \left(\frac{2(1.228)}{1.69^2}\right)^{(0.51-1.97)/2} \left(\frac{1}{1.69^{1.97-0.51}}\right) = 0.519. \tag{4.3.18}$$

These parameter values satisfy the boundary conditions of the generalised Sheth-Tormen model. For completeness we also include the inverse transformation i.e., the transformation of the parameters of the NT mass function to the generalised Sheth-Tormen mass function, as follows:

$$\begin{aligned} h &= 1 - 2b \\ d &= 1 - 2t \\ g &= \frac{a\delta_c^2}{2} \\ e &= \exp\left[\frac{\ln\left(c\delta_c^{2(b-t)a^{b-t}}\right)}{1-2t}\right]. \end{aligned}$$

Now that we have a more convenient functional form of the NT mass function, we can compare it to the Sheth-Tormen mass function (with two free parameters a and p) and discriminate appropriately using the log razor ratio. Using the same approach as before, we compute the log

razor ratio for the Jeffreys' prior as well as a flat prior. The Jeffreys' prior, as well as the likelihood, for the NT mass function is much more computationally intensive to compute (compared to the Sheth-Tormen mass function) due to the extra two parameters.

We use the log razor ratio as before to discriminate between the NT and Sheth-Tormen model. We have already derived the prior distribution and likelihood for the Sheth-Tormen model in section 4.1. In order to compute the razor for the NT model, we begin by deriving the Fisher information matrix that is used in the Jeffreys' prior. Using (4.2.1), with $\boldsymbol{\theta} = (a, b, c, t)$ and $p(\nu|\boldsymbol{\theta}, \mathcal{M}) = T(\nu)$, we have,

$$J_{ij}(\boldsymbol{\theta}, \mathcal{M}) = \begin{bmatrix} J_{aa} & J_{ab} & J_{ac} & J_{at} \\ J_{ab} & J_{bb} & J_{bc} & J_{bt} \\ J_{ac} & J_{bc} & J_{cc} & J_{ct} \\ J_{at} & J_{bt} & J_{ct} & J_{tt} \end{bmatrix},$$

since the Fisher information matrix is symmetric about the diagonal, i.e., $J_{ij} = J_{ji}$. The Jeffreys' prior distribution for the NT model is then given by (4.2.3). For the flat prior distribution, we use the functional form

$$p(\boldsymbol{\theta}|\mathcal{M}_{NT}) = \frac{1}{\Delta a \Delta b \Delta c \Delta t} \tag{4.3.19}$$

where $\mathcal{M}_{NT}$ denotes the NT model. The likelihood is determined by (3.7.1), with the Kullback-Liebler distance given in (4.2.5), hence the likelihood is now a function of four parameters $\boldsymbol{\theta} = (a, b, c, t)$. We look at the results for two cases, one in which the simpler model (Sheth-Tormen) is the fiducial model, and the other in which the NT mass function is the fiducial model.

If we consider the case when the Sheth-Tormen mass function is the fiducial model, the num-

ber of particles that is needed to strongly favour the Sheth-Tormen mass function is considerably more, specifically $N > 20000$ when competing against the NT mass function (refer to figure (4.10)) as compared to the Press-Schechter versus Sheth-Tormen case which required approximately 80 particles (refer to figure (4.9)).

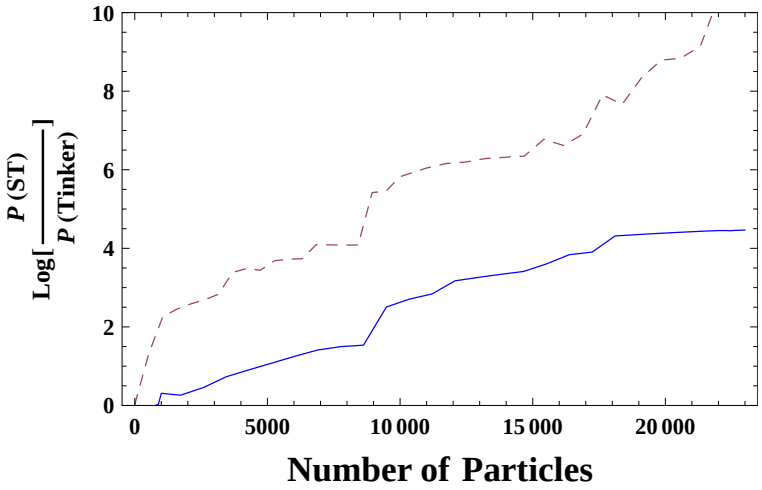

Figure 4.10: The log razor ratio for the Sheth-Tormen mass function against the NT mass function when $\mathcal{M}_{\mathcal{ST}}$ is the fiducial model. The dashed line represents the Jeffreys' prior case and the solid line is for a flat prior.

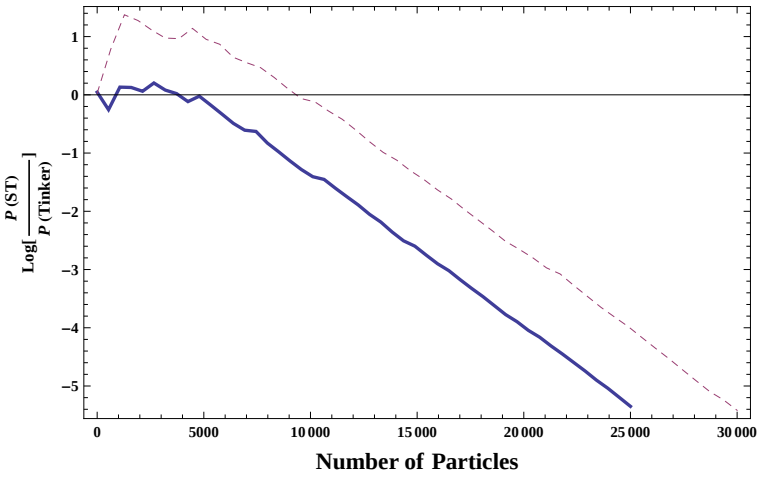

Figure 4.11: Plot of the log razor ratio for the NT mass function versus the Sheth-Tormen mass function. The fiducial model used is the NT mass function. We consider two prior cases, namely, the Jeffreys' prior (shown by the dashed line) and the flat prior (solid line). A log razor ratio of -5 indicates a 'decisive' preference for the NT mass function (refer to table 3.1).

