Studying the dark universe with galaxies

DE-ICAZA-LIZAOLA, MIGUEL,ANGEL,CUITLAHUAC

Studying the dark universe with galaxies

Miguel Angel C de Icaza Lizaola

A thesis presented for the degree of
Doctor of Philosophy

Institute for Computational Cosmology
The University of Durham
United Kingdom

September 2021

Studying the dark universe with galaxies

Miguel Angel C de Icaza Lizaola

Abstract

This work presents two different but connected projects that study the dark components of the universe. We show the first full shape analysis of the eBOSS Luminous Red Galaxy (LRG) sample, which has an effective redshift of $z_{eff} \sim 0.72$ and used the data from the 14th data release of SDSS (DR14). Amongst other parameters, we constrain the growth rate of the universe to have the value $f(z_{eff})\sigma_8(z_{eff}) = 0.454 \pm 0.134$, Our results are in full agreement with the current Λ-Cold Dark Matter cosmological model under the Planck cosmology. This study was followed up with a later data release that has found comparable results (DR16 Gil-Marín et al., 2020).

The second project uses sparse regression methods (SRM) to model the stellar masses of galaxies inside the EAGLE hydrodynamical simulation as a function of the properties of their host dark matter halos, without using prior knowledge of the underlying physics. Our model is designed to be an accurate and simple equation of the host halo properties, which makes it modifiable if one is interested in fitting to a set of statistics. An advantage of SRMs is that they are designed to remove unnecessary terms, our method discarded all parameters related to the angular momentum of the host halo, suggesting that they are not required to explain the stellar mass halo mass relation to the accuracy considered. Using an appropriate formulation of input parameters, our methodology can model satellite and central galaxies at the same time using a simpler model than when they are treated separately. Our models accurately reproduce the stellar mass function and the correlation function of EAGLE galaxies, which makes them an encouraging approach for the construction of realistic mock galaxy catalogs to interpret results from galaxy surveys.

Supervisors: Peder Norberg, Richard Bower and Shaun Cole.

Acknowledgements

Well, it has now been four years and a pandemic since I moved to the UK and started this project. It has been a truly enjoyable time but also a long and difficult process that has been most definitely made easier thanks to the help and support of several colleges and friends. Now, at the end of my stay in this country and with this university, I want to give my most sincere thanks to the people that have been here and made this whole experience enjoyable:

First and foremost I want to thank Peder Norberg, who has been an invaluable support during my stay in Durham. From the first days and our struggles trying to fund my Ph.D. in a time where Mexico was not offering scholarships, to the long discussions and the analysis of all the projects I have been a part of (even those where he was not officially involved in, but for which I needed someone to discuss with). And for much more help, including the herculean task of proofreading this manuscript, for which I cannot be grateful enough.

To Richard Bower who invited me to this cool project he had been thinking about doing, where we tried to make a computer learn physics, which ended up becoming the skeleton of the second half of my Ph.D., for all of his insights into how statistics and astronomy mingle and for always having a new crazy idea every time that we get stuck.

To Shaun Cole for always having the knowledge to point us in the right direction even when nothing seemed to make sense.

To Carlton Baugh for all of his support during my CDT times, and for always having me in mind when there was a talk happening that needed a speaker.

To my Mexican mentors and colleges Mariana Vargas and Sebastien Formenteau for all of the long hours of work we did together for the eBOSS collaboration. And for all the extra time they put out of their busy schedules whenever I needed some extra help.

To all the people that I worked with during my internship in P&G, with special thanks to Stefan Egan and Arturo Martinez.

To all of my colleges from the JUNE collaboration, Kevin Fong, Tristan Caulfield, Ian Vernon, Julian Williams, and Frank Krauss.

To all of the friends that I made along the way. The Durham Ph.D. crew: Aidan, Christoph, Christos, Joseph, Maura, Maria, Andrew, Alice, Vicky, Myles, Sergio and Amy for all the fun. To my 4-year office mate Giorgio. To the climbing crew: Amy, Niamh, Dan (strong, weak, and barbarian), Rob, Tom, Elis, Pheobe, Natalia, Olie, Christian and Hannah, you guys are all legends. Now that I am making this list I am astonished at how many cool people I met in this country. Finally, I want to give a very special shout to Arnau and Carol.

To my parents back in Mexico, Miguel and Julieta, and to my sister Julieta who never let me forget that I am not alone in this, and that no matter what they always have my back.

Finally, I cannot thank Ellen Sirks enough for all of the support that she has given me in these last three years. For being stuck with me through all the pandemic and transforming our tiny apartment into our home. And for being such an inspiring academic and a loving partner.

Contents

Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, Department of Physics, Durham University, United Kingdom. An early and preliminary analysis of the work presented in Chapter 3 appeared in my Master's thesis at the Institute of Astronomy of UNAM, Mexico.

No part of the other chapters has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Some of the work presented in this thesis has been published in journals - the relevant publications are listed below.

- Chapter 3: M. Icaza-Lizaola, M. Vargas-Magaña, S. Fromenteau, S. Alam, B. Camacho, H. Gil-Marin, R. Paviot, Ashley Ross, Donald P. Schneider, Jeremy Tinker, Yuting Wang, Cheng Zhao, Abhishek Prakash, G.Rossi, Gong-Bo Zhao, Irene Cruz-Gonzalez, Axel de la Macorra, 2020, MNRAS, 492(3), 4189-4215*
- Chapter 5: Icaza-Lizaola, Richard G Bower, Peder Norberg, Shaun Cole, Matthieu Schaller, Stefan Egan, 2021, MNRAS, 507(3), 4584-4602
- Chapter 6: Icaza-Lizaola, Richard G Bower, Peder Norberg, Shaun Cole, Matthieu Schaller, 2021-2, (Submitted for publication)

Minor changes where made to some of the content of chapters 3 and 5 with respect to their public versions. These changes are:

*My PhD supervisors Shaun Cole and Peder Norberg, agreed to not be co-authors of this paper due to complications related to the eBOSS publication policy., however, they were significantly involved in the discussion and the writing-up of this work.

- Changes to some text elements so that the notation is homogenized across the whole document.
- Clarifications to queries from the examiners, this are introduced as footnotes numerated with roman numerals.
- Correction of occasional typographic errors found during the reviewing process.
- Some figures and tables from chapters 3 and 5 show very minor modifications, this is done so that they follow the format constraints of this manuscript.
- A couple of paragraphs related with the parameters forecasts made by the eBOSS collaboration at the end of chapter 3 where removed from this work as suggested by the examiners.

List of Figures

List of Tables

Nomenclature

AGN active galactic nucleus

BAO baryon acoustic oscillations

BOSS Baryon Oscillation Spectroscopic Survey

CDM cold dark matter

CMB cosmic microwave background

DESI Dark Energy Spectroscopic Instrument

DM dark matter

eBOSS Extended Baryon Oscillation Spectroscopic Survey

ELG emission line galaxy

EZ effective Zeldovich

HDM hot dark matter

HOD halo occupation distribution

LPT Lagrangian Perturbation Theory

LRG luminous red galaxy

LSS large scale structure

MCMC Monte-Carlo Markov chain

QPM quick particle mesh

RSD redshift space distortions

SDSS Sloan Digital Sky Survey

SN-Ia supernova Ia

Chapter 1

Introduction: the ΛCDM model

1.1 The universe is expanding

It is humbling to remember that at the beginning of the last century there was no evidence of any structure outside of the Milky Way, the universe as we knew it consisted of only nearby stars and a handful of fuzzy gas clouds with a spiral shape. Thomas Wright (1750) from County Durham had suggested that some of these clouds might be distant worlds external to our own, so far away that none of their stars could be distinguished. This idea was later popularised by Immanuel Kant (1755) who coined the term *island universe* for these clouds. The idea remained as a conjecture for many years until the 1920s when telescopes finally became accurate enough to observe details inside these clouds. One of them, the Andromeda nebulae, was close enough to the Milky Way to accurately observe the period of Cepheid stars within it. Edwin Hubble (1924) used these observations to measure the distance to the stars and compare it to the distance from other stars outside of this fuzzy cloud. The conclusion was that these nebulae were much farther away than all other stars in the sky, which proved the *island universe* hypothesised of Wright and Kant. This was the birth of extragalactic astronomy.

Just five years later, Hubble took on the task of collecting the distances to as many of these island universes, now called galaxies, as he could. He also measured the velocity at which these galaxies were moving with respect to us (v). To measure v he used a phenomenon known as redshift: when a light-emitting object moves towards us, the distance traveled by the object in the interval of time between emitting two pulses of electromagnetic radiation is added to the wavelength with which we detect this wave. As a consequence, the object appears bluer than it is. Equivalently when a light-emitting object is moving away from us, the object appears to be redder. The magnitude used to quantify this phenomenon is called redshift (z) and is defined as:

$$z = \frac{\lambda_{\mathbf{detected}} - \lambda_{\mathbf{emitted}}}{\lambda_{\mathbf{emitted}}}, \tag{1.1}$$

where $\lambda_{\mathbf{emitted}}$ is the original wavelength of the object and $\lambda_{\mathbf{detected}}$ is the wavelength that we measure. The redshift of an object is fully determined by the speed, v, at which the object is moving with respect to us and it is not hard to show that (Peacock, 1999):

$$1 + z_{pec} = \sqrt{\frac{1 + \frac{v_{LOS}}{c}}{1 - \frac{v_{LOS}}{c}}}. \tag{1.2}$$

And if we ignore relativistic effects the relation reduces to:

$$v \simeq cz, \tag{1.3}$$

where c is the speed of light. Note that in an expanding universe the redshift of an object is proportional to our distance to the object. And therefore all objects of the same redshift constitute a time snapshot in our lightcone, which means that the redshift of an object indicates the epoch of the universe that we are observing. With this in mind, one can use redshift to refer to a specific moment in the past.

A very unexpected conclusion from the Hubble distance and redshift catalog was that all galaxies that were gravitationally unbound to our galaxy were receding from us. And that the recession velocity had a linear correlation with the distance,

r, to the object. This relation is called the Hubble law. If we name the slope of the linear relationship as the *Hubble constant* (H), then the Hubble law can be expressed as:

$$v(t) = \frac{dr}{dt} = Hr(t). \tag{1.4}$$

The expansion of the universe seems to be the same in any direction that we point our telescopes at. Observations and statistics measured are comparable regardless of the sky direction being considered. This is usually referred to as the universe being *isotropic* with respect to us.

The observed isotropic expansion of the universe raises a philosophical question as there is no reason to believe that our existence should depend on the structure of the universe on the largest scales. So it seems like an incredible coincidence that we just happened to be created on a planet in the galaxy at the center of the expansion. A seemingly more reasonable assumption is that we do not occupy any special place and that the universe is isotropic for any observers anywhere. This is normally referred to as the Copernican principle, under it we conclude that the recession that we observe should also be seen by any other observers in any other galaxy. A universe in which all observers see all distant galaxies receding from them can be referred to as an *expanding* universe.

A corollary from having a universe that is isotropic everywhere is that the universe should have a constant density everywhere. This is usually referred to as the universe being *homogeneous*. Of course, both of these assumptions only hold at sufficiently large scales (typically hundreds of Mpc). For example it is evident that on the scale of our solar system the universe is not isotropic and it does not have a constant density. To understand how isotropy everywhere implies a homogeneous universe we can consider two observers in different regions of the universe and imagine that we draw spheres around them in a way that both spheres intersect each other. Due to isotropy, spheres around an observer should be regions of equal density of galaxies. Therefore the places where both spheres intersect have to have the same density. Given that we can do the same mental experiment with observers at any point and for radii larger than the scale at which the cosmological principle holds, we conclude that the universe should have the same density everywhere.

A universe that is homogeneous and isotropic everywhere is said to follow the cosmological principle, this idea has observational support in the discovery of the cosmic microwave background (CMB) radiation introduced below, and by the fact that galaxy surveys have not found any coherent structure larger than a few hundreds of Mpc (e.g. Zaninetti, 2018; Higuchi et al., 2020) to this date.

In an expanding universe, the distance between two arbitrary objects increases with time (if they are not gravitationally bound). Therefore, it is convenient to define a distance measurement that is independent of time. The standard approach is to define a *comoving distance* x:

$$r(t) = a(t)x, \tag{1.5}$$

where $a(t)$ is called the *scale factor* and is defined in such a way that $a(0) = 1$. The scale factor can also be used as a measurement of time, a given value of the scale factor $a_i < 1$ would correspond to a time in the past where the distance between two objects was a_i times their value at present. Given that redshift is also a measurement of time, cosmological models should have a one-to-one function correlating the two concepts: the relation between the redshift z of an object and the scale factor at the time when the object's light was emitted is given by:

$$a(t) = \frac{1}{1+z}. \tag{1.6}$$

Georges Lemaitre (1931) used the idea of an expanding universe to hypothesise that at some point the universe was compacted into a very hot and dense plasma that he called the *primordial atom*, given that Lemaitre was a priest it is likely that this idea was based on the work of Roberto Grosseteste's (1168-1253) work, a medieval bishop and natural philosopher who suggested that all life and matter was compacted into an infinitesimal atom at the beginning of times. Today, this hypothesis is better known as the Big Bang theory, and it is accepted that this primordial atom cools itself off as it expands and all the structures that we see today should have evolved from tiny fluctuations.

The strongest argument in favor of Lemaitre's hypothesis came in 1965 with the discovery by Arno Penzias and Robert Woodrow Wilson of faint microwave radiation with a temperature of 2.7K, reaching us from every direction in the sky. This radiation is known as the cosmic microwave background (CMB) and it is a relic of a time when the hot universe was cooling off.

In a very hot and compact universe, it is difficult to distinguish between particles and photons due to how energetic matter particles are. At first, electrons are moving too fast to be trapped by the electric potential of the hydrogen nucleus and therefore atoms could not be formed, in its place there was a plasma of electrons and baryons (protons and neutrons). Photons inside this plasma could not travel long distances without being scattered by the charged particles of the plasma via Thomson scattering. By a redshift of around $z = 1100$, the universe cooled down enough for electrons and protons to combine into the first hydrogen atoms. At this point, photons fall out of thermal equilibrium with matter and are free to travel long distances. This is known as the period of decoupling. Given that photons before this time could not travel in a given direction for a long time, the photons from the end of the period of decoupling are the oldest photons that we can hope to detect. The temperature predicted for these photons if they were to reach us today is 2.7K i.e. these photons are the CMB radiation!

The CMB is very homogeneous but there are slight fluctuations in the temperature of about $\Delta T/T = 2 \times 10^{-5}$ of its value (after correcting for variations generated by the motion of our galaxy with respect to the CMB frame, and for variations due to absorption and emission of radiation from gas clouds inside the Milky Way). The temperature of the CMB at a given spot in the sky should be correlated with the density: denser regions have a stronger gravitational pull and therefore electrons fall into them faster, then faster electrons emit hotter photons. These denser regions are the first tracers of structure: as the universe expanded and cooled these structures grew by attracting particles into them.

We started this chapter remembering how the beginning of the last century came with the birth of extragalactic astronomy and with the realisation that our universe was much much bigger than just our island universe. Astronomers realised that up to that point most of their knowledge corresponded to one galaxy out of many. During the last century, there were two more realisations of similar consequences. The first is that we do not know what most of the matter content of the universe is made off. The second is that we do not know what most of the energy content of the universe is.

1.2 The universe expansion is accelerated

In 1988 a team led by Adam Riess and Brian Schmidt, along with a competing team led by Saul Perlmutter, collected information on the luminosity of many supernova Ia (SN-Ia) in distant galaxies (Riess et al., 1998). These supernovae can be used to make measurements of their distance that do not require to assume any theoretical model of cosmology (Phillips, 1993), they are therefore referred to as standard candles. Shockingly it was concluded that these supernovae were systematically fainter than what was expected at the time. There was also a correlation between how unexpectedly faint a galaxy was and its redshift, i.e. the farther a galaxy was, the fainter than expected it was. Given that the distance measurements extracted from redshifts assume the theory of relativity to be correct, it was concluded that, if we wanted to keep the theory of relativity as it is we had to propose that the universe is expanding with positive acceleration. At the same time, galaxy surveys like the APM survey (Dalton et al., 1997) found that their observations were better fitted by flat universes with a positive acceleration, although at the time it was still not clear if the universe had any curvature or not. In 2011, Adam Riess, Brian Schmidt and Saul Perlmutter were jointly awarded the Nobel Prize in Physics for providing evidence of the accelerated expansion of the universe.

An accelerated expansion is hard to understand in a universe dominated by matter, as gravitational interactions should in principle slow the expansion. Therefore, it was hypothesised that there should be another component of the universe that opposes gravity, a sort of pressure term that pushes galaxies away from one another, the nature of this energy (if it exists at all) is still unknown and it has received the name of *Dark Energy.* This was not a new concept at the time: before the discovery of the accelerated expansion of the universe there was no reason to assume that the universe was not static, this presented an issue as gravity was an unopposed attracting force. This led Einstein himself to add a pressure term to his equations, (Harvey, 2012). He stated that this new term could be thought of as empty space taking the role of homogeneously distributed *negative masses.*

Since 1988 different independent analyses have arrived at similar conclusions to the one from Adam Riess. Studies of the CMB (de Bernardis et al., 2000; Spergel et al., 2003) spectrum and studies of the acoustic scale (Eisenstein et al., 2005) (defined below), both concluded that theoretical models built with general relativity need dark energy to agree accurately with observations. Additionally, these studies predict that the missing energy should correspond to around 70% of the energy content of the universe.

1.2.1 What could dark energy possibly be?

As we mentioned above one possibility suggested by Einstein is that dark energy is a property of space itself. Empty space is not the same as nothingness and it has several properties. Perhaps one such property is that it contains energy by some process that we still do not understand. Given that this energy is a property of space itself, it would not be diluted as the universe expands and it should remain constant. This hypothesis is commonly referred to as a *cosmological constant* and is usually represented by the letter Λ. However, it is not clear why this energy should exist at all. An idea from quantum mechanics involves the uncertainty principle. Here the lowest possible state that a vacuum has at a given time called the zero-point energy (ZPE), is not exactly zero, as the uncertainty principle states that there is a limit to accuracy at which we can determine the energy (E) of a system at a given time t $\delta E \delta t \sim h$. The idea is that this ZPE could provide the required energy to explain dark energy (Milonni and Eberlein, 1994). This idea is related to the hypothesis made by Hendrick Casimir who suggested that an attractive force should be measurable between two plates of conducting material in the vacuum due to the presence of non-zero electric fields. This phenomenon (now known as the Casimir effect) has been measured in laboratory experiments (e.g. Lamoreaux, 1997). However, when computing how much energy this effect could generate in the empty space between galaxies, the results ended up being 120 orders of magnitude too big (e.g. Carroll, 2001)!

Another hypothesis suggests that dark energy might be a new undiscovered type of energy scalar field named *quintessence* (e.g. Carroll, 1998) but to this day no new exotic energy has been detected.

Finally, another possibility is that Einstein's theory of relativity is incomplete or flat-out incorrect. Several alternative models have been proposed (e.g. Sotiriou and Faraoni, 2010; Fang et al., 2008; Joyce et al., 2015), but to this date, none of them have been compelling enough to replace general relativity.

For the time being, there is one thing that we need to do to understand this problem better: acquire more and better data. One of the most promising data that we can collect comes from measuring the distance to each of the members of a large catalog of galaxies, this data enables us to map the evolution of dark energy over time and put further constraints on its properties to rule out models.

1.2.2 Large Scale Structure surveys

Arguably the most promising type of data used to explore dark energy in recent years has been large scale structure (LSS) surveys, which provide measurements of redshifts from hundreds of thousands to millions of galaxies mapping out to high accuracy the structures of the Universe. Given that redshifts can be transformed into distance measurements within a given cosmological model, these surveys provide 3-dimensional maps of the Universe. Surveys tend to probe structures within a given redshift range and therefore they are a map of the structure of the universe as it was at the times associated with their redshifts.

Many large scale surveys have been developed in the last decades, some of the most noteworthy ones are the *Two-degree-Field Galaxy Redshift Survey* (Colless et al., 2001; Cole et al., 2005) better known as the 2dFGRS Survey, that was carried out on the *Australian Astronomical Telescope* that measured the redshift of 232,155 galaxies, the Sloan Digital Sky Survey (SDSS) (York et al., 2000; Eisenstein et al., 2005) that has collected spectra for more than three million astronomical objects at the Apache point observatory in new Mexico, the Baryon Oscillation Spectroscopic Survey (BOSS) (Eisenstein et al., 2011; Dawson et al., 2012), wich is a part of SDSS and observed 1.5 million galaxies and 150000 quasars, and its successor, the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) (Dawson et al., 2016) survey that targeted galaxies that complemented BOSS adding around 1 million new objects. eBOSS is of particular importance to this work as in chapter 3 we present the first full shape analysis of its luminous red galaxy sample for its fourteenth data release (Abolfathi et al., 2018; Icaza-Lizaola et al., 2020).

At present, several surveys are being developed, in the next decade, we will acquire unprecedented amounts of data. Some of these surveys are the Euclid space telescope (Laureijs et al., 2011) that is scheduled launch for in 2022 and is expected to measure the redshift of around 50 million objects, The Rubin Observatory Large Synoptic Survey Telescope (LSST) (Ivezić et al., 2019) which is expected to start

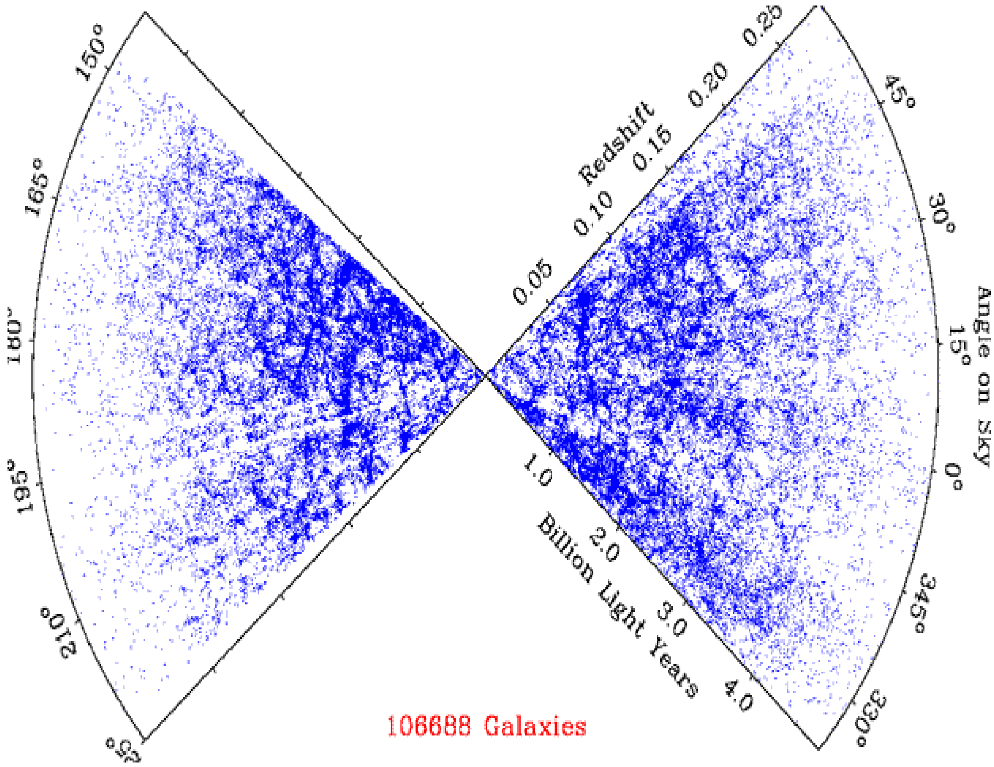


Figure 1.1: The picture shows the light-cone of the 2dFGRS survey, each dot correspond to a galaxy observed by the survey as one moves away from the centre of the cone one explores further redshifts. The image shows how the distribution of galaxies is not random: there is structure in the ways galaxies are distributed. Image credit:Colless et al. (2001) and the 2dFGRS Galaxy Redshift Survey

in 2024 and will measure the redshift of billions of galaxies [*], and the Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016) that saw its first light in 2019 and is currently ongoing, DESI intends to measure the redshift of 35 million galaxies and quasars.

One of the main advantages of LSS surveys is that they allow us to extract information about the clustering patterns of matter. This is shown in figure 1.1, which shows the light-cone of the 2dFGRS survey. The picture shows how galaxies follow certain patterns and includes regions that are very densely populated and others that are more empty, i.e. the distribution of galaxies is not random.

[*] To achieve this huge amount of redshifts LSST will use photometric redshifts, which is a less accurate but faster methodology to compute redshifts. Images measured by the LSST will not have the resolution to distinguish emission or absorption lines and instead relies upon using the fluxes measured with different filters to estimate the redshift of an object.

Measuring the clustering of a survey requires a statistical tool that computes the probability of finding a pair of galaxies at a given distance, r, from each other. The statistical tool used in this work is the *correlation function* ($\xi(r)$), which provides the *excess probability* of finding a pair of galaxies separated from each other by a distance r. Let dV_1 and dV_2 be two volumes separated by a distance r. Our universe follows the cosmological principle i.e. it is homogeneous and isotropic. Note that in a homogeneous universe the correlation function should be independent of the position of vectors $\vec{r}_1$ and $\vec{r}_2$ that determine the position of the volumes dV_1 and dV_1, this is because the probability of finding galaxies inside this volumes should be independent of where in the universe these volumes are located.

In an isotropic universe the correlation function should be independent of the orientation of the vector $\vec{r} = \vec{r}_1 - \vec{r}_2$ connecting dV_1 and dV_1. The universe is isotropic but in a survey, we do not observe galaxies in an isotropic way due to an observational effect known as *redshift space distortions* (RSD), this effect will be discussed in detail in section 1.2.4.1, but for now let us define the correlation function in an isotropic space, in section 2.1 we will discuss the changes on the correlation function $\xi(r)$ when the isotropy hypothesis is no longer valid.

Let $\bar{\rho}$ be the average density of survey objects per unit volume, then by definition we can use the correlation function $\xi(r)$ to compute the probability dP of finding a pair of objects at a separation s from each other. A straightforward definition of the correlation function is then given by the following equation:

$$dP = \bar{\rho}^2[1 + \xi(r)]dV_1 dV_2. \tag{1.7}$$

Let us define the overdensity field δ as

$$\delta(\vec{r}) = \frac{\rho(\vec{r}) - \bar{\rho}}{\bar{\rho}}, \tag{1.8}$$

where $\rho(\vec{r})$ is the density at point $\vec{r}$. If we note that the probablity dP can be writen as the average product of particles in both volumes $dP =< n_1 n_2 >$ and that $\rho_i = n_i/dV_i$, then equation 1.7 can be rewriten as:

$$\xi =< \delta(r_1)\delta(r_2) > . \tag{1.9}$$

Note that if we had a completely random distribution of tracers then $dP = \bar{\rho}^2 dV_1 dV_2$ and therefore $\xi(r) = 0$, i.e. $\xi(r)$ tells you how much more likely it is to get a pair of tracers divided by a distance r when compared to a random distribution of tracers. This helps us come up with a trick to measure the correlation function of an LSS survey, first, we look at all the galaxies in the survey and count out how many of them are separated by a distance r, hereafter $DD(r)$. Then we place galaxies randomly in the same volume and count how many of these random galaxies are separated between them by r, hereafter $RR(r)$. Then if both $DD(r)$ and $RR(r)$ are suitably normalised the correlation function of a survey can be estimated as

$$1 + \xi(r) = \frac{DD(r)}{RR(r)}. \tag{1.10}$$

In principle, one could estimate the expected number of random pairs analytically as the integral over the survey volume of $N^2/dV1dV2$ where N is the number of particles in the whole volume of the survey. However when looking at real surveys there are several observational systematic effects that affect separate areas of the survey differently, so it is common practice to populate the actual survey with random tracers in a way that such systematic effects are considered (see section 3.2.3.3).

There are several ways of measuring the correlation function that follows the same idea behind equation 1.10. In 3.5.2 we present the Landy-Szalay estimator, which reduces the variance of the methodology.

We have mentioned that the CMB fluctuations inform us about how overdensities looked at the moment of recombination. These overdensities should have evolved into the structure that we see today in LSS surveys. A theoretical cosmological model should predict how the structure of the universe evolves, therefore by combining observations of the CMB at the time of recombination and observations of LSS surveys in the recent universe, we can build a complete and powerful data set to test our cosmological models. In what follows we present the *acoustic scale*, which is a quantity related to the size of overdensities at recombination that can be measured from the correlation function of a LSS survey.

1.2.3 Baryon Acoustic Oscillations

The correlation functions from LSS surveys encodes the required information about a very important measurement, the *acoustic scale* of baryon acoustic oscillations (BAO). As with measurements of SN-Ia, acoustic scales can be used to test the expansion rate of the universe.

There are small fluctuations of temperature in the CMB that correspond to denser regions of the universe at the time of recombination. As we will see in section 1.3, most of the matter content of the universe does not interact with photons, only baryons do and they represent only a small percentage of all mass. Let us refer to this non-baryonic component as dark matter (DM).

Figure 1.2 shows four schematics that represent different snapshots on the evolution of an overdensity, going from before recombination to $z = 10$. Given that the perturbation is thought to be adiabatic at first all components are coupled together, this is shown in the top-left plot, which represents a very small perturbation in a smooth background.

Very early in the evolution of the universe neutrinos decouple from all other species and start streaming freely away, this is shown in the top-right plot. This happens before recombination and hence photons are still coupled with baryons. When baryons fall in the potential well of an overdensity, their temperature increases due to compression. This increases the radiation pressure, which in turn generates acoustic waves of baryons and photons traveling outwards. These are shown as the blue and red curves of the top-right plot. These waves are called baryon acoustic oscillations.

After decoupling photons are finally free to stream away. This is shown in the third panel of figure 1.2. At this time the BAO waves are frozen in their place, without radiation pressure to push them away any longer. The result is a clump of dark matter in the centre of an over-density and a shell of baryons separated from the clump by the distance traveled by the wave at the time of decoupling, which is about 150 Mpc in comoving coordinates. This radius is the so-called *acoustic scale*.

As time passes both the DM at the centre and the baryonic shell begin to mix due to their gravitational pull, both perturbations continue to grow as they attract matter from their surroundings. After some time the two curves look more similar as the spherical shell attracts more dark matter and the centre more baryons. This is shown in the last two panels of figure 1.2. At later times the acoustic scale gets imprinted into the DM overdensity.

Places in the universe with a high density of baryons are where galaxies are more likely to form. This means that there should be an excess probability of finding galaxy pairs separated from each other by the acoustic scale, as galaxies are likely to form both in the centre of the overdensity and on the spherical shell. This is shown in the last panel of figure 1.2 that shows that the mass profile of baryons should be much larger at 150 Mpc from the centre than at, say 100 Mpc or 200 Mpc. All of this leads us to conclude that the acoustic scale should be measurable as a peak on the correlation function (as the correlation function is the excess probability of finding galaxy pairs separated by a given distance).

The first measurement of a correlation function with enough accuracy to show the size of the acoustic scale was done by Eisenstein et al. (2005) that used the Sloan Digital Sky Survey to measure the distance to 50,000 galaxies at a redshift between 0.16 to 0.4. This is shown in figure 1.3, where one can observe a bump

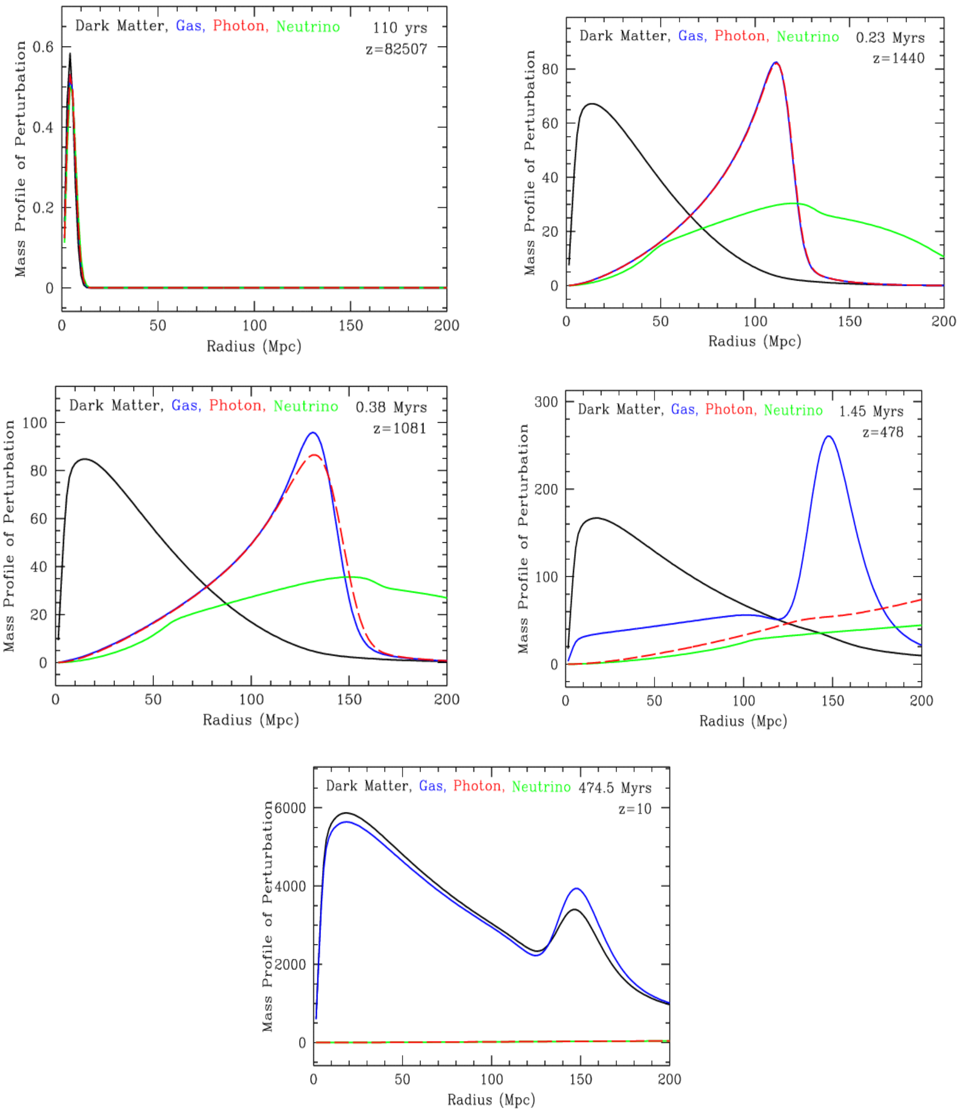


Figure 1.2: These five diagrams represent different snapshots of the evolution of an overdensity at different redshifts. The x-axis of all plots shows the comoving radius of the overdensity (with the zero value corresponding to its centre) while the y-axis shows the fractional density perturbation times the radius square ($\delta(\vec{r})r^2$) note that the fact that it is a fractional density means that the plots do not reflect the fact that there is more more DM than baryonic matter. All plots show four lines, the blue line shows the overdensity in the baryonic matter, the black line in the non-baryonic matter (the so-called dark matter that we introduce later in section 1.3), the green line shows neutrinos, and the red line photons. Image credit: Daniel Eisenstein (https://lweb.cfa.harvard.edu/∼deisenst).

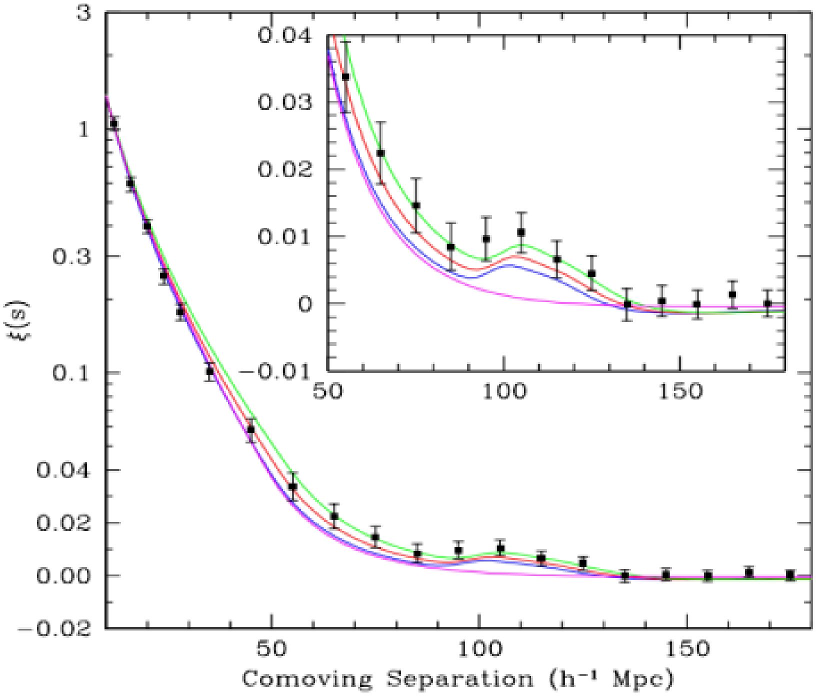


Figure 1.3: The correlation function from the Sloan digital sky survey, that measures the excess probability of finding a pair of galaxies separated from each other by a comoving distance s. The excess probability has a peak at a separation of around $100 Mpc/h$, this corresponds to the acoustics scale. The black dots are the measurements from the SDSS LRG sample and the errors are computed using mock catalogs, we will describe this method of error estimates in more detail in section 2.5. The different coloured lines represent different cosmological models with the magenta model being a model with no dark energy at all and therefore no acoustic peak, the rest of the lines correspond to models with different ratios of mass to dark matter. Plot taken from Eisenstein et al. (2005)

on the correlation function at a separation of around $100 Mpc/h$, this value is a measurement of the size of the acoustic scale. From this point on, BAO analysis has become a staple of modern LSS surveys.

1.2.4 Growth rate of structure.

Up to this point we have talked about constraining dark energy using measurements of distances, either via standard rulers (BAO) or via standard candles (SN-Ia). Both of these measurements provide constraints on the expansion history of the universe, i.e. they measure how much the universe has expanded at different cosmic epochs.

One of the shortcomings of BAO/SN-Ia studies is their inability to test the validity of a cosmological model. Linder (2005) showed that an appropriate choice of an equation of state*, can make different models have the same expansion history.

This brings up the need for a new independent parameter that can break this degeneracy. It is common practice to use a parameter known as the growth rate of structure f (Kaiser, 1987; Percival and White, 2009). To understand what the growth rate is, we need to start by defining how overdensities grow as the universe expands.

Under the paradigm of linear growth of perturbations (e.g. Knobel, 2013, chapter 2), all growing solutions of the overdensity field $\delta(\vec{x}, a)$ (defined in equation 1.8) can be expressed as:

$$\delta(\vec{x}, a) = A(\vec{x})D(a) \tag{1.11}$$

Assuming linear theory, the value of an overdensity at a given scale factor a is related to its value at present by the function $D(a)$, and this relation is independent of the position $\vec{x}$ of the overdensity

$$\delta(\vec{x}, a) = D(a)\delta(\vec{x}, a = 1). \tag{1.12}$$

The function $D(a)$ is appropriately named the *linear growth function*, and it quantifies the amount of structure that has been formed within a given model at a given scale factor a (or equivalently redshift z) with respect to the structure today. An example is shown in figure 1.4. The figure shows the evolution of two models, one with dark energy and the other without. The model with dark energy has a pressure force opposing the gravitational collapse, which means that structure evolves slower than in the model without dark energy. As a consequence, a given measured

*An equation of state provides the relation between density and pressure in a given cosmological model.

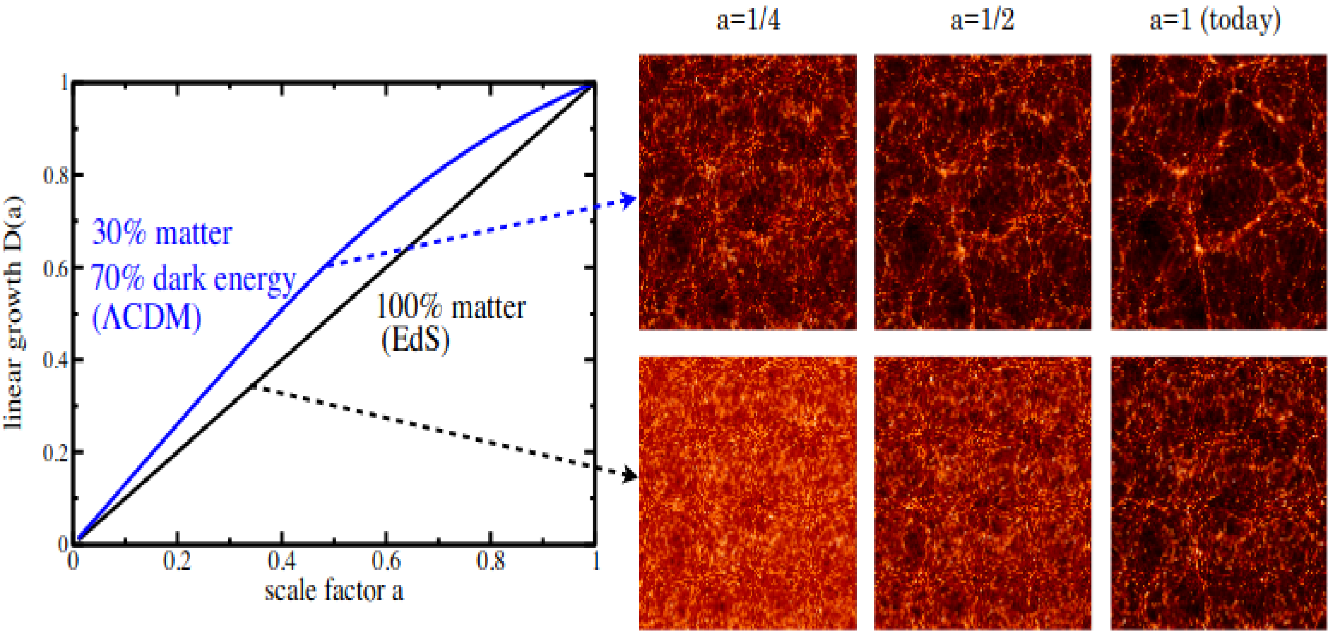


Figure 1.4: The left panel shows the linear growth as a function of the scale factor for two different models, the blue line corresponds to a model with both dark energy and matter, the black line corresponds to a model with only matter. Given that the model with dark energy includes a pressure force opposing gravity it takes longer for the structure to collapse into its present form, this is shown in the panel on the right were to have approximately the same structure today the model with dark energy needed to start collapsing much earlier. Figure from (Huterer et al., 2015)

value of $\delta(\vec{x}, a = 1)$ at present would have different formations histories in different cosmological models. More precisely, the model with dark energy would have to start generating structure much earlier than the model without.

This indicates that if we know the expansion history of the universe (for example by using BAO analysis), one can use measurements of the linear growth function to test if it fits accurately within a given cosmological model.

Measuring the linear growth function requires us to use an observable property correlated to it, the standard approach is to use the aforementioned growth rate of structure, defined as

$$f(a) = \frac{d \ln D(a)}{d \ln a} \tag{1.13}$$

This quantity measures the speed at which structure evolves, and should be correlated with observations of the infall velocities of tracers of matter like galaxies falling into structure.

Let us start by writing the continuity equation as a function of the overdensity field δ as

$$\frac{\partial \delta}{\partial t} + \frac{1}{a} \vec{\nabla} \cdot [(1+\delta)\vec{v}] = 0, \tag{1.14}$$

where $\vec{v}$ is the peculiar velocity of the object (the velocity of the object in comoving coordinates). Under the assumption that $\delta << 1$, and using equation 1.12 to note that $\delta(\vec{x}, a_1) = \delta(\vec{x}, a_2) D(a_1)/D(a_2)$ we rewrite equation 1.14 as

$$\vec{\nabla} \cdot \vec{v} = -a \frac{\partial \delta}{\partial t} = -a\delta \frac{\dot{D}}{D} = -aH\delta f \tag{1.15}$$

which shows that the growth rate is related to the divergence of the peculiar velocity of objects. Therefore measurements of these velocities can be translated into approximations of f. Fortunately, statistical information about these velocities at a given scale factor can be obtained by a phenomenon known as redshift space distortions (RSD).

1.2.4.1 Redshift space distortions

When measuring the distance of an object using redshifts we assume that the shift in the spectra of the object is due to the expansion of the universe. While this is a good first approximation, it does not consider that galaxies have their peculiar motions due to gravitational interactions within their local environment. If these motions have a non-zero velocity component when projected along our line-of-sight

(v_{LOS}) then this peculiar motion should also generate a shift in the spectrum, and the redshift that we measure should be a superposition of the two effects. Let us call *cosmological redshift* (z_{cosmo}) the shift in the spectra that is a consequence of the expansion of the universe, then the redshift that we observe (z_{obs}) is related to it through the following expression:

For an arbitrary number of sequential shifts, the total redshift is given by (for the derivation see Harrison, 1974):

$$1 + z_{obs} = \Pi_i(1 + z_i). \tag{1.16}$$

Including shifts due to the expansion of the universe and due to the peculiar velocity, we get:

$$1 + z_{obs} = (1 + z_{cosmo})(1 + z_{pec}), \tag{1.17}$$

where z_{pec} is the shift due to the peculiar velocity. For $v \ll c$, eq. 1.2 can be approximated to $1 + z_{pec} = (1 - \frac{v_{LOS}}{c})^{-1}$. Plugging this into eq. 1.17, then gives:

$$1 + z_{obs} = (1 + z_{cosmo})(1 - \frac{v_{LOS}}{c})^{-1}. \tag{1.18}$$

A conclusion from this is that there is a small difference between the distance to an object that we compute using the Hubble law ($|\vec{s}|$) and the actual distance to an object ($|\vec{r}|$). Both of these vectors are related by the following expression (e.g. Hamilton, 1998):

$$\vec{s} = \vec{r} + \frac{v_{LOS}}{aH}\vec{z}, \tag{1.19}$$

where $\vec{z}$ is a unitary vector in the line-of-sight direction. This effect is known as redshift space distortions (RSD). From now on we will refer to the set vectors that describe the real distance to an object as the *real* ($\vec{r}$) space, and to the offset vectors that account for the peculiar velocities as the *redshift* space ($\vec{s}$).