The second case we look at is when the fiducial model is the NT mass function (refer to figure 4.11). In this case the more complicated model i.e., the model with more parameters and hence more degrees of freedom, is favoured. However this only occurs at large N, i.e., $N > 25000$. At low N, the simpler model is favoured, but not enough to strongly rule out the NT mass function. This slight preference for the simpler model at low N is mainly due to the Occam's razor effect as discussed in chapter 4.2. At low N, the extra parameters in the NT mass function will down-weight the razor which penalises the more complicated model for the prior space that is not utilised. The prior distribution significantly affects the number of particles required to decisively prefer one model over another. In this case, the razor for the NT model is stronger when using a flat prior, compared to the razor when using the Jeffreys' prior distribution. This occurs for the same reason as discussed in chapter 4.2.2, where in this case, the likelihood of the NT model uses more volume of the flat prior than the Jeffreys' prior distribution.

When we use more complex models the number of particles required to discriminate between models increases. The more free parameters in a model allow the model to better fit observational data, therefore two models that have many free parameters offer a reasonable fit. In order

to distinguish between these models, we would require more data. We have seen that N can range from 80 (Sheth-Tormen vs Press-Schechter) to more than 25 000 (NT vs Sheth-Tormen). These numbers are small when compared to large N-body simulations which contain more than a billion particles [21], however, this project considers more idealised conditions and focuses on the effectiveness of the Bayesian methodology for distinguishing between cosmological mass functions. Factors that would result in an increased number of particles being required are an increased dust limit and inclusion of redshift dependence in the cluster mass function. More realistic modelling of these factors will result in a more practical size of simulation required for model selection. Moreover, we have not included the uncertainty due to baryonic physics which is expected to affect our results. The inclusion of baryonic physics is expected to increase the uncertainty in the mass function by more than 10% [22]. This will result in a larger number of particles being required and hence, increase the size of the simulation. We have also assumed the cluster mass function to be independent of redshift. For simulations, we have the mass function at $z = 0$ everywhere, however for real data, redshift evolution must be taken into account. The dependence on redshift will be incorporated in the linear growth factor, $D(z)$, which will describe the evolution of the cosmological mass function with redshift. Our approach leaves room for incorporating more realistic effects but the methodology used thus far is promising in distinguishing between different cosmological mass functions.

CHAPTER 5

Conclusion

The first part of this thesis reviewed the fundamental theory of cosmology, specifically structure in the universe and the cosmological model. The theory is based on the cosmological principle, which refers to the universe being homogeneous and isotropic on large scales. We argued that knowledge of the formation and properties of galaxy clusters can provide important information in cosmology, and, specifically, constrain the nature of dark energy. Galaxy clusters have long time scales and the demographics vary with redshift. Therefore, they can be used to constrain cosmological parameters such as the normalisation of density perturbations, the density of the universe and the dark energy equation of state. The mass function of galaxy clusters is an important ingredient for probing cosmology and is based on the statistics of large-scale structure. We introduced three nested models of the mass function, namely the Press-Schechter ($\mathcal{M}_{\mathcal{PS}}$), Sheth-Tormen ($\mathcal{M}_{\mathcal{ST}}$) and normalisable Tinker (NT) models.

Chapter 3 reviewed the theory of probability, placing an emphasis on Bayes' theorem and

the probability of the Bayesian evidence for assigning merit to a model. In light of data and using the odds ratio, the preference of one model over another was quantified. The Jeffreys' scale was used as a guideline with $\ln |B_{01}| \geq 5$ indicating a 'decisive' preference. This methodology was applied to two nested models using mock data to illustrate its effectiveness. A part of this chapter investigated the role of the Occam's razor in penalising a single parameter model for the wasted parameter space of the prior distribution. It was discovered that even though a more complicated model provides a better fit to the data, the Bayesian statistics approach incorporates a natural way of penalising a more complicated model if the prior space is not utilised by the likelihood. The theory behind the Fisher information matrix to describe the likelihood around its maximum likelihood estimators was examined. The Fisher information matrix may also be used as a lower bound for the uncertainty on the parameter estimates. This bound holds irrespective of the method used to obtain those estimates, hence it is a good forecast of the minimum uncertainty of the estimates. Finally the role of the Bayesian razor in model selection was explained. The razor uses the Fisher information matrix as a component of its prior distribution, specifically for the Jeffreys' prior.

In chapter 4 the Bayesian razor was applied to different models of the mass function, which involved assigning a probability distribution to the respective mass function. It was determined that the log razor ratio, which was used to determine which mass function would be preferred over another, varies with the number of particles in a simulation. The razor ratio always favoured the model that is used as the fiducial model for a large enough number of particles. One could ask why bother, given that the fiducial model is always the favoured model? The answer would be that these techniques allow one to decide, given a simulation, whether it is worth bothering with a more complicated mass function. The first two models that were compared were the Press-Schechter and Sheth-Tormen mass function. It was discovered that less particles were needed

to strongly favour one model over another when the fiducial model chosen is Sheth-Tormen, as compared to when Press-Schechter was the fiducial model. Increasing the minimum cluster mass limit resulted in more particles being required to discriminate models. This was due to the lack of low mass clusters in the data making the log razor ratio dependent on more particles in high mass clusters to discriminate mass functions.

The razor ratio for two different prior distributions, namely a flat prior and the Jeffreys' prior, was compared. The razor was found to be proportional to the volume of the prior that is used by the likelihood, i.e., the more volume used by the likelihood, the more informative the prior. The razor ratio for the NT model against the Sheth-Tormen model was analysed therafter, and from our analysis it was found that the likelihood of the Sheth-Tormen and NT models used more volume of their respective flat prior distributions as compared to the Jeffreys' prior. There was a significant increase in the number of particles required as compared to model selection with simpler models. Models with more free parameters offer a better fit to simulations than simpler models, hence comparing models with many free parameters will require more data in order to decisively discriminate between these models.

Calculating the razor for higher dimensional models required more time since it is more computationally intensive than the simpler model case. There are possible numerical techniques that could make the computation more efficient, which will be studied in more detail in the near future. Other possible issues to study in the future are incorporating redshift dependence into the mass function for real data and calibrating the minimum cluster mass limit according to simulations. The results in this work did not include the uncertainty due to baryonic physics. The mass function is exponentially sensitive to errors in the calibration of the mass-observable relationship and this relationship requires precise baryonic physics. This is expected to increase the

uncertainty in our result by 10% [22]. In order to incorporate this uncertainty, we shall consider using a Gaussian distribution to represent the uncertainty and convolving this distribution with our original probability distribution for the mass function. Future observational data will probe the number of clusters in a particular survey. Therefore it would be more realistic to relate the number of particles to the number of clusters, an investigation that we will undertake in future work.

APPENDIX A

The Cramer-Rao inequality: proof for simple case

We denote the elements of the Fisher information matrix by

$$F_{\theta_i\theta_j} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle, \tag{A.0.1}$$

where $L = L(x, \boldsymbol{\theta})$ is the likelihood of a particular distribution that is dependent on some data denoted by x, and a set of parameters denoted by $\boldsymbol{\theta}$. We use the Schwarz inequality to prove the Cramer-Rao inequality. Suppose we have a distribution consisting of two parameters with estimators θ_1 and θ_2, then by taking the expectation with respect to the likelihood, we have

$$\left\langle \left(\Delta\hat{\theta}_1 + \lambda\Delta\hat{\theta}_2\right)^2 \right\rangle \geq 0, \tag{A.0.2}$$

where $\Delta\hat{\theta}_\alpha \equiv \hat{\theta}_\alpha - \theta_{0\alpha}$ is the difference between the estimate and the true value. Evidently equation (A.0.2) holds for any λ. It is a minimum if we choose $\lambda = -\frac{\left\langle \Delta\hat{\theta}_1 \Delta\hat{\theta}_2 \right\rangle}{\left\langle (\Delta\hat{\theta}_2)^2 \right\rangle}$ from which we obtain the Schwarz inequality:

$$\left\langle \left(\Delta\hat{\theta}_1\right)^2 \right\rangle \left\langle \left(\Delta\hat{\theta}_2\right)^2 \right\rangle \geq \left\langle \Delta\hat{\theta}_1 \Delta\hat{\theta}_2 \right\rangle^2. \tag{A.0.3}$$

We treat the simple case of one parameter, and the data x. For an unbiased estimator, $\hat{\theta}$, of a parameter whose true value is θ_0,

$$\left\langle \hat{\theta} - \theta_0 \right\rangle = \int (\hat{\theta} - \theta_0) L(x, \theta) dx = 0. \tag{A.0.4}$$

Differentiating with respect to θ we obtain,

$$\int (\hat{\theta} - \theta_0) \frac{\partial L}{\partial \theta} dx - \int L(x, \theta) dx = 0. \tag{A.0.5}$$

The last integral is equal to unity, therefore we write

$$\left\langle (\hat{\theta} - \theta_0) \frac{\partial \ln L}{\partial \theta} \right\rangle = 1, \tag{A.0.6}$$

and the Schwarz inequality yields

$$\left\langle (\hat{\theta} - \theta_0)^2 \right\rangle \geq \frac{1}{\left\langle (\frac{\partial \ln L}{\partial \theta})^2 \right\rangle}. \tag{A.0.7}$$

We use the following result to obtain the Cramer-Rao inequality:

$$\begin{aligned} \int \frac{\partial \ln L}{\partial \theta} L \, dx &= \int \frac{1}{L} \frac{\partial L}{\partial \theta} L \, dx \\ &= \int \frac{\partial L}{\partial \theta} dx. \end{aligned} \tag{A.0.8}$$

We differentiate $\int L(x, \theta) dx = 1$ twice with respect to θ to obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int \frac{\partial L}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int \frac{\partial \ln L}{\partial \theta} L \, dx \\ &= \int \frac{\partial^2 \ln L}{\partial \theta^2} L \, dx + \int \left(\frac{\partial \ln L}{\partial \theta} \right)^2 L \, dx \\ &= \int \left[\frac{\partial^2 \ln L}{\partial \theta^2} + \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] L \, dx, \end{aligned} \tag{A.0.9}$$

where the second and third step in (A.0.9) follows from (A.0.8). This result implies

$$\left\langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right\rangle = - \left\langle \frac{\partial^2 \ln L}{\partial \theta^2} \right\rangle , \tag{A.0.10}$$

therefore substituting (A.0.10) into (A.0.7), we obtain the Cramer-Rao inequality given by

$$\left\langle (\hat{\theta} - \theta_0)^2 \right\rangle \geq - \frac{1}{\left\langle \frac{\partial^2 \ln L}{\partial \theta^2} \right\rangle} = F_{\theta\theta}^{-1}. \tag{A.0.11}$$