The RSD effect will alter the apparent distances of objects differently depending on their peculiar motions and on the component of these motions along the line-of-sight, this is exemplified in Figure 1.5 that shows the spherical collapse of an overdensity in different scenarios, let us think of the lines shown as places where one or many galaxies could be. The observer is thought of as being on the bottom of the figure and therefore the distortion is stronger along the vertical axis while the horizontal axis stays unchanged. The top panel shows how the distortion looks when the distortion given by $\frac{v_{LOS}}{aH}\vec{z}$ is small when compared to the radius of the object (R). In this case, the object looks squished in redshift space, and the circular overdensity looks like an ellipse. As the overdensity continues to collapse it accelerates, eventually reaching the point where $\frac{v_{LOS}}{aH}\vec{z} = R$. In this case, the overdensity looks as if it collapsed into a flat circle, eventually the overdensity collapses very fast and $\frac{v_{LOS}}{aH}\vec{z} > R$: this happens in the nonlinear regime and gives rise to the so-called *fingers of god* effect (e.g. Hamilton, 1998).

By studying RSD we can get information about the peculiar velocities of galaxies by considering the differences in clustering along the line of sight and its perpendicular direction, which can then, in turn, be correlated with the growth rate f by equation 1.15 and with the correlation function by equation 1.9. Chapter 2 will go into details on the procedure that we followed to get constraints on the growth rate by studying RSD in a galaxy survey. For now, let us move into introducing the second significant realisation that astronomers did at the end of the last century: the universe is much more massive than what we observe.

1.3 The universe is massive

Astronomers have known for a long time that galaxies should contain objects that do not emit light. For example, the discovery of the planet Neptune by Urbain Le

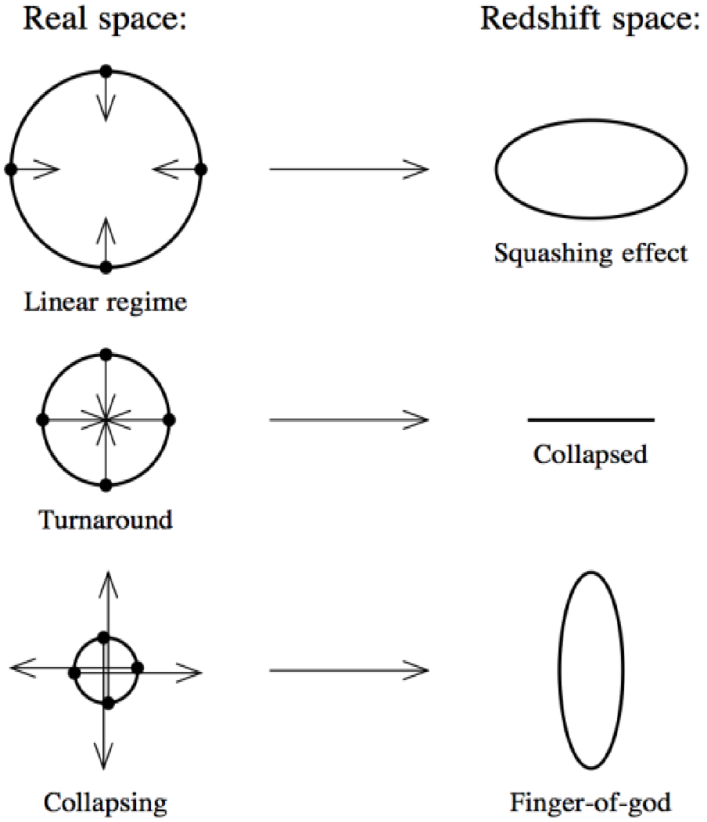


Figure 1.5: Schematic drawings of how circular overdensities look in real and redshift space. The circles on the left represent real space and the ellipses on the right their equivalent shape in redshift space, the observer is located at the bottom of the image, and therefore the overdensity is not distorted in the horizontal direction. Image credit: Hamilton (1998)

Verrier and John Couch Adams (1846) that proposed its existence to explain the anomalies in the motions of Uranus, the discovery of faint companion stars in 1844 (Bessel, 1844)[*], the hypothesis that obscured regions of the sky could be blocked by dark clouds (Secchi, 1877) or the theoretical prediction of black holes in 1784 (Michell, 1784)[†] have all been around for more than 100 years. However, until the

[*] By looking at the proper motions of Sirius, Friederich Bessel realised that the star should have a companion influencing the motions of the star with its gravitational pull. Today we know that this companion star is the white dwarf Sirius B, the companion of which outshines it, so that it could not be distinguished with the telescopes of that time.

[†] Mitchel believed in Newton's corpuscular theory of light. He reasoned that light particles should also be affected by gravity and as such light should slow down due to the gravitational pull of stars. This led him to conclude that there could exist stars so massive that the escape velocity required to leave the star exceeded the speed of light. He coined the term *dark stars* to refer to these objects.

early 1900s it was assumed that the total mass of these dark objects should be small compared to that of visible matter.

The first attempt to try to use dynamic estimates of the motions of stars within a galaxy to predict the ratio between dark and visible matter was done by Lord Kelvin (Thomson and Kelvin, 1904). His task was continued by Poincaré (Poincare, 1906) who used Kelvin's estimates to conclude that within the Milky Way the mass of the non-luminous component of the galaxy should be smaller, or at the most of the order of magnitude, than the mass of visible matter (he observed the velocity dispersion of local stars where it is very difficult to detect the effect of dark matter). He also coined the term DM to refer to the non-luminous component of mass within the universe (Bertone and Hooper, 2018).

In 1933, Fritz Zwicky (Zwicky, 1933) used redshifts to measure the velocity dispersion of galaxies within the Coma cluster and compared it to what one would expect by looking at the mass of all galaxies within the cluster. Astonishingly, he concluded that galaxies move so fast that the cluster should not be able to remain gravitationally bound and most galaxies should just fly away into space. One of the many hypotheses suggested explaining Zwicky's findings was that the dark matter within the cluster should be much more massive than the luminous matter.

Zwicky's results were the cause of a 40-year long debate about the reason for the discrepancy between observed and dynamical masses in galaxies. The debate came to an end in the late seventies with studies of the rotation curves of galaxies (Metropolis et al., 1953; Rubin et al., 1980, e.g.). These studies showed that the rotation velocities of stars and gas orbiting galaxies remained constant at a radius that was very far from the centre of the galaxies, a radius at which the density of visible matter had declined. This suggested that the mass of the galaxy should be far larger than what one would expect from visible matter and also that the distribution of matter within a galaxy should be very different for the luminous and the dark components.

Similar conclusions about the amount of this dark matter have been found by several other independent experiments. Some noteworthy ones use galaxy clusters (e.g. Allen et al., 2011), the CMB (e.g. Planck Collaboration et al., 2016b)*, gravitational lensing (e.g. Taylor et al., 1998) and many more. The consensus of all of these experiments is that dark matter should account for around 85% of all matter in the universe.

1.3.1 What could dark matter be?

When dark matter was first proposed, one expected it to be non-luminous or faint baryonic matter such as undiscovered large numbers of objects like neutron stars (who are non-luminous in the optical range although they can be detected by x-rays), black holes, white dwarfs stars, very faint red dwarfs or brown dwarfs, or unassociated planets. The term massive astrophysical compact halo object (MACHO) (e.g. Alcock et al., 2000) was coined to refer to all massive objects that emit little or no light and drift through interstellar space. However, it has become increasingly clear that this is not the case. On one hand, studies done on the CMB agree remarkably well with models that suggest that most matter should be presented in a form that does not interact with photons (e.g. Planck Collaboration et al., 2016b) and should correspond to around 4% of the energy content of the universe. While analyses done using gravitational lensing have discarded MACHOs as a viable candidate of dark matter (e.g. Tisserand et al., 2007). And diffuse gas clouds should still be visible when illuminated by stars. The final nail in the coffin to baryonic dark matter comes from the chemical abundance of elements in the universe, the theory of big bang nucleosynthesis predicts that if baryons were more than 4% or 5% of the energy density of the universe, then heavier elements like helium or lithium should be much more common (e.g. Coc and Vangioni, 2017).

Another possibility to explain at least some of the observations that require dark

[*] The theoretical models of the CMB Power spectrum require a massive non-baryonic component in order to agree with observations.

matter points to the fact that our laws of gravity have only been tested accurately on scales of the order of magnitude of stars and solar systems, so perhaps the measured dark matter is an indication that we should modify our gravitational laws when working on larger scales. Some modified models of gravity have been proposed (e.g. Milgrom, 1983), but it is difficult to build one single modified model that explains all observations. To this date, the dominant hypothesis is that there are one or more undiscovered non-baryonic particles inside galaxies. These particles should be massive and therefore exert a gravitational pull on objects, but they should not interact with photons (or interact very weakly), which might explain why they are so hard to detect.

Today the most common hypothesis is that dark matter should consist of non-baryonic particles that do not interact through the electromagnetic force, these are usually referred to as weakly interacting massive particles or WIMPS (e.g. Smith and Lewin, 1990). The name emphasises that while these particles should be blind to electromagnetic radiation, it is still plausible that they interact through the weak nuclear force. Several non-baryonic particles have been proposed as dark matter candidates. A standard way to split these models up is by the speed at which their particles moved in the early universe (Bond et al., 1984), with the fastest particle candidate models refereed to as hot dark matter (HDM) models and the slowest ones as cold dark matter (CDM) models. The speed of particles at this time is important as it determines the lower limit of overdensity sizes: very fast particles would spread out from small overdensities into the surrounding under-dense regions, which would mean that the overdensity would disappear. In the CDM paradigm, the original overdensities are small at first and merge to create larger structures as time passes, known as a *hierarchical* structure formation scenario. In the HDM paradigm, the opposite happens and the universe should start with very large structures that later break to form smaller ones. The CDM paradigm ended up agreeing the best with observations (Blumenthal et al., 1984) with Davis et al. (1985) ruling out HDM in favor of CDM by comparing galaxy clustering measurements with predictions from simulations built in the CDM paradigm. It has also been observationally concluded that small galaxies form first followed by clusters later on (e.g. Gilman et al., 2019).

To summarise, while there is much mystery surrounding dark matter we have good reasons to suspect that DM should consist of non-baryonic particles that do not interact with photons and moved slowly in the early universe. Theoretical models of the universe that define dark matter in this way and include a cosmological constant Λ to describe dark energy are known as ΛCDM models. To this day they are the most tested models of cosmology, agreeing remarkably well with most observations (although recently there has been newly discovered tensions at the 1% level between the ΛCDM model and a set of cosmological observations (e.g. Perivolaropoulos and Skara, 2021)).

The ΛCDM model is a beautiful and relatively simple model that can explain a large array of cosmological observations. However, as with any good model in science, it needs to be put to the test. One of the most efficient methods to do this is to ask a computer to create virtual universes for us.

The virtual universes or *simulations* that are usually created can be divided into simulations that include baryons in their computations and those that do not. DM makes up most of the mass component of the universe, and it is significantly easier to model than baryons as it only interacts gravitationally with other particles. More complicated simulations that include baryonic matter are usually referred to as *hydrodynamical* simulations and are much more complicated to produce.

1.3.2 N-body simulations

What we actually mean when we say *simulating the universe* is finding the solutions to the equations of motions for a set of particles that represent the matter of the universe. Let us define a set of N particles of mass m_i $(0 < i < N)$. Given that these are DM particles, they should interact with each other only through gravity, and their dynamics are represented by the following differential equation

$$\vec{F}_i(t) = m_i \frac{d^2\vec{x_i}(t)}{dt^2} = -\sum_{j\neq i} \frac{Gm_i m_j(\vec{x_i}(t) - \vec{x_j}(t))}{(\|\vec{x_i}(t) - \vec{x_j}(t)\|^2 + \epsilon^2)^3/2}, \tag{1.20}$$

where $\vec{F}_i(t)$ is the force felt by the i^{th} particle due to its gravitational interaction with all other particles, G is the gravitational constant, $\vec{x_i}(t)$ the position of the i^{th} particle at time t and ϵ is a softening length, added to avoid computational artifacts when the distance between two particles is small. Note that the equation is only valid in Newtonian dynamics, which is a fair approximation on cosmological scales, however, other astrophysical problems, like modelling particles close to the event horizon of a black hole, require to switch to a general relativity paradigm (e.g. Baker et al., 2006).

Given that for a small time step, $\vec{v_i}(t + dt) \sim \vec{v_i}(t) + (\vec{F}(t)/m_i)dt$ and $\vec{x_i}(t + dt) \sim \vec{x_i}(t) + \vec{v}(t)dt$, the solution to these equations can be numerically approximated for a given set of initial conditions $[\vec{x_i}(t = 0)$,$\vec{v_i}(t = 0)]$. The solutions are the values of the position and velocity of all N-particles at different time steps, this type of simulations are refereed to as *N-body* simulations (e.g. Springel et al., 2001; Springel et al., 2005; Baugh et al., 2019).

In practice solving equation 1.20 or a very large amount of particles and with sufficiently small time-steps is computationally infeasible. And several methods to approximate the solutions are used instead. Two of the more noteworthy are:

- Tree codes (e.g. Barnes and Hut, 1986), where precise calculations are only done in very dense regions of the simulation.
- Particle-mesh codes (e.g. Klypin and Holtzman, 1997), where the gravitational potential is computed over a grid in the simulation volume, and particles are assigned the force applied to their corresponding grid cell.

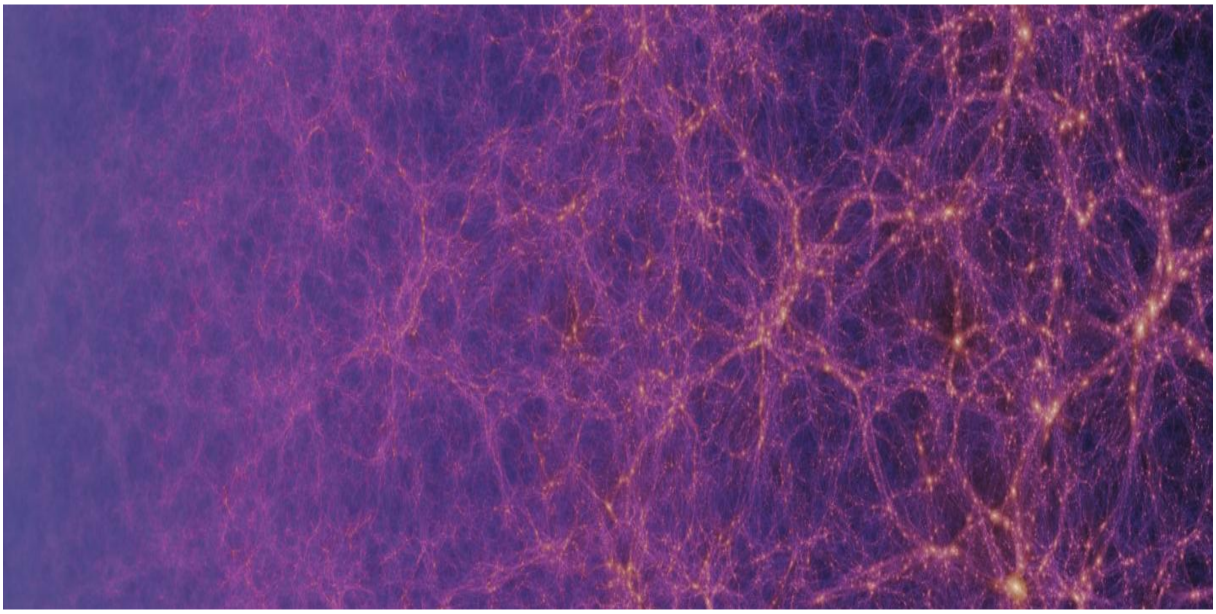


Figure 1.6: This image shows the position of particles in an N-body simulation as a function of time, with the left edge of the image showing the beginning of the simulations and representing the early universe. As one moves to the right, the simulation evolves and structure grows in complexity. The image is colour-coded by particle density. Image Credit: The EAGLE project/Stuart McAlpine.

Figure 1.6 shows the evolution of dark matter density field within the EAGLE simulation (Schaye et al., 2015; Crain et al., 2015). At the beginning of the simulation, the universe is more or less homogeneous with matter somewhat uniformly distributed, however as time passes by, particles clump together by gravitational interactions as denser regions attract particles in neighboring regions to form larger and larger structures.

These clumps of matter are called *halos* and correspond to the denser places in the universe. It is inside these halos where models predict that galaxies form, and therefore tracing and defining these halos is of the utmost importance when studying N-body simulations. One of the most common ways of defining halos is through the *friend of friends* (FoF) algorithm (Huchra and Geller, 1982). Under this paradigm a halo is composed of the collection of all particles that can be linked together using an inter-particle length that is typically 0.2 times the mean inter-particle separation (e.g. Davis et al., 1985).

Under the CDM paradigm, halos merge to form larger and larger halos, also shown in figure 1.6, one useful tool to track the evolution of matter is to store the merging history of halos in such a way that one can relate a halo at a given time with its progenitors, catalogs of halo merging histories are adequately named dark matter halo *merger trees*.

One conclusion drawn after analyzing the merger trees of N-body simulations is that the remnants of small halos accreted into larger ones form self-bound substructures that orbit around the halo core (e.g. Gunn and Gott, 1972; Giocoli et al., 2008), these substructures are usually referred to as *subhalos* and are predicted to be the hosts of satellite galaxies within a galaxy cluster.

So far we have focused on N-body simulations that only include DM particles. However, if we want our simulations to be able to reproduce galaxies we need to include baryons as well. These are usually called hydrodynamical simulations (e.g. Cen and Ostriker, 1992; Pearce et al., 1999; Springel and Hernquist, 2002; Schaye et al., 2015), and are far more complex than simulations that only include DM. This is because baryons interact with each other through the electromagnetic force as well which leads to physical effects like pressure and radiative cooling not present in DM only simulations. The standard approach to make these simulations is to treat baryons as if they were a fluid and study them with hydrodynamical equations (hence the name).

On top of this many baryonic processes like star formation, feedback from supernovae explosions, and feedback from supermassive black holes (e.g. Thacker and Couchman, 2000; Marri and White, 2003; Oppenheimer and Davé, 2008; Booth and Schaye, 2009), happen on scales smaller than the resolution of even the most accurate hydrodynamical simulation, therefore they need to be added by hand at so called subgrid scales. Given to how much more complex hydrodynamical simulations are they are usually run in volumes much smaller than the ones used to run dark matter only simulations.

Simulations are by definition of a finite volume, which means that they have boundaries. If one is not careful about these boundaries one can bias the galaxy evolution around the edges, for example, halos near an edge will grow less than halos near the centre as there would not be enough particles to fall in the halo potential well in the direction of the simulation edge. One common approach is to use periodic boundary conditions in which objects near one side of the simulation edge affect objects near to the opposite side as if both sides were connected, this is similar to treating the simulation as if it were a flat projection of the earth on a map, where Alaska and Japan seem to be in opposite ends while in reality, they are close to each other.

The finite size of a simulation limits the capability of the simulation to study certain phenomena, for example, simulations underestimate the correlation function, even at scales much smaller than the simulation length (Gelb and Bertschinger, 1994; Bagla and Prasad, 2006; Bagla and Ray, 2005). Given that hydrodynamical simulations are in general much smaller than dark matter only simulations, there is an incentive for developing methods that can learn the relations between baryonic mass and dark matter from a hydrodynamical simulation. These methods can learn how to predict properties of galaxies (e.g. their stellar mass, their metalicity, or their luminosity), as a function of the characteristics and the history of their host halos. These relations can then be used to populate the halos of a large N-body simulation with the appropriate galaxies. This would result in galaxy catalogs that are not affected by small simulation volumes. In chapters 5 and 6 we introduce a method that learns to predict the stellar mass of galaxies from the properties of their host halo using the data of a hydrodynamical simulation as input.

1.4 Thesis road-map

In this work, we study the relationships between models and observations or simulations as a tool to analyze the dark components of the universe: dark matter and dark energy. This is done in two different projects.

The first project does a *full shape* analysis of the so-called luminous red galaxy luminous red galaxy (LRG) sample of eBOSS fourteenth data release (Abolfathi et al., 2018). A full shape analysis studies the clustering of the galaxies on an LSS survey taking into account redshift space distortions, and considering that discrepancies between the true cosmology of the universe and the one selected for the analysis will result in a distortion of the clustering signal.

Chapter 2 introduces the necessary constituents that one needs to develop to do a full shape analysis. This chapter is thought of as an introduction that presents the background required to follow the discussion of our eBOSS work. Then chapter 3 presents our analysis and results. These results include a constraint on the value of the growth rate of the universe presented in equation 1.13.

Our second project introduces a novel methodology to predict the stellar mass of galaxies from the properties of their host halos. Three chapters are dedicated to this project. First, chapter 4 introduces the different astronomical concepts that are required to follow the discussion of the rest of the project. This chapter is also thought of as an introductory chapter, here we do not discuss how the methodology works, but instead we motivate building our method, we also present different astronomical concepts that will be crucial in the decisions taken in the design of our method and revise the different approaches that have been historically used to populate dark matter halos with galaxies. In chapter 5 we introduce and test our methodology, here we introduce our method in a sample of relatively massive halos and avoid including any subhalo. This is due to subhalos having complicated evolutionary paths after merging, as opposed to central halos that tend to grow monotonically with time. This chapter can be thought of as a *proof of concept*, where we ran our method in a simpler sample than the one we ultimately envisioned. Then, in chapter 6 we expand the method to include satellite galaxies and a smaller halo cut.

Finally, chapter 7 summarises all the work presented here and discusses the next steps that I intend to explore in my future research.

Chapter 2

Redshift space distortions modelling and fitting

The objective of the next two chapters is to present the redshift space distortions (RSD) analysis of the eBOSS DR14 data for luminous red galaxies. The actual analysis is left for chapter 3, while this chapter introduces the different steps that one needs to follow to perform an accurate RSD analysis.

A RSD analysis can be summarised as a methodology for fitting the free parameters of a cosmological model to a data set. This model should include RSD effects and depend on parameters like the growth rate f defined in section 1.2.4. The end goal of the analysis is to explore the parameter space and find the regions of high likelihood to obtain cosmological parameter constraints within a given model framework, e.g. the ΛCDM model.

The quantity being modeled is the correlation function $\xi(s)$ defined in section 1.2.3. The correlation function can also be measured from the galaxies inside an LSS survey, using e.g. equation 1.10 or the Landy-Szalay estimator that we will present in 3.5.2. A RSD analysis statistically compares the model prediction of the correlation function with the ones measured from the data.

In chapter 1 (§1.2.2) we mentioned that a survey does not observe galaxies in an isotropic way due to redshift space distortions. Section 2.1 discusses these distortions and presents a new parametrisation of $\xi(s)$ that takes into account these anisotropic distortions by performing a multipole expansion.

There are four constituents needed to make the RSD analysis possible, which each will be discussed in turn in this chapter:

1. A data set from which to measure the correlation function. Different LSS surveys choose different types of galaxies, called *tracers* of matter for BAO and RSD analysis. Section 2.2 presents a summary of some of the considerations when selecting a tracer and introduces some of the most common tracers of matter used in RSD analysis.

2. A theoretical model of the correlation function that is sensitive to RSD parameters. Section 2.3 summarises some of the more common and general models, while in section 3.4 we introduce the CLPT-GSRSD model developed by Wang et al. (2014); Reid and White (2011); Carlson et al. (2013), which is the model used in our RSD analysis.

3. A set of free parameters that give a degree of freedom to the model, allowing for the possibility of the model being affected by a poorly selected cosmological model. Section 2.4 introduces the Alcock-Paczynski parameters used in this work.

4. An estimate of the error in the correlation function from our data set. As discussed in section 2.5, this is usually done using collections of hundreds of simulated data sets known as *mock catalogs*. Some of the most widely used methodologies for making these catalogs are briefly presented in section 2.5.

These four constituents can be used to estimate a likelihood that quantifies the accuracy of a model. The likelihood should be a function of the free parameters of the model which include the cosmological parameters that one is trying to fit. Therefore one can use these estimates to explore the parameter space and find the regions of higher likelihood. The standard methodology to make this exploration is presented in section 2.6.

2.1 The two-point correlation function

In the previous chapter, we introduced the correlation function as the excess probability of finding two galaxies separated by a distance r between them, most often written as:

$$dP = \bar{\rho}^2[1 + \xi(r)]dV_1 dV_2. \tag{2.1}$$

Let us note that this equation determines $\xi(r)$ as a function of the real-space distance r. Given that in real space the universe is homogeneous and isotropic the correlation function can be written as function of the separation r only. However, in redshift space, the clustering of objects should also depend on their position with respect to the line of sight. This is shown in figure 2.1, where one can see that the correlation function in redshift space is not isotropic.

With this in mind we define the parameter

$$\mu = \cos(\theta), \tag{2.2}$$

where θ is the angle between the line of sight and the line connecting dV_1 and dV_2. The two-point correlation function of our analysis hence depends on two spatial variables, r and μ.

In practice the method for computing the correlation function works by dividing the correlation function into 2-dimensional bins in the [r,μ] space and then counting how many pairs of galaxies are found in each bin: this is the $DD(r, \mu)$ of equation 1.10. This leads to a technical issue as the larger the number of bins, the better one can trace the [r,μ] space, but the smaller the number of tracer pairs one would have per bin and the more susceptible one will be to statistical noise. Therefore

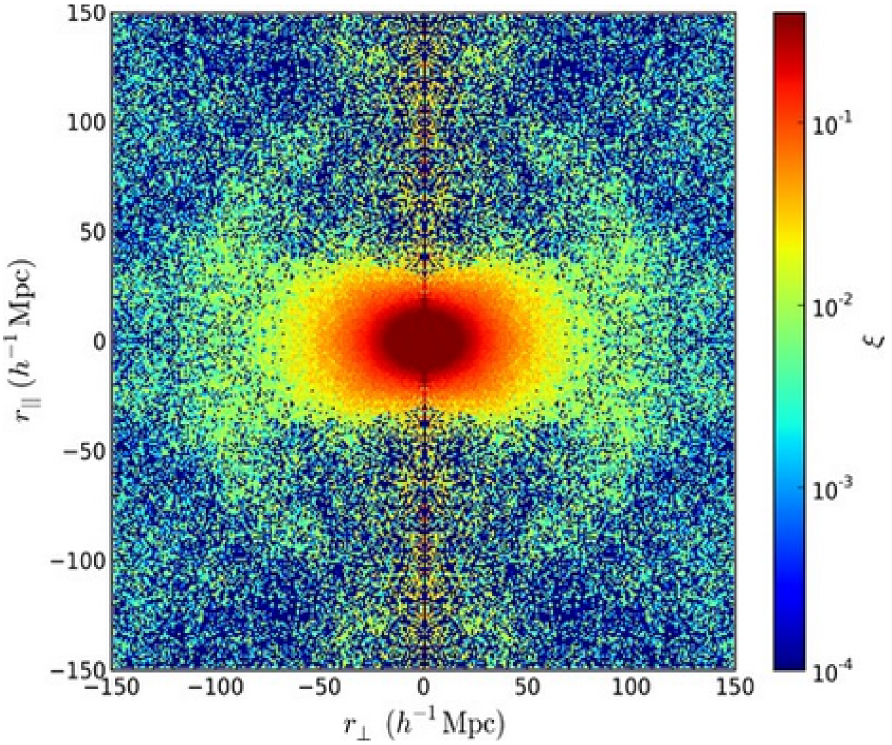


Figure 2.1: Two-point correlation function of the DR11 CMASS galaxies in BOSS. The colour coding of the plot shows the amplitude of the correlation function when divided into parallel ($r_{||}$) and perpendicular ($r_\perp$) axis with respect to the line of sight . Image Credit: Samushia et al. (2014)

there is an incentive to use an optimal set of bins for which the statistical noise is under control while preserving as much angular bin information as possible.

The standard approach for dealing with this issue (Kaiser, 1987; Hamilton, 1992) is to decompose the correlation function using Legendre polynomials

$$\xi(r, \mu) = \sum_{l=0}^{\infty} \xi_l(r) L_l(\mu), \tag{2.3}$$

where L_l is the l^{th} Legendre polynomial, and $\xi_l(r)$ is the l^{th} multipole of the 2-point correlation function, and therefore:

$$\xi_l(r) = \frac{2l+1}{2} \int_{-1}^{1} \xi(r, \mu) L_l(\mu) d\mu. \tag{2.4}$$

Note that $\xi(r, \mu)$ is an even function with respect to μ as RSD anisotropies under the small angle approximation are expected to be symmetric with respect to the line-of-sight. Considering that Legendre polynomials are (anti-)symmetric functions[*], we conclude that $\xi_l(r) = 0$ for all odd values of l.

In chapter 3 we work with the first three non zero multipoles of the 2-point correlation function (see section 3.6 for an in-depth discussion on the multipoles used in our eBOSS analysis). These multipoles are related with the following Legendre polynomials: $L_0(\mu) = 1$, $L_2(\mu) = \frac{1}{2}(3\mu^2 - 1)$ and $L_4(\mu) = \frac{1}{8}(35\mu^4 - 30\mu^2 + 3)$.

2.2 LSS tracers in cosmological surveys

It seems reasonable to expect that galaxies trace the underlying distribution of dark matter, in the sense that massive galaxies should probably inhabit inside large DM halos. However it has been shown that galaxies are not unbiased tracers of the underlying distributions of matter, this can be seen for example by correlations in surveys of different types of galaxies having different amplitudes (Peacock and Dodds, 1994; Oliver et al., 1996; Peacock, 1997), which indicates that galaxies are biased tracers of the underlying matter density field. However, it has also been shown that on large scales, like the ones needed for BAO analysis ($\sim$ 150 Mpc), the bias is expected to be linear (Coles, 1993; Scherrer and Weinberg, 1998). As a consequence, on those scales the correlation function of different types of galaxies should differ in their amplitude but only weakly in their shapes (see e.g. Peacock and Dodds (1994); Peacock (1997)). This indicates that if our goal is to estimate the shape of the correlation function one can select the galaxy type that suits its observational considerations the best. As we will see in section 2.3, the amplitude of the bias between these tracers and the underlying DM can later be added as a free parameter to the theoretical model to be fitted. The type of galaxy selected

[*] Legendre polynomials have definite parity, i.e. they are either even or odd. Mathematically this is expressed as $L_n(-x) = (-1)^n L_n(x)$

as a target for an LSS survey is usually called a *tracer*, to emphasise the fact that these galaxies are tracing the underlying DM distribution.

There are several considerations when selecting the tracer to be used in an LSS survey. For starters an LSS survey should have enough volume and density of galaxies to measure the BAO signal in the correlation function accurately, which requires a volume of at least $\sim 1\mathrm{Gpc}^3$ (Blake and Glazebrook, 2003). The number density of tracers determines the signal to noise of the measurements, the smaller the scale of the correlation function that one aims to measure the more galaxies it will need (Feldman et al., 1994). For the scales needed for BAO analysis a number density of around $\sim 1 \times 10^{-4} h^3 \mathrm{Mpc}^{-3}$ is optimal (e.g. Drinkwater et al., 2010).

Another consideration to account for is how efficient the photometric data, used for selecting the targets of the survey, will be at selecting the relevant tracers. A tracer that can be easily extracted from the photometric data ensures, among other things, that most of the observational time of the survey will be spent observing actual tracers instead of false candidates that will have to be discarded from the sample. An example of a feature of a tracer that makes it easy to be selected is the 4000 Armströng (Å)* break, which allows easy detection of galaxies that are not forming stars. This is very useful when selecting e.g. LRGs as we will discuss below. Another thing to consider is how easy it is to compute the redshift of an object, objects with strong, easily distinguishable features in their spectra require less exposure time to measure their redshifts accurately, this is the case for the strong emission lines of star forming galaxies or the clear H and K absorption features of LRGs.

Another feature to consider is the amplitude of the resulting correlation function, which as we have mentioned varies from tracer to tracer, this amplitude is determined by how clustered the galaxies selected by a given survey are. Tracers with a small amplitude in the power spectrum would make the analysis more susceptible

*The strength of the 4000 Å break is often defined as the ratio of the flux density before and after the 4000 Åbreak.

to shot noise.

In what follows we will introduce some of the more common tracers of matter that have been used in LSS surveys designed for BAO or RSD analysis.

- Magnitude limited galaxy surveys were the first type of data set used for BAO and RSD analysis. These surveys are done at low redshifts where the Malmquist bias[*] is less of an issue and standard galaxies can be observed with enough density. Some of the examples of RSD analysis made with magnitude limited surveys is the one made with the SDSS main galaxy sample (Howlett et al., 2015) that selected targets at $z < 0.2$, and the RSD analysis of the 2dFGRS survey (Peacock et al., 2002). Other upcoming surveys will also have magnitude limited surveys of galaxies at low redshift, that will be used for RSD analysis, e.g. the bright galaxy survey (BGS) of DESI (Zarrouk, 2021).

- LRGs are among the most massive galaxies and are normally associated with the centres of galaxy clusters (Kauffmann et al., 2004), where mergers and interactions with other objects are common, which explains their massive size. The red colour is a consequence of LRGs having long formed all their stars and depleted their gas content. LRGs can be selected from photometric data efficiently due to having a strong 4000 Armströng (Å) break which is common in galaxies with a relative lack of young blue stars (Bruzual A., 1983). A strong 4000 Å break also allows fast and reliable redshift measurements. For example, the more distant LRGs of the SDSS sample did not need a larger average observational time than closer and apparent magnitude brighter galaxies of the SDSS main sample to obtain accurate redshift measurements (Eisenstein et al., 2001).

  LRGs were the tracer selected by the SDSS survey for the first detection ever made of the BAO peak (Eisenstein et al., 2005). Other surveys like BOSS

[*] The Malmquist bias refers to the fact that intrinsically brighter objects are detected preferentially over intrinsically fainter objects due to their apparent brightness

(Dawson et al., 2012) ($z < 0.7$) and eBOSS ($0.6 < z < 1.0$) (Icaza-Lizaola et al., 2020; Ross et al., 2020) also used LRGs to make clustering analysis. Chapter 3 presents the first RSD analysis done with the eBOSS LRG sample of the fourteenth data release. Future surveys like DESI plan to include LRGs as one of their samples between redshifts of 0.3 and 1.0 (e.g. Zhou et al., 2020).

- An emission line galaxy (ELG) is a galaxy with strong emission lines. Such objects are primarily either active galactic nuclei (AGNs) or star-forming galaxies. Given the strength of their emission lines, they do not require large exposure times to measure their redshifts accurately. They are most common in redshifts around $0.5 < z < 2$ which are the epoch of higher star formation rate in the universe (Madau and Dickinson, 2014). The first survey that used ELGs for BAO analysis was the WiggleZ (Drinkwater et al., 2010) survey and it produced the first measurement of the BAO with a considerable sample of galaxies at a redshift larger than $z > 0.31$ (WiggleZ targeted redshifts between $0.2 < z < 1$). Other surveys have since used ELGs as matter tracers e.g. eBOSS targeted ELGs between $0.7 < z < 1.1$ (Raichoor et al., 2017). And several future surveys are expected to have an ELG sample, e.g. DESI (Raichoor et al., 2020), that will explore ELGs in the redshift range $0.6 < z < 1.6$ and EUCLID (e.g. Merson et al., 2018) that will explore the $1 < z < 2$ range. Several considerations make ELGs good candidates tracers. First, they have strong emission lines which make the redshift measurement easy, secondly, these galaxies have strong emission in the UV spectra at the desired redshifts which makes them easy to select from photometric data. And finally, they can be observed over similar redshifts as LRGs, therefore both samples can help constrain systematic effects arising from the tracer selected.

- Quasars are some of the brightest objects in the universe and can be seen to very large distances, and are usually selected as tracers for RSD analyses at large redshifts ($z \gtrsim 2$). Quasars are AGNs that are so bright that they outshine their host galaxies which become undetectable (Kembhavi and Narlikar, 1999). Therefore they look like an unresolved source of light, which gives them their name of *quasi-stellar objects* or quasars for short. Their huge brightness allows them to be seen at very large distances and they are amongst the farthest objects that we have observed. Quasars were much more common in the past, peaking in density around $z \approx 2$ (Schmidt et al., 1995) which is unofficially referred to as the *era of quasars*.

The first quasar survey to be used for LSS analysis was the 2dF QSO survey (Croom et al., 2001) that targeted quasars at $z < 3$. Other important quasar surveys are the BOSS survey (Ross et al., 2012) that included all objects that were visually inspected and confirmed to be quasars, and the eBOSS survey that is centred in the redshift range around the era of quasars $0.9 < z < 2.2$ (Myers et al., 2015). Future surveys like DESI (Yèche et al., 2020) will also use quasars to study RSD at high redshifts.

2.3 Theoretical models that account for RSD

Cosmological perturbation theory (Fry, 1984; Bharadwaj, 1994) studies the evolution of structure in the universe. The theory models how small perturbations evolve through gravity and makes predictions of the overdensity field $\delta(\vec{x}, t)$. Given that an accurate model of $\delta(\vec{x}, t)$ can be used to model the correlation function (equation 1.9), this becomes an appropriate framework to model the correlation function. Perturbation theory usually focuses on matter as the only relevant component and assumes that gravity is the only relevant interaction on all scales except the smaller ones, where baryonic physics becomes important. The standard approach is to treat matter as a pressureless dust that is characterised at any time by its mass density $\rho(\vec{x}, t)$ and its peculiar velocity $\vec{v}(\vec{x}, t)$, and then to use hydrodynamical equations like the Euler, Poisson, and continuity equations (e.g. Hui and Bertschinger, 1996) to evolve these quantities. In general, a direct evaluation of $\rho(\vec{x}, t)$ is not feasible due to the nonlinear coupling between parameters[*].

These coupling terms however become negligible when $\delta \ll 1$ and the velocities are small (Peebles, 1980). The ranges when this happen are known as the linear regime. This regime is particularly important as it corresponds to most of the perturbation's lifetime. Today the linear regime corresponds to scales larger than $20 - 40\,h^{-1}$ Mpc (depending on the tracer).

One difficulty when comparing a predicted correlation function from perturbation theory to observations is the fact that we do not see the actual matter distribution of the universe. Instead, we observe galaxies, which are biased tracers of the underlying matter distribution. The bias function, B, relates the density of galaxy tracers, $\delta_g(\vec{x})$, to the underlying matter distribution, $\delta(\vec{x})$, so that $\delta_g(\vec{x}) = B(\delta(\vec{x}))$. The general approach to deal with this issue is to add free parameters that determine the bias function as part of the methodology that needs to be fitted by the model. For example, the parameters F' and F'' that are fitted by our methodology in chapter 3 are related to a first and second-order approximation of the bias function in the so-called Lagrangian space that will be introduced below. It is common to make a linear approximation of the bias function so that $\delta_g(\vec{x}) \approx b\ \delta_m(\vec{x})$. The constant b is usually referred to as the *linear bias.*

Note that this means that the correlation function of the underlying matter $\xi(\vec{r_1}, \vec{r_2}) = \langle \delta(\vec{r_1})\delta(\vec{r_2}) \rangle$ and the correlation function of matter tracers $\xi_M(\vec{r_1}, \vec{r_2}) = \langle \delta_M(\vec{r_1})\delta_M(\vec{r_2}) \rangle$ should be related by the following expression in the linear regime:

$$\xi(\vec{r_1}, \vec{r_2}) \sim b^2 \xi_M(\vec{r_1}, \vec{r_2}). \tag{2.5}$$

[*] The mechanics of matter tracers in dense environments are affected by the density in neighboring regions e.g. galaxies inside a galaxy cluster are gravitationally disturbed by other bodies inside the cluster which makes predicting the dynamics of the tracer an N-body problem without an analytic solution. However, when a matter tracer has just started its gravitational collapse into a cluster and is still reasonably far from the centre of the cluster, all the interactions can be approximated by a single gravitational pull, in the direction of the centre of mass of the cluster.

In what follows I will present a brief derivation of one of the most famous RSD models: the Kaiser formula (Kaiser, 1987), a RSD model that is valid in the linear regime. The Kaiser model is a perfect example of a RSD model that computes the Fourier transform of the correlation function in redshift space, as a function of both the growth rate, f, and the linear bias, b. Let us start our derivation by noting that the number of galaxies in redshift and real space should be conserved,

$$n^s(\vec{s})d^3s = n(\vec{r})d^3r, \tag{2.6}$$

where the superscript s denotes quantities in redshift space. The relation between the vector in real space $\vec{r}$ and in redshift space $\vec{s}$ is given by equation 1.19:

$$\vec{s} = \vec{r} + V_{los}\hat{r}, \tag{2.7}$$

here we have assumed that the position vector $\vec{r}$, points in the line-of-sight direction, and we have defined the normalised peculiar velocity vector $\vec{V} = \vec{v}/aH$. The peculiar velocity vector is $\vec{V} = (V_{los}\hat{r} + V_{\perp}\hat{r}_{\perp})$, where V_{los} and $V_{\perp}$ are the parallel and perpendicular components of the peculiar velocity with respect to the line-of-sight and $\hat{r}_{\perp}$ is a unit vector perpendicular to $\vec{r}$ (and therefore to the line-of-sight).

We can use equation 1.8 to write $\delta^s(\vec{s}) = (n^s(\vec{s}) - \bar{n}^s(\vec{s}))/\bar{n}^s(\vec{s})$. Plugging this definition and equation 2.7, into equation 2.6 it can be rewritten as:

$$\delta^s(\vec{s}) + 1 = \frac{r^2\bar{n}(\vec{r})}{(r + V_{los})^2\bar{n}(\vec{r} + V_{los}\hat{r})}(1 + \frac{\partial V_{los}}{\partial r})^{-1}[1 + \delta(\vec{r})] \tag{2.8}$$

This equation is valid in all regimes. In the linear theory, it should be true that $V_{los} < V \ll r$ which states that the peculiar velocity correction to the distance is small. Also in the linear regime the density perturbations should be small and therefore $\delta \ll 1$. Equation 1.15 states that $\vec{\nabla}\cdot\vec{V} = -\delta f$, and therefore $\partial V_{los}/\partial r \ll 1$ given that f is independent of scale. Therefore equation 2.8 is simplified to[*]

$$\delta^s(\vec{s}) = \delta(\vec{r}) - \frac{\partial V_{los}}{\partial r}. \tag{2.9}$$

Now from equation 1.15 we can write V_{los} as

$$V_{los} = -f\frac{\partial}{\partial r}\nabla^{-2}\delta(\vec{r}), \tag{2.10}$$

where ∇^{-2} is the inverse function of the Laplacian operator, and we conclude that

$$\delta^s(\vec{s}) = (1 + f\frac{\partial^2}{\partial r^2}\nabla^{-2})\delta(\vec{r}). \tag{2.11}$$

This equation could be used to compute the correlation function in redshift space as a function of their correlation in real space. However, when working with RSD in the linear regime it is common to work in Fourier space due to the k modes evolving independently (although consider e.g. Fisher (1995); Reid and White (2011) for a treatment of the linear scales in configuration space). Let us define $P(k)$ as the Fourier transform of the correlation function in real space (usually referred to as the *power spectrum*).

And let us write equation 2.11 in Fourier space

$$\hat{\delta}^s(\vec{k}) = (1 + f\mu_{\vec{k}}^2)\hat{\delta}(\vec{k}), \tag{2.12}$$

where $\mu_{\vec{k}}^2 = k_z^2/k^2$ and we have used the fact that in Fourier space the operator $\frac{\partial^2}{\partial r^2}\nabla^{-2}$ becomes k_z^2/k^2. Then the power spectrum in redshift space is given by the following relation

[*] Here, and in the rest of this derivation we are also assuming the plane-parallel approximation, see Hamilton (1998) for a discussion of these equations without that approximation.

$$P^s(k,\mu) = (1 + f\mu^2)^2 P(k,\mu), \tag{2.13}$$

Finally, we add the linear bias relation $\delta_g = b\delta$ to arrive to the standard expression of the Kaiser formula

$$P^s_g(k,\mu) = b^2(1 + f\mu^2)^2 P(k). \tag{2.14}$$

Let us note that so far we have worked in linear scales, unfortunately, some of the strongest cosmological constraints come from smaller distance scales that are where the linear scale is less valid. One common approach to go beyond linear theory is to add a random motion to the predictions of linear theory, which is known as the streaming model (Reid and White, 2011; Reid et al., 2012). These random motions are taken from a distribution of velocities that is dependent on the scale. Following e.g. Peacock (1998) and Scoccimarro (2004), one can consider models that modify the Kaiser formula to predict $P^s_g(k,\mu)$ beyond linear theory using a Gaussian streaming model, where, as the name suggests, the random motions are taken from a Gaussian distribution.

So far we have discussed perturbation theory in what is usually referred to as Eulerian coordinates, where one tries to determine the density and velocity fields, i.e. $\rho(\vec{x},t)$ and $\vec{v}(\vec{x},t)$ respectively. However, one of the most successful approaches to model the density field in redshift space is to make a change of coordinates and instead model the displacement field $\vec{\Psi}$ defined as the vector connecting the position $\vec{q}$ of a particle at $t = 0$ with its position at the later time t:

$$\vec{x}(\vec{q},t) = \vec{q} + \vec{\Psi}(\vec{q},t). \tag{2.15}$$

Note that any particle is uniquely identified by $\vec{q}$ and that the field $\vec{\Psi}(\vec{q},t)$ is sufficient to specify its evolution. This approach is called Lagrangian Perturbation

Theory (LPT) (Zel'Dovich, 1970; Matsubara, 2015). The model that we present in section 3.4 is an example of a LPT model.

The standard approach to model $\vec{\Psi}(\vec{q},t)$ is to write it as a perturbative solution:

$$\vec{\Psi}(\vec{q},t) = \vec{\Psi}^{(1)}(\vec{q},t) + \vec{\Psi}^{(2)}(\vec{q},t) + \vec{\Psi}^{(3)}(\vec{q},t) + \vec{\Psi}^{(4)}(\vec{q},t) + ..., \tag{2.16}$$

where $\vec{\Psi}^{(n)}(\vec{q},t)$ is of the order of $(\vec{\Psi}^{(1)}(\vec{q},t))^n$. The solution to the linear term is usually referred to as the Zeldovich approximation (Zel'Dovich, 1970), and it is given by the following equation (Bernardeau et al., 2002):

$$\nabla_q \cdot \vec{\Psi}^{(1)}(\vec{q},t) = -D(t)\delta(\vec{q}) \tag{2.17}$$

where $D(t)$ is the growth function from equation 1.12 (here we are using the variable t to measure time instead of the scale factor a), that quantifies how much structures have grown in a given cosmological model. Note that if the collapse is irrotational (as one would expect in the linear regime of gravitational collapse), the divergence of the displacement field should contain all of the required information about the gravitational collapse of matter. Matsubara (2015) discusses analytical approaches to go beyond linear theory. Analytical solutions of LPT are known to the fourth-order (Rampf and Buchert, 2012). We revisit the LPT approach in chapter 3 when the LPT model used in Icaza-Lizaola et al. (2020) is discussed.