APPENDIX B

Fisher information matrix for the cosmological mass function

We assume that not all particles of mass are found in clusters. Those that are not in clusters are said to be in dust. We have some lower mass limit, ν_d, such that there exist no clusters with mass $\nu < \nu_d$. The probability distribution of a model, $\mathcal{M}$, of a mass function including dust, is given by

$$p(\vec{\nu}|\boldsymbol{\theta}, \mathcal{M}) = f_d(\boldsymbol{\theta}, \mathcal{M})^{N_d} \prod_{k=1}^{N_c} p(\nu_k|\boldsymbol{\theta}, \mathcal{M}), \tag{B.0.1}$$

where

$$f_d(\boldsymbol{\theta}, \mathcal{M}) = \int_0^{\nu_d} d\nu \, p(\nu|\boldsymbol{\theta}, \mathcal{M}), \tag{B.0.2}$$

is the fraction of particles in dust and N_d and N_c is the number of particles in dust and clusters in a given simulation. These are constrained by $N_d + N_c = N$. The Fisher information matrix is defined as

$$J_{ij} = \left\langle -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln p(\vec{\nu}|\boldsymbol{\theta}, \mathcal{M}) \right\rangle, \tag{B.0.3}$$

where $\langle\rangle$ is the expectation with respect to $p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})$. Substituting (B.0.1) into (B.0.3) we get

$$\begin{aligned} J_{ij} &= \left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(N_d \ln f_d(\boldsymbol{\theta},\mathcal{M}) + \sum_{k=1}^{N_c}\ln p(\nu_k|\boldsymbol{\theta},\mathcal{M})\right)\right\rangle \\ &= \left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j} N_d \ln f_d(\boldsymbol{\theta},\mathcal{M})\right\rangle + \left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{k=1}^{N_c}\ln p(\nu_k|\boldsymbol{\theta},\mathcal{M})\right\rangle. \end{aligned} \tag{B.0.4}$$

We will study each term individually. Looking at the first term, we see that N_d does not depend on $\boldsymbol{\theta}$ and can be taken outside of the derivative i.e.,

$$\left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j} N_d \ln f_d(\boldsymbol{\theta},\mathcal{M})\right\rangle = \left\langle N_d \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_d(\boldsymbol{\theta},\mathcal{M})\right\rangle. \tag{B.0.5}$$

The expectation of N_d however depends on $\boldsymbol{\theta}$. Since a particle is either in dust or not, N_d follows a binomial distribution with probability of being in dust given by $f_d(\boldsymbol{\theta},\mathcal{M})$. Hence,

$$\langle N_d\rangle = \sum_{N_d=0}^{N} p(N_d)N_d = N f_d(\boldsymbol{\theta},\mathcal{M}). \tag{B.0.6}$$

Therefore (B.0.5) becomes

$$\left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j} N_d \ln f_d(\boldsymbol{\theta},\mathcal{M})\right\rangle = N f_d \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_d(\boldsymbol{\theta},\mathcal{M}), \tag{B.0.7}$$

since $f_d(\boldsymbol{\theta},\mathcal{M})$ is independent of $\vec{\nu}$. The second term in (B.0.4) is slightly more complicated as it has both a discrete and continuous random variable i.e. N_c and $\vec{\nu}$. We take the expectation over both of these variables so that

$$\left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{k=1}^{N_c}\ln p(\nu_k|\boldsymbol{\theta},\mathcal{M})\right\rangle = \sum_{N_c=0}^{N} p(N_c)\sum_{k=1}^{N_c}\int_{\nu_d}^{\infty} d\nu_k \frac{p(\nu_k|\boldsymbol{\theta})}{f_c}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\nu_k|\boldsymbol{\theta}), \tag{B.0.8}$$

where $f_c(\boldsymbol{\theta},\mathcal{M}) = 1 - f_d(\boldsymbol{\theta},\mathcal{M})$ is the fraction of particles that are in dust and serves as a normalising factor for the measure in the expectation integral. Since each variable ν_k is independent

and identically distributed (B.0.8) becomes

$$\begin{aligned}\left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{k=1}^{N_c}\ln p(\nu_k|\boldsymbol{\theta},\mathcal{M})\right\rangle &= \frac{1}{f_c(\boldsymbol{\theta},\mathcal{M})}\sum_{N_c=0}^{N}p(N_c)\sum_{k=1}^{N_c}\int_{\nu_d}^{\infty}d\nu\,p(\nu|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\nu|\boldsymbol{\theta}) \\ &= \frac{1}{f_c(\boldsymbol{\theta},\mathcal{M})}\sum_{N_c=0}^{N}p(N_c)N_c\int_{\nu_d}^{\infty}d\nu\,p(\nu|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\nu|\boldsymbol{\theta}).\end{aligned} \tag{B.0.9}$$

Using a similar argument for $\langle N_d\rangle$, the expectation for $\langle N_c\rangle = \sum_{N_c=0}^{N}p(N_c)N_c = Nf_c(\boldsymbol{\theta},\mathcal{M})$. Equation (B.0.9) is then just

$$\left\langle \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{k=1}^{N_c}\ln p(\nu_k|\boldsymbol{\theta},\mathcal{M})\right\rangle = N\int_{\nu_d}^{\infty}d\nu\,p(\nu|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\nu|\boldsymbol{\theta}). \tag{B.0.10}$$

Finally combining (B.0.7) and (B.0.10) we get

$$J_{ij}(\boldsymbol{\theta},\mathcal{M}) = N\left[f_d\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_d(\boldsymbol{\theta},\mathcal{M}) + \int_{\nu_d}^{\infty}d\nu\,p(\nu|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\nu|\boldsymbol{\theta})\right]. \tag{B.0.11}$$

APPENDIX C

The Kullback-Liebler distance for the cosmological mass function

The Kullback-Liebler distance between two probability distributions, say $a(\vec{\nu})$ and $b(\vec{\nu})$ is given by

$$D(a||b) = \int d\vec{\nu}\, a \ln\left(\frac{a}{b}\right), \tag{C.0.1}$$

so if we take our set of assumed true parameter values to be $\mathbf{A}$, and our arbitrary model $\mathcal{M}$ with parameter values to be $\boldsymbol{\theta}$, the Kullback-Liebler distance measure becomes

$$D\left(p(\vec{\nu}|\mathbf{A},\mathcal{M})||p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})\right) = \int d\vec{\nu}\, p(\vec{\nu}|\mathbf{A},\mathcal{M}) \ln\left[\frac{p(\vec{\nu}|\mathbf{A},\mathcal{M})}{p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})}\right].$$

This is just the expectation of $\ln\left[\frac{p(\vec{\nu}|\mathbf{A},\mathcal{M})}{p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})}\right]$ with respect to $p(\vec{\nu}|\mathbf{A},\mathcal{M})$. Hence,

$$\begin{aligned} D\left(p(\vec{\nu}|\mathbf{A},\mathcal{M})||p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})\right) &= \left\langle \ln\left[\frac{p(\vec{\nu}|\mathbf{A},\mathcal{M})}{p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})}\right]\right\rangle \\ &= \langle \ln p(\vec{\nu}|\mathbf{A},\mathcal{M})\rangle - \langle \ln p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M})\rangle . \end{aligned} \tag{C.0.2}$$

Recall that the probability distribution for the mass function is given by

$$p(\vec{\nu}|\boldsymbol{\theta},\mathcal{M}) = f_d(\boldsymbol{\theta},\mathcal{M})^{N_d}\prod_{k=1}^{N_c} p(\nu_k|\boldsymbol{\theta},\mathcal{M}), \tag{C.0.3}$$

where $f_d(\boldsymbol{\theta},\mathcal{M})$ is defined in (B.0.2) in appendix B. The discrete random variables, N_c and N_d, are the number of particles in clusters and dust respectively. These random variables follow a binomial distribution. Henceforth we drop $\mathcal{M}$ since the probability distribution holds for a particular model. Substituting (C.0.3) in (C.0.2) we obtain

$$\begin{aligned} D\left(p(\vec{\nu}|\mathbf{A})||p(\vec{\nu}|\boldsymbol{\theta})\right) &= \left\langle N_d \ln f_d(\mathbf{A}) + \sum_{k=1}^{N_c} \ln p(\nu_k|\mathbf{A}) \right\rangle - \left\langle N_d \ln f_d(\boldsymbol{\theta}) + \sum_{k=1}^{N_c} \ln p(\nu_k|\boldsymbol{\theta}) \right\rangle \\ &= \langle N_d \ln f_d(\mathbf{A}) - N_d \ln f_d(\boldsymbol{\theta})\rangle + \left\langle \sum_{k=1}^{N_c} \left(\ln p(\nu_k|\mathbf{A}) - \ln p(\nu_k|\boldsymbol{\theta})\right) \right\rangle \\ &= \left\langle N_d \ln \frac{f_d(\mathbf{A})}{f_d(\boldsymbol{\theta})} \right\rangle + \left\langle \sum_{k=1}^{N_c} \ln \frac{p(\nu_k|\mathbf{A})}{p(\nu_k|\boldsymbol{\theta})} \right\rangle \equiv I_1 + I_2. \end{aligned} \tag{C.0.4}$$

Let us study the first integral, I_1. From (B.0.2) in appendix B, the factor $\ln \frac{f_d(\mathbf{A})}{f_d(\boldsymbol{\theta})}$ in the expectation is independent of ν. The expectation of N_d however, is given by

$$\langle N_d \rangle = \sum_{N_d=0}^{N} p(N_d) N_d = f_d(\mathbf{A}) N, \tag{C.0.5}$$

where we are taking the expectation with respect to the distribution of N_d. Hence I_1 becomes

$$I_1 = N f_d(\mathbf{A}) \ln \frac{f_d(\mathbf{A})}{f_d(\boldsymbol{\theta})}. \tag{C.0.6}$$

The second integral is given by

$$I_2 = \left\langle \sum_{k=1}^{N_c} \ln \frac{p(\nu_k|\mathbf{A})}{p(\nu_k|\boldsymbol{\theta})} \right\rangle. \tag{C.0.7}$$

We take the expectation over N_c and ν since they are both random variables. N_c is a discrete random variable and ν is a continuous random variable. We therefore have the the following result,

$$I_2 = \sum_{k=1}^{N_c} p(N_c) \sum_{k=1}^{N_c} \int_{\nu_d}^{\infty} d\nu_k \, \frac{p(\nu_k|\mathbf{A})}{f_c(\mathbf{A})} \ln \frac{p(\nu_k|\mathbf{A})}{p(\nu_k|\boldsymbol{\theta})}, \tag{C.0.8}$$

where $f_c(\mathbf{A}) = 1 - f_d(\mathbf{A})$ serves as a normalising factor for the measure in the expectation integral. Since each ν_k is independent and identically distributed we have

$$\begin{aligned} I_2 &= \sum_{k=1}^{N_c} p(N_c) \frac{1}{f_c(\mathbf{A})} N_c \int_{\nu_d}^{\infty} d\nu \, p(\nu|\mathbf{A}) \ln \frac{p(\nu|\mathbf{A})}{p(\nu|\boldsymbol{\theta})} \\ &= N \int_{\nu_d}^{\infty} d\nu \, p(\nu|\mathbf{A}) \ln \frac{p(\nu|\mathbf{A})}{p(\nu|\boldsymbol{\theta})}, \end{aligned} \tag{C.0.9}$$

where the last step follows from the fact that $\langle N_c \rangle = N f_c(\mathbf{A})$. Hence, we have

$$D\left(p(\vec{\nu}|\mathbf{A})||p(\vec{\nu}|\boldsymbol{\theta})\right) = N \left[f_d(\mathbf{A}) \ln \frac{f_d(\mathbf{A})}{f_d(\boldsymbol{\theta})} + \int_{\nu_d}^{\infty} p(\nu|\mathbf{A}) \ln \frac{p(\nu|\mathbf{A})}{p(\nu|\boldsymbol{\theta})} d\nu \right]. \tag{C.0.10}$$

Bibliography

[1] C. Deffayet, "Cosmology on a brane in Minkowski bulk," *Physics Letters B*, vol. 502, pp. 199–208, Mar. 2001, arXiv:hep-th/0010186.