2.4 The Alcock-Paczynski effect

So far we have considered RSD as the only type of distortions apparent from the way we measure structures in the universe. However, other phenomenon can affect our distance measurements, for example, a poor choice of cosmological model that is used to compute distances of galaxies.

A LSS survey measures two observables: the redshifts and angular positions of objects. Then one uses these measurements to infer a distance, which requires us

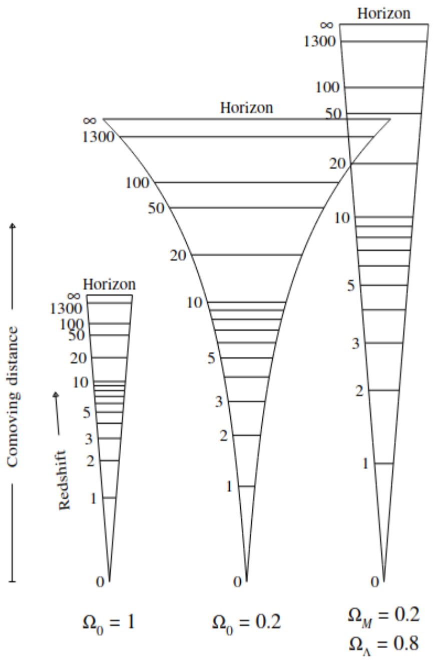


Figure 2.2: Each light-cone represents an aperture of fixed angle for three different cosmological models. The model on the left represents a model with no dark energy, the model in the middle shows a model of a universe that is not flat, and the model on the right shows the ΛCDM model. The plot shows how different cosmological models measure radial and angular distances differently. This figure is taken from Hamilton (1998).

to assume a cosmological model. If this model were not the *true* cosmology of the universe, then the distance measured would be wrong and the correlation functions that we measure would be distorted.

To understand this let us look at figure 2.2 where the light-cone of three universes with different cosmological models of fixed aperture are shown. We can see that the cosmological model determines both the comoving distance, which at small z is given by $r = cz/H(z)$ due to the Hubble law (using equation 1.3), and the so-called angular diameter distance $D_A(z)$, defined as $D_A = D/\partial\theta$ for a light source of diameter D that subtends an angle $\partial\theta$ on the sky. In a flat universe, D_A is related to the comoving distance r through the relation $D_A = r/(1+z)$.

Let us refer to our selected cosmology as the *fiducial* cosmology, to differentiate it from the true cosmology of the universe. Then let us predict how a spherical shell of galaxies of diameter D would look if someone would try to map it using the fiducial cosmology that is significantly different from the true cosmology. The distance between galaxies at different ends of the shell and along the line of sight will be determined by the Hubble law to have a value D^{fid}, where the superscript fid emphasis that the measurement is done in the fiducial cosmology. Then the angle that the object would need to subtend in the sky in order for the shell to look like a sphere will be $\partial\theta^{fid} = D_A^{fid}/D^{fid}$, which, as shown by figure 2.2, is different from the angle that one would observe (as this angle is determined by the true cosmology).

This indicates that if the cosmology selected was not the true cosmology of the universe the sphere would look like an ellipsoid. Given that we know that the distribution of matter is isotropic on large scales (due to the cosmological principle) finding different clustering signals along the line-of-sight and transverse directions (after correcting for RSD) will suggest that the fiducial cosmology selected is different from the true cosmology. Let us note that as a consequence the BAO signal will have an offset along the line of sight and along its perpendicular direction.

Alcock and Paczynski (1979) proposed to include any possible distortions in the data as free parameters of the model to be fitted alongside the rest of the free parameters (in our case our RSD parameters). In the rest of this section we introduce a common parametrisation, which is usually referred to as the Alcock-Paczynski parameters.

Let us first consider the case where one does not analyze the differential clustering along the line-of-sight and transverse directions separately (for example in a BAO analysis without RSD). In this case, the anisotropic shift between the line-of-sight and perpendicular directions would not be measured but there will still be a shift in the position of the BAO peak due to the fact that the predicted comoving distance r^{fid} will be different from the true distance. As a consequence, the BAO peak will be displaced.

We define the spherically average distance to an object as

$$D_V(z) = \left[(1+z)^2 D_A^2 \frac{cz}{H(z)}\right]^{1/3}. \tag{2.18}$$

This is the average of three comoving distance measurements, two of them computed using the formula $r = D_A(1+z)$ (this accounts for the fact that there are two radial directions perpendicular to the line of sight) and the other computed along the line of sight as $r = cz/H(z)$. One can then add the following Alcock-Paczynski parameter to the list of free parameters explored by the methodology,

$$\alpha = \frac{D_V(z) r_s^{fid}}{D_V^{fid}(z) r_s}, \tag{2.19}$$

where r_s is the BAO scale. Let us note that if our parameter space exploration suggests that the best fit models are such that α deviates significantly from $\alpha = 1$ it would suggest that the fiducial cosmology used to build the model and to analyse the data needs to be revised.

When one analyses the differences in clustering along the line-of-sight and transverse directions, like in a RSD analysis, one can instead use the following Alcock-Paczynski parameters.

$$\alpha_\perp = \frac{D_A(z) r_s^{fid}}{D_A^{fid}(z) r_s}, \ \alpha_\parallel = \frac{H(z)^{fid} r_s^{fid}}{H(z) r_s}, \tag{2.20}$$

where again, if the best-fit parameters have significant deviations from $\alpha_\parallel = 1$, or $\alpha_\perp = 1$, it would indicate that the fiducial cosmology needs to be revised.

Let us also note that the values of $D_A^{fid}(z)$ and $H(z)^{fid}$ used in an analysis are known, therefore an estimate of the regions of high likelihood of the parameters $\alpha_\perp$ and $\alpha_\parallel$ corresponds to an estimate of $D_A(z) r_s^{fid}/r_s$ and $H(z) r_s/r_s^{fid}$. These estimates are a sought-after result of a RSD analysis and are considered, along with an estimate of the growth rate f, the main results of a RSD analysis.

2.5 Galaxy mocks

We stated that one of the necessary constituents of a RSD analysis is an estimate of the uncertainty on the multipoles of the correlation function measured from the observational survey. It has been shown (e.g. Norberg et al., 2009) that methodologies that measure the error from the data itself like the jackknife and the bootstrap methods (Efron, 1982) have limitations in their use in a LSS analysis.

Another possible approach to estimate the uncertainty of the multipoles is to generate artificial data sets to estimate the expected uncertainty in the correlation function. We will refer to a collection of artificial sets as *mock catalogs*, where each artificial data set is called a mock.

The number of mocks required to make an accurate estimate prediction depends on the size of the observational data set and on the accuracy of the error estimate required by the analysis. Modern LSS surveys require several hundreds of mocks, for example, Percival et al. (2014) suggests that at least 600 mocks are needed to analyze the clustering of the BOSS samples.

These mocks are not only useful to estimate the errors of the data but they can also be used to test the methodology of a RSD analysis. The idea is to plug in the mock catalogs, which were built with a known set of cosmological parameters, into the fitting methodology. If one can not recover the expected value of the parameters within the predicted error (that is obtained by an exploration of the parameter space as discussed in section 2.6) this might indicate a problem. These tests represent an important part of our RSD analysis in eBOSS, and are presented in section 3.6.

One possible approach is to generate halo catalogs using N-body simulations, and then populate these halos with galaxies. Unfortunately, it is not easy to produce hundreds of accurate N-body simulations, given how computationally expensive each of them is[*]. Therefore, there is a strong incentive to generate approximate methods to create cheaper mocks.

Over the last decades, there have been several attempts to generate cheap and reasonably accurate mocks that have been used in LSS analysis.

One of the cheaper and faster methods to create mock catalogs is to generate so-called log-normal mocks (Coles and Jones, 1991), where one assumes that the probability density function (PDF) of the galaxy or halo density fields should follow a log-normal distribution. This assumption is based upon the observation that the PDF of the log of density fields, $\ln(1+\delta)$, measured from N-body simulations roughly matches a Gaussian PDF (e.g. Coles and Jones, 1991). The standard approach is to feed these mocks with an input correlation function $\xi(r)$ and transform it into log space $\xi_{\ln}(r) = \ln(1+\xi(r))$, this new correlation function can then be used to compute a log-normal density field $\delta_{\ln}$. Log-normal mocks have the advantage that their statistics are completely determined by the input two-point correlation function.

This is shown in the first panel of figure 2.3 where the correlation function in real space of several mocks is compared to a reference halo catalog extracted from the BigMultiDark simulation. We can tell that the log-normal mocks (yellow lines) make accurate predictions of the two-point correlation function.

Note that no velocity field was assigned to the log-normal mocks. This could in principle be done (e.g. White et al., 2013) but it was not a part of the analysis done in Chuang et al. (2015), therefore there is no estimate of log-normal mocks in redshift space.

The displacement of matter through time is not predicted by log-normal mocks and as a consequence the observed pattern of the cosmic web is not properly reproduced. This can be seen in the fourth panel of figure 2.3, which shows that log-normal

[*] Although in recent years simulation catalogs have became more efficient and codes like the Abacus Summit (Garrison et al., 2021) are capable of realising hundreds of N-body simulations of the required volume for DESI like surveys.

mocks make a poor model of the three-point correlation function[*].

More complex mocks that address this problem can be made with methods that try to predict the gravitational motion of particles using perturbation theory. Two codes have pioneered this method, mainly the PT Halos (Manera et al., 2012; Manera et al., 2015) and the PINOCCHIO (Monaco et al., 2002) algorithms.

PT Halos uses second-order LPT to create a field of DM particles and identifies halos in this field using an FoF algorithm with a linking length that is optimally selected following the procedure of Manera et al. (2012).

PINOCCHIO uses an ellipsoidal collapse model[†] solved in third-order LPT, to compute the time at which the elements of an initial linear density field collapse.

The accuracy of both PINOCCHIO and PT Halos is limited by the accuracy of the perturbation theory that they use, which is not very accurate in the highly non-linear regime, which results in limited success in modelling halos hosting galaxies. This is shown in the third panel of figure 2.3, where one can see that both models struggle to reproduce the power spectrum of the reference halo catalog at large values of k (that correspond to small perturbation sizes).

To improve the description of smaller scales, a set of methodologies that try to fit the mocks to statistics of a target simulation have been developed. In what follows we introduce some of the more well-known methodologies.

- The quick particle mesh (QPM) methodology (White et al., 2013) consists in running fast N-body simulations with low accuracy, that are later calibrated to reproduce statistics of a full N-body simulation built with the Tree-PM code (White, 2002). The simulations are built using the particle mesh methodology described in chapter 1.3.2, which gives them their name. The simulation speed is achieved by reducing the number of time steps used.

[*] $\zeta = \langle \delta(\vec{r_1})\delta(\vec{r_2})\delta(\vec{r_3}) \rangle$.

[†] Similar principle to a spherical collapse model but assuming that the original overdensity is not symmetric along the main axis.

In essence, the QPM simulation makes a trade-off between speed and accuracy, as each time step requires computing the position and velocity of the particles and requires memory to store the results. QPM methods use around ten time steps, to be compared to, e.g., the 11,000 time-steps required by the Millennium simulation (Springel et al., 2005), which give us an idea of how fast these methods can be.

The resulting models do not have enough resolution to resolve halos inside of them so some particles are selected as halos with a probability dependent on their density (particles in denser regions are more likely to be halos). As with log-normal mocks the sampling function from where these probabilities are taken is given by a Gaussian in $\ln(1+\delta)$ space, then the mean of the Gaussian is fitted to reproduce the large-scale bias of halos in the Tree-PM simulation. The mass of the halo is also a function of density and they are assigned in such a way that they reproduce the Tree PM mass function. Finally, galaxies are assigned to halos using a method known as Halo Occupation Distribution. We will describe this further in section 4.3, but in short the method estimates the probability $P(N|M)$ of finding N galaxies of a given type or mass in a DM halo of mass M. The halo occupation distribution (HOD) used is in QPM mocks described in (Tinker et al., 2012)

- The effective Zeldovich (EZ) methodology (Chuang et al., 2015) uses the Zeldovich approximation of equation 2.17 to make a first-order approximation of the DM density field at each grid point. Then this density field is populated with matter tracers (halos or galaxies) by mapping a target density field of tracers into the grid following the methodology described in Chuang et al. (2015). The target density field is measured from either a full N-body or a LSS survey. Once these first tracers are added to the grid several other steps are done to improve the accuracy of the resulting mock. These steps include adding scatter to the relation, including density thresholds/saturation parameters that modify the amplitude of the resulting power spectrum, adding a parameter that enhances the BAO signal, and adding a parameter that modifies the shape of the initial power spectrum. All of these steps add several free parameters to the model, which are fitted so that the resulting mock reproduces the statistics, including the 2-point and 3-point correlation functions, of the target observations/N-body simulations.

- PATCHY also uses LPT to evolve the position of particles by modelling the displacement field ($\Psi(\vec{q}, t)$). In this case, it uses a LPT model known as Augmented Lagrangian Perturbation Theory (Kitaura and Heß, 2013) (instead the Zeldovich approximation is used by EZ mocks). The model works by using a spherical collapse model ($\Psi_{SC}(\vec{q}, t)$) to model the linear scales and second-order LPT ($\Psi_{2LPT}(\vec{q}, t)$) to add a correction on smaller scales. The density field is given by

$$\Psi(\vec{q}, t) = \Psi_{SC}(\vec{q}, t) + \Psi_{2LPT}(\vec{q}, t) \tag{2.21}$$

  After evolving the particles one has a catalog of evolved particles at time t, from where one can compute the DM density. The next necessary step is to determine the bias relation between galaxy density and DM density. This is done by fitting a bias model to the two- and three-point statistic of a target distribution.

  The target distribution comes from the Multi Dark N-body simulation Halo catalog that is filled with galaxies using a Halo Abundance Matching (HAM) algorithm calibrated to a LSS survey as described in Rodríguez-Torres et al. (2016).

- The *Comoving Lagrangian acceleration* (COLA) mocks (Tassev et al., 2013) uses second-order LPT to compute the displacement of ($\vec{r}(\vec{q}, t)_{2LPT}$). As stated earlier, 2LPT solves the largest scales accurately but struggles with the smaller scales. To tackle this issue, COLA uses a particle mesh with few time steps (similar to QPM mocks) to compute the residual displacement ($\vec{r}(\vec{q}, t)_{\text{residual}}$) of particles with respect to the trajectory computed

> with LPT. The accuracy at small scales is controlled by the number of time steps used in the particle mesh, with the authors proposing 10 steps starting at a redshift of $z = 9$. The trajectory of a particle is then computed as $\vec{r}(\vec{q}, t) = \vec{r}(\vec{q}, t)_{LPT} + \vec{r}(\vec{q}, t)_{\text{residual}}$.

Figure 2.3 shows how the statistics of EZ, PATCHY and COLA mocks compare with the reference halo catalog extracted from the BigMultiDark simulation (unfortunately QPM mocks were not part of this study). The plot shows how these models agree well with the reference halo in all four comparisons, making them the most reliable mocks of the ones presented so far.

Once one has built a set of hundreds of mock simulations one can compute the two-point correlation functions of each of them individually. These estimates can be used to compute their covariance matrix, which we use as the estimate of the uncertainty of the measured correlation function. The covariance matrix is estimated using equations 3.21 and 3.22 of section 3.5.

2.6 Exploring the parameter space

In the previous sections, we have introduced the four key components required for a RSD model comparison with observational data.

First the values of the multipoles of the correlation function are computed from an observational data set. A LSS survey provides the distance (in the form of redshifts) to N galaxies, that we use to measure the correlation function $\xi(s, \mu)$ in R distances $s_i = [s_1, .., s_R]$ and M angular parameters $\mu_i = [\mu_1, .., \mu_M]$. Then one can use equation 2.4 to compute the multipoles of the correlation function. Let us suppose that one computes l multipoles, where $[0, 2, ..., 2L = l]$ and $L\epsilon\mathbb{N}$ (let us remember that $\xi_{l_i}(s) = 0$ for odd values of l_i), then we define the data vector $\vec{D} = [\xi_0(s_1), .., \xi_0(s_R), .., \xi_l(s_1), .., \xi_l(s_R)]$.

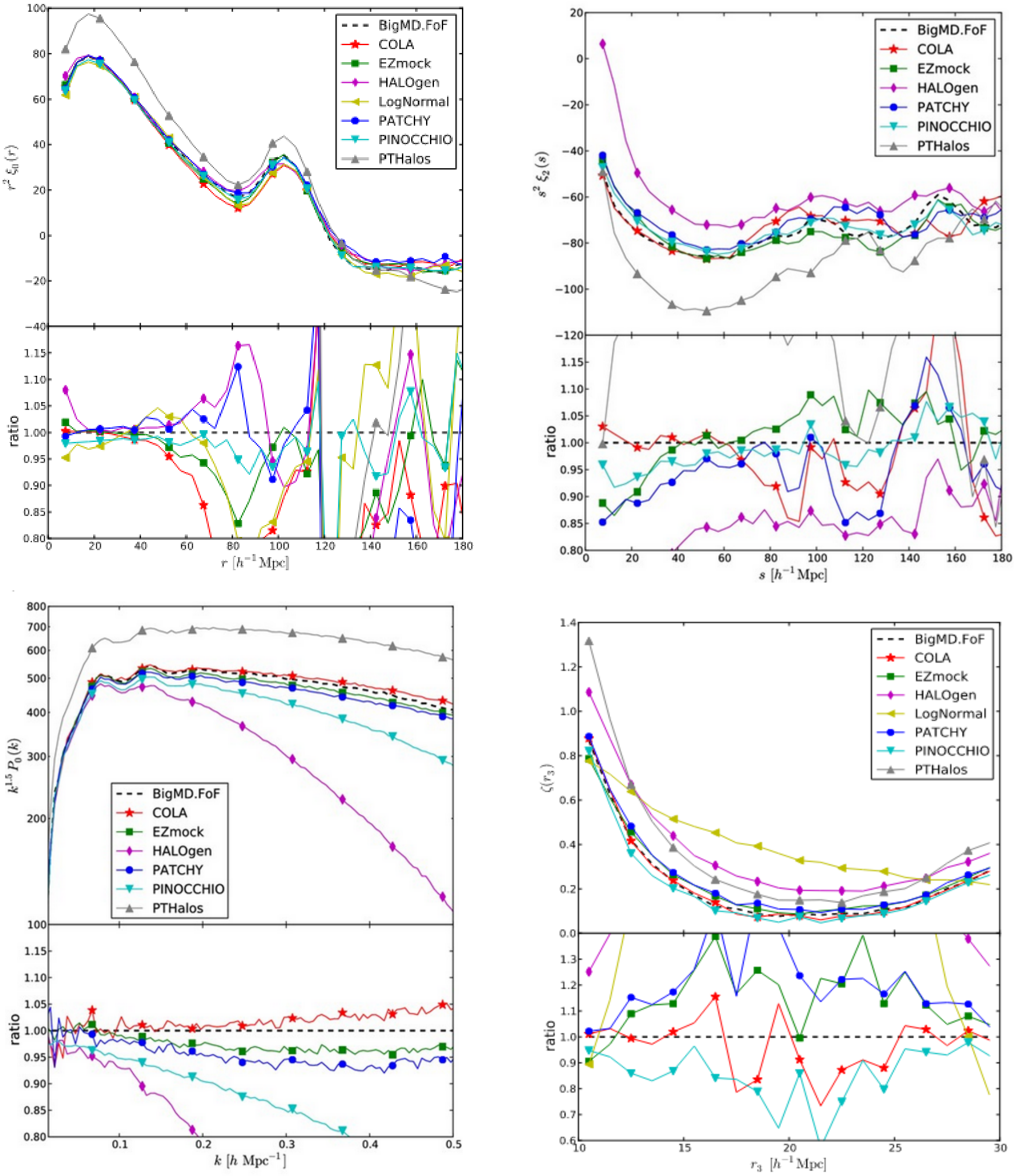


Figure 2.3: Performance results of different correlation statistics for several mock simulations represented by the coloured lines. The dashed lines show the statistics extracted from the BigMultiDark simulation (Klypin et al., 2016), which is a full N-body simulation and is used as a reference catalog. The top-left panel shows the correlation function in real space. The top-right panel shows the quadrupole of the correlation function in redshift space (noting that the differences between the mocks are more pronounced in the quadrupole terms compared to the monopole terms.). The third panel shows the monopole of the power spectrum in redshift space. Finally, the last panel shows the three-point correlation function in real space. Image credit: (Chuang et al., 2015)

The second constituent is a theoretical model of the used multipoles that depends on a set of free parameters P that should include the Alcock-Paczynski parameters and the growth rate f amongst other parameters. We define the vector of model predictions of the correlation function $\vec{M}(P) = [\xi_0^M(P, s_1), .., \xi_0^M(P, s_R),$ $.., \xi_l^M(P, s_1), .., \xi_l^M(P, s_R))]$ where the superscript M in $\xi_{l_i}^M$ emphasises the fact that we refer to a modeled value of the correlation function, built using the parameters P.

Finally, the last constituent is an estimate of the errors on the observed multipoles measured from mock catalogs and used to compute a covariance matrix, $C_{i,j}$. If the vector $\vec{D}$ has K elements, then $i, j < K$.

We define the χ^2 (Lupton, 1993) merit function of our model as

$$\chi^2(D, P) = [M(P) - D]^T C^{-1} [M(P) - D]. \tag{2.22}$$

χ^2 is an estimate of the goodness of the fit and has a small value when the model with parameters P predicts the data accurately.

In the rest of this chapter, we present a methodology to explore the parameter space and find the regions P of the space that is more likely to explain an observed data D, this is usually referred to as the *posterior* probability $\mathcal{L}(P \mid D)$. The exploration is done using a sampling methodology known as a Monte-Carlo Markov chains (MCMCs) (e.g. Knox et al., 2001; Dunkley et al., 2005). However, before presenting this method I introduce the equations used to compute the posterior at a given point in parameter space. Let us define the likelihood $\mathcal{L}(D \mid P)$ as the probability of observing the data D given our model built with parameters P. The Bayes theorem indicates that the posterior and the likelihood are related as

$$\mathcal{L}(P \mid D) = \frac{\mathcal{L}(D \mid P)\mathcal{L}(P)}{\mathcal{L}(D)}, \tag{2.23}$$

where $\mathcal{L}(P)$ is known as the *prior* and expresses our knowledge on the parameter space, and $\mathcal{L}(D) = \int_\Theta \mathcal{L}(D \mid P)dP$is known as the marginal probability. As we will see, the MCMC method that we use to explore the parameter space with only computes the ratio of the posterior between different points, i.e. $(\mathcal{L}(P_i \mid D)/\mathcal{L}(P_j \mid D))$ and therefore there is no need to compute $\mathcal{L}(D)$ (as it is canceled out).

If we asume that the likelihood should be a multivariate Gaussian then:

$$\mathcal{L}(D \mid P) = \frac{1}{\sqrt{(2\pi)\mathrm{Det(C)}}} e^{-\frac{1}{2}[M(P)-D]^T C^{-1}[M(P)-D]}. \tag{2.24}$$

This can be rewritten using equation 2.22 as

$$\mathcal{L}(D \mid P) \sim e^{-\chi^2(D,P)/2}, \tag{2.25}$$

noting that a large value of $\mathcal{L}(D \mid P)$ corresponds to a small value of χ^2.

It is common practice to introduce the priors $\mathcal{L}(P)$ using an N-dimensional interval I in parameter space, where $\mathcal{L}(P) = 1$ if $P \epsilon I$ and $\mathcal{L}(P) = -\infty$ otherwise. This is usually referred to as using *flat* priors. In summary, the posterior definition at a given point P is computed using the following expression

$$\mathcal{L}(P \mid D) \sim \begin{cases} e^{-\chi^2(D,P)/2}, & \text{if } P \epsilon I \\ -\infty, & \text{otherwise} \end{cases} \tag{2.26}$$

Exploring the parameter space of a RSD analysis can be a challenging task, this is because the number of points in the parameter space that need to be explored can be large, and the evaluation of the likelihood at each point can be relatively slow.

There are several reasons that contribute to this issue: the complexity of the theoretical models can make them relatively slow to execute[*]; the relatively large number

[*] In the analysis presented in chapter 3 estimating the vector $M(f,P)$ takes around three seconds for a set value of the parameters.

of parameters needed and the correlations between some of these parameters[*]; systematic errors within the data can create degenerate solutions[†], which can make the relevant space very large to explore.

With this in mind, it is useful to use an algorithm where the regions of high likelihood are explored more carefully than regions of small likelihood. A common approach is to use MCMCs. This is the method selected for the RSD analysis of the final data releases of some of the most popular surveys, e.g. the BOSS main galaxy sample (Alam et al., 2017) and the eBOSS final data release analysis of Quasars (Neveux et al., 2020), ELGs (de Mattia et al., 2021) and LRGs (Bautista et al., 2021), and it is the method that we use in the analysis of chapter 3.

MCMC chains are designed to find the regions of high posterior probability in the parameter spaces. However, they should agree with the results of methods that acquire less information faster. For example minimisation routines like the Powell methodology (Press et al., 2002) can find the best-fit parameters (that correspond to the point of maximum likelihood in the parameter space) significantly faster than what it would take to run a full MCMC chain. This can be used as a computationally cheap consistency test of a MCMC chain (that should find the same best-fit parameters).

Minimisation methods can also be used to explore the parameter space when one has several cosmological data sets, for example, when one has built a set of N mock catalogs, one can run a minimisation routine on all N mocks individually and obtain N estimates of the parameters. Given that one knows the values of the parameters P used to build the mocks, it would be expected that the mean estimated value of those P parameters from those N mocks would agree well with the assumed values of those P input parameters. This can be used to test the full RSD analysis methodology. Such analysis is presented for our methodology when discussing figure 3.12 in the next chapter.

[*] Our analysis required six free parameters (see section 3.4).

[†] See e.g. section 3.10.1.

2.6.1 MCMC chains

In the rest of this section we introduce the MCMC algorithm. The algorithm works by walking in the parameter space computing the likelihood at every step that it takes, making a chain of likelihoods in which each point is correlated to the former.

In order to decide in which direction to move the algorithm first takes a random trial step. Let us denote as $\mathcal{L}(P_i \mid D)$ and $\mathcal{L}(P_{i+1} \mid D)$ the likelihoods at the current location and at the trial position. If $\mathcal{L}(P_{i+1} \mid D) > \mathcal{L}(P_i \mid D)$ then the algorithm chooses the trial position as its new destination. If $\mathcal{L}(P_{i+1} \mid D) < \mathcal{L}(P_i \mid D)$ then the algorithm has a probability $\mathcal{L}(P_{i+1} \mid D)/\mathcal{L}(P_i \mid D)$ of accepting the trial step as its new position, and a probability $1 - \mathcal{L}(P_{i+1} \mid D)/\mathcal{L}(P_i \mid D)$ of rejecting it. This ensures that every point in the parameter space could eventually be part of the chain, but that most of the running time will be spent exploring the region of high likelihood.

One important configuration parameter of a MCMC chain is the step size of the algorithm. Note that a very small step size would correspond to a very large acceptance rate which makes the method more similar to a random walk, while on the other hand a very large step size would correspond to a low acceptance rate, both of which will make the space exploration very inefficient[*]. The standard approach to compute the step size is to run one first chain with your *best guess* of the step size and use this chain to compute the covariant matrix of your parameters which can then be used to compute the optimal step size. Following Dunkley et al. (2005), as a rule of thumb an optimal MCMC should have an acceptance rate between 0.2 and 0.4.

[*] Note that in both cases the chain would theoretically eventually converge, as a random walk would eventually explore the whole space and a small acceptance rate will work as a very slow MCMC.

2.7 Conclusions

So far we have introduced the five different constituents that are needed to do a RSD analysis. These are:

- A data set with redshifts of galaxies that are members of a pre-determined family of tracers of matter. From these galaxies, one computes the multipoles of the 2-point correlation function.
- A theoretical model that predicts these multipoles as a function of a set of free parameters. These parameters usually include the growth rate and parameters that determine the bias of tracers with respect to the underlying DM distribution.
- One can also add a set of free parameters that allows for the possibility that one selects for the analysis a cosmological model that is different from the *true* cosmology of the universe. This can be done with the Alcock-Paczynski parametrisation. A deviation from the expected values when doing a likelihood exploration in the parameter space is a red flag that suggests that the fiducial cosmology needs to be revised. Let us also note that an estimate of the Alcock-Paczynski parameters is equivalent to an estimate of $H(z)$ and $D_A(z)$.
- An estimate of the covariance matrix of the multipoles, which is usually done by building a set of mock catalogs with hundreds or even thousands of simulated universes.
- A methodology that can combine all of the previous components to estimate the likelihood of the model for a set of cosmological parameters. This methodology can then be plugged into a MCMC chain to explore the parameter space and find the regions of high likelihood.

In the next chapter, we present the first RSD analysis done with the fourteenth data release of eBOSS. We introduce the different constituents used in the analysis in detail and explore the parameter space of both the data and of a set of mock catalogs, the latter is used as a test of the methodology. The final result is an estimate of the growth rate $f(z)$ and of the distance measurements $H(z)$ and $D_A(z)$ at the mean redshift of the survey ($z = 0.72$).

Chapter 3

Structure growth rate measurement from the anisotropic eBOSS LRG correlation function in the redshift range $0.6 < z < 1.0$

3.1 Introduction

[†,‡] The standard cosmological model (ΛCDM) accurately describes most observations. However, the acceleration of the expansion of our universe requires the existence of a dominating source of exotic energy, i.e., the Dark Energy. This

[†] This chapter presents the RSD analysis of eBOSS DR14 and published as Icaza-Lizaola et al. (2020). While this work was led by myself and Dr. Mariana Vargas, it was a collaborative effort with other members of the eBOSS community. I was not involved in the realisation of some parts of the work presented here that were carried out by other members of the collaboration. This is the case for the survey footprints and the survey masks presented in section 3.2.2, the building of the catalogs presented in section 3.2.3, and the comparison of the results of our fits with those obtained using a BAO analysis without RSD that is presented in section 3.6.3.

[‡] Throughout this chapter we use the notation $\vec{r}$ to refer to distances in redshift space, this is in contrast with the previous two chapters where the symbol $\vec{s}$ was used and $\vec{r}$ was reserved for the real space distances. This is done to respect the notation in the published manuscript.

energy remains undetected to this day, which has led to many searches for an alternative explanation. One possibility is to modify the geometric part of Einstein's equations, which corresponds to changing the General Relativity (hereafter GR) equations rather than invoking a new component in the stress-energy tensor. Within the paradigm of GR, it is common to add a cosmological constant, Λ, coupled to the metric.

Another way to reproduce cosmological observations is to modify the gravity model. Various alternative gravitational models have been studied during the past 50 years which can be grouped in different families. Extra-field theories, such as $f(R)$ (Sotiriou and Faraoni, 2010), Tensor-Scalar theories, extra-dimension theories, such as DGP (Fang et al., 2008), braneworld, and string gravity models, and higher-order theories such as the Galileons model (Joyce et al., 2015) are some of them.

All modified gravity models must recover the GR results at the local scale (i.e., for high density) where GR has been strongly tested; this is generally solved by invoking screening mechanisms. Therefore, any modification has to appear in the context of weak gravity and large scales; this is the reason why cosmology, and more particularly Large-Scale-Structures (LSS) observations, is the appropriate framework for this study.

Cosmological constraints on the theory of gravity are primarily produced from LSS observations, the most important of these being Supernovae (Riess et al., 1998; Perlmutter et al., 1999), Baryon Acoustic Oscillations (Eisenstein et al., 2005; Alam et al., 2017) and weak lensing (Sheldon et al., 2004), and from the early universe through Cosmic Microwave Background observations, when the density contrast was of the order of $\sim 10^{-5}$ (Planck Collaboration et al., 2016a).

Large-scale peculiar velocities, combined with standard clustering, are a unique framework to distinguish between the various models of gravity. However, obtaining precise relative velocity measurements at large scales (>10 h^{-1} Mpc) is challenging. The Kinetic Sunyaev-Zel'dovich effect is a possibility (Mueller et al., 2014) but requires measurements of massive galaxy clusters with high precision on the SZ signal estimation. Conversely, we can directly use the imprint of these velocities on the redshift measurement through the RSD in the anisotropic correlation function of galaxies (or other tracers of the dark matter) (Kaiser, 1987; Hamilton, 1992; Cole et al., 1995; Peacock et al., 2001; Cabré and Gaztañaga, 2009; Alam et al., 2015; Satpathy et al., 2017; Zarrouk et al., 2018). The measured redshift is the sum of the Hubble flow, the Doppler effect due to the peculiar velocities of the observer and the observed galaxies, and a small contribution from gravitational redshift. If the peculiar velocities are randomly distributed (i.e. from satellite galaxies inside clusters), then they only contribute as a noise. They are, however, correlated with the density field, revealing cosmological information, in particular allowing us to distinguish between dark energy models or deviations from GR. The Redshift Distortion introduces anisotropies in the galaxy-galaxy two-point correlation function, particularly if we stack the information around over-densities, where these tracers live. Performing an anisotropic study, i.e., using the angle with respect to the line-of-sight as a statistical breakdown, we can detect the coherent deformations of the 3D two-point correlation function predicted by the Kaiser (1987) effect.

BAO and supernova measurements are constraints on the expansion history of the universe. However, it has been shown that an appropriate choice of the equation of state $w(a)$ can allow different cosmological models to have the same expansion history (Linder, 2005). In order to break this degeneracy one can complement expansion history observations with the clustering history of the structures through the measurement of the linear growth rate:

$$f(a) = \frac{d \ln D(a)}{d \ln a}, \tag{3.1}$$

where $D(a)$ is the linear growth factor as a function of the scale factor a, and it quantifies the degree of structure at that time. In this paper we extend the growth rate f measurement from previous surveys to an effective redshift of $z = 0.72$ using the luminous red galaxies (LRG) sample from the extended Baryon Oscillation

Spectroscopic Survey (eBOSS; Dawson et al. (2016)).

The paper is organised as follows: Section 3.2 presents the data, Section 3.3 describes the mock catalogs used for the estimation of the covariance matrix and for our systematics checks. Section 3.4 presents the modelling of the RSD signal as well as the parametrisation used for the Alcock-Paczynski test. Section 3.5 describes the methodology followed in our analysis. Section 3.6 presents our analysis, using mock catalogs, of the systematic effects associated with our methodology. The results for the eBOSS-CMASS sample are presented in Section 3.7. Finally, the cosmological implications of this work are reviewed in Section 3.8.

3.2 Data

Our sample of spectroscopic data was collected during the first two years of eBOSS (Dawson et al., 2016), which is the cosmological component of the fourth generation of the Sloan Digital Sky Survey (SDSS-IV; Blanton et al. (2017)). All of our spectra were obtained by the Sloan 2.5m telescope using two multi-object spectrographs (Smee et al., 2013) at Apache Point Observatory in New Mexico, USA (Gunn et al., 2006). All of these data belong to the SDSS Data Release 14 (Abolfathi et al., 2017), of which we analyze the LRG Sample. The LRG targets were selected based on updated photometric data from SDSS I/II/III imaging (Fukugita et al., 1996; Gunn et al., 1998) for which the calibration of the photometric data was updated following the procedure presented in Schlafly and Finkbeiner (2011). The target selection process also used infrared photometry data from the Wide-Field Infrared Survey Explorer (WISE; Wright et al. (2010)). The WISE satellite observed the entire sky using four infrared channels respectively centred at 3.4 μm (W1), 4.6 μm (W2), 12 μm (W3), and 22 μm (W4). The eBOSS LRG sample uses the W1 and W2 bands. Given that stars have different properties than galaxies in infrared (particularly due to the galactic dust), the WISE data allow us to reduce the stellar contamination, it is also useful for extending the redshift range with

Table 3.1: Characteristics of the LRG data catalogs used. The upper panel corresponds to the BOSS CMASS sample from DR12, the lower to the eBOSS LRG DR14 sample. $N_{\rm star}$ and $N_{\rm qso}$ are the number of objects whose spectra were determined to be stars or quasars instead of LRGs. $N_{\rm zfail}$ is the number of objects whose redshift measurement was not reliable, and $N_{\rm cp}$ the number of objects without spectra due to close pair effects. The last line reports the number of galaxies and the effective volume of our final sample, which is a combination of the CMASS and eBOSS samples.

CMASS LRG Sample DR12							
Catalog	Area (deg^2)	Total redshifts					
CMASS-BOSS NGC	1011.15	26149	-	-	-	-	-
CMASS-BOSS SGC	788.09	20290	-	-	-	-	-
eBOSS LRG DR14 Sample							
Catalog	$N_{\rm gal}$	$N_{\rm star}$	$N_{\rm qso}$	$N_{\rm cp}$	$N_{\rm zfail}$	$A_{\rm eff}\,[\rm deg^2]$	$V_{\rm eff}\,[\rm Gpc^3]$
eBOSS NGC	45826	2897	18	2263	4957	1033.4	0.356
eBOSS SGC	34292	4273	18	1687	4366	811.6	0.262
Total	80118	7170	36	3950	9323	1844.0	0.618
eBOSS-CMASS	126557						0.900

respect to BOSS. The target selection follows the algorithm described in Prakash et al. (2016).

3.2.1 eBOSS-CMASS Sample

Our eBOSS DR14 LRG Sample includes data of the first two years of the eBOSS program combined with the BOSS CMASS data (Alam et al., 2017) which overlaps with the eBOSS footprint in a redshift range of $0.6 < z < 1.0$. This approach allows construction of a more complete sample without decreasing the median redshift.

The eBOSS-CMASS sample is composed of 80118 galaxies from eBOSS and 46439 from CMASS, yielding a total of 126,557 galaxies. The numbers separated by Galactic hemisphere are listed in table 3.1. The sky coverage in the North Galactic Cap (hereafter NGC) is 1011.15 deg^2 and 788.09 deg^2 in the South Galactic Cap (hereafter SGC), giving a total solid angle of 1844.0 deg^2. The effective volume of eBOSS is 0.618 Gpc^3 which increases up to 0.9 Gpc^3 when considering the eBOSS-CMASS sample.

Figure 3.1 shows the number density of the sample as a function of redshift for

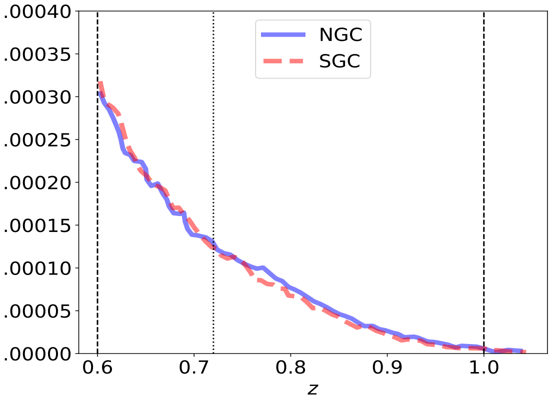


Figure 3.1: Number density of the LRG sample as a function of the redshift for both hemispheres, the solid blue lines correspond to the North Galactic Cap (NGC) and the dashed red lines to the South Galactic Cap (SGC); the dashed vertical lines indicate the redshift cuts applied. The median redshift of the sample is $z = 0.72$ and is represented by the vertical dotted line.

both hemispheres, the solid blue lines correspond to the NGC and the dashed red lines to the SGC; the dashed vertical lines indicate the redshift cuts applied for our analysis. The median redshift of the sample is $z = 0.72$, which is represented by the vertical dotted line.

3.2.2 Footprint and Masks

The left and right panels of Figure 3.2 display the sky coverage of the galaxy sample for the NGC and SGC respectively, where the colour scale indicates the targeting completeness defined as:

$$C = \frac{N_{\rm gal} + N_{\rm qso} + N_{\rm star} + N_{\rm cp} + N_{\rm zfail}}{N_{\rm target}}, \tag{3.2}$$

where

- $N_{\rm gal}$ is the number of galaxies with good quality eBOSS spectra.
- $N_{\rm cp}$ is the number of galaxies without spectra due to the fiber collision effect. Two fibers cannot be closer than $62''$ on a given plate.

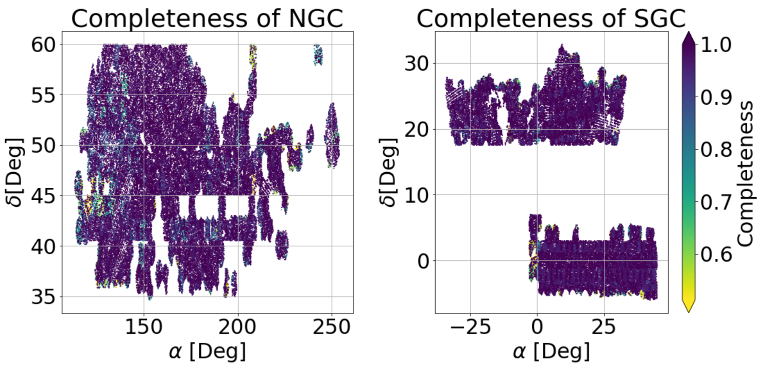


Figure 3.2: Angular mask of the North Galactic Cap NGC (left) and the South Galactic Cap SGC (right). The colour indicates the targeting completeness of the DR14 LRG sample in a given area of the sky, which is computed using equation 3.2. Regions of low targeting completeness (where $C < 0.5$) were not included in the final sample.

- N_{star} denotes the number of observed objects which are spectroscopically confirmed to be stars.
- N_{zfail} denotes the number of objects for whom the measured spectra lacks sufficient qualities to provide a confident redshift measurement.

The targeting completeness is computed by sector, and the mean completeness is 96.3% (where the NGC has an average completeness of 95.9% and the SGC 96.9%). We only use data from regions with a completeness higher than 0.5 (this value is smaller than the completeness used in BOSS).

Certain areas in the sky have to be excluded from the final data sample. The maps of these excluded regions are known as veto masks and have to be removed from our random catalogs as well. The veto masks used in eBOSS were:

- The Collision priority veto mask that excludes regions that are closer than $62''$ from an already observed target, as any object inside this radius would not be observed due to fiber collision.
- The Bright veto mask which excludes regions around stars that are part of the Tycho catalog (Høg et al., 2000) with Tycho BT magnitudes larger than 6 and lower than 11.5. The excluded radius is magnitude-dependent and it goes

from $0.8'$ to $3.4'$. An additional mask excludes regions around bright galaxies and other objects (Rykoff et al., 2014); it is also magnitude-dependent and goes from a radius of $0.1'$ to $1.5'$.

- The Bad fields veto mask excludes regions of the sky with bad photometry. If the local sky is badly determined (as occasionally happens in regions with complex backgrounds), the core of an object can be strongly negative.
- The Extinction mask excludes regions where the Galactic extinction is such that $E(B-V) > 0.15$ or where the seeing FWHM is larger than $2.3''$, $2.1''$, and $2.0''$ in g, r, and i bands, respectively.
- The Centre Focal Plane mask excludes LRG targets that lie within $92''$ of the centre of the telescope focal plane, where a centre post holds the plate and prevents fibers from being assigned.

The total masked area is 12.3% for the NGC and 18.2% for the SGC.

3.2.3 Catalog for LSS analysis

Two data catalogs that differ in their treatment of the photometric systematics and of spectroscopic incompleteness were used to create the sample for our study. The first is a BOSS-like catalog where traditional weighting schemes are applied, described in Ross et al. (2017), to the data. The second is denoted as the "official catalog", and it was used in Bautista et al. (2018) for performing the BAO analysis. Here some improvements with respect to previous analysis were implemented: the forward modelling of the randoms for the spectroscopic incompleteness and the multilinear regression and subsampling of the randoms for the photometric systematics.

In this section, we briefly review both methodologies, first describing the different treatments of the photometric systematics, and then the procedures used for dealing with the redshift incompleteness. Finally, we summarise the weights applied to the data for both cases and also the subsampling techniques used in the random catalogs in each case.

3.2.3.1 Correcting for Photometric Systematics

Here we will give a brief description of the two methodologies for correcting photometric systematics:

- Iterative method ("BOSS-like") was developed in Ross et al. (2017). The basic idea is to include the systematics in an iterative way and estimate at each step the associated weights. For the eBOSS LRG sample, we studied the correlation of the mean density as a function of seven potential observational systematics related with SDSS photometry: stellar density, i-band depth, z-band sky flux, z-FWHM, and r-band extinction[*]. We followed the iterative method starting with the main systematics reported in previous works. Figure 3.3 displays the mean density of data, $N_{\rm gal}$, normalised by the random number density, $N_{\rm ran}$, as a function of six of the seven systematics considered in the analysis before and after corrections. The most significant weights are those due to stellar density ($w_{\rm star}$), followed by the r-band extinction ($w_{\rm ext}$), airmass ($w_{\rm air}$), and z-band sky flux ($w_{\rm sky}$). The systematics related with the WISE maps did not have any strong correlation requiring correction, thus we decided not to include them in the weight estimation. We calculate a weight for each galaxy that takes in account a linear relationship for each potential systematic [i].

$$w(\text{sys}) = \frac{1}{\text{mx} + \text{b}}. \tag{3.3}$$

[*] Additionally we explored two additional maps derived from WISE photometry: one for the median number of single-exposure frames per pixel in the WISE W1 band (denoted as WISE W1 Cov Med) and the median of accumulated flux per pixel in the WISE W1 band (denoted by WISE W1 Med).

[i] $mx + b$ represents a linear fit to N_{gal}/N_{ran} to curves like the ones from figure 3.3, where the y-axis shows N_{gal}/N_{ran} and the x-axis a systematic that is being modeled for. We make one fit for each systematic considered in this work.

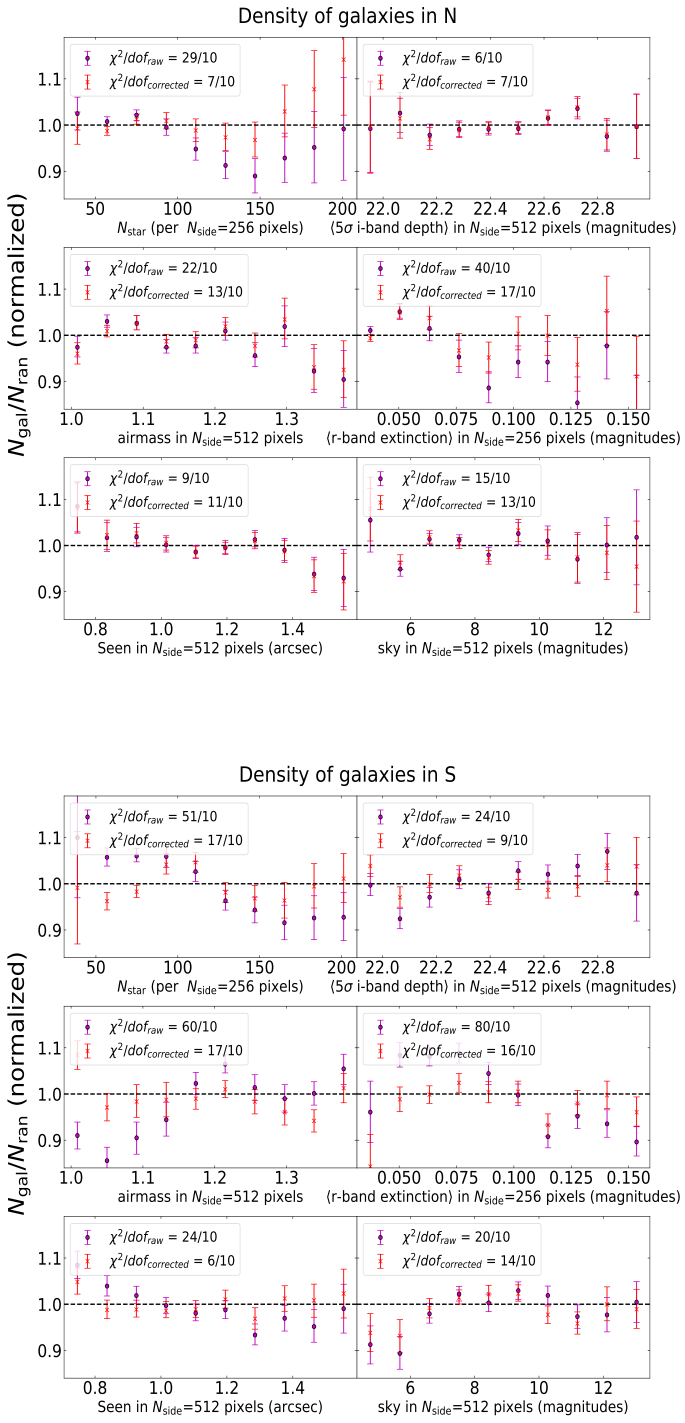


Figure 3.3: We show the mean density of data, $N_{\rm gal}$, normalised by the random number density, $N_{\rm ran}$ as a function of six of the seven systematics considered in the analysis. The most significant weights were those due to stellar density ($w_{\rm star}$), r-band extinction ($w_{\rm ext}$), airmass ($w_{\rm air}$), and z-band sky flux ($w_{\rm sky}$).

The total systematic weight $w_{\rm systot}$, is defined as

$$w_{\rm SYSTOT} = w_{\rm star}\, w_{\rm ext}\, w_{\rm air}\, w_{\rm sky} \tag{3.4}$$

- Multi-regression Method: We followed the same methodology presented in Bautista et al. (2018), where the correlation between systematic maps and density were computed using a multilinear regression of the seven systematic maps instead of the iterative method. The advantage of this method is that it does not assume the systematics are independent, as does the iterative method. Additionally, in the official catalogs, instead of using weights associated with galaxies, the randoms are subsampled following the correlation found with the multi-regression method; the subsampling of the randoms or the weighting scheme of the galaxies should yield the same results; the main differences observed in the catalogs should be derived uniquely from the Iterative/Multi-regression methodologies.

Figure 3.4 presents the multipoles for the eBOSS sample (NGC and SGC separated), comparing the iterative and multilinear regression methods. The monopole from both hemispheres without corrections shows a large spurious correlation at large scales that is reduced when either of the methods for correcting the observational systematics is applied. There is an excellent agreement in both methods for correcting photometric systematics. The SGC does show slightly better performance using the multi-regression method.

3.2.3.2 Correcting for Spectroscopic Completeness

Previous analyses on the eBOSS LRG sample reported that fluctuations in the (S/N) significantly affect the probability of obtaining a confident redshift (see Figure 5 of Bautista et al. (2018)). Additionally, the probability of obtaining a confident redshift varies across the focal plane, decreasing near the edges (see Figure 6

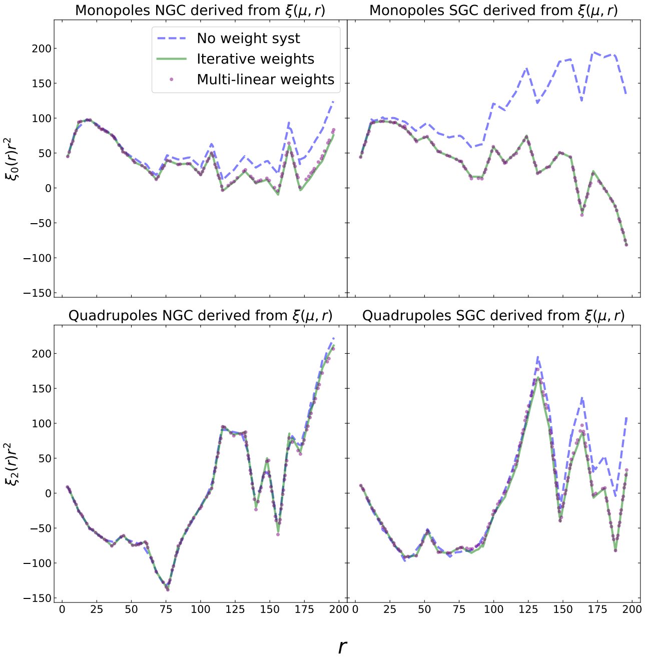


Figure 3.4: Multipoles for eBOSS sample (NGC left and SGC right) comparing the iterative and multilinear regression methods. The monopole from both hemispheres without corrections shows a large spurious correlation at large scales in monopole that is reduced when either of the methods for correcting the observational systematics is applied.

of Bautista et al. (2018)). We define the failure rate as:

$$\eta = \frac{N_{\text{gal}}}{N_{\text{zfail}} + N_{\text{gal}}}, \tag{3.5}$$

where the failure rate in eBOSS LRGs sample is 10%, which is significantly higher than previous surveys; for example, in BOSS the failure rate was only 1.8% (This is due to eBOSS targeting fainter galaxies than BOSS).

The variations of the failure rate across the focal plane could bias the clustering measurements. In order to account for the effect of this redshift incompleteness, we applied two methods to mitigate the effect on the clustering measurements; in particular, we studied how the two techniques affect the RSD analysis.

- Nearest-neighbor up-weighting. The procedure followed in BOSS (Reid et al., 2012) was to upweight the nearest neighbor with a good redshift and spectroscopic classification in its target class, within a sector. It has been shown that this method introduces structure into the monopole at small scales, and also modifies the quadrupole amplitude, which could potentially affect the growth rate measurements.

- Forward-Modeling. This approach uses a probabilistic model that depends on the the position of its fiber in the focal plane and the overall signal-to-noise ratio of the plate. The model for failures is then applied to the random sample by subsampling, mimicking the patterns retrieved in the model. For more details about this modelling we refer the reader to Bautista et al. (2018).

3.2.3.3 Data Weights

We now specify the weights applied for each catalog and the randoms treatment.

- w_{SYSTOT}. As described previously, these weights account for the fluctuations of the observational conditions that can impact the clustering signal. For the BOSS-like method these weights are computed as described in the Iterative Method.

- w_{FKP}. These weights are used for both set of catalogs. They serve to optimise clustering signal-to-noise ratio for a survey with density varying with respect to the redshift. Also known as FKP weights (Feldman et al., 1994), they are defined as:

$$w_{\mathrm{FKP}} = \frac{1}{1 + \bar{n}(z) P_0}, \tag{3.6}$$

  where $\bar{n}(z)$ is the average comoving density of galaxies as a function of redshift and P_0 is the value of the power spectrum at scales relevant for our study ($k = 0.14 h\,\mathrm{Mpc}^{-1}$). For the eBOSS LRG sample we adopt a value of $P_0 = 10^4 h^{-3}\,\mathrm{Mpc}^3$, which is the same value used in the final BOSS CMASS clustering measurements.

- w_{CP}. This weight accounts for the fiber collisions and is used for both catalogs. Targets missed due to fiber collisions do not happen randomly on the sky; they are more likely to occur in overdense regions. For mitigating this effect we followed the up-weighting technique described previously.
- w_{NOZ}. This weight accounts for the redshift failures. For the BOSS-like method this weight is computed for each galaxy following the up-weighting technique described in the previous section.

For the official catalogs these weights are set to 1, as the spectroscopic incompleteness is modeled to subsample the randoms as described in the previous section.

3.3 Mocks

We use three different sets of mock catalogs in our analysis. The first is a collection of 1000 Quick Particle Mesh (QPM) mocks (White et al., 2013), which will be used for computing the covariance matrices and for doing several systematic tests. The second one is a set of 1000 Effective Zeldovich approximation method (EZ) mocks Chuang et al. (2015), that are used to test variance of our fitting methodology. The third catalog is a set of 84 high-fidelity mocks called CutSky-Mocks Alam et al. (2017). These catalogs will be necessary for testing the accuracy of the model used.

3.3.1 QPM Mock Catalogs

We use 1000 realisations of QPM mocks using the following cosmology $\Omega_{\mathrm{M}} = 0.29$, $h = 0.7$, and $\Omega_{\mathrm{b}}h^2 = 0.02247$. A Halo Occupation Distribution (HOD) framework is adopted for populating halos with galaxies following the 5-parameter method described in Tinker et al. (2012) but taking into account the HOD parameters tuning to the DR14 eBOSS LRG sample in Zhai et al. (2017).

The same boxes were used for generating NGC and SGC mocks, thus there should be a small correlation between them (particularly in the large modes). In order to

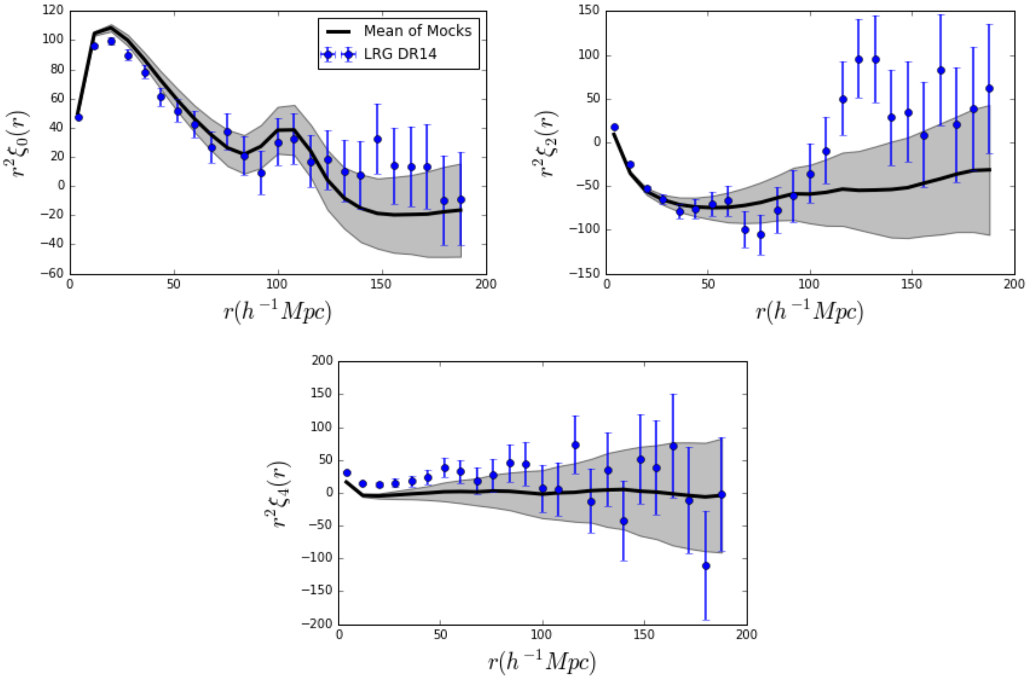


Figure 3.5: The Black solid lines are the mean of our 1000 QPM mocks for the Monopole (left), Quadrupole (centre), and Hexadecapole (right); the shaded regions are the 1-σ variations. The blue dots represent the data points and the associated error bars and are equal to the 1-σ variation shown in the shaded contours.

mitigate this effect, we combined mocks produced by different realisations of the NGC and the SGC. The mask that we applied to the mocks will be described in Section 3.2.2.

Our QPM mocks are needed for two reasons: to compute an estimate of the covariance matrix and to test our methodology. Figure 3.5 shows the mean of the mocks compared with the data; the solid lines represent the mean of the mocks correlation function and the blue dots the data correlation function multipoles with their associated error bars. There is a good agreement between the data and the mocks for scales larger that 30 h^{-1}Mpc; at smaller radii a mismatch appears, which might be related to the resolution of the mocks.

3.3.2 EZ Mock Catalogs

EZ simulations are light-cone mock catalogs created following the Effective Zeldovich methodology described in Chuang et al. (2015). In order to construct the eBOSS+CMASS sample, the CMASS and eBOSS mocks are calibrated and generated separately and then combined. The CMASS mocks are constructed in four redshift bins: (0.55, 0.65), (0.65, 0.7), (0.7, 0.8), and (0.8, 1.0025), while the eBOSS mocks are constructed at five redshift bins: (0.55, 0.65), (0.65, 0.7), (0.7, 0.8), (0.8, 0.9), and (0.9, 1.05). The fiducial cosmology is a flat ΛCDM model with $\Omega_{\mathrm{M}} = 0.307115$, $h = 0.6777$, $\sigma_8 = 0.8225$, $\Omega_{\mathrm{b}} = 0.048206$ and $n_{\mathrm{s}} = 0.9611$. We will use these mocks to test the variance of the fitting methodology.

3.3.3 N-Series Cut Sky Mocks

Our N-Series Cut Sky Mock library contains 84 mocks generated with N-body simulations that where done using GADGET2 (Springel, 2005). Our mocks have the angular and radial mask of BOSS NGC DR12 based on simulations with 2048^3 particles in a volume of $(2.6\,h^{-1}\mathrm{Gpc})^3$ corresponding to resolution particle mass about $1.5 \times 10^{11}\mathrm{M}_{\odot}h^{-1}$. We used these mock catalogs to test the theoretical systematics related to our modelling methodology. The N-Series cosmology is $\Omega_{\mathrm{M}} = 0.286$, $h = 0.7$, $\Omega_{\mathrm{b}} = 0.047$, $\sigma_8 = 0.820$, and $n_{\mathrm{s}} = 0.96$.

3.4 Modeling Redshift Space Distortions

In order to model the different multipoles of the two-point correlation function, we use the combined Convolutional Lagrangian Perturbation Theory (CLPT) and Gaussian Streaming RSD (CLPT-GSRSD) formalism, developed by Wang et al. (2014), Reid and White (2011), and Carlson et al. (2013). In this section we briefly describe this theoretical framework.

3.4.1 CLPT

CLPT provides a non-perturbative resummation of Lagrangian perturbation to the two-point statistic in real space for biased tracers. The starting point for the Lagrangian framework is the relation between the Lagrangian coordinates $\vec{q}$ that are related to the Eulerian coordinates $\vec{x}$ as:

$$\vec{x}(\vec{q},t) = \vec{q} + \vec{\Psi}(\vec{q},t), \tag{3.7}$$

where $\Psi(\vec{q},t)$ is the displacement field at each time t. The two-point correlation function is expanded in its Lagrangian coordinates considering the tracer X, in our case the Luminous Red Galaxies, to be locally biased with respect to the initially Cold Dark Matter overdensity $\delta(\vec{q})$. The expansion is performed over different orders of the Lagrangian bias function $F[\delta(\vec{q})]$, defined as:

$$1 + \delta_X(\vec{q},t) = F[\delta(\vec{q})]. \tag{3.8}$$

The Eulerian contrast density field is computed convolving with the displacements:

$$1 + \delta_X(\vec{x}) = \int d^3 F\left[\delta(\vec{q})\right] \int \frac{d^3k}{(2\pi)^3} e^{i\vec{k}(\vec{x}-\vec{q}-\vec{\psi}(\vec{q}))}. \tag{3.9}$$

Assuming that the expectation value of the n^{th} derivative of the Lagrangian bias function F is given by:

$$\langle F^n \rangle = \int \frac{d\delta}{\sqrt{2\pi}\sigma} e^{-\delta^2/2\sigma^2} \frac{d^nF}{d\delta^n}, \tag{3.10}$$

the two-point correlation function is obtained by evaluating the expression $\xi_X(\vec{r}) = \langle \delta_X(\vec{x})\delta_X(\vec{x}+\vec{r}) \rangle$ corresponding to Eq 19 of Carlson et al. (2013) and which can be simplified using the bias expansion as in their Eq. 43:

$$1 + \xi_X(\vec{r}) = \int d^3q M(\vec{r},\vec{q}), \tag{3.11}$$

where $M(\vec{r},\vec{q})$ is the kernel of convolution taking into account the displacements and bias expansion up to its second derivative term. The bias derivative terms are computed using a linear power spectrum (LPS). The LPS that we used was computed using the code CAMB (Lewis et al., 2000) for a fixed cosmology described as the fiducial cosmology of our analysis.

As we are interested in studying RSD, we also must model the peculiar velocity's effect on the clustering statistic. CLPT can compute the pairwise velocity distribution $\vec{v}_{12}$ and the pairwise velocity dispersion σ_{12}. This calculation is done following the formalism of Wang et al. (2014) which is similar to the one described above but modifying the kernel to take into account the velocities rather than the density:

$$\vec{v}_{12}(r) = (1 + \xi_X(\vec{r}))^{-1} \int M_1(\vec{r}, \vec{q}) d^3q, \tag{3.12}$$

and

$$\sigma_{12}(r) = (1 + \xi_X(\vec{r}))^{-1} \int M_2(\vec{r}, \vec{q}) d^3q. \tag{3.13}$$

The kernels $M_{1,2}(\vec{r}, \vec{q})$ also depend on the first two derivatives of the Lagrangian bias $\langle F' \rangle$ and $\langle F'' \rangle$, which are free parameters, in addition to the growth rate f, for our model. Hereafter we eliminate the brackets for the Lagrangian bias terms to have a less cumbersome notation in the following sections.

3.4.2 CLPT-GSRSD

While CLPT generates more accurate multipoles than the Lagrangian Resummation Theory (LRT) from Matsubara (2008) and the linear theory, we still require better performance to study the smaller scales of our quadrupoles. This represents an issue that is particularly important when doing RSD measurements as the peculiar velocities are generated by interactions that occur on the scales of clusters of galaxies.

In order to achieve the required precision, we map the real space CLPT models of the two-point statistics into redshift space following the Gaussian Streaming Model (GSM). This formalism was proposed by Reid and White (2011). Here, the pairwise velocity distribution of tracers is assumed to have a Gaussian distribution that is dependent on both the separation of tracers r and the angle between their separation vector and the line-of-sight μ.

The methodology of using CLPT to model the necessary inputs of a GSM was implemented by Wang et al. (2014). Its predictions are computed via the following integral:

$$\begin{aligned} 1+\xi_X(r_\perp, r_\parallel) = \int & \frac{1}{\sqrt{2\pi(\sigma_{12}^2(r,\mu)+\sigma_{\mathrm{FoG}}^2)}}[1+\xi(r)] \\ & \times \exp -\frac{[r_\parallel - y - \mu v_{12}(r,\mu)]^2}{2(\sigma_{12}^2(r,\mu)+\sigma_{\mathrm{FoG}}^2)} dy, \end{aligned} \tag{3.14}$$

where, as stated , $\xi_X(r)$, $v_{12}(r)$, and $\sigma_{12}(r)$ are computed from CLPT.

Reid and White (2011) demonstrated that GSM can predict accuracies of $\approx 2\%$ when DM halos are used as tracers. However, not all LRGs are central halo galaxies; approximately 20% of them are satellite galaxies with a peculiar velocity respect to their host halo. Therefore, we need to consider a contribution to the velocity dispersion due to the Fingers of God (FoG) effects on non-linear scales. We have addressed this point by adding the σ_{FoG} parameter to Eq. 3.14.

To summarise, given a fiducial cosmology, our model has four free parameters $[f, F', F'', \sigma_{\mathrm{FoG}}]$. The cosmology determines the LPS used in the model. The following subsection describes how we include variations of the cosmological parameters around the fiducial values using the Alcock-Paczynski Test.

3.4.3 Including the Alcock-Paczynski Effect

We described above the model for the RSD signal given a fixed fiducial cosmology that determines the LPS to be used. However, we can extract additional information by measuring the galaxy clustering along the line-of-sight and perpendicular to the line-of-sight, and we can extract geometrical information via the Alcock-Paczynski (AP) test (Alcock and Paczynski, 1979). In this work, for extracting

AP information, we use the parametrisation described in Xu et al. (2012), Vargas-Magaña et al. (2014), and Anderson et al. (2014), which derives measurements of the isotropic dilation of the coordinates parametrised by α and the anisotropic warping of the coordinates parametrised by ϵ [*]. We remind the connection with the other parametrisation, that we will further use for comparison with previous works, is given by :

$$\begin{aligned} \alpha &= \alpha_{\perp}^{2/3}\alpha_{||}^{1/3}, \\ 1+\epsilon &= \left(\frac{\alpha_{||}}{\alpha_{\perp}}\right)^{1/3}. \end{aligned} \tag{3.15}$$

where $\alpha_{\perp}$ and $\alpha_{||}$ are defined in terms of dilation in the transverse and line-of-sight directions.

3.5 Methodology

3.5.1 Fiducial Cosmology

Our analysis is performed using the following fiducial cosmology:

$$\begin{aligned} \Omega_{\rm M} &= 0.31, \\ \Omega_{\Lambda} &= 0.69, \\ \Omega_{\rm k} &= 0, \\ \Omega_{\rm b}h^2 &= 0.022, \\ \Omega_{\nu}h^2 &= 0.00064, \\ w &= -1, \\ w_{\rm a} &= 0, \\ h &= 0.676, \\ n_{\rm s} &= 0.97, \\ \sigma_8 &= 0.8. \end{aligned}$$

[*] Note that $\alpha = 1$ and $\epsilon = 0$ for the mocks, if we use their natural cosmology as the fiducial cosmology for the analysis.

Table 3.2: Expected values of cosmological parameters for the QPM mocks and Fiducial Cosmology at different redshift ranges/model. The units for $H(z)$ are $\mathrm{km\,s^{-1}Mpc^{-1}}$) and (Mpc) for $D_A(z)$.

Model	z-range	z_{eff}	$f(z)$	$\sigma_8(z)$	$f(z)\sigma_8(z)$	$H(z)$	$D_{\mathrm{A}}(z)$
QPM	$[0.6, 1.0]$	0.72	0.806	0.557	0.449	-	-
Fiducial	$[0.6, 1.0]$	0.72	0.819	0.550	0.450	101.94	1535
Nseries	$[0.43, 0.7]$	0.5	0.740	0.637	0.471	-	-

This fiducial cosmology is different from the ones used to compute the mocks; this additional bias in our methodology has to be considered. This extra bias will be defined in Section 3.6.

3.5.2 2PCF Estimator

The following section will describe the methodology used to compute the two-point clustering statistics of the DR14 LRG sample described in Section 3.2.

We are interested in constraining RSD parameters. Therefore, we must study the clustering of galaxies in two directions: the one parallel to the LOS, where peculiar velocities of infalling galaxies generate RSD, and its perpendicular direction, where no distortion occurs. We decompose the vector $\vec{r}$, which represents the distance between two galaxies, into two components: $r_{||}$ parallel to the LOS and $r_{\perp}$ that is perpendicular to it:

$$r^2 = r_{||}^2 + r_{\perp}^2. \tag{3.16}$$

Let us remember from equation 2.2 that if θ denotes the angle between the galaxy pair separation and the LOS direction, then $\mu = \cos\theta$ and we have the relation:

$$\mu^2 = \cos^2\theta = \frac{r_{||}^2}{r^2}, \tag{3.17}$$

and our two direction parameters will be $[r, \mu]$.

The 2D-correlation function $\xi(r, \mu)$ is computed using the Landy-Szalay estimator (Landy and Szalay, 1993):

$$\xi(r, \mu) = \frac{DD(r, \mu) - 2DR(r, \mu) + RR(r, \mu)}{RR(r, \mu)}, \tag{3.18}$$

where $DD(r,\mu)$, $RR(r,\mu)$, and $DR(r,\mu)$ are the number of pairs of galaxies which are separated by a radial separation r and angular separation μ. The three symbols represent the data-data, random-random, and data-random pairs, respectively.

The multipoles are Legendre moments of the 2D-correlation function $\xi(r,\mu)$, and can be computed using the following equation:

$$\xi_\ell(r) = \frac{2\ell+1}{2}\int_{-1}^{+1} d\mu\, \xi(r,\mu)\, L_\ell(\mu), \tag{3.19}$$

where $L_\ell(\mu)$ is the ℓ-th order Legendre polynomial.

We will focus primarily on the monopole, the quadrupole, and the hexadecapole ($\ell = 0$, $\ell = 2$, and $\ell = 4$).

The pair-counts were computed using the public code CUTE (Alonso, 2012). However, there are three corrections to be considered when using the LS equation (3.18):

- The number of galaxies in the Data catalogs (N_D) is approximately 50 times smaller than the ones in our random catalogs (N_R). Therefore the Random and Data pairs should be compared as

$$\frac{DD(r,\mu)}{RR(r,\mu)} \times \frac{N_R(N_R-1)}{2} \times \frac{2}{N_D(N_D-1)}.$$

- The number of galaxies in the SGC ($N_{D,S}$) is smaller than those in the NGC ($N_{D,N}$). Therefore the total number of pairs should be added as:

$$DD(r,\mu) = \frac{2(DD_N(r,\mu) + DD_S(r,\mu))}{(N_{D,N}(N_{D,N}-1) + N_{D,S}(N_{D,S}-1))}.$$

- Each galaxy has a particular weight w_i as described in Section 3.2. Hence, the total number of galaxies in any catalog is weighted as

$$N^w = \sum w_i.$$

3.5.3 Fitting

Unless stated otherwise, we will be using 13 bins of $8h^{-1}$Mpc in width, in the interval between [$28h^{-1}$Mpc, $124h^{-1}$Mpc]. Given that we will be working with either the first two non-zero multipoles or the first three (depending on the test), the analysis will have a total of either 26 or 39 bins.

We will now compare our measured two-point statistics with those predicted by our model and try to find the best-fit model parameters.

In order to identify best-fit parameters, we minimise the χ^2 function,

$$\chi^2 = (\vec{m} - \vec{d})^T C^{-1} (\vec{m} - \vec{d}) \tag{3.20}$$

where $\vec{m}$ is the vector formed by the model predictions, and $\vec{d}$ is the equivalent vector observed from our data. Examining Eq. 3.20 reveals that the smaller the value of χ^2, the more similar $\vec{m}$ is to $\vec{d}$.

The sample covariance is defined as:

$$C_{S_{ij}} = \frac{1}{N_{\text{mocks}} - 1} \sum_{m=1}^{N_{\text{mocks}}} (\xi_i^m - \bar{\xi}_i)(\xi_j^m - \bar{\xi}_j), \tag{3.21}$$

where N_{mocks} is the number of mocks, and $\bar{\xi}_i$ is the average of the i^{th} bin.

We scale the inverse sample covariance matrix, C_s^{-1}, using Eq. 17 of Hartlap et al. (2007):

$$C^{-1} = C_s^{-1} \frac{N_{\text{mocks}} - N_{\text{bins}} - 2}{N_{\text{mocks}} - 1}. \tag{3.22}$$

This procedure corrects for the fact that the matrix in Eq. 3.21 is a biased estimate of the true inverse covariance matrix C^{-1}.

Figure 3.6 shows the covariance and correlation matrix computed from 1000 QPM mocks. Most of our error arises from the elements on the diagonal (variance of a given bin), but there is also a significant contribution coming from elements outside of the diagonal (covariance between different bins).

In order to identify the best-fit parameters, we minimise the χ^2 function.The minimisation of the χ^2 is done using the Powell algorithm (Press et al., 2002). This

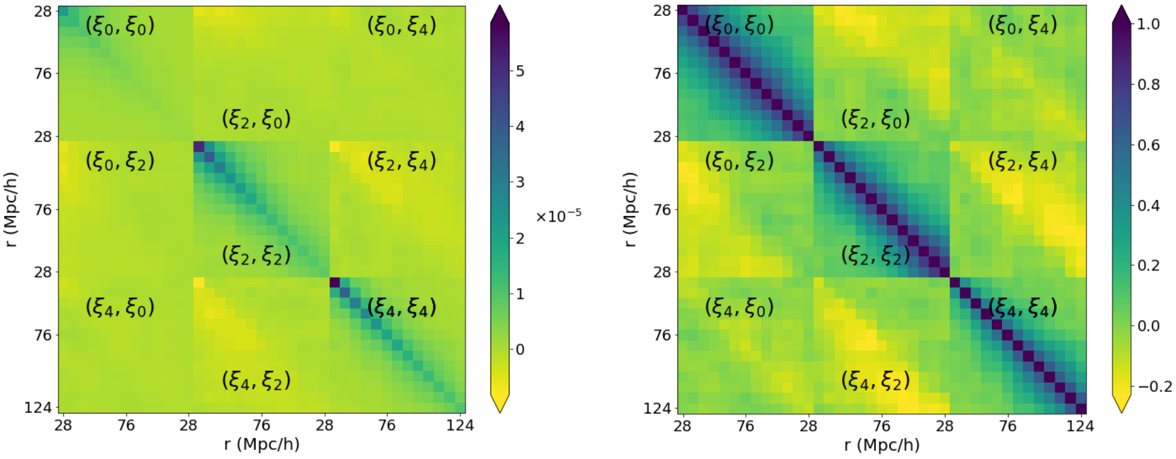


Figure 3.6: Density map of the Covariance matrix (left) computed from our 1000 QPM Mocks simulations. The matrix has 13 bins in r and 3 multipoles. The right panel presents the correlation matrix defined as $Corr_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$. The normalisation is done such that the diagonal is always unity, and it shows how much covariance (off-diagonal) there is compared to our variance (diagonal).

algorithm will find a unique solution if the parameter space is gaussian, which should be a fair assumption when fitting our mocks. This method is adequate for our work as it does not require us to compute the gradient of the CLPT-GSRD model with respect to the model parameters, which would be challenging. Due to the nature of the algorithm it is not necessary to specify any prior, just some starting points, if our assumption about the parameter space being somewhat gaussian is correct then any starting point should work fine and one that is close to the Best Fit should reduce the running time.

The estimate of the errors on our fits will be computed using MCMC chains, but we will only do this analysis for our data sets (Section 3.7) and not for the mocks.

3.6 Testing for systematic uncertainties

3.6.1 Testing Accuracy of GSRSD Hexadecapole Model with High-Resolution Simulations

This section is dedicated to testing the performance of the methodology developed in Section 3.5. Here, we will use the N-Series CutSky mocks described in Section 3.3 to check the reliability of the CLPT model with regards to recovering the cosmological parameters. These high resolution mocks are built with the BOSS-CMASS properties that allow us to study the accuracy of the model. We will run our fitting methodology on these high-fidelity mocks in order to test if their fiducial parameters can be recovered. The N-Series CutSky mocks have been used previously in the literature for testing the monopole- and quadrupole-only methodologies.

We fit our N-series CutSky mocks twice, the first using only the monopole and the quadrupole, and the second including the hexadecapole. The fits are done following the methodology described in Section 3.5.3, but here we will be using 21 bins of $5h^{-1}$ Mpc in width, in the interval between [$27.5h^{-1}$ Mpc, $127.5h^{-}1$ Mpc]. We decided to choose a smaller bin-size to facilitate comparisons with other previous results. The sample covariance matrix used to perform these fits is computed using the QPM-BOSS CMASS sample re-scaled to match the mocks volume, and provides our error estimate (the covariance matrix obtained from N-Series would be quite noisy given the limited number of realisations available). The pair-counts of our mocks were computed using the mocks cosmology to transform angular positions and redshifts into comoving coordinates. To be consistent, the CLPT-GSRSD input template was also computed using the cosmology of the mocks.

The expected values of the linear growth rate of f (f_{exp}) are reported in Table 3.2 for the natural cosmology of the mocks. We define the bias of the growth rate estimation b_f as:

$$b_f = \langle f_{\mathrm{measured}} \rangle - f_{\mathrm{exp}} \tag{3.23}$$

We use the measurement of the Eulerian bias $b = 2.3$ performed by Zhai et al. (2017) as reference. This estimate was computed using our same sample with the addition of an HOD model.

The left panel of Figure 3.7 shows the mean of the multipoles; the error bars are the diagonal terms of the covariance matrix divided by $1/\sqrt{N_{\mathrm{Mock}}}$. The different colours (and line-styles) represent the best fit model for the mean of the mocks using the fiducial range at different minimum scales of the fits when the cuts are applied to all multipoles.

The model using the hexadecapole fitted at the full range (blue lines) does not match the hexadecapole of the mean of the mocks accurately at any scale, this can be seen in the corresponding residual plot (bottom panel of the figure) where the value of the residuals is close to 50% of the value of the model, this is very large when compared to the residuals of the monopole and quadrupole that are around 10% and 2% respectively (second and third panels of the figure). Increasing the minimum range of the fit mostly affects the quadrupole at large scales and has little effect on the monopole and hexadecapole at any scale. By comparing the residuals of the quadrupole(third panel of the figure) of the full-range fit and the reduced-range fit we can tell that the full range fits adjusts the quadrupole better (i.e. the blue solid line is a better fit than the green dotted one).

The right panel of Figure 3.7 displays a similar exercise to the one in the left panel, except this time we only cut the minimum scale of the hexadecapole while leaving the other two multipoles in the full range. The changes on the quadrupole are now less severe than when varying all multipoles. By looking at the residual plot of the quadrupole (third panel of the figure) we see that the model considering the hexadecapole in the full range (green dash-dotted line) matches the quadrupole slightly better than the model using only the monopole and quadrupole (purple solid line), but it does not improve the other multipoles significantly. It is not clear that including the hexadecapole improves the fits significantly when compared to the monopole and quadrupole only case.

Table 3.3 reveals that the bias in f is slightly larger when we include the hexadecapole in the full range than when we only use the monopole and the quadrupole ($b_f = 0.005$ compared to $b_f = 0.004$). However, the bias in ϵ is smaller when the hexadecapole is left out of the fits ($b_\epsilon = 0.002$ compared to $b_\epsilon = 0.0005$). The bias in alpha is the same for both cases ($b_\alpha = 0.001$). The right panel of Figure 3.7 shows that the best fit model for both cases are very similar, only showing small differences in the quadrupole at the scales in the range [80,110] h^{-1}Mpc.

Reducing the range for all multipoles (second block of the table) increases the biases in f and ϵ. If by contrast we constrain the range only for the hexadecapole (third block), we reduce the bias in f and α to $b_f = 0.001$ and $b_\alpha < 0.001$, respectively, leaving the bias value for ϵ unchanged.

In summary, there is no clear preference between the case with the 3 multipoles and just considering monopole and quadrupole. There is a trade-off between the biases in ϵ and f: the smaller bias in f is obtained when using the hexadecapole while the smaller bias in ϵ comes from using only the monopole and quadrupole. As there is not a clear trend we will explore the hexadecapoles impact on the LRG sample analysis further.

Figure 3.8 displays the model behavior for variations of the parameters, and is included to explain the different trends observed with mocks when using multipoles up to order $\ell = 2$ compared to $\ell = 4$. We also indicate the variations in our model predicted by changes of $\sim 20\%$ in the input parameters, that correspond to deviations of $\Delta f = 0.15$, $\Delta\alpha = 0.2$, and $\Delta\epsilon = 0.2$ around the fiducial cosmology expected value. The error bars were obtained from the diagonal of the mocks covariance matrix. The variations in ϵ have a large impact on the predicted hexadecapole at all scales (middle curve), while the variations of the hexadecapole due to variations on α and f are significantly smaller and of a comparable order of magnitude. This behavior explains why the fits are driven by ϵ when the hexadecapole is included. Considering that the error bars between 20 and 60 h^{-1}Mpc are smaller, their constraining power is significantly larger.

Table 3.3: Results from fitting the mean of N-series Mocks. The expected values for the N-series mocks are $f(z = 0.5) = 0.740$, $\alpha = 1.0$ and ϵ= 0.0. The fits are done over bins of $5h^{-1}$ Mpc each so that the full range of each multipole (27.5 h^{-1} Mpc, 127.5 h^{-1} Mpc) will have 21 bins.

$\xi_0 + \xi_2$ with cuts in all multipoles								
Model	Range	F'	F''	f	α	ϵ	$\sigma_{\rm FoG}$	χ^2/d.o.f
$\xi_0 + \xi_2$	27.5-127.5	0.999	0.637	0.736	1.001	5e-4	1.076	68.5/36=1.90
$\xi_0 + \xi_2 + \xi_4$ with cuts in all multipoles								
$\xi_0 + \xi_2 + \xi_4$	27.5-127.5	1.003	1.034	0.745	1.001	-0.002	1.770	91.2/57=1.60
$\xi_0 + \xi_2 + \xi_4$	37.5-127.5	1.014	1.708	0.735	1.001	-0.003	2.239	84.0/51=1.65
$\xi_0 + \xi_2 + \xi_4$	42.5-127.5	1.022	1.870	0.731	0.999	-0.004	0.530	78.7/48=1.64
$\xi_0 + \xi_2 + \xi_4$	47.5-127.5	1.027	3.149	0.721	0.997	-0.004	1.018	70.8/45=1.57
$\xi_0 + \xi_2 + \xi_4$ with a cut in hexadecapole only								
$\xi_0 + \xi_2 + \xi_4$	37.5-127.5	1.010	1.543	0.742	1.000	-0.002	2.793	86.15/55=1.57
$\xi_0 + \xi_2 + \xi_4$	42.5-127.5	1.011	1.649	0.741	1.000	-0.002	2.938	86.18/54=1.60
$\xi_0 + \xi_2 + \xi_4$	47.5-127.5	1.012	1.697	0.741	1.000	-0.002	2.984	86.28/53=1.62

As stated before, even if our results using the hexadecapole do not show significant biases, figure 3.7 shows that the model obtained using the cosmology of the mocks does not accurately match the mean of the hexadecapole mocks at any scale, in particular at the lower scales that have more weight in the likelihood. This mismatch in the hexadecapole is pushing ϵ to higher values and as a consequence the correlated parameters follow. Therefore, the accuracy of the model at all scales is critical for not biasing the fitted parameters.

We now analyze the individual mocks for three cases: 1) fitting the complete range [27.5,127.5] h^{-1}Mpc using monopole and quadrupole, 2) fitting the complete range [27.5,127.5] h^{-1}Mpc using monopole, quadrupole, and hexadecapole, and 3) fitting the complete range [27.5,127.5] h^{-1}Mpc for monopole and quadrupole and reducing the range to [47.5,127.5] h^{-1}Mpc for the hexadecapole. Figure 3.9 show the results of the individual fits in all three cases and for the four parameters of interest [$f\sigma_8$, $b\sigma_8$, α, ϵ], as well as their respective best fit distributions histograms. The coloured dashed lines indicate the mean of the best fits, and the dotted line represents the expected value of the parameters. Table 3.4 presents the results of the individual fits for the parameters of interest.

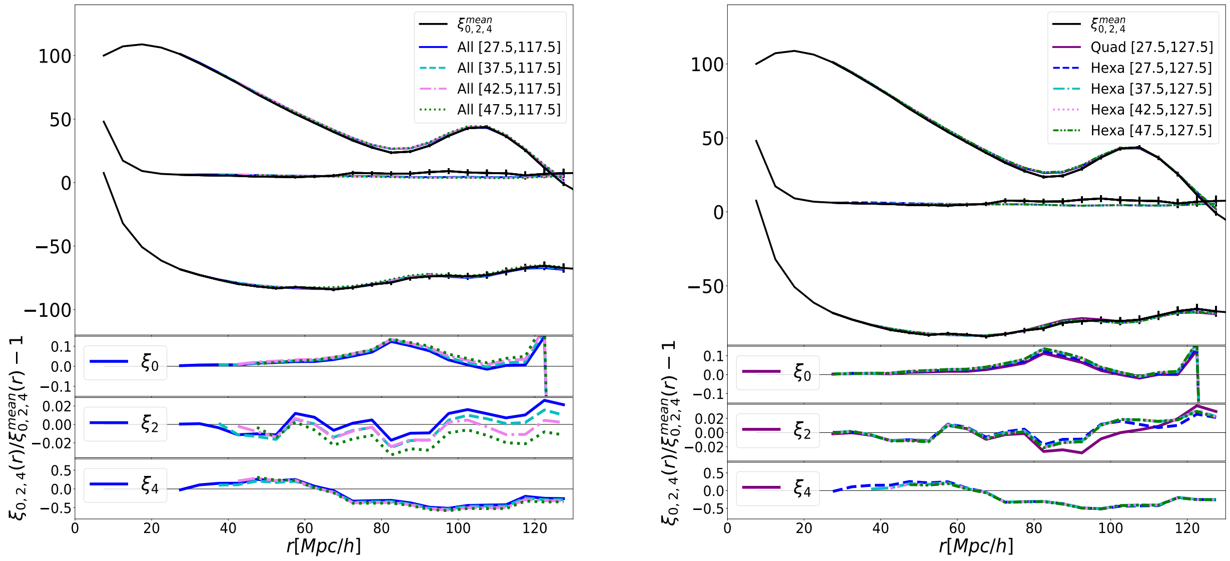


Figure 3.7: The mean of the mocks is shown as the black line in both plots. The error bars are computed from the re-scaled QPM mocks covariance. The left panel shows the best fit models from different lower ranges of the multipoles. In the right panel only the lower range of the hexadecapole is varied. The error plots show the quotient between the best fit model and the mean of the mocks. For all cases there are residuals in the hexadecapole, the smaller residuals are obtained by the monopole + quadrupole.

The monopole- and quadrupole-only fits show a bias in the estimation of the three parameters of $|b_{f\sigma_8}| = 0.003$, $|b_\alpha| = 0.002$, and $|b_\epsilon| = 0.0004$. The standard deviation of the distributions are $S_f = 0.051$, $S_\alpha = 0.014$, and $S_\epsilon = 0.019$ respectively; the expected values are within the dispersion. Thus the significance of the biases are $0.5\sigma, 1.1\sigma$ and 0.2σ. These numbers are in agreement with the test performed for the BOSS sample and these numbers are comparable with the results obtained in Alam et al. (2017) *. The full range hexadecapole fits show a lower bias in the f parameters, with a value of $|b_{f\sigma_8}| = 0.0007$, $|b_\alpha| = 0.00008$ and $|b_\epsilon| = 0.0002$ respectively. The standard deviation of the distributions decreases for f, α and ϵ, with $S_f = 0.037$, $\alpha = 0.013$ and $S_\epsilon = 0.010$. The significance of the biases decreases significantly to $< 0.1\sigma$, 0.1σ, and 0.2σ respectively. Constraining the

*The BOSS analysis only reported the result of $\Delta f\sigma_8$ for the N-Series Mocks Challenge.

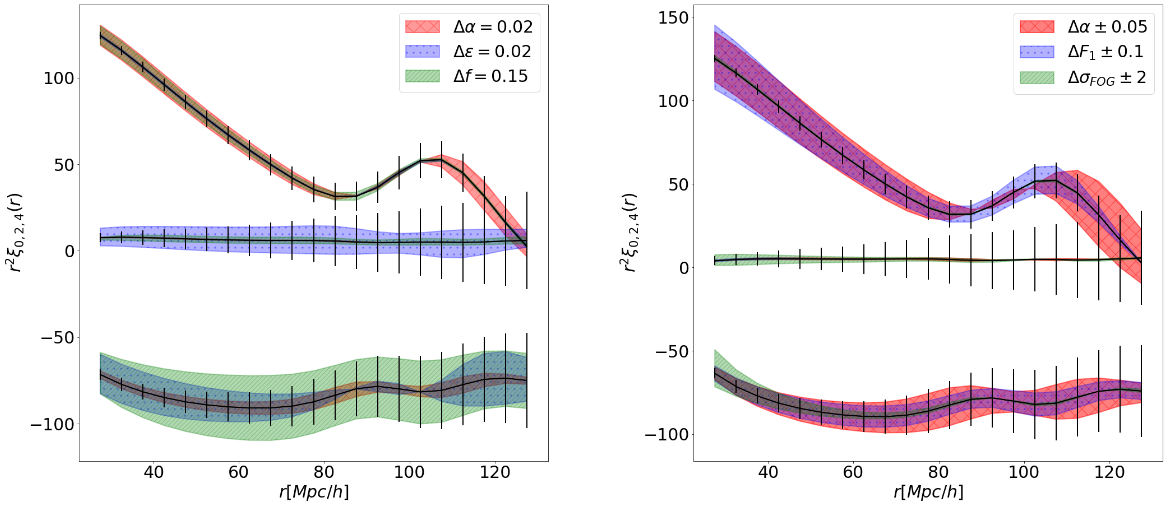


Figure 3.8: Left panel: Model variations for Δf, $\Delta\alpha$, and $\Delta\epsilon$ compared to the error bars coming from covariance: the behavior for the monopole (top), the hexadecapole (middle) and the quadrupole (bottom). The variations in ϵ have a large impact on the hexadecapole, while the variations in α and growth rate are of the same order of magnitude for the hexadecapole. This behavior explains why fits are driven by ϵ when the hexadecapole is taken into account. Right panel: Model variations for $\Delta F'$, $\Delta\alpha$, and $\Delta\sigma_{\mathrm{FoG}}$ compared to the errors produced by covariance: the behavior for the monopole (top), the hexadecapole (middle) and the quadrupole (bottom).

range of hexadecapole fits, produces biases of $|b_{f\sigma_8}| = 5e-4$, $|b_\alpha| = 0.001$, and $|b_\epsilon| = 5e-4$, while also decreasing the standard deviation of the distributions compared with the monopole and quadrupole fits, $S_{f\sigma_8} = 0.042, S_\alpha = 0.013$ and $S_\epsilon = 0.014$, giving a significance of the biases of $0.1\sigma, 0.9\sigma$, and 0.4σ, respectively.