[2] P. Binétruy, C. Deffayet, U. Ellwanger, and D. Langlois, "Brane cosmological evolution in a bulk with cosmological constant," *Physics Letters B*, vol. 477, pp. 285–291, Mar. 2000, arXiv:hep-th/9910219.

[3] R. Maartens, "Dark Energy from Brane-world Gravity," in *The Invisible Universe: Dark Matter and Dark Energy* (L. Papantonopoulos, ed.), vol. 720 of *Lecture Notes in Physics, Berlin Springer Verlag*, pp. 323–+, 2007.

[4] M. Kunz and D. Sapone, "Dark Energy versus Modified Gravity," *Physical Review Letters*, vol. 98, pp. 121301–+, Mar. 2007, arXiv:astro-ph/0612452.

[5] A. Albrecht, G. Bernstein, R. Cahn, W. L. Freedman, J. Hewitt, W. Hu, J. Huth, M. Kamionkowski, E. W. Kolb, L. Knox, J. C. Mather, S. Staggs, and N. B. Suntzeff,

"Report of the Dark Energy Task Force," *ArXiv Astrophysics e-prints*, 2006, arXiv:astro-ph/0609591.

[6] P. Astier, "Can luminosity distance measurements probe the equation of state of dark energy," *Physics Letters B*, vol. 500, pp. 8–15, Feb. 2001, arXiv:astro-ph/0008306.

[7] G. M. Voit, "Tracing cosmic evolution with clusters of galaxies," *Review of Modern Physics*, vol. 77, pp. 207–258, Apr. 2005, arXiv:astro-ph/0410173.

[8] B. P. Koester, T. A. McKay, J. Annis, R. H. Wechsler, A. Evrard, L. Bleem, M. Becker, D. Johnston, E. Sheldon, R. Nichol, C. Miller, R. Scranton, N. Bahcall, J. Barentine, H. Brewington, J. Brinkmann, M. Harvanek, S. Kleinman, J. Krzesinski, D. Long, A. Nitta, D. P. Schneider, S. Sneddin, W. Voges, and D. York, "A MaxBCG Catalog of 13,823 Galaxy Clusters from the Sloan Digital Sky Survey," *Astrophys. J.*, vol. 660, pp. 239–255, May 2007, arXiv:astro-ph/0701265.

[9] C. L. Sarazin, "X-ray emission from clusters of galaxies," *Reviews of Modern Physics*, vol. 58, pp. 1–115, Jan. 1986.

[10] A. E. Hornschemeier, B. Mobasher, L. P. Jenkins, N. A. Miller, C. A. Kilbourne, M. W. Bautz, and D. M. Hammer, "Deep X-ray (and Multiwavelength) Survey of the Coma Cluster of Galaxies," in *Bulletin of the American Astronomical Society*, vol. 38 of *Bulletin of the American Astronomical Society*, pp. 1192–+, Dec. 2006.

[11] R. A. Sunyaev and Y. B. Zeldovich, "Small-Scale Fluctuations of Relic Radiation," *Astrophys. Space Sci.*, vol. 7, pp. 3–19, 1970.

[12] D. E. Johnston, E. S. Sheldon, R. H. Wechsler, E. Rozo, B. P. Koester, J. A. Frieman, T. A. McKay, A. E. Evrard, M. R. Becker, and J. Annis, "Cross-correlation Weak Lensing of

SDSS galaxy Clusters II: Cluster Density Profiles and the Mass–Richness Relation," *ArXiv e-prints*, Sept. 2007, 0709.1159.

[13] K. Umetsu, "Cluster Weak Gravitational Lensing," *ArXiv e-prints*, Feb. 2010, 1002.3952.

[14] R. Mandelbaum, U. Seljak, T. Baldauf, and R. E. Smith, "Precision cluster mass determination from weak lensing," *Mon. Not. R. Astron. Soc.*, pp. 683–+, May 2010, 0911.4972.

[15] M. D. Gladders, H. K. C. Yee, S. Majumdar, L. F. Barrientos, H. Hoekstra, P. B. Hall, and L. Infante, "Cosmological Constraints from the Red-Sequence Cluster Survey," *Astrophys. J.*, vol. 655, pp. 128–134, 2007, arXiv:astro-ph/0603588.

[16] A. Mantz, S. W. Allen, H. Ebeling, and D. Rapetti, "New constraints on dark energy from the observed growth of the most X-ray luminous galaxy clusters," *Mon. Not. R. Astron. Soc.*, vol. 387, no. 3, pp. 1179–1192, 2008.

[17] J. P. Henry, A. E. Evrard, H. Hoekstra, A. Babul, and A. Mahdavi, "The X-ray cluster normalization of the matter power spectrum," *Astrophys. J.*, vol. 691, no. 2, pp. 1307–1321, 2009.

[18] E. Rozo, R. H. Wechsler, E. S. Rykoff, J. T. Annis, M. R. Becker, A. E. Evrard, J. A. Frieman, S. M. Hansen, J. Hao, D. E. Johnston, B. P. Koester, T. A. McKay, E. S. Sheldon, and D. H. Weinberg, "Cosmological constraints from the sloan digital sky survey maxbcg cluster catalog," *Astrophys. J.*, vol. 708, no. 1, pp. 645–660, 2010.