Figure 3.10 shows the summary of the analysis for our three cases in the same format as the results reported in Alam et al. (2017): the points correspond to the mean of the results obtained from fitting our 84 SkyCut with the BOSS mask mocks, the three quantities shown are (from left to right) the mean of $\Delta f = f - f_{\mathrm{exp}}$, $\Delta\alpha = \alpha - \alpha_{\mathrm{exp}}$, and $\Delta\epsilon = \epsilon - \epsilon_{\mathrm{exp}}$, and the error indicated is the standard deviation of our fits. The panels contain the results from: 1) the fits with monopole and quadrupole (left), 2) the fits also including the hexadecapole (middle), and 3) the fits using multipoles up to $\ell = 4$ and using a constrained range on the hexadecapole

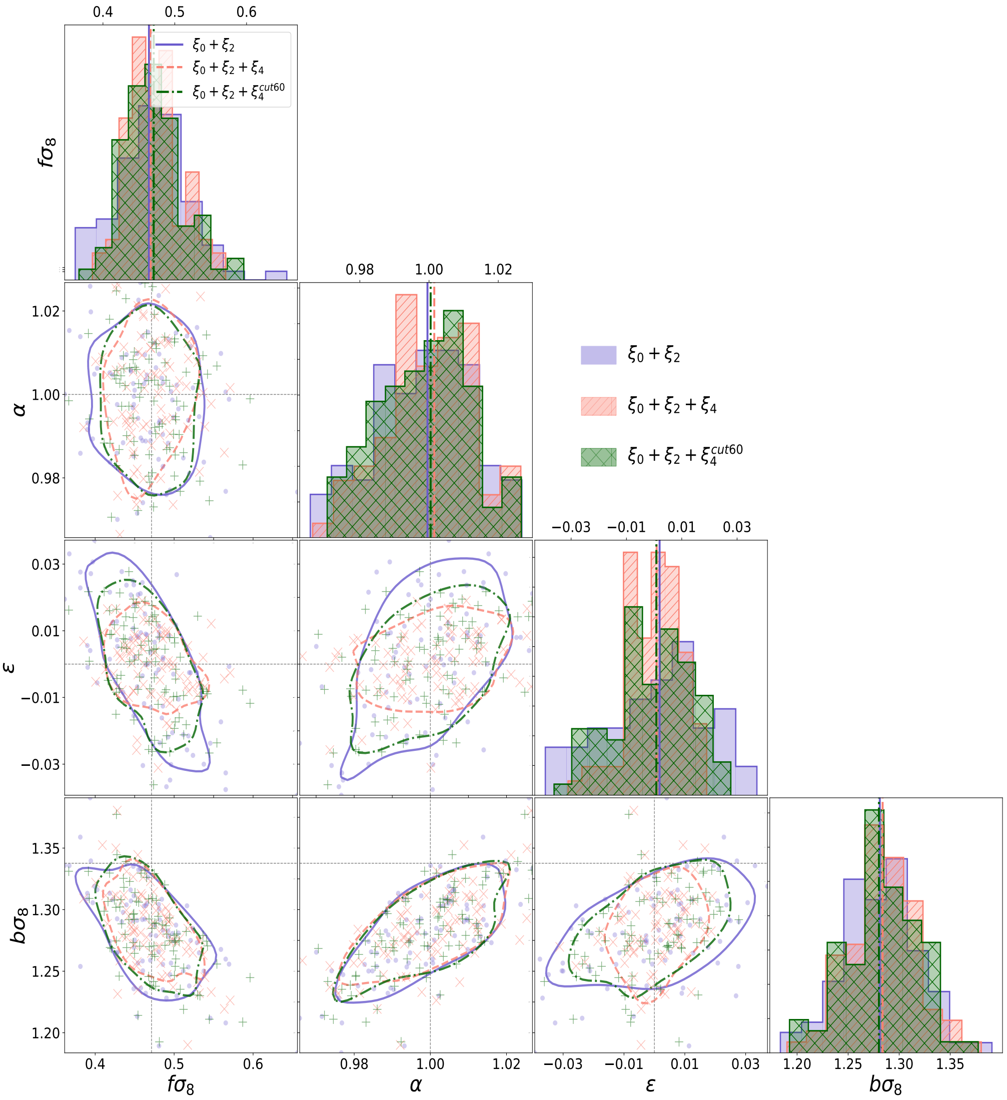


Figure 3.9: Results from the best fits of all of the individual mocks for the four parameters of interest [f, b, α, ϵ]. Also shown are their respective best fits distributions histograms and the 1σ confidence region. The dotted black lines represent the expected value of each parameter. The coloured lines in each histogram indicate the mean value of that parameter found by our fits. We present three cases: 1) fitting the complete range [27.5,127.5] h^{-1}Mpc using monopole and quadrupole (blue dots), 2) fitting the complete range [27.5,127.5] h^{-1}Mpc using monopole, quadrupole, and hexadecapole (red x's), and 3) fitting the complete range [27.5,127.5] h^{-1}Mpc for monopole and quadrupole and reducing the range to [47.5,127.5] h^{-1}Mpc for the hexadecapole (green crosses).

Table 3.4: Results from fitting the 84 N-Series sky mocks with our fiducial methodology. The columns denoted by $\widetilde{x}$ are the mean, S_x denotes the standard deviation, and the bias (defined by equation 3.23) is denoted by b_x, with $x = f$, α, ϵ.

Results for Fiducial Methodology with N-series Sky mocks										
Model	$\widetilde{f\sigma_8}$	S_f	b_f	$\widetilde{\alpha}$	S_α	b_α	$\widetilde{\epsilon}$	S_ϵ	b_ϵ	N_{mocks}
$\ell_{\rm max} = 2$ [27.5,117.5]	0.459	0.051	-0.003	0.998	0.014	-0.002	4e-4	0.019	4e-4	81
$\ell_{\rm max} = 4$ [27.5,117.5]	0.471	0.037	-7e-5	1.0	0.013	-8e-5	1e-4	0.010	2e-4	83
$\ell_{\rm max} = 4$ [47.5,117.5]	0.471	0.042	-5e-4	0.999	0.013	-0.001	-5e-4	0.014	-5e-4	84

(right). We also include the result for the growth rate obtained by BOSS and reported in Alam et al. (2017) (far right value of the left panel).

These results suggest that the most accurate results (smaller parameter biases in all parameters normalised by the dispersion) are obtained using the multipoles up to $\ell = 4$ in the full range.

However, we would like to highlight that we noticed that the best model of the hexadecapole does not accurately match the mean of the mocks (the value of the residuals is close to 50% of the value of the model compared to 10% and 4% for the monopole and quadrupole respectively). For the individual fits, given the large error bars on the hexadecapole, this mismatch does not bias our individual measurements, but produces small bias in the best fit of the mean in ϵ. Bearing all of this in mind we choose to analyze both cases (with and without hexadecapole), but we will take a conservative approach and report the monopole and quadrupole only analysis as our final result of this work. Based on the results from N-Series we adopt $\sigma^{SYS}_{f\sigma_8} = 0.004$, $\sigma^{SYS}_{\alpha} = 0.001$, and $\sigma^{SYS}_{\epsilon} = $ 5e-4 as an estimate of the potential bias of f, α and ϵ.

3.6.2 Testing Systematics with eBOSS-Mocks

We will dedicate this section to test the variance of the methodology developed in Section 3.4. This analysis will be done using two sets of approximative mocks,

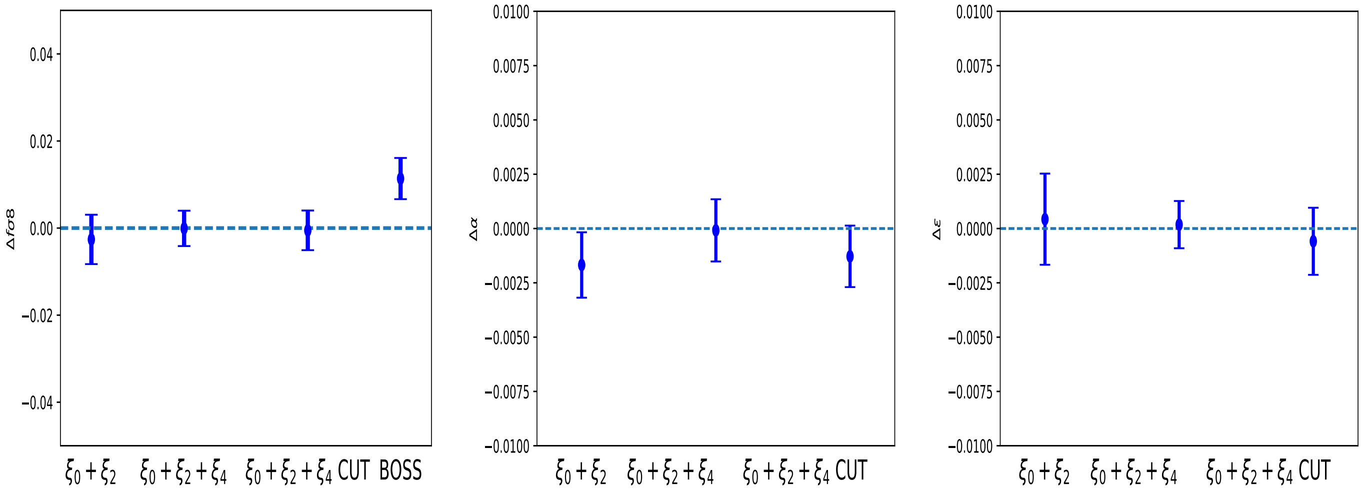


Figure 3.10: Systematic errors in RSD and AP parameters from using different multipole combinations in the fit. From left to right: mean of $\Delta f = f - f_{\rm exp}$, $\Delta\alpha = \alpha - \alpha_{\rm exp}$, and $\Delta\epsilon = \epsilon - \epsilon_{\rm exp}$. These measurements were obtained from fitting N-Series sky mocks using two configurations: 1) multipoles up to order $\ell = 2$ and 2) multipoles up to order $\ell = 4$. The left panel includes the result from the previous work (Alam et al., 2017). The less significant biases are obtained by the monopole + quadrupole fits. Including the hexadecapole reduces the bias and variance producing more significant bias in f and larger biases in α and ϵ.

the QPM and EZ described in Section 3.3, both built with the same properties of our eBOSS sample. The mocks were calibrated to match the data, however, these approximative mocks lack the accuracy to study the biases of our methodology. As shown in Figure 3.11, the QPM and EZ mock have a small mismatch in the monopole at small scales. Additionally, both seem to systematically underestimate the hexadecapole. Bearing this in mind, our estimates of the bias will only come from the results of the N-Series Cut-sky mocks obtained in the last section*, we proceed to quantify the dispersion of the fitting methodology. Our specific goal is to estimate the dispersion expected around the parameters of interest of our model. This will be done by applying the fitting methodology from section 3.5.3 to 100 of our individual QPM an EZ mocks, which will give us 100 estimates of the best fit values.

*The N-series mocks provided an estimate of the biases on a sample that is similar to that of BOSS-LRG; the mean redshift was slightly lower than the one from the eBOSS LRG sample considered in this work but it had similar clustering properties, i.e. the galaxy bias.

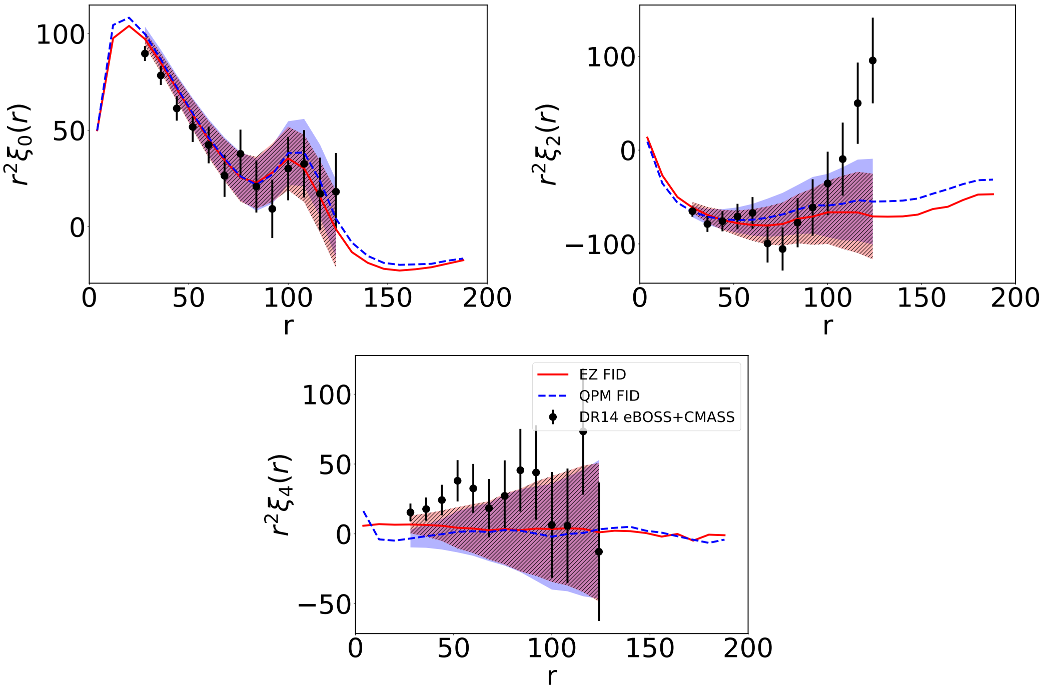


Figure 3.11: Mean QPM (dashed blue line) and EZ (solid red line) mocks and DR14 correlation function (black dots), all computed in the fiducial cosmology. QPM and EZ mock underestimate the hexadecapole. The monopole (left) shows small mismatches between the mocks and the data at small scales, the quadrupole data (centre) presents a large correlation in the large scale quadrupole that lies outside the 1σ variation observed in the mocks, the hexadecapole data (centre) has a larger amplitude than the one predicted by the mocks.

We test two cases: 1) Considering only the multipoles up to $\ell = 2$ (skipping the hexadecapole), and therefore following the methodology used in previous analysis performed with the LRG sample, which we will refer to as "$\xi_0 + \xi_2$". We also consider the effect of extending the multipoles up to $\ell = 4$ and using the full range for all multipoles, which we will refer to as "$\xi_0 + \xi_2 + \xi_4$".

We used the cosmology used for the QPM mocks generation in order to compute their comoving coordinates, and the fiducial cosmology for computing those of the EZ mocks. Table 3.5 summarises the results from our fits. The first block corresponds to the monopole + quadrupole fits using the QPM/EZ mocks; the second block describes the analysis adding the hexadecapole to our fits. The dispersions obtained from our two sets of mocks when only using the monopole and the quadrupole in the fits (first block of Table 3.5) are fairly consistent for all of the parameters of interest: $S_{f\sigma_8}^{\mathrm{QPM}} = 0.113$ and $S_{f\sigma_8}^{EZ} = 0.122$, $S_{\alpha}^{\mathrm{QPM}} = 0.039$ and $S_{\alpha}^{\mathrm{EZ}} = 0.043$, and $S_{\epsilon}^{QPM} = 0.053$ and $S_{\epsilon}^{\mathrm{EZ}} = 0.044$. The dispersion is also consistent with previous results found on the anisotropic LRG DR14 BAO analysis from Bautista et al. (2018) where: $S_{\alpha}^{\mathrm{BAO}} = 0.048$ and $S_{\epsilon}^{\mathrm{BAO}} = 0.055$.

In order to compare with the previous results from BOSS reported in Alam et al. (2017), we need to rescale the variance using the differences in volume between the two samples; the effective volume of our sample is 0.9 Gpc^3 while BOSS-CMASS accounts for 4.1 Gpc^3 in the [0.5,0.75] redshift slice. The CMASS sample reported the following standard deviations for the [0.5,0.75] redshift slice: $S_{f\sigma_8}^{\mathrm{BOSS}} = 0.058$, $S_{\alpha}^{\mathrm{BOSS}} = 0.016$, and $S_{\epsilon}^{\mathrm{BOSS}} = 0.022$ (Table 6 of Alam et al. (2017)), we can scale them roughly to the eBOSS volume using $S_X^{\mathrm{eBOSS}^2} = (S_X^{\mathrm{BOSS}^2} \times$ 4.1 Gpc^3)/0.9 Gpc^3, yielding the following scaled dispersions: $S_{f\sigma_8}^{\mathrm{BOSS}} = 0.124$, $S_{\alpha}^{\mathrm{BOSS}} = 0.034$, and $S_{\epsilon}^{\mathrm{BOSS}} = 0.047$. These values are in agreement with the dispersion obtained with our QPM/EZ mocks.

Now, let us examine the fits that include the hexadecapole. The dispersion obtained from the two sets of mocks is also consistent for the parameters f and ϵ: $S_{f\sigma_8}^{\mathrm{QPM}} = 0.090$ and $S_{f\sigma_8}^{\mathrm{EZ}} = 0.089$, and $S_{\epsilon}^{\mathrm{QPM}} = S_{\epsilon}^{\mathrm{EZ}}0.050$ and $S_{\alpha}^{\mathrm{QPM}} = S_{\alpha}^{\mathrm{EZ}} = 0.028$. Also we observe the dispersion in all parameters decreases when considering the hexadecapole as expected.

Figure 3.12 shows the distribution of the differences between the parameters of interest and their expected values on a mock-by-mock basis, i.e. $\Delta f\sigma_8 = \langle f\sigma_8 - f\sigma_{8\mathrm{exp}}\rangle$, $\Delta\alpha = \langle\alpha - \alpha_{\mathrm{exp}}\rangle$, $\Delta\epsilon = \langle\epsilon - \epsilon_{\mathrm{exp}}\rangle$, and for $b = 1 + F'$, $\Delta b = \langle b\sigma_8 - b\sigma_{8\mathrm{exp}}\rangle$, for both the analyses using multipoles up to $\ell = 2$ and up to $\ell = 4$. Reviewing the monopole + quadrupole fits (in blue dots) reveals that both sets of mocks show a well-behaved distribution that is centred close to zero and is symmetric. From the hexadecapole fits (red x's), we also observe symmetric distributions centred around zero, however, especially the 1D distributions for the $f\sigma_8$ and ϵ parameters are slightly shifted.

These shifts in the distributions when considering the hexadecapole are related to the QPM/EZ mocks poor precision and to the fact that the model and the mean multipoles present mismatches, the following paragraphs will briefly show these mismatches. As we will see, the biggest mismatch between mock and model occurs in the hexadecapole for the QPM mocks and in the quadrupole for the EZ mocks.

Figure 3.13 shows a comparison between the mean of the mocks and the model templates built with the true cosmology of the mocks[*] denoted by “Model GS $f(z = 0.72)$". The left panel shows the comparison between the mean of the QPM mocks and its model template and the equivalent comparison for the EZ mocks is in the right panel. The growth rate used for building the model in the right panel is at the effective redshift of the mocks.

The figure reveals that the mean of the QPM mocks does not match the model with the cosmology used for their generation (gray solid line), which is evident in the quadrupole residuals. However, a template using a growth rate corresponding to a lower redshift ($z = 0.56$) is a better match with the mean of the mocks (red dotted line); this model is denoted by “Model GS $f(z = 0.56)$" and is shown with red dotted lines.

From this analysis we can draw the following conclusions. First, the GSRSD model cannot match the multipoles of the QPM mocks, as they show a mismatch in the mean of the mocks and the model for the quadrupole, giving rise to a higher value than the input value of the simulations. Second, the model of the hexadecapole is systematically larger than the mean of the mocks, and in particular any conclusion about the bias of the hexadecapole cannot be extracted from the fits of the QPM mocks.

The right panel of Figure 3.13 shows an equivalent comparison between mean and model template using the EZ mocks. As for the QPM mocks we also see a mismatch, but this time between the small scales of the quadrupole: the template with the

[*] the cosmology used for building the mocks

cosmology and redshift of the EZ mocks (red dotted line) does not match with the mean quadrupole of the mocks (black solid line).

It is interesting to notice that the mean hexadecapole matches the template.

One possible explanation to why EZ mocks seems to be more accurate at fitting the hexadecapole might be that, as mentioned in section 2.5, EZ mocks fits their free parameter to the 3-point correlation function (among other observables).

The three point statistics might have a small effect on the 2-point estimation, given that the skewness has a small effect on the variance and the mean estimations. Given that the amplitude of the hexadecapole is small, a slight modification might have a stronger effect on this polynomial than on the monopole and quadrupole.

We also notice that the mismatch in the quadrupole behaves differently for different scales, the scales lower than 50 h^{-1} Mpc are overestimated and the scales larger than 50 h^{-1} Mpc are underestimated. Thus, the EZ mocks seem to not be reproducible by the model. Apparently, the template with the mocks cosmology fits the mean better, but the template is not capable of fitting all of the scales of the quadrupole and the hexadecapole simultaneously.

We can summarise the results of this section as follow: 1) the dispersion obtained from both sets is consistent with each other and with previous results from BOSS (Alam et al., 2017) and from the DR14 BAO group (Bautista et al., 2018). 2) both sets of eBOSS mocks lack the accuracy to study the biases of our methodology: the QPM mocks seem to slightly overpredict the quadrupole expected by the GSRSD model and are not a good match to the hexadecapole. The EZ mocks have a better match to the hexadecapole, but can not match the quadrupole at small scales (lower than 50 h^{-1} Mpc).

3.6.3 Comparison of AP parameters results with BAO-only fits

In this section, we compare our results to those obtained in Bautista et al. (2018), which is a previous analysis using this same sample. The left panel of Figure 3.14

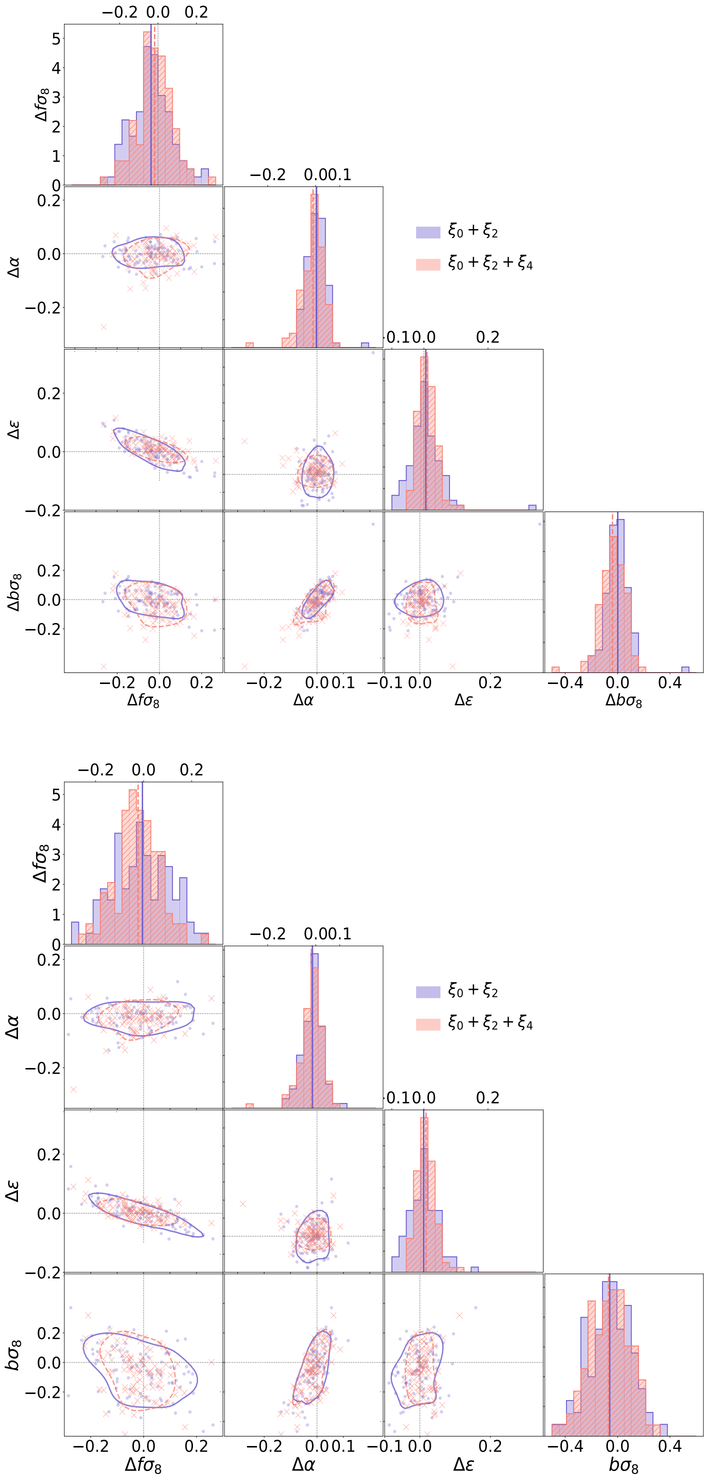


Figure 3.12: Scatter triangle plots comparing fits for full shape fits using $\xi_0 + \xi_2$ (blue dots) and $\xi_0 + \xi_2 + \xi_4$ (red x's) for QPM (up) and EZ (down) mocks. We show the difference of the best fit values with respect to the expected values for each of the parameters of interest. The means are indicated as solid lines for the two cases explored. The dotted lines indicate the expected values, which are zero for all cases.

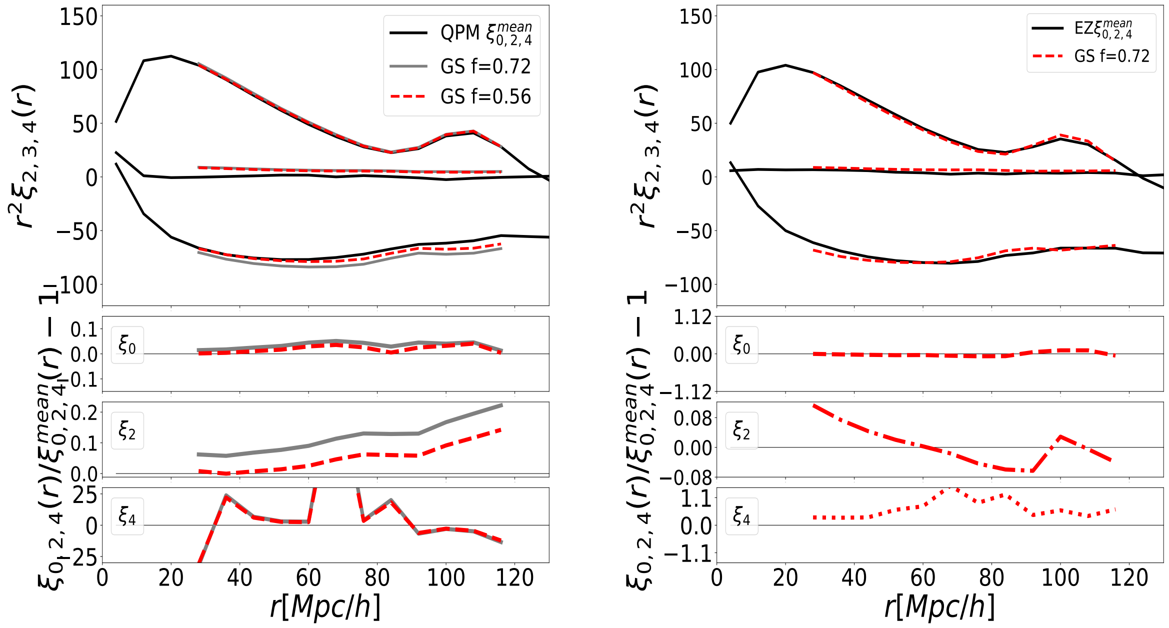


Figure 3.13: Mean QPM/EZ mocks vs. Template with Mock Cosmology. The error bars are smaller than the size of the points. For QPM mocks, the template with the mocks cosmology does not match the mean of the mocks (black line), this is evident in the quadrupole and hexadecapole residuals. For the EZ mocks, the template with the mocks cosmology fits the mean better, but the template does not match all of the scales of the quadrupole and the hexadecapole simultaneously.

Table 3.5: Results from fitting the 100 QPM/EZ mocks for FS analysis. We include the analysis for both cases using the hexadecapole in addition to the monopole and quadrupole. The columns denoted by $\widetilde{x}$ are the mean, and the S_x denotes the standard deviation. The variables are the difference of the parameters of interest compared to their expected values on a mock-by-mock basis, i.e. $\Delta f\sigma_8 = \langle f\sigma_8 - f\sigma_{8\mathrm{exp}}\rangle$, $\Delta\alpha = \langle\alpha - \alpha_{\mathrm{exp}}\rangle$, $\Delta\epsilon = \langle\epsilon - \epsilon_{\mathrm{exp}}\rangle$, for both the analysis using multipoles up to $\ell = 2$ and using multipoles up to $\ell = 4$.

Monopole-Quadrupole fits								
Model	$\widetilde{\Delta f\sigma_8}$	$S_{\Delta f\sigma_8}$	$\widetilde{\Delta\alpha}$	$S_{\Delta\alpha}$	$\widetilde{\Delta\epsilon}$	$S_{\Delta\epsilon}$	χ^2/d.o.f	N_{mock}
FS-QPM MQ	−0.036	0.113	0.003	0.039	0.006	0.053	1.0	97
FZ-EZ MQ	−0.007	0.122	0.009	0.043	0.001	0.044	1.0	91
Including Hexadecapole								
FS-QPM	−0.018	0.090	−0.011	0.050	0.009	0.028	1.1	84
FS-EZ	−0.024	0.089	0.005	0.050	0.008	0.028	1.0	97

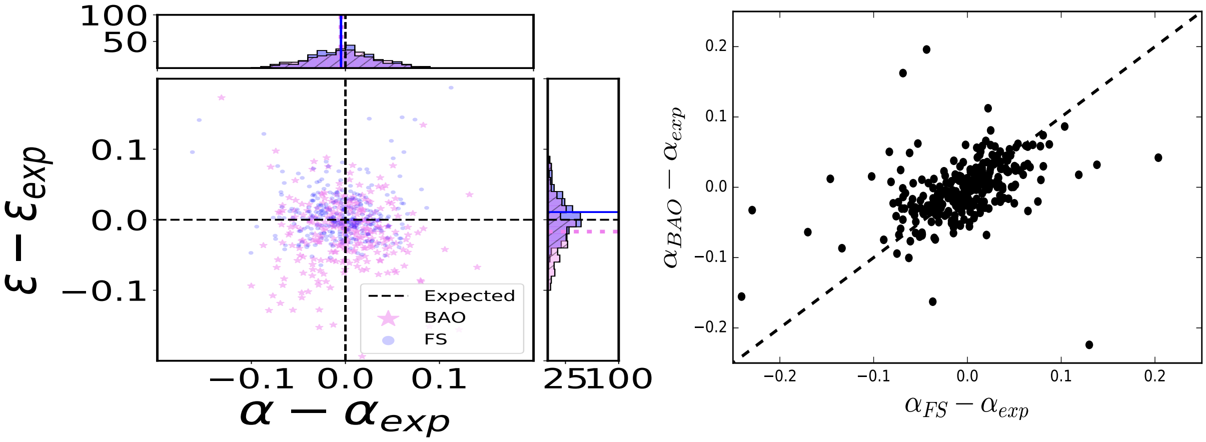


Figure 3.14: Left panel: A comparison of the BAO fits and full shape using $\xi_0 + \xi_2$ for the mocks. Right panel: comparison of best fits in isotropic dilation parameter for FS and BAO for the mocks. The dispersion for the anisotropic warping, ϵ, from BAO fits is slightly larger compared to the FS best fits. FS analysis breaks some degeneracies in ϵ and reduces its dispersion.

shows the difference between our QPM FS fits to the combined sample and the expected value compared to those from the anisotropic BAO parameters, the later taken from Bautista et al. (2018). The dispersion for the anisotropic warping, ϵ, from BAO fits is slightly larger compared to the FS best fits. In an RSD analysis other parameters that affect the quadrupole are included (most significantly the growth rate f), so it is not surprising that FS analysis breaks some degeneracies in ϵ and reduces its dispersion. There is also a small shift in the isotropic dilation parameter, α, when comparing the FS analysis best fits to those coming from BAO. The left panel of Figure 3.14 shows the scatter plot for α, with a Pearson correlation factor of $r = 0.5$. There are several differences in the fitting methodology between these two fits. Obviously the modelling of the signal is different in BAO and in our RSD+AP model, but in addition the fitting range used in BAO is wider in its r-range that is extended to 180 Mpc/h while our FS analysis is constrained to r-values lower than 130 Mpc/h. Also, the binning used in BAO is 5 Mpc/h in width, while this work is using bins with a width of 8 Mpc/h.

3.6.4 Testing the Impact of Spectroscopic Incompleteness

To test the effect of redshift incompleteness in our clustering, we consider three cases: the first is our mock catalogs with no redshift failures. Then, we study the effect of the two mitigations techniques described in Section 3.2.3.2. The redshift failures are added to the mocks by associating a position in the plate to each galaxy, then the catalog of binned probabilities is used to mimic the effect of the redshift failures observed in our data. The second case explored is the up-weighting methodology, and finally, for the third case, the forward-modelling technique.

Figure 3.15 displays the impact of different mitigation methods on the average of all 1000 mock catalog correlation functions. The three lines represent the case without redshift failure corrections and the up-weighting and forward modelling corrections. While the monopole is equally well recovered in all three cases, the quadrupole shows a clear shift (i.e., bias) at all scales when using the up-weighting method. The forward-modelling corrections recover the expected values for scales smaller than $r = 140$ Mpc/h, but show slight discrepancies at larger scales.

Table 3.6 lists the results of the best-fit parameters found by fitting all 100 QPM mocks using both correction schemes. We compared the results of the mocks where redshift failures are applied and corrected by one of the two mitigation techniques with the case where no redshift failures are considered. We report the difference of the mean of the best fits as an indicator of the systematic bias related to the spectroscopic completeness denoted by Δf, $\Delta\alpha$, and $\Delta\epsilon$; we also report the the dispersion S_x, where $x = f, \alpha,\ \epsilon$. We observe that the up-weighting technique differs from the case without redshift failures by $|\Delta f| = 0.016$ $(\Delta f/(S_f/\sqrt{N_{sim}})$ (0.7σ), $|\Delta\alpha| = 0.001$ (0.1σ), $|\Delta\epsilon| = 0.003$ (0.7σ). When using the forward modelling, the systematic error reduces to $|\Delta f| = 0.004$ (0.1σ), $|\Delta\alpha| =< 0.001$ $(<0.1\sigma)$ and $|\Delta\epsilon| = 0.005$ (0.8σ). There is an increase of the dispersion for the case of the up-weighting technique in the parameters f and ϵ, which decreases for f for the forward modelling scheme but is still larger when compared to the case without

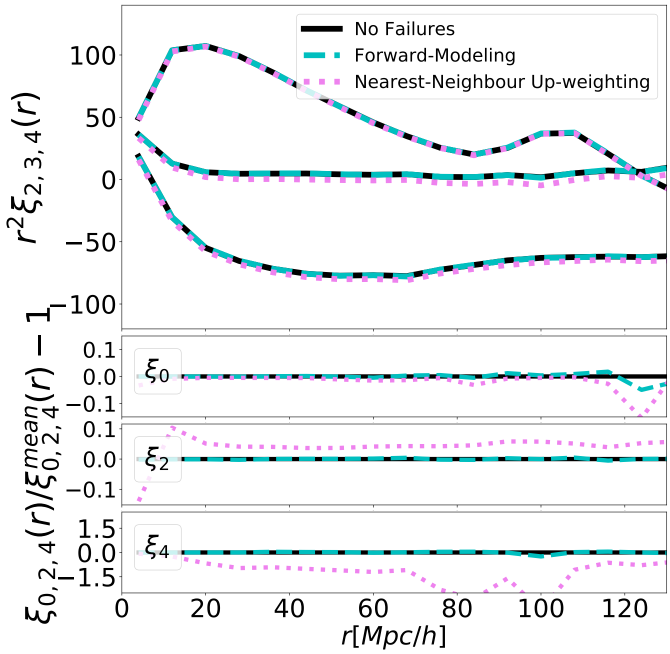


Figure 3.15: Impact of the redshift completeness on the multipoles and the effect of the mitigation techniques for correcting potential biases. The monopole (top), the quadrupole (bottom) and hexadecapole (middle) are presented in three cases: without redshift failures, correcting by the up-weighting technique, and correcting using the Forward modelling technique. While the monopole is well recovered by the two correction techniques, the quadrupole/hexadecapole shows a clear shift (i.e., bias) at all scales when corrected with the up-weighting method. The forward-modelling recovers the expected values for scales smaller than $r = 140$ Mpc/h.

redshift failures, but increases the shift by 0.002 on ϵ. In any cases the biases are less than 1σ. Given these results, we conclude that the forward modelling scheme performs slightly better than the up-weighting scheme. Therefore, in the rest of our analysis, we will adopt the forward modelling scheme for correcting the redshift failures.

3.7 Results on the LRG DR14 sample

We performed the analysis on the eBOSS-CMASS sample combining the NGC and SGC (if not otherwise stated). The covariance matrices used in our fits were rescaled by a factor of 0.9753 in order to account for the slight mismatch between

Table 3.6: Testing for Redshift Failures. Fitting results from 100 QPM mocks using two different techniques for mitigating the redshift failures. We compared the results of the the mocks where redshift failures are applied and corrected by one of the two mitigation techniques to the case where no redshift failures are considered. We report the difference of the mean of the best fits as an indicator of the bias related to the spectroscopic completeness denoted by Δf, $\Delta\alpha$, $\Delta\epsilon$ and we also report the dispersion by ΔS_x, where $x = f, \alpha, \epsilon$.

Testing Impact of Mitigation Techniques for Redshift Failures.									
Mitigation Methodology	Δf	S_f	$\Delta\alpha$	S_α	$\Delta\epsilon$	S_ϵ	$\Delta F'$	$\Delta F''$	$\Delta\sigma_{FOG}$
No Fiber Collisions	-	0.232	-	0.113	-	0.050	-	-	-
Forward Modeling	+0.003	0.252	<-0.001	0.116	+0.005	0.061	0.005	0.394	-0.271
Fiber Weights	-0.016	0.250	-0.001	0.112	+0.003	0.054	<0.001	-0.141	-0.234

the footprint area of the data and of the mocks.

Before running a full Monte Carlo Markov Chains (MCMC) analysis, we will compute the Best fit parameters using the minimisation methodology of the last chapter, which will help us understand how susceptible our models are to changes in the distance range of our analysis.

While these results were not expected to provide any information on the confidence contours of our parameters (as an MCMC would), they give an idea of the maximum likelihood values. The main reason for performing these tests is that our MCMC analysis in its current implementation is prohibitively time-consuming; we simply can not afford to run all the tests on our data using a full MCMC approach(as we will see in appendix 3.10.1, our models can be degenerated when using broad biases, how large the biases can be depends on the range of the bins and on the error sizes. These degeneracies make the convergence significantly slower). Further development needs to be done in order to reduce the time of convergence of our final analysis. These maximum likelihood tests can also be used as a check of the robustness of our MCMC results.

Our first test compares the robustness of the fit against variations in the maximum fitting range (the maximum distance in h^{-1}Mpc where the correlation function is measured). This test is particularly important in our analysis. Figure 3.5 shows that the quadrupole estimates made with the data show large correlations at scales larger than 100 h^{-1}Mpc, which are outside the variance observed in the mocks. This anomalous correlation at large scales affects the capability of our model to fit the data multipoles. We suspect this behavior could be related to an unknown systematic or a statistical fluctuation. Given that we could not identify any systematic that affects the quadrupole, and that we can not exclude a large fluctuation, we also analyzed the behavior of the fits when those large scales are eliminated in all multipoles with $\ell \geq 2$. Our main result, however, is quoted with the complete range. If this behavior is repeated in the DR16 analysis, that will indicate a systematic error that needs to be analyzed properly to provide non-biased results. If the origin of this correlation is a statistical fluctuation, this feature will probably be diluted with an increase in volume.

Thus, before exploring the likelihood surface, we performed some maximum likelihood fits using a variety of ranges. The fiducial case uses the complete range between [28,124]h^{-1}Mpc for the multipoles up to $\ell = 0, 2, 4$. We also tested some variants of this range to investigate the impact of cutting the large scale of each multipole on the best fits.

Table 3.7 lists the results of the best fits for the fiducial cases and several variants and figure 3.16 shows how the best fit models compare to the data. From table 3.7, we conclude that reducing the range of the fit from 128 to 92 h^{-1}Mpc improves the goodness of the fit for both the monopole+quadrupole fit and the monopole+ quadrupole+hexadecapole fit. The χ^2/d.o.f., a measurement of the goodness of a fit, reduces from 2.1 to 1.35 for the $\ell_{max} = 2$ case (However, there is no reduction when we reduce the range for both monopole and quadrupole where χ^2/d.o.f. stays at 2.09), and from 1.81 to 1.16 if we eliminate large scales for all $\ell = 0, 2, 4$ (It stays the same if we only limit hexadecapole, 1.14 if we restrict the range of the large scales for both $\ell = 2, 4$ but not for the monopole). By using the complete range we increase the discrepancy between the fits using different order multipoles

Table 3.7: Best Fits from Maximum Likelihood Fits for different scenarios: using the fiducial ranges for the multipoles up to $\ell = 2$ (first line), using multipoles up to $\ell = 4$ (second line), and systematically excluding the large scales for the different multipoles considered in the fits (lines three to seven).

Best Fits from Maximum Likelihood for LRG sample DR14								
Varying maximun range and ℓ								
model	range (h^{-1}Mpc)	F'	F''	f	α	ϵ	σ_{FOG}	χ^2
$\xi_0 + \xi_2$	[28,124][28,124]	1.005	0.74	0.905	0.947	-0.026	0.009	42.4/20=2.1
$\xi_0 + \xi_2 + \xi_4$	[28,124][28,124][28,124]	1.05	-2.7	0.694	0.965	0.038	-1.51	59.81/33=1.81
$\xi_0 + \xi_2$	[28,124][28, 92]	0.91	-3.28	0.710	0.935	0.050	2.48	24.3/18=1.35
$\xi_0 + \xi_2$	[28,92][28, 92]	0.753	-3.73	0.589	0.874	0.088	4.21	25.13/12=2.09
$\xi_0 + \xi_2 + \xi_4$	[28,124][28,124][28,92]	1.07	-2.58	0.690	0.969	0.038	0.96	52.0/29=1.79
$\xi_0 + \xi_2 + \xi_4$	[28,124][28, 92][28,92]	0.937	-2.96	0.571	0.92	0.092	5.07	28.6/23=1.14
$\xi_0 + \xi_2 + \xi_4$	[28,92][28, 92][28,92]	0.73	-3.79	0.508	0.858	0.120	6.49	24.34/21=1.16

(i.e. $\ell_{max} = 2$ versus $\ell_{max} = 4$). The difference in the best fit parameters for the growth rate f is 0.211 for the complete ranges (row one minus two), and it becomes 0.139 when reducing the quadrupole and hexadecapole ranges to [28,92]h^{-1}Mpc (row three minus six). Similar trends occur with ϵ, where the differences in the best fit values range from 0.064 to 0.042. These trends indicate that the large correlation observed in the quadrupole is not properly modeled by our CLPT-GSRSD template. When we exclude the large scales of the monopole, there is a significant shift in both f and ϵ, with f shifting from 0.905 to 0.589 and ϵ from -0.026 to 0.088. These shifts are expected when eliminating the large scales on the monopole. Finally, excluding the large scales on the hexadecapole affects the f fits, and mildly affects the ϵ fits, as the quadrupole and hexadecapole capacity to break the degeneracy between f and ϵ is derived from the BAO scales. The goodness of the fit, χ^2/d.o.f., improves when removing the large scales, due to the incapability of modelling the anomalous correlation; this approach loses all the information encoded in the BAO in the quadrupole and hexadecapole. Consequentially, the results for the AP parameters are degraded and potentially biased. We will perform the MCMC exploration for the same four cases for completeness, but we will quote the full range as our final result.

As stated, we used an MCMC methodology for exploring the likelihood surface,

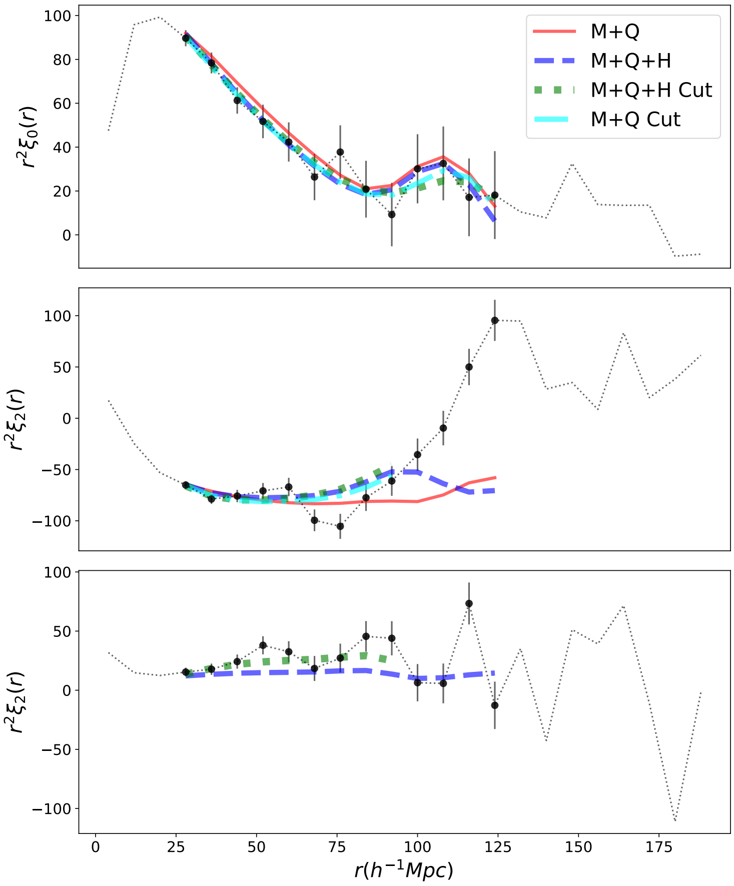


Figure 3.16: The maximum likelihood model for four cases: 1) using multipoles up to $\ell = 2$ in the fiducial range, 2) using multipoles up to $\ell = 4$ for the fiducial range, 3) using multipoles up to $\ell = 2$ but restricting the quadrupole range to [28,92] h^{-1}Mpc, and 4) using multipoles up to $\ell = 4$ but restricting the range of the quadrupole and hexadecapole to [28,92] h^{-1}Mpc.

which was done using the Monte Python public code (Audren et al., 2013). We use flat priors for our parameters; the range of these priors is presented in Table 3.8. We run two different chains in the combined data set (NGC+SGC). The first is with the monopole and quadrupole only ($\xi_0 + \xi_2$) and using the fiducial distance range. The second chain also runs with the monopole and quadrupole ($\xi_0 + \xi_2$), but restricting the range in the quadrupole to [28,92] h^{-1}Mpc (the monopole stays in the same range of [28,124] h^{-1}Mpc). Table 3.9 displays the results from the MCMC analysis. Our final measurement was performed on the *combined* sample, which includes the NGC and the SGC, and was done using the fiducial methodology (i.e., 8 h^{-1}Mpc bins on the fiducial range). The first block reports the final result of this

Table 3.8: Flat priors ranges on the parameters of the model.