[19] A. Vikhlinin, A. V. Kravtsov, R. A. Burenin, H. Ebeling, W. R. Forman, A. Hornstrup, C. Jones, S. S. Murray, D. Nagai, H. Quintana, and A. Voevodkin, "Chandra cluster cosmology project III: cosmological parameter constraints," *Astrophys. J.*, vol. 692, no. 2, pp. 1060–1074, 2009.

[20] A. Mantz, S. W. Allen, D. Rapetti, and H. Ebeling, "The Observed Growth of Massive Galaxy Clusters I: Statistical Methods and Cosmological Constraints," *ArXiv e-prints*, 2009, arXiv:0909.3098.

[21] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce, "Simulations of the formation, evolution and clustering of galaxies and quasars," *Nature*, vol. 435, pp. 629–636, 2005, arXiv:astro-ph/0504097.

[22] R. Stanek, E. Rasia, A. E. Evrard, F. Pearce, and L. Gazzola, "Massive Halos in Millennium Gas Simulations: Multivariate Scaling Relations," *ArXiv e-prints*, 2009, arXiv:0910.1599.

[23] R. Opher and A. Pelinson, "Constraints on dark energy from the observed density fluctuations spectrum and supernova data," *ArXiv Astrophysics e-prints*, 2005, arXiv:astro-ph/0505476.

[24] W. H. Press and P. Schechter, "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation," *Astrophys. J.*, vol. 187, pp. 425–438, 1974.

[25] A. Jenkins, C. S. Frenk, S. D. M. White, J. M. Colberg, S. Cole, A. E. Evrard, H. M. P. Couchman, and N. Yoshida, "The mass function of dark matter haloes," *Mon. Not. R. Astron. Soc.*, vol. 321, pp. 372–384, 2001, arXiv:astro-ph/0005260.

[26] A. E. Evrard, T. J. MacFarland, H. M. P. Couchman, J. M. Colberg, N. Yoshida, S. D. M. White, A. Jenkins, C. S. Frenk, F. R. Pearce, J. A. Peacock, and P. A. Thomas, "Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations," *Astrophys. J.*, vol. 573, pp. 7–36, 2002, arXiv:astro-ph/0110246.

[27] R. K. Sheth and G. Tormen, "Large-scale bias and the peak background split," *Mon. Not. R. Astron. Soc.*, vol. 308, pp. 119–126, 1999, arXiv:astro-ph/9901122.

[28] J. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. Warren, G. Yepes, S. Gottlöber, and D. E. Holz, "Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality," *Astrophys. J.*, vol. 688, pp. 709–728, 2008, arXiv:0803.2706.

[29] P. Coles, *From cosmos to chaos : the science of unpredictability Probabilities*. 2006.

[30] B. I. and Y. B. Zeldovich, "An essay towards solving a problem in the doctrine of chances," *The Philosophical Transactions of the Royal Society*, vol. 53, pp. 370–418, 1763.

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

[32] J. Skilling, "Nested Sampling for General Bayesian Computation," 2004, http://www.inference.phy.cam.ac.uk/bayesys/.

[33] P. Mukherjee, D. Parkinson, and A. R. Liddle, "A Nested Sampling Algorithm for Cosmological Model Selection," *Astrophys. J. Lett.*, vol. 638, pp. L51–L54, 2006, arXiv:astro-ph/0508461.

[34] D. Parkinson, P. Mukherjee, and A. R. Liddle, "Bayesian model selection analysis of WMAP3," *Physical review D*, vol. 73, no. 12, p. 123523, 2006, arXiv:astro-ph/0605003.

[35] B. Nikolic, "Fitting and Comparison of Models of Radio Spectra," *ArXiv e-prints*, 2009, arXiv:0912.2317.

[36] J. Väliviita and T. Giannantonio, "Constraints on primordial isocurvature perturbations

and spatial curvature by Bayesian model selection," *Physical review D*, vol. 80, no. 12, p. 123516, 2009, arXiv:0909.5190.

[37] F. Feroz, M. P. Hobson, and M. Bridges, "MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics," *Mon. Not. R. Astron. Soc.*, vol. 398, pp. 1601–1614, 2009, arXiv:0809.3437.

[38] R. van Haasteren, "Bayesian evidence: can we beat MultiNest using traditional MCMC methods?," *ArXiv e-prints*, 2009, arXiv:0911.2150.

[39] M. Bridges, A. N. Lasenby, and M. P. Hobson, "WMAP 3-yr primordial power spectrum," *Mon. Not. R. Astron. Soc.*, vol. 381, pp. 68–74, 2007, arXiv:astro-ph/0607404.

[40] M. Kunz, R. Trotta, and D. R. Parkinson, "Measuring the effective complexity of cosmological models," *Phys. Rev. D*, vol. 74, no. 2, p. 023503, 2006, arXiv:astro-ph/0602378.

[41] B. A. Bassett, P. S. Corasaniti, and M. Kunz, "The Essence of Quintessence and the Cost of Compression," *Astrophys. J.*, vol. 617, pp. L1–L4, 2004, arXiv:astro-ph/0407364.

[42] T. D. Saini, J. Weller, and S. L. Bridle, "Revealing the nature of dark energy using Bayesian evidence," *Mon. Not. R. Astron. Soc.*, vol. 348, pp. 603–608, 2004, arXiv:astro-ph/0305526.

[43] M. Beltran, J. Garcia-Bellido, J. Lesgourgues, A. R. Liddle, and A. Slosar, "Bayesian model selection and isocurvature perturbations," *Phys. Rev. D*, vol. 71, no. 6, p. 063532, 2005, arXiv:astro-ph/0501477.

[44] R. Trotta, "Applications of Bayesian model selection to cosmological parameters," *Mon. Not. R. Astron. Soc.*, vol. 378, pp. 72–82, 2007, arXiv:astro-ph/0504022.

[45] V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions," 1996, arXiv:adap-org/9601001v1.