Measurements with LRG sample DR14.	
f	[0.0,2.0]
F'	[0.0,3.0]
F''	[-15,15]
σ_{FOG}	[0,40]
α	[0.8,1.2]
ϵ	[-0.2,0.2]

work, the monopole and quadrupole-only fits. The second block is for the $\xi_0 + \xi_2$ fits when excluding the large scales of the quadrupole. The third block lists the results for the Galactic hemispheres separately, this is shown for identification of any residual systematics in the data; we will discuss these results at the end of this subsection. The fourth block is quoted as a reference and it shows the fits of the BAO-only analysis done with this same sample in Bautista et al. (2018). MCMC chains using the hexadecapole are included in the final block for completeness and discussed in the Appendix 3.10.2 as a robusteness test, but it is not part of our main results.

Figure 3.17 shows the likelihood surfaces for the two runs over the $\xi_0 + \xi_2$, one chain is in the fiducial range and the other is eliminating the large scales for the quadrupole, i.e. [28,92] h^{-1}Mpc. The latter is added for completeness, but as stated before, our final result will be quoted using the full range. The figure contains the $1-2\sigma$ confidence contours for the growth rate $f\sigma_8$, the linear bias $b\sigma_8$, the dilatation parameter α, and the warping parameter ϵ, together with their marginalised 1D distributions. The 1-σ regions are fully contained inside our priors for both cases. However, the 2-σ regions are cut by our prior to large values of ϵ and small values of α, our reasons for not using larger priors on the Alcock-Paczynski parameters will be discussed in the Appendix 3.10.1.

The results of $f\sigma_8$ and α are consistent within 1σ, for both ranges. However, given the anti-correlation between the ϵ and f parameters and the fact that the quadrupole is dominated by ϵ at the larger scales, when including the larger bins of the quadrupole the ϵ is driven from its expected value and as a consequence f shifts as well. When the last three bins of the quadrupole are avoided we achieve a significant improvement in the goodness of the fit towards ~ 1; the price paid for this approach is to eliminate the BAO information. This increases the degeneration of the parameters and biases the results, thus we lose information that constrains ϵ, which in turn causes the contour areas to become larger, providing more freedom to the fitter to move f to lower values (Figures 3.17).

Finally when comparing the results for the $\xi_0 + \xi_2$ with the fit using $\xi_0 + \xi_2 + \xi_4$ we find agreement within $1 - \sigma$ for $f\sigma_8$ and α, but the ϵ values have a $1.3 - \sigma$ difference. We should notice that tighter priors were used for hexadecapole because a bi-modality appears using the priors defined in table 3.10.1, more discussion about the results and the prior selection for the hexadecapole is provided in the Appendix 3.10.2.

Figure 3.18 displays the best-fit anisotropic models compared to the data for our fiducial choice of analysis. As expected, the monopole and quadrupole are visually good fits for the data, except for the large scales of the quadrupole where the correlation becomes strongly positive (scales larger than 90 h^{-1}Mpc).

To finalise this section, we analyze separately the North Galactic Cap (NGC) and the South Galactic Cap (SGC). Figure 3.19 displays the 1σ and 2σ confidence contours obtained from running an MCMC analysis separately on both hemispheres. They are computed using our standard priors quoted in Table 3.8. The contours in both galactic caps are poorly defined and the 1σ interval in ϵ and α are sharply cut by our imposed priors.

As discussed in appendix 3.10.1 our methodology has difficulty on fixing the Alcock-Paczynski parameters to a unique value given the size of our errors compared to the strength of the BAO signal, this leads to unphysical values of ϵ and α (and therefore f due to their correlation with ϵ) being accepted by the MCMC chain and affects all of the constraints of our parameters.

Table 3.9: Results for the DR14 LRG sample. The first block is for our fiducial methodology, using the fiducial range for the $\xi_0 + \xi_2$ fit. The second block is for the $\xi_0 + \xi_2$ fits when excluding the large scales of the quadrupole. The third block shows the fits separating the hemispheres NGC and SGC and using $\xi_0 + \xi_2$ in the fiducial range. The fiducial value for the $\sigma_8(z_{eff=0.72}) = 0.55$. The eulerian bias is defined as $b = 1 + F'$., The second section of the table shows the BAO only results from Bautista et al 2018. The range of the fits is between [32,182] and they use a bin size of 8 Mpc/h. The third section includes our results using the hexadecapole.

Measurements with LRG sample DR14 Official Version.

Case	$f\sigma_8$	$b\sigma_8$	$< F'' >$	σ_{FOG}	α	ϵ
$\xi_0 + \xi_2$ [28,124][28,124]	$0.454^{+0.119}_{-0.140}$	$1.110^{+0.116}_{-0.100}$	$2.245^{+3.849}_{-4.35}$	$3.713^{+2.987}_{-2.31}$	$0.955^{+0.055}_{-0.05}$	$0.000^{+0.090}_{-0.050}$
$\xi_0 + \xi_2[28, 124][28, 92]$	$0.337^{+0.121}_{-0.110}$	$1.088^{+0.101}_{-0.100}$	$-1.19^{+4.002}_{-2.900}$	$5.027^{+2.721}_{-2.870}$	$0.930^{+0.050}_{-0.050}$	$0.083^{+0.059}_{-0.06}$
$\xi_0 + \xi_2$ NGC	$0.598^{+0.150}_{-0.190}$	$1.262^{+0.121}_{-0.150}$	$4.372^{+3.657}_{-5.810}$	$3.008^{+2.740}_{-1.940}$	$1.103^{+0.066}_{-0.100}$	$-0.05^{+0.085}_{-0.040}$
$\xi_0 + \xi_2$ SGC	$0.359^{+0.168}_{-0.16}$	$1.119^{+0.169}_{-0.12}$	$0.328^{+1.725}_{-1.96}$	$4.783^{+3.732}_{-3.00}$	$0.929^{+0.087}_{-0.07}$	$0.077^{+0.081}_{-0.07}$

Measurements BAO-only with LRG sample DR14 from Bautista et al 2018.

Case	range	$\alpha_\perp$	$\alpha_{\|\|}$	corr	α	ϵ
Anisotropic	26-178	$1.01^{+0.08}_{-0.05}$	$0.82^{+0.09}_{-0.08}$	-0.39	$0.942^{+0.048}_{-0.024}$	$-0.067^{+0.033}_{-0.022}$

Measurements with LRG sample DR14 including hexadecapole.

Case	$f\sigma_8$	$b\sigma_8$	<F”>	σ_{FOG}	α	ϵ
$\xi_0 + \xi_2 + \xi_4$ [28,124][28,124][28,124]	$0.31^{+0.09}_{-0.09}$	$1.19^{+0.10}_{-0.10}$	$-1.1^{+3.2}_{-3.3}$	$5.8^{+3.3}_{-3.2}$	$0.986^{+0.047}_{-0.046}$	$0.091^{+0.046}_{-0.048}$

Given that the errors are larger for the North and the South separately that in the combined sample, the degeneration is stronger. This leads to several regions being accepted to within 1-σ that would otherwise be rejected due to their inability to reproduce the BAO peak. More data will tend to reduce this behavior and that is in fact what we see in the combined sample, where the errors are smaller.

Figure 3.20 presents the data multipoles for the NGC (blue points) and SGC (red x's); the error bars correspond to their 1-σ variance from the sample covariance matrix computed using the QPM mocks. The blue solid line represents the fits made by our MCMC analysis in the NGC, the red dashed line is the analog for the South Galactic Cap. These fits are done using the mean values obtained by our MCMC chains, which we use as our estimates of the best fits. The NGC and SGC have a significant difference in the clustering amplitude at small scales, and the

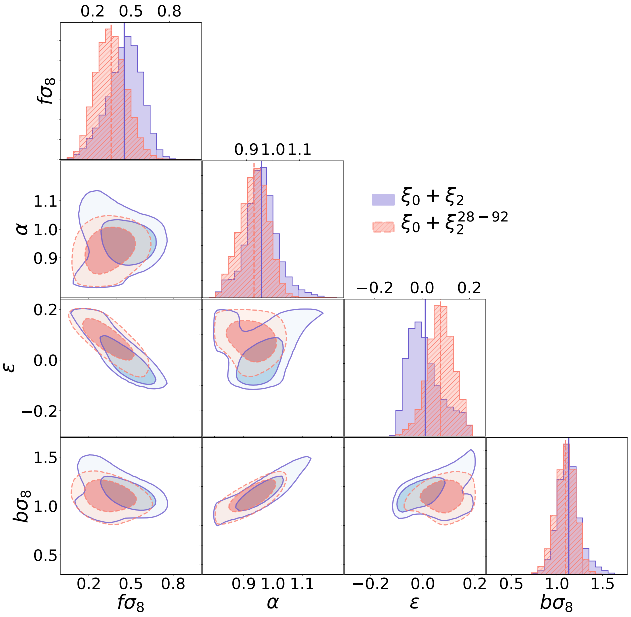


Figure 3.17: The shaded regions show the $1-2\sigma$ confidence surfaces found by our MCMC chains for the RSD-AP parameters for the $\xi_0+\xi_2$ space in the fiducial range (blue solid) and when excluding the large scales in quadrupole (red dashed). The confidence contours for the growth rate $f\sigma_8$, the linear bias $b\sigma_8$, the dilatation parameter α, and the warping parameter ϵ are indicated, along with their 1D distributions. The dashed lines of each histogram are the mean values found by the MCMC chain.

peak is shifted in one hemisphere compared with the other (it is not well defined in either of the hemispheres). Both models reasonably reproduce the multipoles; this is especially true in the smaller scales which are the ones with more weight in the fit (due to their smaller variance). There is a difference in the multipole amplitude between both galactic caps, as a consequence the contours in Figure 3.19 are displaced among each other, the results for the combined sample surrounds the regions where both contours intercept (see table 3.9).

3.8 Cosmological Implications

Table 3.10 presents our final constraints on the growth rate $f\sigma_8$, the angular diameter distance $D_A(z)$, and the Hubble parameter $H(z)$ including the statistical

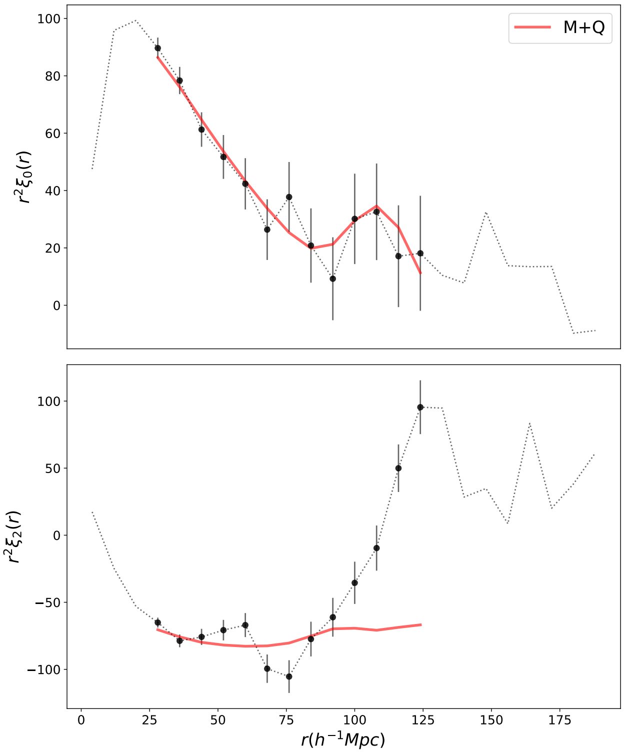


Figure 3.18: This plot shows how our data (black dots) compares to our model (red solid line). The model is built using the values from the first line of table 3.9, that were computed using multipoles up to $\ell = 2$ and our combined sample (NGC+SGC) in the fiducial range. The model is visually a good fit for the data, except for the large scales of the quadrupole.

and the systematic error*. Our fiducial cosmology was used to convert the best-fit dilation parameters $\alpha_{||}$ and $\alpha_{\perp}$, into distance measurements. The table includes the same variants of the methodology quoted in table 3.9 for the combined sample, and the values are in agreement with each other within 1σ.

Our final constraint, the logarithmic growth of structure multiplied by the amplitude of dark matter density fluctuations, is $f(z_{eff})\sigma_8(z_{eff}) = 0.454{\pm}0.134$. Using the Alcock-Paczynski dilation scales allowing us to constrain the angular diameter distance and the Hubble distance we arrive to: $D_A(z_{eff}) = 1466.5 \pm 133.2\,(r_s/r_s^{fid})$ and $H(z_{eff}) = 105.8 \pm 15.7\,(r_s^{fid}/r_s)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ where r_s is the sound horizon at the end of the baryon drag epoch and r_s^{fid} is its value in the fiducial cosmology at an effective redshift $z_{eff} = 0.72$. These measurements correspond to relative

*The systematic error is based on the results from N-Series.

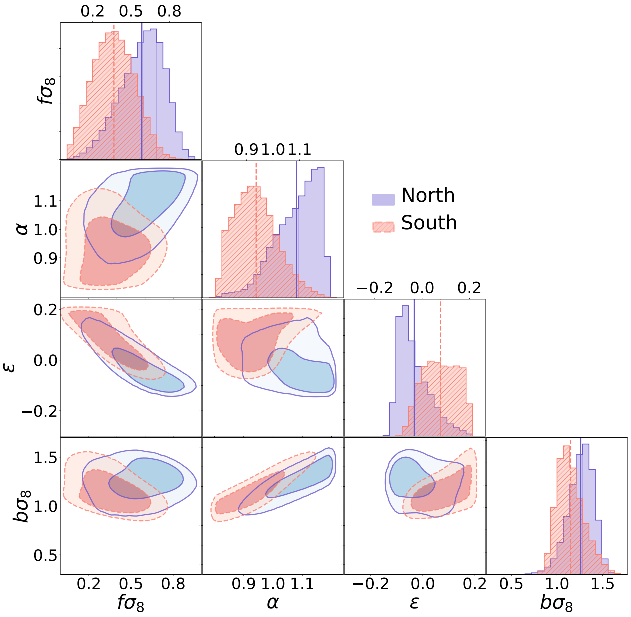


Figure 3.19: Equivalent to Figure 3.17 but presenting the MCMC chains for the RSD-AP parameters in the NGC (blue solid) and in the SGC (red dashed); both of them are in the fiducial range.

errors of 29.4%, 9.1%, and 14.9%, respectively considering the systematic error.

Bautista et al. (2018)'s analysis with the DR14 LRG sample reported a low statistical power of the current sample, and generated anisotropic BAO results yielded slightly worse results than isotropic fits. Further data releases from eBOSS should increase the statistical significance of our measurements.

Figure 3.21 presents our measurements compared with previous results from SDSS-III-BOSS DR12 from both galaxies (Alam et al., 2017) and Lyman-alpha quasars (du Mas des Bourboux et al., 2017; Bautista et al., 2018), the eBOSS quasar measurements from Gil-Marín et al. (2018); Zarrouk et al. (2018); Hou et al. (2018)*, and the Main Galaxy Sample (MGS) from SDSS-II-DR7 (Ross et al., 2015). Our measurements are consistent with previous analyses and the ΛCDM model.

Our measurements with the CMASS-eBOSS sample are correlated with the CMASS measurements. The correlation coefficient between the two measurements was

*Figure 3.21 quotes the (Gil-Marín et al., 2018) result; however, the three measurements from the different analyses were shown to be fully consistent.

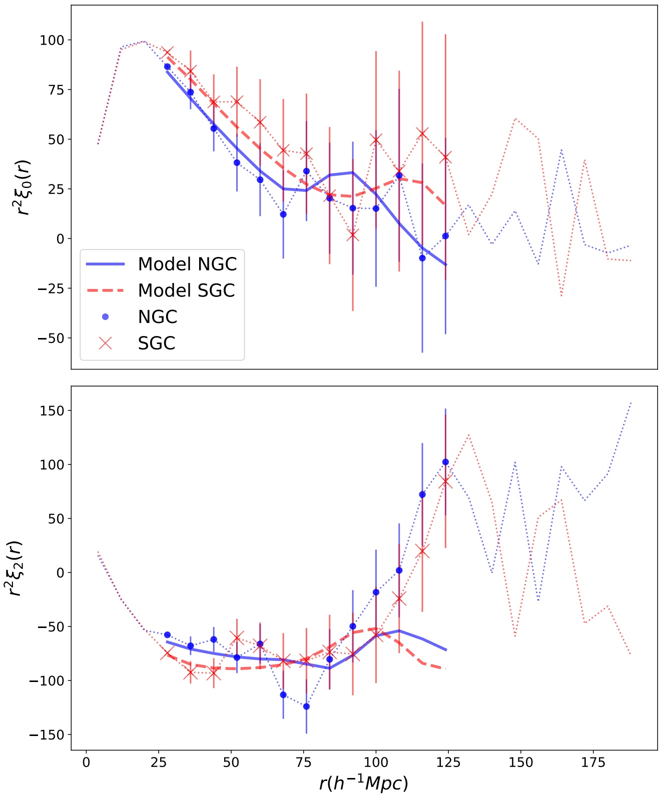


Figure 3.20: The blue dots and red x's represent the NGC and the SGC measured data points for the Monopole (top) and the Quadrupole (bottom). The errors bars are the standard deviation computed using the QPM mocks. The solid lines indicates the best model found by the MCMC for the NGC/SGC.

Table 3.10: Cosmological constraints on DR14 LRG sample, using $D_A(z = 0.72)^{fid} = 1535, H(z = 0.72)^{fid} = 101$. Systematic error included.

Measurements with LRG sample DR14.						
Model	range (h^{-1}Mpc)	$\alpha_{\|\|}$	$\alpha_{\perp}$	$f\sigma_8$	$D_A(z)(r_s^{fid}/r_s)$	$H(z)(r_s/r_s^{fid})$
$\xi_0 + \xi_2$	[28,124][28,124]	0.954±0.149	0.955±0.083	0.454±0.134	1466.5±133.2	105.8±15.7
BAO-only						
$\xi_0 + \xi_2$	[32,182]	0.72	0.82 ± 0.085	1.1 ± 0.065	-	

roughly estimated to be 0.16 (Bautista et al., 2018); a proper measurement of this correlation will be achieved for the DR16 analysis.

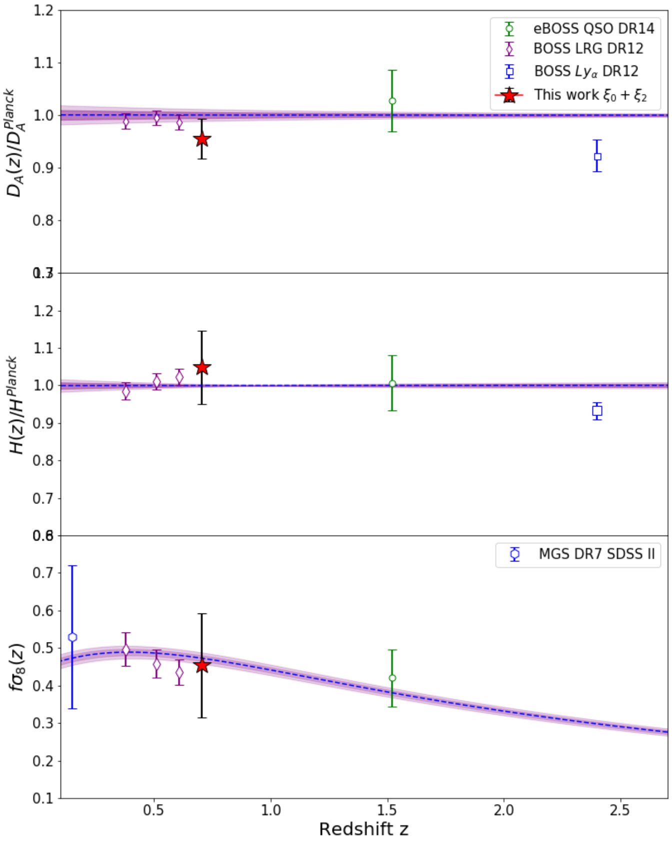


Figure 3.21: Measurements from DR14 eBOSS-CMASS sample using multipoles up to $\ell = 2$ (red start) fitting in the range [24,128] h^{-1}Mpc.

3.9 Conclusions

The RSD effect generates an artificial anisotropy on the clustering of galaxies which can be used to constrain the growth rate, $f(z)\sigma_8$, and the radial and angular distances to the sample (i.e., the $H(z)$ and $D_A(z)$ parameters). We used the LRG sample from the first two years of the eBOSS, denoted as DR14, to measure these parameters at the mean redshift of the survey ($z = 0.72$). We presented the first full-shape analysis of this sample (i.e. modelling Redshift-Space Distortions (RSD) simultaneously with an Alcock-Paczynski (AP) parametrisation), and that should be followed up on and improved on once the full observational time of the eBOSS survey is completed for the final DR16 sample. The measured correlation function was decomposed into the first three non-zero multipoles of its Lagrange expansion, and compared with theoretical predictions made with a Convolution Lagrangian Perturbation Theory (CLPT) model combined with a Gaussian Streaming model (GS). We considered six free parameters, four RSD-parameters $[f, F', F'', \sigma_{FoG}]$ and two AP parameters $[\alpha, \epsilon]$.

We tested our methodology using a set of 84 high-precision N-Series CutSky mocks built with BOSS-CMASS properties. We fitted all individual mocks using two different methodologies: using only multipoles up to $\ell = 2$, and using all multipoles up to $\ell = 4$. The fits using all the multipoles were computed in two different distance ranges, first using the complete [28,124] h^{-1}Mpc range for all of them, then removing the smaller scales of the hexadecapole. This extends on previous works that performed this exploration using only the monopole and the quadrupole. From the individual fits the most accurate results (smaller parameter biases in all parameters normalised by the dispersion) are obtained using the multipoles up to $\ell = 4$ in the full range. Besides the fact we do not find significant biases in the distributions, when fitting the mean we noticed that the model hexadecapole does not accurately match the mean of the mocks, this generates small biases in the fits of the mean ϵ parameter. The reason why this mismatch does not bias our measurements in the individual realisations is because of the larger errors bars we have on the hexadecapole. This behavior is related to the fact that the fits are driven by ϵ when we include the lower bins of the hexadecapole. The error bars for those lower scales are smaller, and therefore their constraining power is larger. This makes the accuracy of the model at small scales critical.

In order to characterise the statistical properties of the sample, especially its variance, we run our fitting methodology on two different sets of low-precision mocks with eBOSS properties: the QPM and EZ mocks. All of the mocks in both sets are fitted twice, the first considering only the multipoles up to $\ell = 2$, and the second with multipoles up to $\ell = 4$. The dispersion obtained from the two sets of low-precision mocks was fairly consistent in all cases and for all of the parameters of interest. However, the biases and distributions were not consistent with those obtained using high-precision mocks. The discrepancy arises because the GSRSD model can not match the multipoles of QPM/EZ mocks, thus no conclusion about the bias could be extracted from these fits. They were only used as a reference for the variance of the best fits for eBOSS-like mocks.

The tests performed with mocks (high and low precision), demonstrated that the constraining power of the lower bins of the hexadecapole is large due to the smaller error bars of those points. We concluded that including the hexadecapole is desirable; however, it becomes critical to have accurate models, particularly of the small scales of the quadrupole and the hexadecapole. In this work, we adopted the conservative approach of reporting the ξ_0+ξ_2 as our final result and used the hexadecapole results only as a consistency test.

We considered that even if the results with high-precision mocks validated fitting the hexadecapole with our model, the biases observed when fitting the mean and the mismatch in the model hexadecapole for the mean needs further exploration. Additionally, we did not have high-precision mocks with the properties of the eBOSS sample available (higher redshift and lower mean density) and we could not properly study the statistical properties of the fitting methodology with low-precision mocks.

Our final measurement was performed on the “combined" sample, using the fiducial methodology considering only the monopole and quadrupole. We constrained the logarithmic growth of structure $f\sigma_8 = 0.454\pm\ 0.134$, $\alpha_{||} = 0.954\pm\ 0.149$ and $\alpha_{\perp} = 0.955\pm\ 0.083$.

The eBOSS DR14 LRG sample presents a large correlation in the large scale quadrupole that lies outside the 1σ variation observed in the mocks. This feature could be related to an unknown systematic effect or just a large statistical fluctuation.

Given that we could not find any systematic that affects the quadrupole, and that a large fluctuation cannot be excluded, we analyzed the behavior of the fits when we eliminated those large scales in multipoles with $\ell >= 2$ as a robustness test for our main result. Avoiding the latest three bins of the quadrupole achieves a significant improvement in the goodness of the fit to ~1; however the price paid is to eliminate the BAO information, which increases the degeneration of the parameters and biases the results, thus we lose information that constrains ϵ, and the contour regions become larger, giving more freedom to the fitter to move f to lower values.

We quote as our final cosmological constraint the logarithmic growth of structure multiplied by the amplitude of dark matter density fluctuations, $f(z_{eff})\sigma_8(z_{eff}) = 0.454 \pm 0.134$, and the Alcock-Paczynski dilation scales which allow constraints to be placed on the angular diameter distance $D_A(z_{eff}) = 1466.5 \pm 133.2(r_s/r_s^{fid})$ and the Hubble distance $H(z_{eff}) = 105.8 \pm 15.7(r_s^{fid}/r_s)$kms^{-1}Mpc^{-1}, where r_s is the sound horizon at the end of the baryon drag epoch and r_s^{fid} is its value in the fiducial cosmology at an effective redshift $z_{eff} = 0.72$. These measurements correspond to relative errors of 29.4%, 9.1%, and 14.9%, respectively considering the systematic error.

Our results are consistent with previous measurements and with a ΛCDM model using Planck 2018 cosmology.

3.10 Appendix of chapter 3

3.10.1 Selecting priors

As discussed in chapter 3.7, figure 3.17 shows that the 1-σ regions are fully contained inside our priors. However, the 2-σ regions are clearly cut by our prior to large values of ϵ. In this section we will discuss our reasons for not using larger ϵ

priors in our analysis, as we will see prior selection was challenging given the size of our errors.

Figure 3.18 shows a comparison between our final model and the multipoles of the data set, it is clear that the detection of the BAO signal is weak: the error-bars of the monopole have a similar size to the power of the BAO peak. This is problematic as the BAO peak locks the Alcock Paczynski parameters around a specific value. Given the limited capability of our methodology to fix the cosmology our model is vulnerable to being degenerated. As a consequence, we have to be very careful when choosing our priors as a large prior in the Alcock Paczynski parameters will result in degenerated regions contributing significantly to our statistics.

This is shown in figure 3.22 where we have run a second MCMC chain of our fiducial methodology but extending the priors of ϵ to $[-0.3, 0.3]$. These chains were done with a fixed value of $F2 = 0.0$ to save computational time as the goal of is not to obtain precise statistics but to show the effect of larger ϵ priors (F2 contributions to our model corresponds to second-order corrections on small scales, primarily broadening the parameter contours). The priors for the other parameters (i.e. neither ϵ nor F2) stay at the value quoted in table 3.8.

In figure 3.22, the solid-line blue contours show our default results, while the dashed-line red ones show those with the enlarged priors on ϵ.

A second locus is present for large values of ϵ and small values of $f\sigma_8$. This second locus is centred somewhere around $f \approx 0.3$ assuming a nominal σ_8 value consistent with Planck ($\sigma_8(z_{eff}) = 0.55$), which would result in the Alcock-Paczynski parameters switching the cosmology to $\Omega_M(z = 0) \approx 0.03$ (for $\sigma_8(z = 0) = 0.8$ and a flat universe, assuming that $f(z) \approx \Omega_M(z)^{0.6}$). This strongly disagrees with previous constraints made by Planck, that predicts a value of $\Omega_M(z = 0) = 0.315 \pm 0.007$ (Planck Collaboration et al., 2020). Hence, the DR14 data does not allow us to broaden the priors too much, as the accuracy is not yet there in the data to rule out cosmological parameters already strongly rejected by Planck measurements.

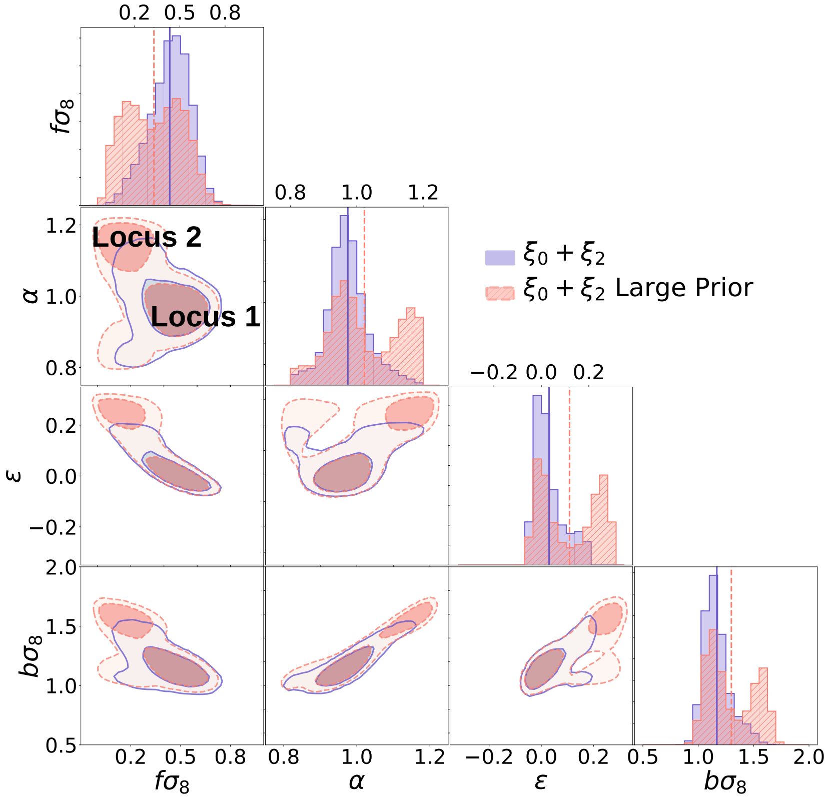


Figure 3.22: This plot is equivalent to Figure 3.17, here we are presenting the MCMC chains of two fits to the RSD-AP parameters in the $\xi_0+\xi_2$ space done with diferent priors in ϵ. The blue solid-line contours use the priors quoted in table 3.8 for all parameters but $F2$ that is set to zero. The red dashed-line contours have larger priors on ϵ which are expanded to $[-3, 3]$ and also set $F2$ to zero.

Figure 3.23 shows why this secondary locus is chosen by our MCMC analysis to be an acceptable fit. The blue line is the median of the models of 100 points chosen randomly from the subset of MCMC points within the locus centred around $\alpha \approx 1$ and $\epsilon \approx 0$ ($l$ocus 1 in figure 3.22). The blue shaded regions indicate the 18^{th} and 84^{th} percentile confidence range. The red line and line-shaded region correspond to models randomly selected from points of our MCMC chain inside locus 2 (top panel of 3.22).

From figure 3.23 we observe that the best fit model within locus 2 do not show a well defined BAO peak. However, statistically, both sets of models are equally good and indistinguishable in terms of their likelihood. DR16 should have smaller errors around the BAO signal which could in principle discard this second solution (locus 2). As we have stated, this second locus is discarded using Planck CMB constraints, therefore we consider reasonable to choose priors on the Alcock Paczynski parameters that keep it out of our statistics. Considering mild Planck CMB constraints, it is reasonable to assume priors on α and epsilon of ± 0.2 around their nominal value, as Planck strongly rejects cosmologies that are beyond that alpha and epsilon range to several sigmas.

Figure 3.24 is included as a robustness test of our methodology, here our fiducial result is compared with a new MCMC chain computed reducing the priors of α to $[0.9, 1.1]$. As in 3.22 $F2$ is set to zero for both chains to save computational time. The plot shows that the α contours are cut by the new priors, nevertheless, the 1-σ contours of both chains are centred around the same values and have a similar shape, the main difference being marginaly reduced size of the contours, which is expected when reducing the priors.

3.10.2 Likelihoods for eBOSS sample using hexadecapole.

In section 3.6.1 we applied our methodology to find the maximum likelihood fits of 84 Nseries high-resolution simulations. We have shown that our methodology provides consistent results with and without hexadecapole information.

As stated in section 3.7, we adopted the conservative approach of reporting the ξ_0+ξ_2 as our final result and using the hexadecapole results just as a consistency test. In this appendix, we show results including the hexadecapole.

We run two different chains that include ξ_4 using the combined data set (NGC+SGC). One using the priors shown in table 3.8, and a second chain with more constraining priors. The main reason for this choice is that when considering the priors quoted in table 3.8 we find a double peak when fitting the full range, which is shown in Figure 3.25. The Figure shows the $1-2\sigma$ confidence contours for the growth rate $f\sigma_8$, the linear bias $b\sigma_8$, the dilatation parameter α, and the warping parameter ϵ,

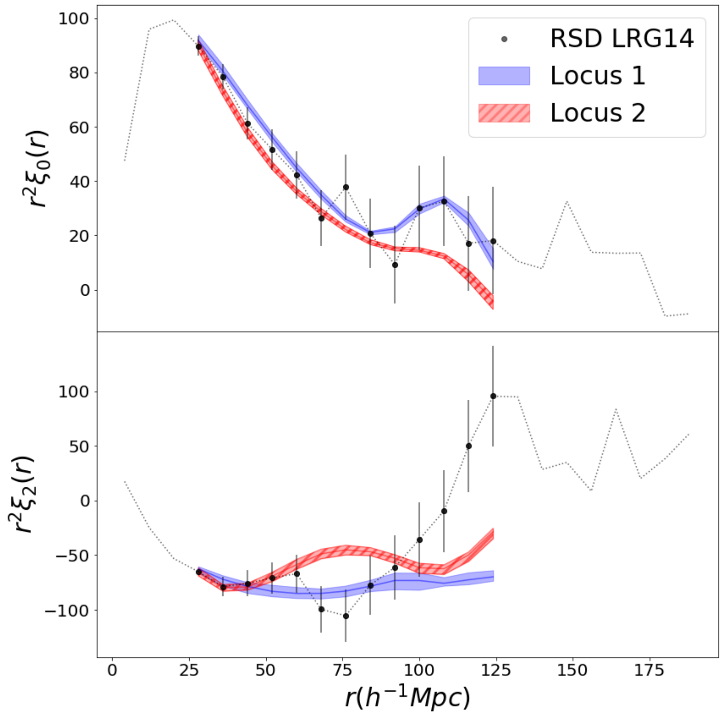


Figure 3.23: This figure shows the median and 1-sigma percentiles for 2 sets of 100 models built with 100 points chosen randomly from the subset of those explored by our MCMC. The blue shaded regions correspond to points inside the peak centred around the expected cosmology ($\alpha \approx 1$ and $\epsilon \approx 0$). The red line and line-shaded region contours are computed with points inside the second peak that appears for large values of ϵ.

together with their 1D distributions. The only difference between both plots are the priors; The red dashed-line represent a chain with the priors of table 3.8, in the blue solid-line contours the priors in α have been reduced to the interval [0.88, 1.12].

As discussed in appendix 3.10.1, this double peaked distribution is a consequence of degenerated solutions not being rejected due to the size of our errors. Following the same procedure done in appendix 3.10.1 we will only analyse the solution that is not in disagreement with mild Planck CMB constraints. In order to try to avoid this second degenerate solution we will reduce the size of our priors in the α parameter to the interval [0.88, 1.12], while the rest of the parameters are fixed to the values of table 3.8, these priors were chosen arbitrarily so that they contain the 1-σ region of the main peak and completely exclude the second. We acknowledge that it is

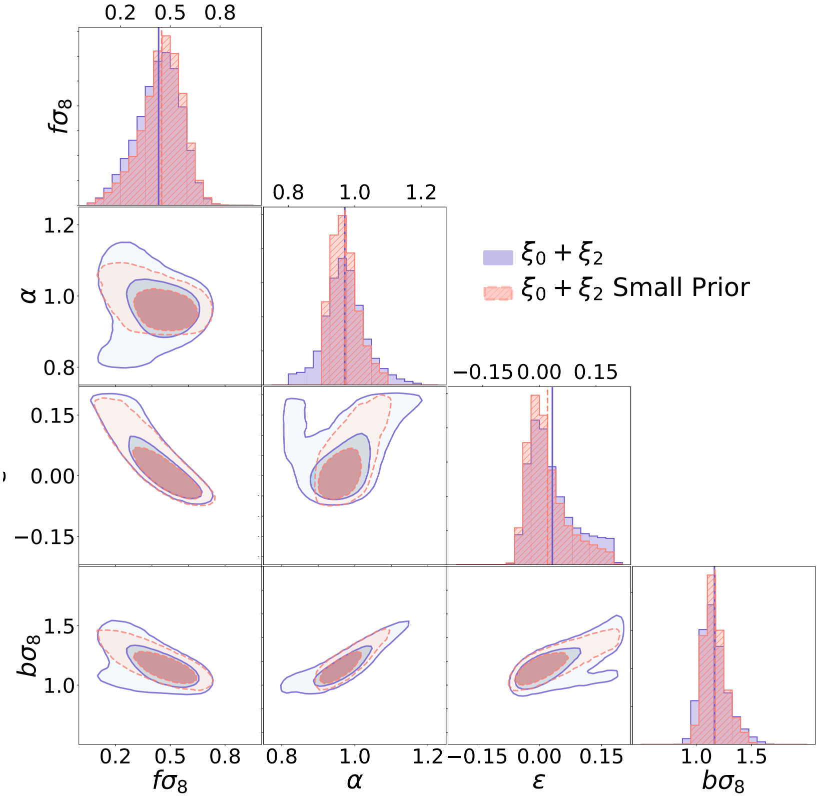


Figure 3.24: This plot is equivalent to Figure 3.17, here we presenting the MCMC chains of two fits to the RSD-AP parameters in the $\xi_0 + \xi_2$ space done with diferent priors in α. The solid-line contours use the priors quoted in table 3.8 for all parameters but $F2$ that is set to zero. The red dashed-line contours have smaller priors on α which are reduced to $[0.9, 1.1]$ and also set $F2$ to zero.

possible for the statistics obtained from this chain to still be slightly distorted by the presence of this second peak or by the position of the more constrained prior, the reduced error bars of DR16 should make the second peak less significant which could allow us to use larger priors.

The statistical results of our parameters are quoted in table 3.11: the first line repeats for comparison purposes the results for monopole and quadrupole only ($\xi_0 + \xi_2$). The rest of the table includes the results using monopole, quadrupole, and hexadecapole ($\xi_0+\xi_2+\xi_4$). The second line uses the full [28,124] h^{-1}Mpc range in all multipoles. In the last line, the monopole is in the [28,124] h^{-1}Mpc range, and the quadrupole and hexadecapole are in the [28,92] h^{-1}Mpc range, we cut out

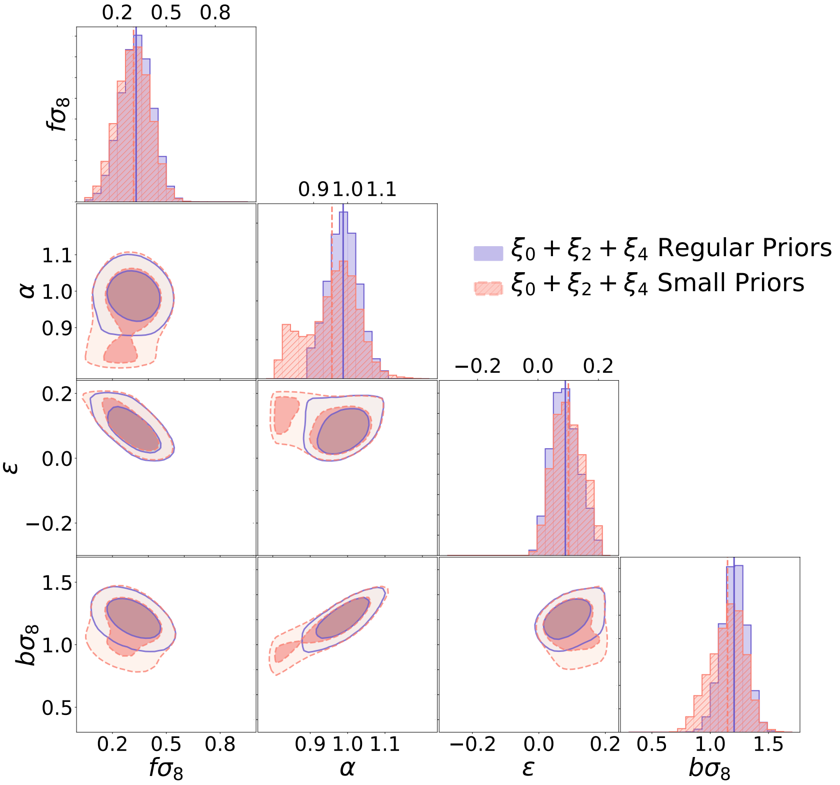


Figure 3.25: The shaded regions show the $1-2\sigma$ confidence surfaces found by our MCMC chains for the RSD-AP parameters using $\xi_0+\xi_2+\xi_4$ in the [28,124] h^{-1}Mpc range. The red dashed-line contours represent a chain with the priors of table 3.8, in the blue solid-line contours the priors in α have been reduced to the interval [0.88, 1.12]. The confidence contours for the growth rate $f\sigma_8$, the linear bias $b\sigma_8$, the dilatation parameter α, and the warping parameter ϵ are indicated, along with their 1D distributions. The dashed lines of each histogram are the mean values found by the MCMC chain .

the large scales for the quadrupole and hexadecapole where potential systematic errors could be present. The results of $f\sigma_8$ and α are consistent in the fiducial ranges within $1-\sigma$, for the two cases, $\xi_0+\xi_2$ and $\xi_0+\xi_2+\xi_4$, but the ϵ values have a $1.3-\sigma$ difference. Figure 3.26 shows the likelihood surfaces of the $\xi_0+\xi_2+\xi_4$ compared with our fiducial methodology ($\xi_0+\xi_2$), they are both in the fiducial [28,124]h^{-1}Mpc range for all multipoles and both chains and they use the priors from table 3.8.

Figure 3.11, show that our hexadecapole data have a stronger amplitude on small scales that both sets of mocks. This mismatch in amplitude could be a problem of the mocks or could be due to a real signal in the data, or could be due to either an undetected systematic error in our data or a statistical fluctuation. If it is the latter then the increase in data with DR16 should reduce this shift. If it is a real cosmological signature it should become more significant in DR16. Regardless of the origin of this larger amplitud, the MCMC fitter prefers a large value of ϵ and a small value of $f\sigma_8$ to fit the amplitud of the data hexadecapole (see Figure 3.26.).

Figure 3.26 shows the results of cutting the large scales for the quadrupole and hexadecapole ($\ell = 2, 4$), the red dashed-line contours come from a chain where the monopole is in the [28,124] h^{-1}Mpc range, and the quadrupole and hexadecapole in the constrained range of [28,92] h^{-1}Mpc. When the last three bins of the quadrupole and hexadecapole are avoided we achieve a significant improvement in the goodness of the fit (we saw this same behavior in section 3.7 when removing the large scales of the quadrupole), we also lose the secondary locus that was present in the full approach whithout having to reduce our priors; however, the price paid is to eliminate the BAO information. This increases the degeneration of the parameters and biases the results, we lose information that constrains ϵ and α, which in turn causes the contour areas to become larger, providing more freedom to the fitter to move f and α to lower values (Figure 3.27).

The main impact of removing the large scales in the hexadecapole fits is in parameters that require the BAO peak to be constrained, as expected. When excluding the large scales, the BAO information is lost, and α is shifted in consequence.

Table 3.11: Results for the DR14 LRG sample. The first block is for our fiducial methodology, using the fiducial range for the $\xi_0 + \xi_2$ fit. The second block is for the $\xi_0 + \xi_2 + \xi_4$ fits in the ranges [28,124], [28,124], and [44,124] h^{-1}Mpc for their multipoles $\ell = 0, 2, 4$. The third block is for the $\xi_0 + \xi_2$ and $\xi_0 + \xi_2 + \xi_4$ fits when excluding the large scales for quadrupole and quadrupole/hexadecapole, respectively. The fiducial value for the $\sigma_8(z_{eff=0.72}) = 0.55$ (0.5495932). The eulerian bias is defined by $b = 1 + F'$.

Measurements with LRG sample DR14 Oficial Version.						
Case	$f\sigma_8$	$b\sigma_8$	<F">	σ_{FOG}	α	ϵ
$\xi_0 + \xi_2$ [28,124][28,124]	$0.454^{+0.119}_{-0.140}$	$1.110^{+0.116}_{-0.100}$	$2.2^{+3.8}_{-4.4}$	$3.7^{+3.0}_{-2.3}$	$0.955^{+0.055}_{-0.05}$	$0.000^{+0.090}_{-0.050}$
$\xi_0 + \xi_2 + \xi_4$ [28,124] [28,124] [28,124]	$0.31^{+0.09}_{-0.09}$	$1.19^{+0.10}_{-0.10}$	$-1.1^{+3.2}_{-3.3}$	$5.8^{+3.3}_{-3.2}$	$0.986^{+0.047}_{-0.046}$	$0.091^{+0.046}_{-0.048}$
$\xi_0 + \xi_2 + \xi_4[28, 124][28, 92][28, 92]$	$0.285^{+0.093}_{-0.094}$	$1.079^{+0.108}_{-0.110}$	$-1.5^{+3.3}_{-3.0}$	$5.5^{+2.6}_{-2.8}$	$0.917^{+0.054}_{-0.056}$	$0.107^{+0.041}_{-0.039}$

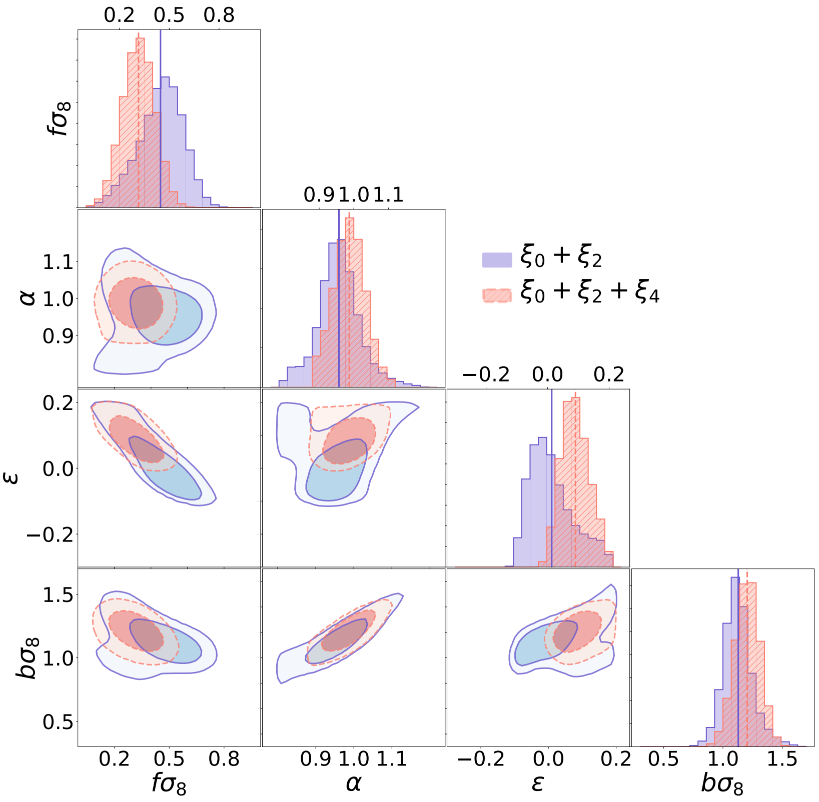


Figure 3.26: The shaded regions show the $1 - 2\sigma$ confidence surfaces found by our MCMC chains for the RSD-AP parameters for two cases: $\xi_0 + \xi_2$ (red dashed-line contours) and $\xi_0 + \xi_2 + \xi_4$ (blue solid-line contours), all multipoles in both models are in the [24,128] h^{-1}Mpc range.