[46] W. L. Freedman, B. F. Madore, B. K. Gibson, L. Ferrarese, D. D. Kelson, S. Sakai, J. R. Mould, R. C. Kennicutt, Jr., H. C. Ford, J. A. Graham, J. P. Huchra, S. M. G. Hughes, G. D. Illingworth, L. M. Macri, and P. B. Stetson, "Final results from the Hubble Space Telescope Key Project to measure the Hubble constant," *Astrophys. J.*, vol. 553, no. 1, pp. 47–72, 2001.

[47] A. Raychaudhuri, "Relativistic Cosmology. I," *Physical Review*, vol. 98, pp. 1123–1126, May 1955.

[48] http://www.jrank.org/space/pages/2262/cold-dark-matter-(CDM).html.

[49] J. A. Fillmore and P. Goldreich, "Self-similar spherical voids in an expanding universe," *Astrophys. J.*, vol. 281, pp. 9–12, 1984.

[50] J. J. Mohr, "Studies of Structure Formation and Cosmology with Galaxy Cluster Surveys," in *AMiBA 2001: High-Z Clusters, Missing Baryons, and CMB Polarization* (L.-W. Chen, C.-P. Ma, K.-W. Ng, & U.-L. Pen, ed.), vol. 257 of *Astronomical Society of the Pacific Conference Series*, p. 49, 2002.

[51] M. Scodeggio, L. F. Olsen, L. da Costa, R. Slijkhuis, C. Benoist, E. Deul, T. Erben, R. Hook, M. Nonino, A. Wicenec, and S. Zaggia, "ESO Imaging Survey. VII. Distant cluster candidates over 12 square degrees," *Astron. Astrophys. Supp.*, vol. 137, pp. 83–92, May 1999, arXiv:astro-ph/9807336.

[52] C. Papovich, I. Momcheva, C. N. A. Willmer, K. D. Finkelstein, S. L. Finkelstein, K. Tran, M. Brodwin, J. S. Dunlop, D. Farrah, S. A. Khan, J. Lotz, P. McCarthy, R. J. McLure,

M. Rieke, G. Rudnick, S. Sivanandam, F. Pacaud, and M. Pierre, "A Spitzer-Selected Galaxy Cluster at z=1.62," *ArXiv e-prints*, Feb. 2010, 1002.3158.

[53] http://www.rssd.esa.int/SA/PLANCK/include/report/redbook/151.htm.

[54] http://dls.physics.ucdavis.edu/.

[55] G. O. Abell, "The Distribution of Rich Clusters of Galaxies," *Astrophys. J. Supp.*, vol. 3, p. 211, 1958.

[56] F. Zwicky, "On the Masses of Nebulae and of Clusters of Nebulae," *Astrophys. J.*, vol. 86, p. 217, 1937.

[57] M. Bradac, P. Schneider, M. Lombardi, and T. Erben, "Strong and weak lensing united I: the combined strong and weak lensing cluster mass reconstruction method," *ArXiv Astrophysics e-prints*, Oct. 2004, arXiv:astro-ph/0410643.

[58] A. Vikhlinin, A. Kravtsov, W. Forman, C. Jones, M. Markevitch, S. S. Murray, and L. Van Speybroeck, "Chandra Sample of Nearby Relaxed Galaxy Clusters: Mass, Gas Fraction, and Mass-Temperature Relation," *Astrophys. J.*, vol. 640, pp. 691–709, Apr. 2006, arXiv:astro-ph/0507092.

[59] P. J. Mancinelli and A. Yahil, "Local Nonlinear Approximations to the Growth of Cosmic Structures," *Astrophys. J.*, vol. 452, pp. 75–+, Oct. 1995, arXiv:astro-ph/9411022.

[60] E. Gaztañaga and J. A. Lobo, "Nonlinear Gravitational Growth of Large-Scale Structures Inside and Outside Standard Cosmology," *Astrophys. J.*, vol. 548, pp. 47–59, Feb. 2001, arXiv:astro-ph/0003129.

[61] S. D. M. White, G. Efstathiou, and C. S. Frenk, "The amplitude of mass fluctuations in the universe," *Mon. Not. R. Astron. Soc.*, vol. 262, pp. 1023–1028, 1993.

[62] J. A. Peacock and A. F. Heavens, "Alternatives to the Press-Schechter cosmological mass function," *Mon. Not. R. Astron. Soc.*, vol. 243, pp. 133–143, 1990.

[63] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, "Excursion set mass functions for hierarchical Gaussian fluctuations," *Astrophys. J.*, vol. 379, pp. 440–460, 1991.

[64] R. G. Bower, "The evolution of groups of galaxies in the Press-Schechter formalism," *Mon. Not. R. Astron. Soc.*, vol. 248, pp. 332–352, 1991.

[65] C. Lacey and S. Cole, "Merger rates in hierarchical models of galaxy formation," *Mon. Not. R. Astron. Soc.*, vol. 262, pp. 627–649, 1993.

[66] R. Stanek, D. Rudd, and A. E. Evrard, "The effect of gas physics on the halo mass function," *Mon. Not. R. Astron. Soc.*, vol. 394, pp. L11–L15, 2009, arXiv:0809.2805.

[67] H. Wu, A. R. Zentner, and R. H. Wechsler, "The Impact of Theoretical Uncertainties in the Halo Mass Function and Halo Bias on Precision Cosmology," *ArXiv e-prints*, 2009, arXiv:0910.3668.

[68] M. Manera, R. K. Sheth, and R. Scoccimarro, "Large-scale bias and the inaccuracy of the peak-background split," *Mon. Not. R. Astron. Soc.*, vol. 402, pp. 589–602, 2010, arXiv:0906.1314.

[69] R. T. Cox, "Probability, Frequency and Reasonable Expectation," *American Journal of Physics*, vol. 14, pp. 1–13, 1946.

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

[71] H. Jeffreys, *Theory of Probability, 3rd ed.* Oxford University Press, 1961.