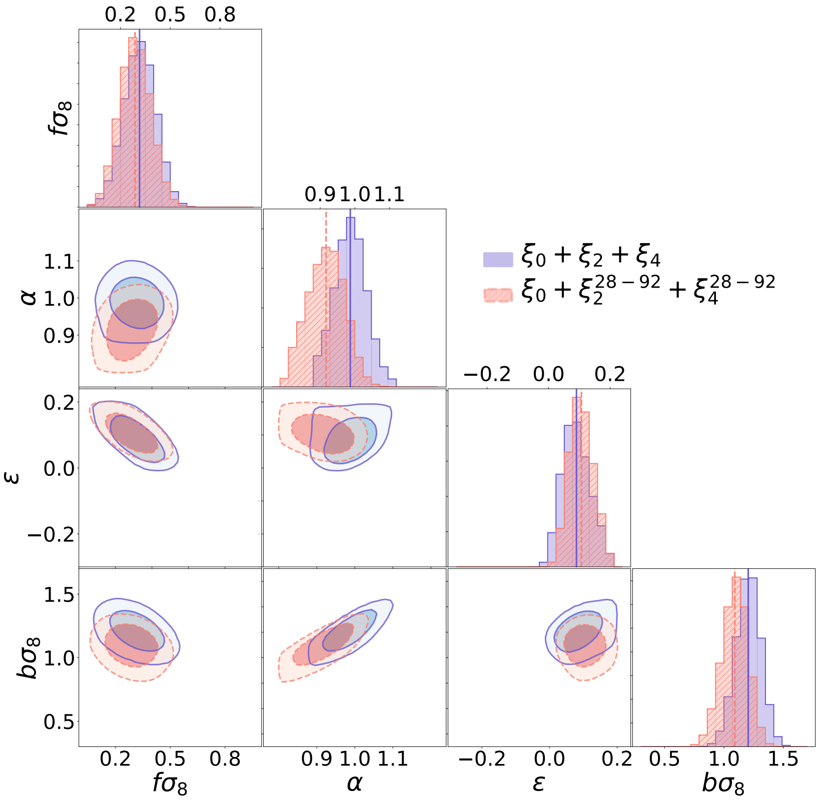


Figure 3.27: Similar to figure 3.26, here the red dashed-line contours represent a fit where the quadrupole and hexadecapole are reduced to the [28,92] h^{-1}Mpc range while the monopole stays in the full range ([24,128] h^{-1}Mpc), the blue solid-line contours show the fit where monopole quadrupole and hexadecapole are in the full range.

Chapter 4

From hydrodynamical simulations to galaxy mock catalogs

As mentioned in the introduction (Chapter 1) this thesis presents two independent research projects both of which study the dark sector of the universe. In the previous two chapters, we discussed the RSD analysis from the eBOSS DR14 survey. There, we found the regions of high likelihood for a set of parameters related with RSD under the paradigm of the ΛCDM model.

In the second part of this thesis, we discuss a novel methodology developed during my Ph.D. that studies the relation between the mass of galaxies and the properties of their host DM halos. The end goal is to populate halo catalogs from, e.g. an N-body simulation, with galaxies of the correct stellar mass.

The development of the method is left for chapters 5 and 6. First, chapter 5 focuses on introducing and testing the methodology on a sample of relatively large central halos as a proof of concept. Chapter 6 expands the method to include satellite galaxies and uses a smaller halo mass cut.

This chapter introduces the different astronomical concepts that are required to follow the discussion of the next two chapters. Section 4.1 motivates the need for methods that populate DM catalogs. Then, section 4.2 summarises the different physical processes that affect the evolution of central and satellite galaxies within a DM halo. In section 4.3 we introduce the most common methods that have been historically used to populate DM catalogs with galaxies. Finally, in section 4.4 we review the next steps for this project.

4.1 The need for methods that populate dark matter only simulations

Modern LSS surveys are designed to measure the redshifts of millions of galaxies, they are arguably the best data sets for testing the standard models of cosmology. However, given the completeness of some of these surveys at relatively low redshifts they can also be used to test galaxy evolution models.

Most of the surveys designed to study galaxy evolution focus on redshifts $\sim$1, as this is the epoch of higher star formation rate in the universe (Madau and Dickinson, 2014). Some of the most noteworthy surveys are the zCOSMOS survey (Lilly et al., 2007) that computed the redshift and observed the morphology of 8500 galaxies at $z \sim 0.8$, the DEEP2 survey (Newman et al., 2013) that measured the redshift of around 50,000 at $z \sim 1$, and the VIPERS survey (Scodeggio et al., 2018), that observed 90,000 galaxies with redshifts around $0.5 < z < 1.2$. These can be supplemented with observations from low redshift surveys to get a more complete picture of galaxy evolution from the star formation epoch at $z \sim 1$ and up to the present. Perhaps the most well-known low redshift survey is the SDSS main galaxy survey (York et al., 2000), which has a median redshift of $z = 0.1$ and has collected the redshifts of around 1 million galaxies.

Other galaxy evolution surveys are being developed at present, for example the 4MOST WAVES (Driver et al., 2019) survey that will study 1.65 million galaxies, 0.9 million galaxies with the WAVES-Wide survey that has a redshift cut at $z = 0.2$, and the rest by the WAVES-Deep magnitude limited survey that is expected to observe galaxies in the $z < 0.8$ range.

Mock catalogs that mimic the survey will be needed for the analysis of these data sets (as with the eBOSS RSD analysis presented in chapter 3). These mocks will be used for several proposes, for example: estimating the errors of certain statistics (like the covariance matrix of the 2-point correlation function described in section 3.5.3) or testing the data analysis methodologies (for example our analysis in section 3.6).

Arguably the best tool to study theoretically the evolution and interactions of galaxies and halos are hydrodynamical simulations. These simulations show that central and satellite galaxies follow different evolutionary paths inside halos. This is a consequence of the differences in their interactions within the cluster. Section 4.2 presents a summary of the different physical processes that make galaxies and halos evolve once merged into a cluster.

The volumes required by mock catalogs used in the analysis of LSS surveys to study galaxy evolution can be very large and are generally of the order of magnitude of $\approx (1[\mathrm{Gpc}/h])^3$ (e.g. Safonova et al., 2021). Hydrodynamical simulations that have enough resolution to include galaxies considered by those surveys and with enough volume are well beyond the capabilities of current hydrodynamical simulation codes. For example the largest EAGLE simulation (McAlpine et al., 2016) the Illustrious simulation (Nelson et al., 2015) and the MassiveBlack-II (Khandai et al., 2015) were all built in boxes with a volume of around $\approx (100[\mathrm{Mpc}])^3$. Newer generation of hydrodynamical simulations like Illustris TNG (Springel et al., 2017) and EAGLE XL (in preparation) have volumes of $\approx (300[\mathrm{Mpc}])^3$ but this will still not be enough to make mock catalogs of the required volume.

With this in mind, there is a strong incentive to build methods that can use hydrodynamical simulations to learn the relationships between galaxies and dark matter, then use these relations to populate N-body dark matter only simulations of the required volume. This is a very complicated process that has been tackled with different methodologies. Section 4.3 summarises some of the most widely used ones making some emphasis on a new type of algorithms that have been developed intensively in the last few years, the so-called *machine learning* techniques. In chapters 5 and 6 we present a novel machine learning methodology that provides a new way of populating halos with galaxies of a given stellar mass.

4.2 The evolution of galaxies and dark matter in halos and subhalos

It is a well known observational result that the properties of galaxies are correlated with their environment (e.g. Blanton and Moustakas, 2009; Boselli and Gavazzi, 2006). Galaxies that are not inside clusters, and therefore relatively isolated, are more likely to have characteristics of younger galaxies, like spiral shapes and bluer colours (a blue galaxy is a star-forming galaxy). These are of course general trends, as it is possible to find blue galaxies inside a cluster. On the other hand galaxies inside clusters are likely to present characteristics of *older* galaxies like a red colour (related with little star formation) and an elliptical shape .

The accepted explanation of this phenomenon states that galaxies in halos should suffer transformations when accreted from their more isolated environments into the cluster. The idea is that if a galaxy was to lose a significant amount of baryons during infall, it would not be able to generate more stars and will therefore become a red elliptical.

Several mechanisms have been proposed to deprive a galaxy of baryonic-mass during infall:

- *Ram pressure stripping* (Gunn and Gott, 1972; Abadi et al., 1999; Vollmer et al., 2001) is the name given to the clash that galaxies experience against the baryonic intercluster medium when infalling into a cluster. From the reference frame of the galaxy this is experienced as a *wind*. This wind opposes the gravitational potential of the galaxy, if the wind is strong enough to overcome the potential it will blow gas away.

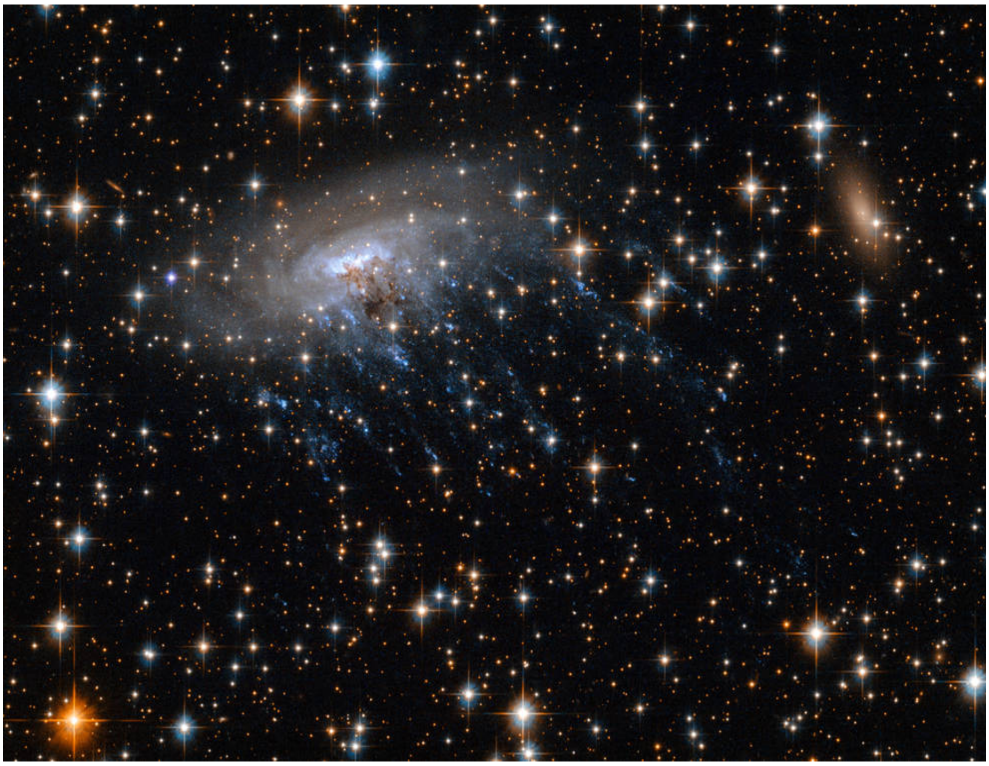


Figure 4.1: Image of the spiral galaxy ESO 137-001 taken by the HST. The galaxy moves through the Abell 3627 cluster and is subjected to ram pressure stripping leaving a trail of baryonic gas at its pass. This is an example of a *jellyfish* galaxy. Image credit: NASA, ESA

The gravitational potential of a galaxy is dependent on the distance to its centre. Hence the outer parts of a galaxy will be the most affected and the galaxy loses its gaseous halo, which is its main source of fuel for stellar formation (Erb, 2008; Hopkins et al., 2008). This process is usually referred to as *starvation* (Larson et al., 1980).

Ram pressure stripping can be observed in action on the so-called jellyfish galaxies (see figure 4.1), which are galaxies moving within the intercluster medium and leaving a trail of visible gas behind them (giving them a shape that resembles that of a jellyfish).

- *strangulation* (Kampakoglou and Benson, 2007; Peng et al., 2015) is the name given to the process where the gravitational potential of the halo strips a galaxy from its baryonic gas due to tidal effects. This removes all gas beyond a certain *tidal radius* defined by the distance at which the gravitational force of the galaxy matches the tidal forces.

Note that the elliptical orbits of particles mean that some might come in and out of the tidal radius. This makes the tidal radius evolve with time as these particles get stripped, making strangulation a complicated phenomena to model.

- *harassment* (e.g. Moore et al., 1996; Mastropietro et al., 2005) is a disruption experienced by both central and satellite galaxies. It happens when there is a close encounter with other galaxies that do not result in a merger but that make the gravitational potential of a galaxy change rapidly. They are usually called *fly-bys* encounters. While these encounters do not necessarily end up having enough force to strip material from a galaxy they can heat the galaxy by transferring energy to the orbiting particles. This has the effect of expanding the galaxy and destroying ordered and cold structures like disks, making galaxies in clusters more likely to be elliptical than any other shape.

Let us recall from our discussion about N-body simulations in section 1.3.2, that within the ΛCDM paradigm, DM clumps into clusters of matter called halos, inside which galaxies form. We discussed also how under the CDM paradigm halos can merge and that the remnants of small halos accreted into larger ones, which we call *central halos*, form self-bound substructures that we call *subhalo* or *satellite halo*. Through the rest of this work, we will refer to the galaxies living inside central and satellite halos as *central galaxies* and *satellite galaxies* respectively.

So far we have focused on what happens to the baryons inside subhalos when falling into halos; however, their dark matter also undergoes significant changes. DM simulations show that subhalos lose mass during their infall processes, this is mostly due to processes related to tidal effects between interacting subhalos (e,g, Mo et al., 2010; van den Bosch et al., 2018).

Tidal heating (e.g. Lynden-Bell, 1967; Green and van den Bosch, 2019) refers to the perturbation of particles in the subhalo at large radii, when having close encounters with other subhalos. This perturbation is equivalent to the harassment phenomena on galaxies. When the tidal force is strong enough, the larger halo will pull dark matter out of a smaller one, commonly known as tidal stripping (e.g. Merritt, 1983; Hayashi et al., 2003; Green and van den Bosch, 2019). This effect is common at close collisions and can be quite effective when the average separations of halos are not much larger than the average subhalo radius. This happens often near the centre of the halo, where the relative velocities are smaller and can lead to the central galaxy absorbing a significant amount of matter from smaller subhalos. Given that dynamical friction[*] makes subhalos fall deeper into the potential well of the central halo, tidal stripping should increase with time making subhalos lose more and more mass as they merge.

The picture presented here results in very different evolutionary tracks for central and satellite halos. Central halos and their galaxies should grow monotonically with time becoming bigger as time passes by accreting matter from their close environment and from satellite galaxies. On the other hand, satellite halos should have grown monotonically while they were isolated, but once they merged their total dark matter content should decrease due to tidal effects. At the same time their galaxies should have stopped growing due to processes like ram pressure stripping, harassment, and strangulation.

The evolution of structure described here can be very complicated to model accurately, but is naturally included in hydrodynamical simulations. This gives us an incentive to try to learn the relations between galaxies and host halos from the hydrodynamical simulations themselves. In the next two chapters we develop a methodology that feeds a machine learning algorithm with data from the EAGLE

[*] Dynamical friction (Chandrasekhar, 1943) refers to the effect of a subhalo losing angular momentum when interacting with a smaller and lighter body: gravity causes the small object to accelerate and gain angular momentum, then conservation of angular momentum and energy dictates that the subhalo should slow down and fall into a lower orbit.

simulation to build a function that predicts the stellar mass of an object based on the properties of its host halo. Before this, we will use the rest of this chapter to review other popular methods that have been used to populate halo catalogs with galaxies.

4.3 History of populating DM catalogs with galaxies

Several methods have been developed that can model the properties of galaxies without having the computational costs and resolution problems that hydrodynamical simulations have. In this section, we present a summary of the most relevant methods that have been used to achieve this goal. In general, these methods model functions that provides a prediction for the bias and stellar properties of galaxies as a function of their host halo properties. Then these models are used to populate halo catalogs built with N-body simulations in a way that reproduces the statistical properties of galaxies observed by galaxy surveys.

Probably the most used methodology is the so-called *semi-analytical* galaxy formation modelling (e.g. White and Frenk, 1991; Kauffmann et al., 1993; Cole et al., 1994; Cole et al., 2000; Bower et al., 2008; Lacey et al., 2016; Baugh et al., 2019). This method consists of building physically motivated equations that model some of the processes that define galaxy properties inside a halo, like gas cooling, star formation, feedback from supernovae, stellar evolution and so on. These equations should be completely determined by the initial conditions of the universe, the assumed dark matter properties of the halo and its local environment, and by a set of free parameters that need to be fitted to observations. While semi-analytical models are not as computationally expensive as hydrodynamical simulations, some of the baryonic physics that they try to model is very complicated and they rely on several simplifying assumptions.

For a set value of the free parameters of a semi-analytical model, one can use the equations of the model to populate a halo catalog with galaxies. Once the galaxy catalog is built one can measure its statistics e.g. the correlation functions or the luminosity functions. These statistics can be compared with those from observational catalogs to compute a likelihood estimate, where values of the free parameter that generate catalogs with realistic correlation and luminosity functions will have a large likelihood. An MCMC algorithm (see 2.6.1) can then use these likelihood estimates to explore the parameter space of the semi-analytical model. This methodology can be used to fit the free parameters of the model to an observational survey. Semi-analytical models fitted in this way have shown good agreement with predictions made using hydro-simulations (e.g. Benson et al., 2001; Helly et al., 2003).

Given the complexity of hydrodynamical simulations and semi-analytical models, there is an incentive to use the catalogs built by these methods and learn empirically the relation between dark matter halos and galaxies. This relation can be characterised using the halo occupation distribution (HOD) function $P(N|M)$ which is defined as the probability of finding N galaxies of a given type or mass in a dark matter halo of mass M. Knowing an accurate measurement of a HOD allows us to populate any N-body simulation with galaxies in a statistically significant way, assuming the halos are statistically independent of their larger scale environment. Note that $P(N|M)$ can be computed from any catalog of galaxies and host halo masses. Efforts have been made to extract the HOD from both hydrodynamical simulations (e.g. White et al., 2001; Yoshikawa et al., 2001) and semi-analytical models (e.g. Kauffmann et al., 1999; Benson, 2001), with the resulting HODs appearing to be in good agreement with each other (e.g. Berlind et al., 2003).

In the last five years or so new methods based on machine learning algorithms have been implemented to study the relation between halos and baryonic matter. Kamdar et al. (2016) and Agarwal et al. (2018) use extremely randomised tree and random forest algorithms respectively to predict galaxy properties based on dark matter halo properties like halo mass, growth rate and the maximum mass a halo achieved during its evolution. The models are trained on data from hydrodynamical simulations and the resulting models are impressively accurate at reproducing the simulation properties. However, the actual universe might have different properties to the hydrodynamical simulations, and it would be hard to modify the models to reproduce actual observations, given that the models built by such machine learning algorithms are *black box* models.

In order to avoid this issue Moster et al. (2021) uses a neural network approach that asks the algorithm to reproduce observed properties like clustering and stellar mass functions instead of the actual galaxy masses of a dark matter halo in an hydrodynamical simulation. The results are accurate, but it may be challenging to extract an interpretation out of this model, as they are only asked to reproduce observed statistics accurately but not individual properties. In this sense this method is similar to halo abundance matching (e.g. Klypin et al., 2013), and inherits some of its problems, the main one being that fitting a simulation to a set of observed statistical properties does not guarantee that it would also reproduce a new statistic that it was not fitted to match.

4.4 Conclusions

In chapters 5 and 6 we present a novel methodology to model the stellar mass of galaxies from the evolutionary path and present-day mass of their host halo. This methodology uses machine learning algorithms to model an equation of state that predicts the stellar masses of galaxies. Our models are determined by a set of free parameters defined by the methodology that determines the size of the contribution of different functional relations between a set of DM properties. In this sense, our methodology is similar to semi-analytical models as it tries to model the physics behind the matching to generate an equation of state. Given that the final model depends on a small set of free parameters this could be fitted to a set of statistics from an observational data set, this possibility will be explored in chapter 7

Chapter 5

A sparse regression approach to modelling the relation between galaxy stellar masses and their host halos

5.1 Introduction

Gravitational collapse in the expanding universe leads to the formation of complex, highly non-linear structures. The force of gravity can be accurately modelled through N-body cosmological simulations. However, observational probes of the universe's structure usually rely on galaxies, bringing into play a much broader range of baryon physics. Unlike dark matter only (DMO) simulations, which only allow interactions through gravity, baryon simulations need to deal with complicated feedback processes and are strongly influenced by events happening at scales much smaller than the size of the simulation grid (Schaye et al., 2010). While this can be mitigated by including sub-grid models of these processes, in the form of sources or sinks of energy and matter, the resulting computational cost of accurate baryonic simulations remains far greater than that of DMO simulations. As a consequence the volume of the universe that can be modelled in this way is limited. A hybrid approach is therefore necessary, in which a large volume DMO simulation is populated with galaxies based on the halo-galaxy relationship found in smaller baryonic simulations. This requires a methodology that can extract robust halo-galaxy relationships making optimal use of the available volume of baryonic simulations. In this paper, we explore whether sparse-regression models, which are a type of machine learning algorithm, provide an attractive approach.

A full reconstruction of the baryonic universe would require us to also model satellite galaxies. These are subject to additional physics such as tidal striping, heating (Merritt, 1983) and other environmental processes. In this work, we focus on developing and presenting our methodology, applying it to model central galaxies. We leave the extension of our methodology to include satellite galaxies for a future work.

It is already well established that there is a strong correlation between the stellar mass (M^*) of a central galaxy and the mass of its host halo (White and Rees, 1978). This relation is known as the Stellar Mass – Halo Mass (SMHM) relation. However, there is a significant scatter in the SMHM relation (e.g. More et al., 2010; Zu and Mandelbaum, 2015) which indicates that the stellar mass of a galaxy may also depend on other factors. Here, we investigate whether the formation history and the angular momentum of the host dark matter halo also play a role. Dependence on formation history is often referred to as assembly bias (Sheth and Tormen, 2004; Gao et al., 2005; Gao and White, 2007; Ramakrishnan et al., 2019). The effect of assembly bias in the EAGLE simulation has been studied in Chaves-Montero et al. (2016), where it was concluded that it might alter the clustering signal amplitude of the sample by up to 20%. It is worth noting, however, that while assembly bias has been detected in several simulations, the efforts made to detect it on observations have been inconclusive to date (e.g. Lin et al., 2016; Tojeiro et al., 2017; Salcedo et al., 2020).

To explore the effect of assembly bias, Matthee et al. (2016) studied the correlation between the residuals in the SMHM relation and different DM properties on EAGLE. They found that the parameters that are most correlated with this residual are those that are determined by the evolution of the halo mass, in particular concentration and halo formation time. They found no other parameter strongly correlated with the residual of the SMHM relation once it was corrected for the halo concentration correlation. Our aim in this thesis is to investigate the optimal measure of halo formation trajectory and to determine whether the prediction of stellar mass can be improved by including the additional halo specific angular momentum. In observations, the angular momentum of a galaxy appears correlated with its stellar mass (Fall and Romanowsky, 2013). However, while there is a correlation between the history of the specific angular momentum of a galaxy and its host halo (Zavala et al., 2016), Danovich et al. (2015) uses cosmological simulations to suggest that the specific angular momentum of gas and dark matter undergo decoupled formation histories and that it is only the final distribution of spin parameters that is similar between baryons and dark matter. Nevertheless, it remains physically plausible that halo specific angular momentum and galaxy formation efficiency may be interconnected in some more complex way.

The aim of this work is to develop a sparse regression approach to find a polynomial equation that relates the stellar mass of a galaxy with the properties of its DM halo. This is a form of Machine Learning (ML). More conventional ML algorithms such as neural networks (e.g. Lecun et al., 2015) and random forests (e.g. Breiman, 2001) are some of the most powerful tools for parameterising a data set. However, algorithms like neural networks or ensemble models (Roberts and Everson, 2001) generate models with virtually no explainability, so that extracting the physics behind the model would be difficult. While the network could predict galaxy properties, it would be hard to gain confidence that the output is physically reasonable. Random forest algorithms work by building a collection of decision trees that are designed to be as uncorrelated as possible. These models are easier to interpret, as one can measure how often a variable was used and how drastically the entropy decreases in each step. A potential issue with some machine learning algorithms is the slow evaluation speed of a final model.

Sparse regression methods (SRM; Hastie et al., 2015) are a set of minimisation algorithms that are efficient at discarding unnecessary free parameters. This makes them very useful at minimizing functions for which one suspects that most free parameters are irrelevant except for a small subset that one is trying to identify. SRM provide a trade-off between including very many free parameters (which would result in over-fitting to random artefacts in the data) and eliminating too many parameters (which would result in a poor description of the data). SRM have been proposed as the appropriate framework to extract the governing equations of a physical system from the data alone with relatively little prior knowledge required of the system's physics (Brunton et al., 2016). A key advantage of the SRM approach is that the small number of retained coefficients are likely to have a clearer physical interpretation. Further more, given that the models produced with SRM can be simple polynomial equations, their evaluation comes with virtually no computational cost.

We apply the sparse regression methodology to model the stellar mass of galaxies in the EAGLE 100 Mpc hydrodynamical simulation (Schaye et al., 2015; Crain et al., 2015). The EAGLE simulation provides a reasonable description of the observed universe (Crain et al., 2015; Artale et al., 2017; Furlong et al., 2015; Trayford et al., 2016), and is ideal for our proposes as parameter values of the DM halos are stored at several redshift slices (McAlpine et al., 2016), which permit us to model assembly bias. While all models presented here were calibrated on EAGLE data, we expect that the methodology can be applied to other hydrodynamical simulations with similar success. One common issue with this type of analysis is the danger of including a selection bias in the independent variables due to dark matter halos in hydrodynamical simulations being affected by baryonic processes that might alter properties like their density profile (e.g. Schaller et al., 2015; Martizzi et al., 2012; Navarro et al., 1996). With this in mind, we use a one-to-one matching (Schaller et al., 2015) between our hydrodynamical simulation and a dark matter-only simulation built using the same properties and initial conditions.

Our work builds on other ML methods that have shown promising developments in the creation of mock catalogs using DM halos. Moster et al. (2021) uses neural networks to populate DM halos from N-body simulations with galaxies. While their goal is similar to ours, the philosophy behind both models is different. Their approach avoids using hydrodynamical simulations and focuses on placing the galaxies inside halos in such a way that it reproduces observed properties of the galaxy populations. While that approach leads to accurate models, it would by construction be hard to extract any physical interpretation out of it. Lucie-Smith et al. (2018) used a random forest algorithm to predict which DM particle in a simulation would end up inside a DM halo of a given mass, while Berger and Stein (2019) used a neural network to build DM halo mocks.

In this work, we focus on the properties of central galaxies. For mock simulations to be compared to observations of large-scale structure surveys they need to be populated with galaxies in such a way that they reproduce the stellar mass function (SMF) and the clustering patterns of galaxies. This would require us to assign both a central and a population of satellite galaxies to each dark matter halo. We discuss the additional challenges of modelling the stellar mass of satellite galaxies at the end of the paper.

This chapter is organised as follows. Section 5.2 introduces the sparse regression methodology used to build our model and includes an example model that illustrates the behavior of the algorithm. Section 5.3 introduces the hydrodynamical simulation from which the input data were extracted and discusses how the data was processed to be used by our algorithm. The details on running the algorithm using the data set are presented in Section 5.4. As sections 5.2 and 5.4 introduce and test our methodology, readers primarily interested in the astrophysical results can go directly to section 5.5. Section 5.5 shows the results of the different configurations in which we run our code, and we discuss the physical interpretation of different terms of our governing equations and compare the stellar mass distribution and clustering statistics to those from EAGLE. Our conclusions and final thoughts, along with a brief discussion on the next steps that we aim to take, are presented in Section 5.6.

5.2 The Sparse Regression Methodology

This section starts by setting the general problem in §5.2.1. This is followed, in §5.2.2, by an introduction to the sparse regression method considered, i.e. the Least Absolute Shrinkage and Selection Operator (LASSO). We explain our minimisation implementation in §5.2.3 and the penalty hyperparameter definition in §5.10. We end in §5.2.5 with a simple example to more clearly illustrate our methodology.

5.2.1 Problem statement

We are interested in finding a function that models a physical property, y', that might be determined by a set of M variables $\vec{x}' = [x'_1, ..., x'_M]$ (we reserve the symbol x for normalised variables - see below). In this work y' is the stellar mass of a galaxy and $\vec{x}'$ a set of present and past properties of its DM halo. We can build a data set of values of y' and their associated $\vec{x}'$ by looking at large catalogs where the value of both has been measured. In this paper we use the output of the EAGLE hydrodynamical simulations (see Section 5.3).

We collect a sample of N galaxies to build a vector $\vec{y}'$, where

$$\vec{y}' = [y'_1, ..., y'_N], \tag{5.1}$$

and an associated matrix $\mathbf{X}'$, with each row $\vec{x}'_\alpha$ $(1 \leq \alpha \leq N)$ representing the list of dependent variables associated with the DM halo of the corresponding galaxy

y'_α:

$$\mathbf{X}' = \begin{bmatrix} \vec{x}'_1 \\ . \\ \vec{x}'_N \end{bmatrix} = \begin{bmatrix} x'_{11} & ... & x'_{1M} \\ . & . & . \\ x'_{N1} & ... & x'_{NM} \end{bmatrix} \tag{5.2}$$

The different columns of matrix $\mathbf{X}'$ correspond to different properties of the DM halo, where each property can have different units and distributions. It is, therefore, necessary to standardise our data. We choose to do this using the mean and standard deviation of the distribution, and define the normalised variable as:

$$\vec{z}_i = \frac{\vec{z}'_i - \mu(\vec{z}'_i)}{\sigma(\vec{z}'_i)}, \tag{5.3}$$

where $\vec{z}'_i$ is now a column of $\mathbf{X}'$ and $1 \leq i \leq M$ and μ and σ are the mean and standard deviation operators. The same normalisation scheme is also applied to our dependent variable: $y = (y' - \mu(y'))/\sigma(y')$. Note that the primed variables refer to natural quantities and non-primed variables to standardised ones that have zero mean and unit variance.

The observed values of the M variables of $\vec{x}_\alpha$ will be used as inputs for a series of functions whose output one hopes to use to predict y_α. These functions can in principle have any desired form, and so we will use a gradient descent algorithm to fit a linear combination of them to $\vec{y}$ (Section 5.2.3). Although other approaches like singular value decomposition Golub (1970) could be used in the hyperbolic case, we wish to ensure that the method is generic.

We consider a library of D functions, and their evaluated values for the observed parameters of the α^{th} galaxy $\vec{f}(\vec{x}_\alpha) = [f_1(\vec{x}_\alpha), ..., f_D(\vec{x}_\alpha)]$. The library of functions that we use in this work consists of:

- A constant term $f^0(\vec{x}_\alpha) = 1$.
- M linear terms of the form $[f^1_1(\vec{x}_\alpha), ..., f^1_M(\vec{x}_\alpha)] = [x_{\alpha 1}, ..., x_{\alpha i}, ..., x_{\alpha M}]$ where $1 \leq i \leq M$.

- $M(M+1)/2$ quadratic terms of the form $[f_1^2(\vec{x}_\alpha)$, ..., $f^2_{M(M+1)/2}(\vec{x}_\alpha)] = [x^2_{\alpha 1}$, ..., $x_{\alpha i}\, x_{\alpha j}$, ..., $x^2_{\alpha M}]$ with $1 \leq i \leq j \leq M$.
- $M(M+1)(M+2)/6$ cubic terms of the form $[f_1^3(\vec{x}_\alpha)$, ..., $f^3_{M(M+1)(M+2)/6}(\vec{x}_\alpha)]$ $= [x^3_{\alpha 1}$, ... , $x_{\alpha i}\, x_{\alpha j}\, x_{\alpha k}$, ..., $x^3_{\alpha M}]$ with $1 \leq i \leq j \leq k \leq M$.

The total number of functions considered is:

$$D = 1 + M + \frac{M(M+1)}{2} + \frac{M(M+1)(M+2)}{6}. \tag{5.4}$$

The number M of DM halo properties that we use depends on the specific parametrisation of the present and past properties of the halo that we select. We consider four different models each with different values of M (Section 5.3.5).

This methodology is able to deal with far more complicated functional forms than the polynomial forms used here. For example, we experimented with exponential decays and step functions. However including these more complicated functions in our initial testing did not improve our models, but increased the computational time so we excluded them from our final fits in this work.

Our goal is to find optimised values of the coefficients $\vec{C} = [C_1, ..., C_D]$ that will make the linear combinations of our D functions a sufficiently accurate model of $\vec{y}$. Specifically, we aim to find the optimised values of $\vec{C}$ such that $\vec{F}(\vec{C}, \mathbf{X}) \approx \vec{y}$, where $\vec{F}(\vec{C}, \mathbf{X})$ is defined as:

$$\vec{F}(\vec{C}, \mathbf{X}) = \mathbf{F}(\mathbf{X})\vec{C}^T = \begin{bmatrix} f_1(\vec{x}_1) & ... & f_D(\vec{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\vec{x}_N) & ... & f_D(\vec{x}_N) \end{bmatrix} \begin{bmatrix} C_1 \\ \vdots \\ C_D \end{bmatrix}. \tag{5.5}$$

We discuss the precise meaning of the approximate equality in the following section. Our aim is to achieve a balance between the accuracy of fitting the data while keeping the model as simple as possible. Clearly there is an underlying assumption that the functions included can be linearly combined into a sufficiently accurate model. In the absence of a detailed understanding of the physical system our approach is to include a large number of functions in our library, spanning the possible range of physical interactions.

5.2.2 Sparse regression

Sparse regression methods aim to minimise the error term $|\vec{F}(\vec{C}, \mathbf{X}) - \vec{y}|$ while discarding any unnecessary functions by setting their associated coefficients C_j to a negligibly small value. This makes them the appropriate framework for our problem as it allows us to include a large number of functions while knowing that all of the unnecessary ones will be discarded by the methodology. The fewer surviving coefficients the easier it is to interpret the solution (i.e. the more explainable it is). The solution will also be less susceptible to over-fitting to random fluctuations in the training data.

One of the most common sparse regression algorithms is LASSO (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996; Tibshirani and Friedman, 2017), where one minimises

$$L = \chi^2 + \lambda P(\vec{C}). \tag{5.6}$$

$P(\vec{C})$ is known as the penalty function and its value should increase with the absolute value and number of coefficients that are not set to zero. The coefficient λ is a hyperparameter of the model and determines the relative magnitude of the penalty term. The value of λ is determined using a k-fold methodology (Hastie et al., 2015), as described in Section 5.2.4.

χ^2 is the normal chi-squared function defined as

$$\chi^2 = \sum_{\alpha=1}^{N} \frac{(F_\alpha(\vec{C}, \mathbf{X}) - y_\alpha)^2}{\sigma_{y_\alpha}^2}, \tag{5.7}$$

where σ_{y_α} is an estimate of the uncertainty in measurement y_α and $F_\alpha(\vec{C}, \mathbf{X})$ is the α^{th} element of $\vec{F}(\vec{C}, \mathbf{X})$. In the absence of the penalty term, L would be the negative of the logarithmic likelihood function (ie., $L = -2 \ln \mathcal{L}$).

In the standard LASSO approach $P(\vec{C})$ is defined as

$$P(\vec{C}) = \sum_{i=1}^{D} \mid C_i \mid . \tag{5.8}$$

We introduce a regularisation term to smooth out the gradient discontinuities that occur when parameters are close to zero,

$$P(\vec{C}) = \sum_{i=1}^{D} \mid C_i \mid e^{-(\epsilon/C_i)^2}, \tag{5.9}$$

where ϵ is a small constant. Note that $\exp(-\,(\epsilon/C_i)^2)$ is very close to zero when $|C_i| \ll \epsilon$ and close to one when $|C_i| \gg \epsilon$. Therefore ϵ determines how close to zero a coefficient C_i needs to go before its contribution to the penalty is negligible. We adopt a value of $\epsilon = 10^{-3}$, which we show in Section 5.4 makes unnecessary coefficients go close enough to zero to be clearly distinguished from the ones that are useful, while keeping a reasonable computational cost (the closer to zero unnecessary coefficients are required to get the longer the minimiser needs to run). We define a cutoff value ν as the threshold between used and discarded parameters: every coefficient larger than ν will be used in our model and all smaller coefficients are discarded. The exact value of ν is presented in Section 5.4.

In equation 5.9 the contribution of each coefficient C_i is independent of the contribution of all other coefficients. This means that there is not a strong penalty for having many small, but larger than ϵ, values of C_i. We found that a more efficient approach at eliminating non-essential coefficients is to consider the contribution of a coefficient, in comparison to all of the other surviving coefficients. This is achieved by the following penalty function $P(\vec{C}) = \sum_{i=1}^{D} \left[\sum_{j\neq i} \mid C_j \mid\right] \mid C_i \mid$. Combining both modifications our penalty function has the following form

$$P(\vec{C}) = \sum_{i=1}^{D} \left[\sum_{j\neq i} \mid C_j \mid e^{-(\epsilon/C_j)^2}\right] \mid C_i \mid e^{-(\epsilon/C_i)^2}. \tag{5.10}$$

This is the form of the penalty function adopted in our algorithm.

The χ^2 is a measure of the goodness of fit, which decreases as the model becomes more accurate. Balancing of the goodness of fit statistic and penalty term makes sparse models robust against over-fitting: an over-fitted model would use many parameters to make a unrealistically good fit, which would make the χ^2 very small but it would also make the penalty term large (as it grows with the number of parameters). The minimum should belong to a model that is as simple as possible, while still being a sufficiently good fit. This is why when using a large library of functions all but a small subset of the coefficients end up being set to zero.

By making some general assumptions, we can estimate that in the optimised solution $P(\vec{C}) = \mathcal{O}(1)$. First we note that $P(\vec{C}) \approx \sum_{i=1}^{D} \left[\sum_{j\neq i} \mid C_i \mid\mid C_j \mid\right] \approx (\sum_{i=1}^{D} \mid C_i \mid)^2$, and that the optimised solution should satisfy $\vec{F}(\vec{C}, \mathbf{X}) = \mathbf{F}(\mathbf{X})\vec{C}^T \approx \vec{y}$ Secondly let us note that, in our case, $F_i(\mathbf{X})$ correspond to third order combinations of elements of $\vec{x}_i$, with each element standardised to be of the order of magnitude of the elements of $\vec{y}$ and therefore $F_{i\alpha}(\mathbf{X}) \approx \mathcal{O}(y_\alpha)$. From here it should be that $\sum_{i=1}^{D} \mid C_i \mid\approx \mathcal{O}(1)$, and consequently that $P(\vec{C}) \approx \mathcal{O}(1)$.

The properties of the simulated galaxies do not have formal measurement errors, but we still expect a random scatter due to the stochastic nature of the formation process. We therefore estimate a constant $\sigma_y^2 = \sigma_{y_\alpha}^2$ (for $1 \leq \alpha \leq N$) using

$$\sigma_y^2 = \sum_{\alpha=1}^{N} \frac{(F_\alpha(\vec{C}, \mathbf{X}) - y_\alpha)^2}{N} \tag{5.11}$$

evaluated at $\vec{C}$ that minimises equation 5.7 when $\sigma_{y_\alpha}^2 = N$. A consequence of using this definition of $\sigma_{y_\alpha}^2$ is that if we then minimise Eq. 5.6 with no penalty ($\lambda = 0$) we find

$$L(\lambda = 0) = N. \tag{5.12}$$

The optimised value of λ should be such that χ^2 and $\lambda P(\vec{C})$ are of comparable size. Given that by $P(\vec{C}) \approx \mathcal{O}(1)$, and that we constructed $\sigma_{y_\alpha}^2$ such that $\chi^2 \approx \mathcal{O}(N)$ then $\lambda \approx \mathcal{O}(N)$. This allows us to estimate the sizes of penalty that we should explore.

5.2.3 Minimisation

We use a gradient descent algorithm to minimise Eq. 5.6. The process starts at an initial point in parameter space and iteratively walks in the direction opposite to the gradient of L with respect to C_i. We use a variation of Arfken (1985), the standard practice for most machine learning methodologies. The size of each step is determined by a parameter η. At every step one computes L, if it is larger at the new position then η is reduced (as it would likely mean that it overshot the minimum). In the opposite case η size is increased if L is smaller at the new position as it is likely that we are still far from the minimum.

The gradient of L from Eq. 5.6 is computed with respect to the vector of coefficients C_i. In the standard methodology one makes a step in the direction of the gradient at the current position. However, we found that this did not perform well in the steep-sided valleys that characterise L. In such valleys, a step will overshoot the minimum, and as a consequence the next step would be in the opposite direction than the previous one but with a slightly smaller step size. Progress along the valley toward the global minimum is then slow. This makes convergence inefficient in high dimensional spaces, as the minimiser tends to jump from one wall of a potential well to the opposite wall at each step instead of following a more direct downwards path.

We achieved performance gains by using the following adaptation of the algorithm for determining the next step of the minimisation. Defining the position of the i^{th} step as $p_i = C_1^i, ..., C_D^i$, the gradient vector

$$-\nabla(L)(p_i) = \frac{\partial L}{\partial C_1}(p_i), .., \frac{\partial L}{\partial C_D}(p_i) \tag{5.13}$$

points downhill towards the nearest local minimum. Since we are only interested in the direction of the gradient and not its magnitude we can normalise the vector as

$$\nabla(\overline{L})(p_i) = \nabla(L)(p_i) \Big/ |\nabla(L)(p_i)| \,. \tag{5.14}$$

We make a first trial step on the downhill direction that takes us to the following position in parameter space

$$p_{i+1/2} = p_i - \eta \nabla(\overline{L})(p_i). \tag{5.15}$$

The direction of the next step p_{i+1} is given by the mean of the gradients at p_i and $p_{i+1/2}$,

$$p_{i+1} = p_i - \eta[\nabla(\overline{L})(p_i) + \nabla(\overline{L})(p_{i+1/2)}]/2\,. \tag{5.16}$$

This swings the direction of travel to align with the valley.

In order to determine if our code has converged we look at the size of steps η taken by the minimiser. A very small step size indicates that we have not moved far for several steps. Our code will run until the step size becomes smaller than some tolerance value. A smaller value of the tolerance means we get closer to the minimum, however, the computational cost of our minimisation is strongly dependent on this tolerance value. We found that a tolerance of the step size of $\eta < 10^{-6}$ produces stable results and manageable low computational cost.

5.2.4 Penalty Hyperparameter

We will use a k-fold methodology to fit the hyperparameter λ. K-fold is a well-known method that is standard practice for fitting hyperparameters in sparse regression (e.g. Hastie et al., 2015). The method works by randomly dividing data into k independent subsets of roughly the same size. Then the hyperparameter λ is sampled in k independent runs, each time one subset is left out of the minimisation and is used to test the model on data it has not seen before. The set left out is called the test set. The rest of the data points are used for running the minimisation algorithm and are referred as the training set. In this work we will use a value of $k = 10$, which is standard practice. Each run explores λ with thirty uniformly spread points in $\log_{10}\lambda$ between $\lambda = 1$ and $\lambda = N$, to which the case of $\lambda = 0$ is added.

The higher the granularity of λ that we explore the more computationally expensive our code becomes. We found by testing that 30 uniformly spread out points in $\log_{10}\lambda$ was enough to find sufficiently smooth curves without a high computational cost. In principle, one could explore larger values of λ. However in our case models with $\lambda = N$ provided already significantly worse fits than models with smaller λ values, which indicates that the penalty was already too large at $\lambda = N$. This is true for all models presented in this paper except for the example of Section 5.2.5 where we needed to run between $\lambda = 0$ and $\lambda = 800$.

In order to quantify the quality of fit for a given $\vec{C}$, we will use the root mean square error (RMSE) defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\alpha=1}^{N}(F_\alpha(\vec{C},\mathbf{X}) - y_\alpha)^2}{N}}. \tag{5.17}$$

When λ is close to zero, the error in the model of the training set would be small as there is no significant penalty and the model is overfitted. Such a model is poor at predicting results in data that it has never seen before and this translates into a large error on the test set (see Figs. 5.2 and 5.5). For very large values of the penalty, the model becomes too simple as coefficients are heavily penalised: a model that is too simple will show large error on both the test and the training sets. When λ is large enough to avoid overfitting but not too large that models become too simple, the RMSE of the test set will reach its minimum (as illustrated in Fig. 5.2). If λ_k is the value of λ where the minimum is for a given k-fold, then $\lambda_\mu = \mu(\lambda_k)$ is an estimate of the optimised value of λ, where μ is the *mean* operator.

It is common practice in sparse regression to choose a value that is larger than λ_μ by one standard deviation, this is the one-standard-error rule from Hastie et al. (2015). This is done to avoid over fitting due to inaccuracies in the methodology. In this work we implement a modified version of the one-standard-error rule. Let us define $\text{RMSE}_k(\lambda)$ as the RMSE of the k^{th} k-fold as function of λ. By construction, $\text{RMSE}_k(\lambda)$ is minimised for $\lambda = \lambda_k$. If $\sigma(\text{RMSE}_k(\lambda_k))$ is the standard deviation of the collection of $\text{RMSE}_k(\lambda_k)$, then for each k-fold we define the optimised value of λ, λ_{min}, as:

$$\lambda_{\text{min}} = \mu \left(\text{RMSE}_k(\lambda_k) + \sigma(\text{RMSE}_k(\lambda_k)) \right) . \tag{5.18}$$

In order to find our surviving coefficients, we run the minimisation algorithm again on the complete data set, setting λ to λ_{min}. With P, the number of coefficients C_i larger than the cutoff value ν, we define our library of surviving functions $F_S(\mathbf{X}) = [F_{S_1}(\mathbf{X}), .., F_{S_P}(\mathbf{X})]$, for which $C_{S_j} > \nu$ with $1 \leq j \leq P$.

The penalty is useful for selecting which coefficients to discard and keep, but once this is done the presence of a penalty term biases all coefficients to smaller values. A penalty rewards smaller coefficients over larger ones, as the size of the penalty increases with the size of the coefficients. Having this in mind, our final model is found by re-running our minimisation algorithm using only the functions in $F_S(\mathbf{X})$ and setting λ to zero, i.e. without penalty.

5.2.5 Example

In this section we introduce a simple example to more clearly illustrate our methodology. We build a matrix $\mathbf{X}'$ as in Eq. 5.2, where each column $\vec{z}'_i$ has thirty points ($N = 30$) and each point is a random number between zero and one. We will use three independent variables, $\vec{z}'_1$, $\vec{z}'_2$ and $\vec{z}'_3$, i.e. $M = 3$. We will also build a dependent variable $\vec{y}'$ as:

$$\vec{y}' = 1.3 + 2\vec{z}'_1 + \mathbf{Noise} \tag{5.19}$$

where the noise comes from a Gaussian distribution centred on zero and with a width of 10% of the standard deviation of $1.3 + 2\vec{z}'_1$. All of our variables in this example have the same order of magnitude, so there is no explicit need to standardise their units. Hence we use y' and $\vec{z}'$ notation within this section.

Before running the model, we do not know the shape of Eq. 5.19. However, let us suppose that we suspect that $\vec{y}'$ should depend on the parameters $\vec{z}'_1$, $\vec{z}'_2$ and $\vec{z}'_3$. As we are uncertain on how to model the dependence between the parameters, we include a large set of functions. In this example, our library of functions includes a constant, linear and quadratic terms only (leaving out cubic terms for simplicity). In total we end up with ten functions ($D = 10$).

For the purpose of illustration, we focus on C_1 and C_3, the parameters associated with the linear functions of $\vec{z}'_1$ and $\vec{z}'_3$. Fig. 5.1 follows the trajectory of the minimiser for our example model and for these two coefficients. From Eq. 5.19 we know that $C_3 = 0$ and $C_1 = 2$. The minimiser starts in an arbitrary position (in the case of this example in $C_1 = C_3 = 1$) and follows the trajectory shown by the blue line. The dashed lines represent the contours of both the χ^2 (elliptic dotted contours) and $P(\vec{C})$ (dashed contours) of Eq. 5.10. A gradient descent algorithm will try to move perpendicularly to these contours, but the modifications in our algorithm allow the path to quickly align to the valley around $C_3 = 0$. Apparent deviations from this motion come from the fact that we are looking at the 2 dimensional projection of a ten dimensional trajectory.

Fig. 5.2 shows the evolution of the RMSE with respect to λ for our example model using the k-fold methodology of Section 5.2.4. For this example, we divide the data in to $k = 5$ folds. The blue dashed line correspond to the training set and the green lines to the test set. The solid lines are the median of each set. We explore the hyperparameter λ between $\lambda = 0$ and $\lambda = 800$. This is different to our nominal λ range which would be between 0 and N otherwise. This is because the ratio $D/N = 10/30$ is much larger in this example than in our nominal set up using the full simulated data set, for which we have hundreds of functions to fit almost 10,000 galaxies (i.e. $D/N \sim 0.01$). This significant change in this ratio of D/N requires a larger penalty to be considered to avoid any overfitting.

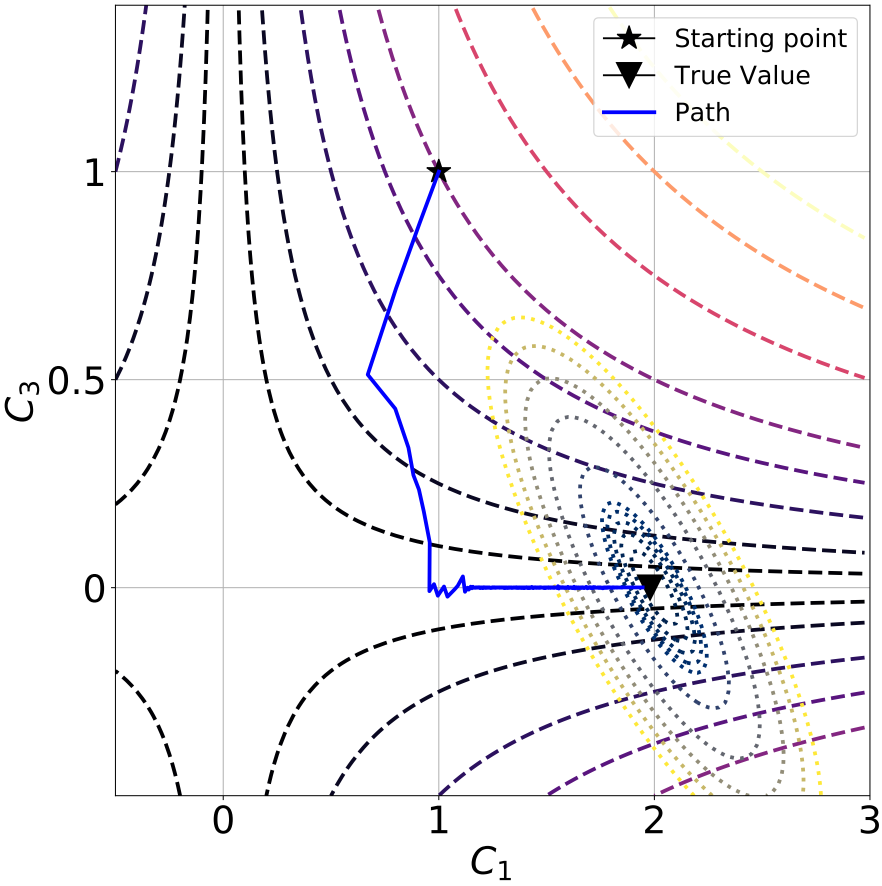


Figure 5.1: Isocontours of the penalty function defined by Eq. 5.10 for the two different coefficients C_1 and C_3 associated with the functions $C_1 f_1^1(\vec{z}'_\alpha) = C_1 z'_{\alpha 1}$ and $C_3 f_3^1(\vec{z}'_\alpha) = C_3 z'_{\alpha 3}$. The dashed hyperbolic and dotted elliptical lines are the isocontours of our penalty function and of the χ^2 statistic respectively. Given that the gradient is perpendicular to the contour lines, the minimisation routine can efficiently move toward the origin of the plot, and also to one of the axes. Hence the code will quickly reach the minimum if either or both coefficients are zero.

When λ is close to zero in Fig. 5.2, the RMSE in the training set (blue line) is small, the model is overfitted and therefore bad at predicting the result in data that it has never seen before. This results in the comparably larger error on the test set (green line). For the largest values of the penalty, the model becomes too simple and the error on both the test and the training set begins to increase.

Around $\lambda \sim 10$ in Fig. 5.2, the fit of the test set improves and the RMSE reaches its minimum. This is where the model is the least susceptible to overfitting while still capturing the important features of the data set. The black dots indicate the

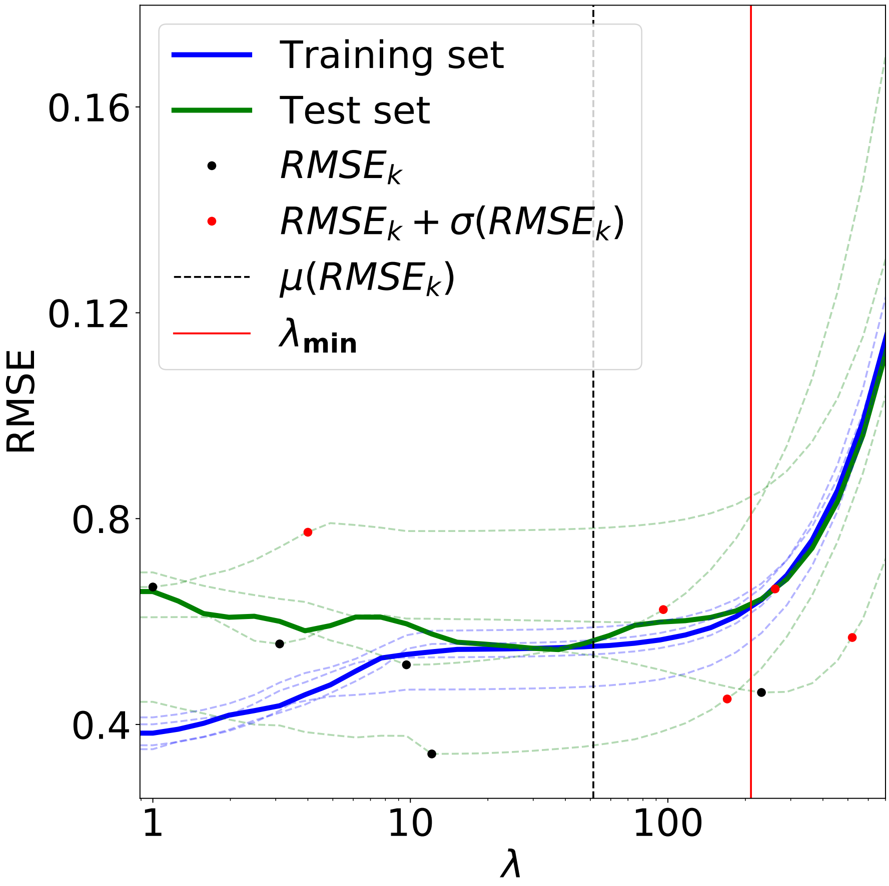


Figure 5.2: Evolution of the RMSE from the best fits of our example model as function of the hyperparameter λ. The blue and green dashed lines represent the RMSE of the training and test sets respectively. The solid lines represent the median of these curves. When λ is close to zero, the training set has a very small error and the test set a comparably larger one: this is due to overfitting of the minimiser and it improves as λ grows. The black dots indicate the minimum RMSE for the test set of each individual k-fold: this is where overfitting was smallest. The black dashed line shows the mean value of the λs of the black dots. The red dots are plotted at the values of λ given by our modified one-standard-error rule. The red line indicates the mean of λs of these red dots and is our estimate of λ_{min} from Eq. 5.18

.

minimum RMSE for the test set of each individual k-fold and the black dashed line shows the mean value of these points, $\mu(\text{RMSE}_k)$. Our optimal value of λ is given by λ_{min}, defined by Eq. 5.18 and shown as a vertical red line.

Fig. 5.3 shows how the best-fit coefficients of our example model evolve for different values of λ. Each curve is the mean curve from our 5 different folds. As stated in

Section 5.2.2, the code will not set parameters exactly to zero but to a very small value which is determined by the parameter ϵ of Eq. 5.10. Fig. 5.3 shows that in this example the value is $\sim 6 \times 10^{-4}$ (this is true for both the example model and the galaxy data set as it only depends on ϵ). Therefore we select a cutoff value of $\nu = 1 \times 10^{-3}$. This is is shown as the grey shaded region in Fig. 5.3.

The coloured lines correspond to the coefficients that were above the cutoff value ν at λ_{min}, and therefore included in the final model. At $\lambda = 0$, all coefficients are above the cutoff threshold due to over-fitting. As λ grows, coefficients drops below the threshold value and at λ_{min} all coefficients other than C_0 and C_1 have been discarded. This is expected as C_0 and C_1 are the non-zero coefficients used in building $\vec{y}'$ according to Eq. 5.19. The grey dashed lines are the coefficients that were below ν at λ_{min} and therefore discarded. The vertical dashed line corresponds to the optimised value λ_{min}.

The final model selected by the algorithm is:

$$\vec{F}(\vec{C}, \mathbf{X}') = 1.27 + 1.98\,\vec{z}'_1 \tag{5.20}$$

Considering that this a fit to data generated using Eq. 5.19 with 10% Gaussian distributed noise, we can conclude that our algorithm generated a sparse and accurate representation of the data.

5.3 Data Set

This section introduces the data set that is used in our analysis. In Section 5.3.1, we introduce the hydrodynamical simulation, which data is used to train our model on. In Section 5.3.2, we describe the selection of the galaxies considered, as well as the method used to address inconsistencies in the classification of galaxies between different snapshots of the simulation. Section 5.3.3 and Section 5.3.4 introduce the set of variables that form $\vec{x}'_\alpha$ of Eq. 5.2: the former introduces all variables associated with the mass of the host halo and the latter with its angular momentum.

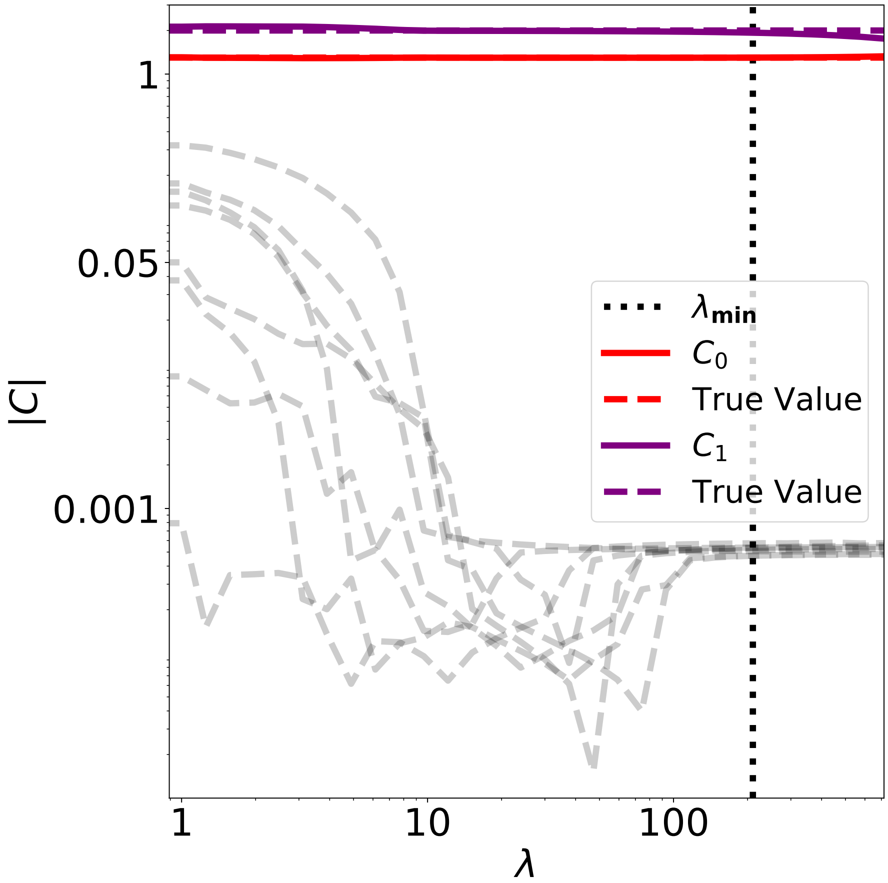


Figure 5.3: Evolution of the absolute value of the best fit coefficient values as function of the hyperparameter λ. The coloured lines show the value of the mean fit of our independent k-fold runs for our surviving coefficients, i.e. those coefficients that are larger than the cutoff value ν at the optimised value λ_{min} of the hyperparameter λ. The dashed lines show the true coefficients used to create the data from Eq. 5.19. The grey dashed lines show the evolution of the values of the coefficients that were discarded in the final model. The grey shaded area represents our cutoff value ν, below which parameters will be taken out of the fit. The dotted black line represents λ_{min}. We note that at λ_{min} all coefficients are set to zero except C_0 and C_1.

Finally §5.3.5 lists the four models considered, which differs from one another by the set of variables used to define $\vec{x}'_\alpha$.

5.3.1 The EAGLE simulation

Hydrodynamical simulations provide powerful insight into the galaxy formation process. The resulting database catalogs DM halos and their connection to baryonic properties such as stellar mass. In this work, we use the Evolution and Assembly of Galaxies and their Environments (EAGLE, Schaye et al., 2015; Crain et al., 2015) simulations, a suite of hydrodynamical simulations built inside cubic periodic volumes. We use the largest of these volumes, corresponding to a box of 100 comoving Mpc of length.

The simulation runs using a modification of the GADGET 3 code described in Springel (2005). The code uses Smooth Particle Hydrodynamics methods to model the mechanics of the baryon fluid. In order to compute the gravitational potential, the code uses a combination of a Particle Mesh (at large scales) and a hierarchical Tree algorithm (at grid and subgrid scales). The details on the modifications can be found in Schaller et al. (2015). The simulations are built using the Planck cosmology (Planck Collaboration et al., 2014).

Baryonic physical processes that cannot be solved directly are implemented into the simulation as sources and sink terms, where energy and matter are either absorbed or injected locally into the simulation. These subgrid models should depend only on the local property of the gas. The subgrid models implemented account for radiative cooling (Wiersma et al., 2009a), star formation (Schaye and Dalla Vecchia, 2008), star formation feedback (Dalla Vecchia and Schaye, 2012), black hole growth (Rosas-Guevara et al., 2015; Springel et al., 2005), Active Galactic Nuclei feedback (Booth and Schaye, 2009) and chemical enrichment (Wiersma et al., 2009b). The uncertain parameters of the subgrid models need to be calibrated, which is done by choosing the values that would reproduce the galaxy mass function at z=0.1, the galaxy size-stellar mass relation and the black hole mass-stellar mass regression. Discussion of the calibration process can be found in Crain et al. (2015).

Haloes are defined using a Friends-of-Friends algorithm (FoF; e.g. Einasto et al., 1984) with a linking length of b=0.2, i.e. all particles that can be linked together with an inter-particle distance smaller than 0.2 times the mean inter-particle distance form a halo. Once the halos have been identified, the SUBFIND algorithm (Springel et al., 2001) identifies the self-bound local overdensities of each FoF group as subhalos. The subhalo that contains the particle with the lowest value of the potential energy will be defined as the central sub-halo.

The simulation information is saved at 29 redshifts from z=20 to z=0 (i.e. 29 snapshots), and is used to build merger trees, which connect a halo to its progenitors at earlier redshifts (Qu et al., 2017). The main progenitor of a halo is defined as the progenitor with the largest mass at all earlier outputs. We use these main progenitors to track the mass evolution of a DM halo (Section 5.3.3). Note that when two halos pass close to each other without merging they could momentarily belong to the same FoF group. As a consequence, the mass and the subhalo chosen as the central may be inconsistent at this snapshot when compared to the one immediately before or after the interaction (Behroozi et al., 2015). We introduce a scheme to clean such issues from the input data in section 5.3.2.

5.3.2 Data selection

Our data set consists of central galaxies inside halos with a mass larger than $M^c_{200} > 10^{11.1} M_\odot$. M^c_{200} corresponds to the mass inside the radius R^c_{200} of a halo, which is the radius within which the density is 200 times the critical density of the universe. The stellar mass of a galaxy is measured as the baryonic mass contained inside a sphere of 30 proper kpc around the centre of potential of the halo.

Baryonic processes inside halos can affect their measured DM properties (Bryan et al., 2013; Schaller et al., 2015). If we run our code using the properties of the

DM found in a hydrodynamical simulation, we risk including biases by fitting the stellar mass using a property that has been modified by the presence of baryons (this modification would be correlated with the stellar mass as the halos with more baryons would be more modified). To avoid this bias, it is common practice to extract all DM input properties from a DM only simulation generated with the same initial conditions, box size and resolution as the full hydrodynamic simulation.

The matching between the hydrodynamic and DM-only simulations is described in Schaller et al. (2015). The 50 most bounded DM particles of each halo in the hydrodynamic simulation are found. If a halo in the DM-only simulation has at least half of those most bound particles it is considered its analog. Using this method, 99% of the halos with $M^c_{200} > 1 \times 10^{11.1} M_\odot$ are matched. We collect information about the host DM halo at different redshifts (Section 5.3.3) and require that our halos are present in all snapshots. With this in mind, we only use galaxies with a progenitor defined at $z = 4$. Our full sample consists of 9521 galaxies.

Inconsistencies between snapshots are a well-known characteristic of the merger trees (Behroozi et al., 2015) created by running the halo finder separately on each snapshot. When two halos interact some of the particles of one can be assigned to the other regardless of where they belonged in past snapshots. One consequence is that small central halos can be considered satellites of a larger halo if they are close to each other at a given snapshot. In EAGLE, M^c_{200} is only computed for central halos, which means that they will not have a value of M^c_{200} at these snapshots.

When this happens we interpolate the value of M^c_{200} in the missing slices using the following methodology: we look for the M^c_{200} value of both the nearest earlier and later redshifts where the halo was still central. We use these values to do a linear interpolation of M^c_{200} in the missing slice. The nearest earlier redshift is always well defined (as at $z = 0$ all of our selected subhalos are central); however, a small subset of galaxies have a non-central progenitor at their largest redshifts, and therefore their nearest latter subhalo is not necessarily well defined. In these cases we select the third to last and second to last halos and perform our interpolation with those. We follow a similar procedure to correct the angular momentum of halos that are not considered central in a given slice. We found that the value of the angular momentum can have drastic variations when compared to its value at the surrounding redshift slices, which is due to the number of particles assigned to the halo changing significantly when it is misclassified as a subhalo.

The black lines of Fig. 5.4 shows the halo mass history relative to the halo mass at redshift zero of four halos from $z = 4$ and to $z = 0$. The figure shows that different halos have very different formation histories. We will explore whether galaxies that have followed different halo formation paths will end up having different residuals in the SMHM relation.

5.3.3 DM Mass

Once we have selected the galaxies in our data set, we define the M parameters of the DM halo that are used to build the matrix X' of Eq. 5.2. The first variable accounted for is the halo mass at redshift zero (or any variable highly correlated with it), as the SMHM relation explains most of the scatter in the stellar mass. We will denote the Halo Mass input variable of a galaxy as M_0' and define it as

$$M_0' = \log_{10}(M_{200}^c(z = 0)/M_\odot). \tag{5.21}$$

We use Eq. 5.3 to standardise the units and denote the halo mass in standardised units as M_0.

There is significant scatter around the SMHM relation due to their varied formation history, therefore we should also add parameters that are good estimators of the mass evolution of the DM host halo. This can be done by adding the halo mass of the main progenitor of a host halo at different redshift slices into our X' matrix.

The EAGLE simulation has 19 snapshots between $z = 0$ and $z = 4$ (McAlpine et al., 2016). However, information between redshift slices that are close to each other is strongly correlated as halos have not evolved significantly. Keeping this in mind and given that the computational cost of running the minimiser increases exponentially with the number of parameters, we only use a subset of the available redshifts. The ten redshifts slices that we use as inputs are $z_{\mathbf{in}}$=[0.0, 0.18, 0.37, 0.62, 1.0, 1.26, 1.74, 2.48, 3.02, 3.98].

Sparse regression methods work best if variables are independent, therefore we will use the ratio between the mass at a given redshift and the mass at redshift zero (so that the significant correlation of the mass at a given redshift and its mass at $z = 0$ is removed). We will denote these variables as $(M_z/M_0)'$ and they are defined as:

$$(M_z/M_0)' = \log_{10}\left(\frac{M_{200}^c(z)}{M_{200}^c(z=0)}\right) \tag{5.22}$$

We then use Eq. 5.3 to standardise the units and form M_z/M_0.

An alternative approach to characterise halo evolution is the formation time (Lacey and Cole, 1994), defined as the time at which a halo has assembled half of its present day mass. We generalise this idea to define five formation criteria ($\mathbf{FC}'$) by finding the redshifts (instead of times) at which the DM halo has assembled 20%, 30%, 50%, 70% and 90% of its mass respectively. The set of all five formation criteria for our sample will be referred to as $\mathbf{FC}'_p$, where p denotes the percentage used. Fig. 5.4 shows a set of horizontal blue lines corresponding to halo mass ratios (Eq. 5.22) of 90%, 70%, 50%, 30% and 20%. The redshifts at which each formation history curve (black solid line) intersects this blue horizontal lines is a visual representation of $\mathbf{FC}'_p$. The redshifts that correspond to a given formation criteria are found by performing a linear interpolation of the halo mass ratios. As with all parameters, a final step is to standardise the units using Eq. 5.3 and to form $\mathbf{FC}_p$.

5.3.4 Specific angular momentum

There is a well known observational scaling relation between the angular momentum of a galaxy and its stellar mass (Fall and Romanowsky, 2013). It is a matter of discussion, however, how much of a role the angular momentum history of a dark matter halo plays in determining the specific angular momentum of its

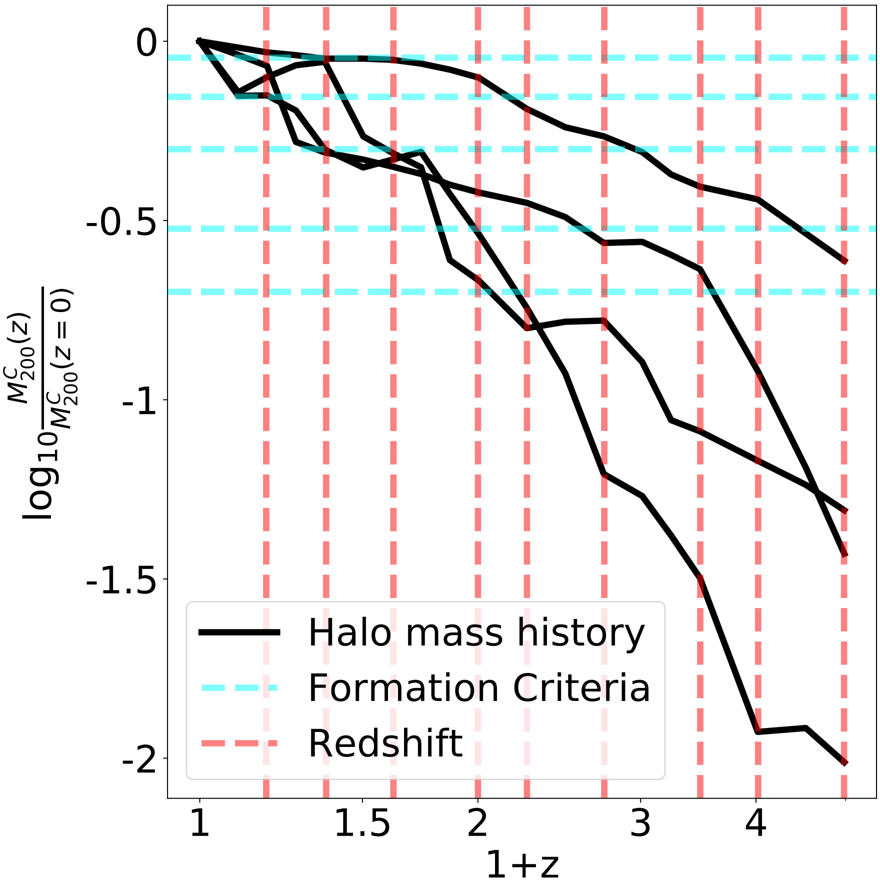


Figure 5.4: Halo mass history of four halos between $z = 4$ and $z = 0$ (black lines), as given by the ratio of the mass at z to its present day value. The vertical red dashed lines indicate the redshifts used in the analysis (i.e. $z_{\mathbf{in}}$). The $(M_z/M_0)'$ parameters are given by the intersection of the red and black lines. The blue horizontal lines correspond to a constant mass ratio of 90%, 70%, 50%, 30% and 20% (from top to bottom). The formation criterion parameters $\mathbf{FC}'_p$ can be visualised as the intersection between the blue and the black lines.

host galaxy. Zavala et al. (2016) finds strong correlations between both parameters using the EAGLE simulation. However, Danovich et al. (2015) suggest that the specific angular momentum of gas and dark matter undergo different formation histories, which would suggest that any correlation between them is a by-product of a third correlation with other parameters like the mass formation history. Having this in mind, we will generate candidate models that also include specific angular momentum input parameters on top of the mass evolution parameters.

Angular momentum evolution is included in our methodology by computing the halo specific angular momentum vector, $\vec{\mathbf{j}}$, defined within a radius R and for each redshift slice z as:

$$\vec{\mathbf{j}}(R,z) = \frac{\sum_i m_i(\vec{r}_i - \vec{r}_c) \times (\vec{v}_i - \vec{v}_c)}{\sum_i m_i}, \tag{5.23}$$

where $\vec{r}_i$ and $\vec{v}_i$ are the position and velocity vectors of each particle within a radius R of the centre of mass, m_i is the mass of the particle, and $\vec{r}_c$, $\vec{v}_c$ are the position and velocity of the centre of mass of the halo. We will use different values of R in order to capture the angular momentum evolution of the full halo and of its centre separately. The values of R that are included in our model are R^c_{200}, $\frac{R^c_{200}}{2}$ and $\frac{R^c_{200}}{5}$, which are all functions of redshift.

The specific angular momentum defined in Eq. 5.23 correlates strongly with the mass of the halo: this is driven by the scaling relations $|\vec{r}| \propto M^{1/3}$ and $|\vec{v}| \sim \sqrt{\frac{GM}{R}} \propto M^{1/3}$. To avoid strongly correlated variables in our parameter set, we define the following specific angular momentum parameter:

$$(S(R,z))' = \log_{10}(|\vec{\mathbf{j}}(R,z)|) - \frac{2}{3}\log_{10}(M^c_{200}(z)/M_\odot), \tag{5.24}$$

where $|\vec{\mathbf{j}}(R,z)|$ is the norm of $\vec{\mathbf{j}}(R,z)$.

Given that the angular momentum is a vector, we need two types of variables to describe it: one capturing its magnitude and the other one its direction. Therefore, we will also include the change in the parameter, Θ', defined as the cosine of the angle between the halo specific angular momentum at redshift z w.r.t. the one at the present time, i.e.:

$$(\Theta(z))' = \frac{\vec{\mathbf{j}}(R^c_{200},z) \cdot \vec{\mathbf{j}}(R^c_{200},0)}{|\vec{\mathbf{j}}(R^c_{200},z)||\vec{\mathbf{j}}(R^c_{200},0)|}. \tag{5.25}$$

Note that by definition $(\Theta(z=0))' = 1$ for all galaxies and hence we only include $(\Theta(z>0))$ in our list of variables. As with all other variables we use Eq. 5.3 to standardise the units and form the scalars $S(R,z)$ and $\Theta(z)$. We form the following library of $\vec{\mathbf{j}}$ parameters for each halo i at each redshift: $S_i(R^c_{200},z)$, $S_i(\frac{R^c_{200}}{2},z)$, $S_i(\frac{R^c_{200}}{5},z)$ and $\Theta_i(z)$. The evolution of our specific angular momentum parameters has significant statistical noise and so it is smoothed across different redshifts using a Gaussian Kernel.

5.3.5 Models

The four models considered in this work are:

1. **Mass ratio**: This model includes values of the halo mass at redshift zero, M_0 (Eq. 5.21), and the halo mass ratios, M_z/M_0 (Eq. 5.22), that parameterise the DM halo mass evolution. With 10 different redshift slices, this gives a total of $M = 10$ input parameters, resulting in a total of $D = 286$ functions to minimise over (Eq. 5.4).

2. **Formation criterion**: In this model, the DM ratios are replaced by the formation criterion $\mathbf{FC}_p$, defined in Section 5.3.3. This model uses, as parameters, 5 values of $\mathbf{FC}_p$ (with $p = [90, 70, 50, 30, 20]$) and the halo mass at redshift zero, M_0, resulting in $M = 6$ and $D = 84$ functions to minimise over.

3. **Mass ratio and $\vec{\mathbf{j}}$**: In this model we add the specific angular momentum parameters $\vec{\mathbf{j}}$ (and more specifically $S(R^c_{200}, z)$, $S(R^c_{200}/2, z)$, $S(R^c_{200}/5, z)$ and $\Theta(z)$ at each of the ten snapshot considered), to the library of free parameters of the mass ratio model. The library of functions contains the linear, quadratic and cubic terms of the Halo mass evolution parameters M_z/M_0. Only the linear terms of the specific angular momentum parameters are included. To include all the quadratic and cubic terms would result in $D = 23426$ functions to minimise over, which at the moment is too computationally expensive for our algorithm. Hence we will only include linear terms for the specific angular momentum parameters, ending up with a total of $D = 326$ functions to minimise over.

4. **Formation criterion and $\vec{\mathbf{j}}$**: This model is similar in spirit to model (iii), but we add the terms of the specific angular momentum parameters, $\vec{\mathbf{j}}$, to

the library of free parameters of the formation criterion model instead. As with model (iii), we consider only the linear terms of the specific angular momentum parameters, ending up with $D = 123$ functions to minimise over.

5.4 Running the Algorithm

In this section we present some specific aspects of applying the methodology presented in Section 5.2 to the data described in Section 5.3. In particular tests of the consistency of the algorithm are considered: we evaluate the impact of the chosen ϵ parameter in Section 5.4.2 and discuss the uncertainty of the parameter models in Section 5.4.3. The model results are presented and discussed in Section 5.5.

5.4.1 Training, holdout and test sets

The data is randomly divided into two, the training set and the holdout set. The training set contains 85% of the data and is used by the algorithm to build the model. The remaining 15% constitute the holdout set and is not used until the model is completed. The final model is applied to the holdout set to test its accuracy by considering data not used in the building of the model and therefore is unbiased to over-fitting.

Note that the holdout data set is different from the test sets used for estimating the optimal value of the hyperparameter λ in the k-fold methodology of Section 5.2.4. The latter constitutes sets drawn from the training set that are systematically kept out of the minimisations done while exploring the λ parameter space and are used to determine λ_{min}. They are part of the methodology for building our model. The holdout set, on the other hand, is kept out of this methodology completely and is used to evaluate the final model once it is built.

5.4.2 Penalty Hyperparameter

This section applies the methodology used for optimizing the hyperparameter λ, as introduced in Section 5.2.4. It discusses the impact of the assumed value for the parameter ϵ, used in the penalty function (Eq. 5.10). From Section 5.2.4, the optimal value of λ, λ_{min}, is determined using a k-fold method with $k = 10$ folds. Each fold runs independently and in parallel on different computer nodes. Fig. 5.5 shows the evolution of the RMSE of the mass ratio model (§5.3.5) as function of the hyperparameter λ. The green and blue dashed lines correspond to the test and training sets respectively. Each test set (green dashed lines) has around 800 points, which is around 10% of our training data set, and the minimisation runs with $D = 286$ free parameters. The green dashed lines show some spread in their amplitudes, which are correlated with their value at $\lambda = 0$. This spread is a consequence of dividing the subsets randomly. Some subsets will contain a larger amount of points that are well predicted by the model and will, therefore, have smaller errors.

As we saw with Fig. 5.2, the RMSE of the training set is smaller when $\lambda \sim 0$ as overfitting makes the model agree unreasonably well with the data it uses for the fitting. In contrast when the model is tested on data it has not seen before, the RMSE is larger, as shown by the comparatively larger error on the test set. As λ increases, the error on each test set decreases and eventually, reaches a minimum (RMSE_k) around $\lambda \sim 100$, as shown by the black dots in Fig. 5.5. This is where the model is least susceptible to overfitting, while still capturing the important features of the data set.

The red dots in Fig. 5.5 show the correction obtained with the one-standard-error rule from Eq. 5.18. The plot shows that these points are to the right of the minimum value of the green dashed lines, however the differences in RMSE between the actual minima and the red dots are small. This means that the resulting models are simpler (and therefore more explainable) and with comparable accuracy. The

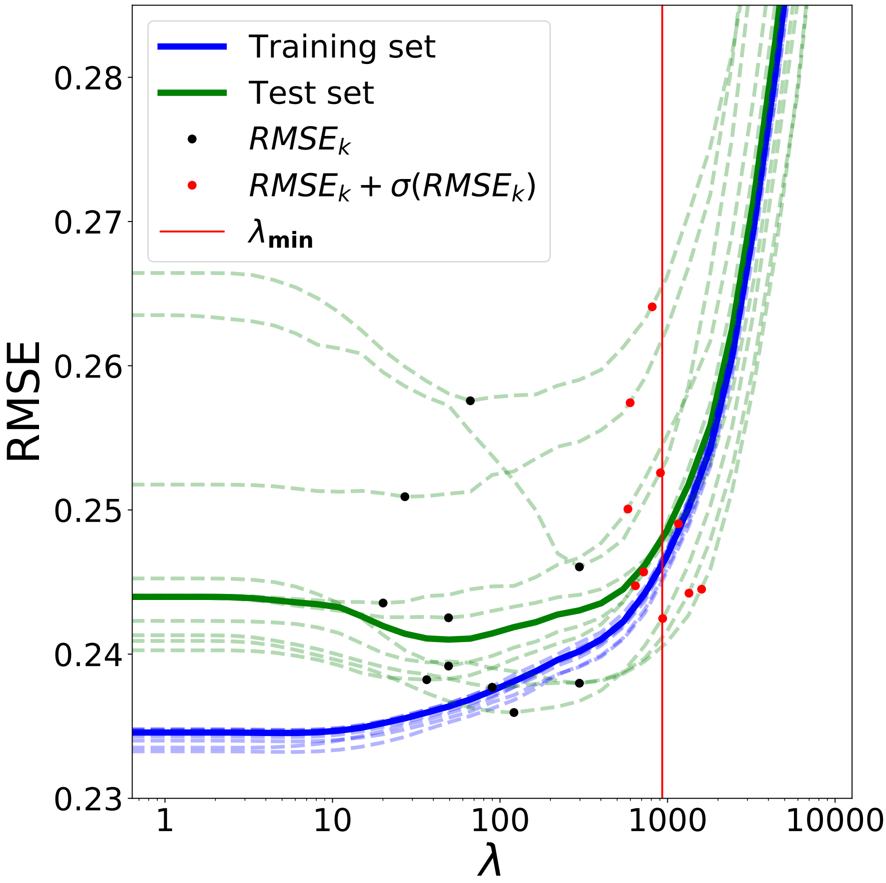


Figure 5.5: Evolution of the RMSE (Eq. 5.17) of the mass ratio model (§5.3.5) as a function of hyperparameter λ for our nominal EAGLE data set (Section 5.3). The blue and green dashed lines represent the training and test sets respectively. The solid lines represent the median of these curves. The black dots show the minimum of the dashed lines (RMSE_k) and the red dots the one-standard-error rule correction from Eq. 5.18. The red solid line corresponds to the mean λ of the red dots and is our estimate of λ_{min}.

red solid line is the optimised value of the hyperparameter, λ_{min}, as estimated using Eq. 5.18.

Fig. 5.6 shows the evolution of the coefficients C_i of Eq. 5.5 of the mass ratio model (§5.3.5) as a function of the hyperparameter λ. The vertical black dotted line shows the value of λ_{min} found by our algorithm. Each coloured line corresponds to a coefficient that is above the cutoff value ν at λ_{min}, with ν represented as the boundary between the white and grey regions of the plot. The grey dashed lines correspond to the coefficients that are below ν at λ_{min} and therefore discarded.

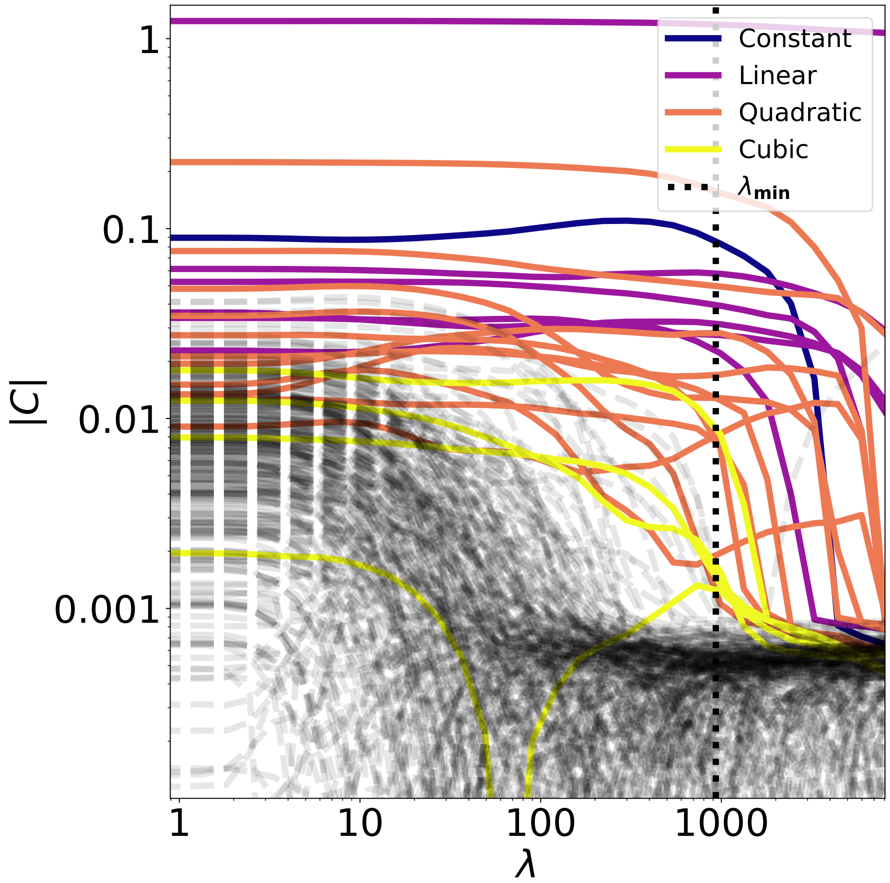


Figure 5.6: Evolution of the best fit value of each coefficient of the mass ratio model as a function of the hyperparameter λ. The coloured lines show the values of the accepted coefficients and the black dashed lines represent the rejected coefficients. The vertical dotted black line shows the value of λ_{min} and the grey shaded area represents the region bellow the cutoff value ν: all coefficients above the shaded region at λ_{min} are retained by the model and represented by coloured lines.

The figure shows that coefficients that have been discarded have a value of around 0.0005 or lower (shown by the average value of the black dashed lines at large values of λ), given that our cutoff value is 0.001 there is a distinct separation between the chosen coefficients and those discarded.

Different C_i coefficients are fitted by the minimiser with different orders of magnitude*. Therefore we need to make sure that the value of the parameter ϵ of the penalty term (Eq. 5.10), which determines how close to zero unnecessary paramet-

*As our input variables are not Gaussian, several parameter values are above one standard deviation. In a standardised space this will mean that they will be larger than one. As a consequence linear, quadratic and cubic coefficients will require different scales to make similar contributions.

ers get in the minimisation, is such that discarded coefficients are well below the cutoff value ν, and are close enough to zero that they can be separated from useful coefficients.

Nominally we use a value of $\epsilon = 10^{-3}$, which, as shown in Fig. 5.6, corresponds to the minimiser setting unused parameters to a value as small as $\approx 6 \times 10^{-4}$. This is comparable to the findings of the example presented in Section 5.2.5. A very small value of ϵ increases the computational time significantly given that parameters need to be driven further toward zero. Our choice represents a value of ϵ that is small enough to get parameters close enough to zero while not being so small that the code becomes too expensive to run.

To test what impact the value chosen for ϵ has, we consider the formation criterion model (§5.3.5). This model has less free parameters than the mass ratio model ($D = 84$ versus $D = 286$) and hence requires significantly less computational time, enabling an adequate ϵ parameter space to be explored. Fig. 5.7 shows the resulting coefficients after running our full algorithm using five different values of ϵ using the formation criterion model. The coloured lines show the parameters that are above the cutoff value ν in the model built with $\epsilon = 10^{-3}$ (our standard value) and represent the variables that where chosen by the algorithm. The grey dashed lines correspond to the values rejected at $\epsilon = 10^{-3}$. The cutoff value ν depends on how close parameters get to zero and therefore it is a function of ϵ. For the propose of illustration, we set $\nu = \epsilon$.

For larger values of ϵ, there is no clear cut between discarded coefficients and most of the cubic and quadratic terms end up in our model. On the opposite end at $\epsilon = 5 \times 10^{-4}$, all accepted coefficients are significantly greater than the cutoff value. While the difference between useful and useless coefficients is clearer at $\epsilon = 5 \times 10^{-4}$, our standard configuration with $\epsilon = 10^{-3}$ seems to work just as well while being significantly faster to run.

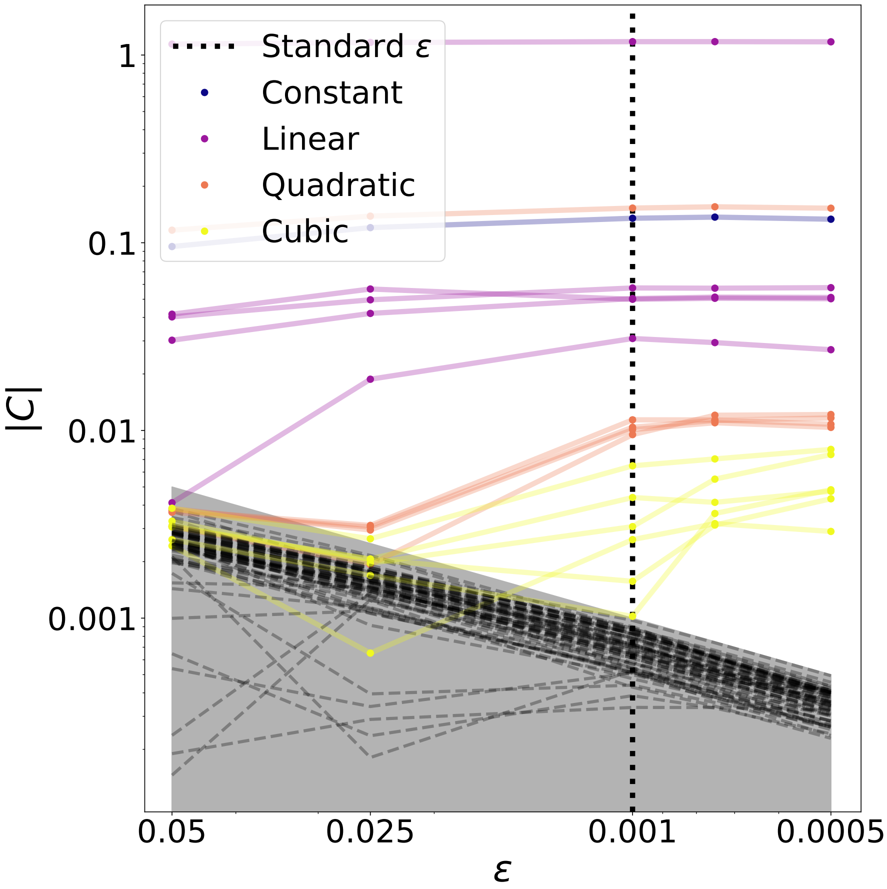


Figure 5.7: Best fit coefficients for the formation criterion model for five different values of the ϵ parameter (Eq. 5.2.3). This parameter determines how close to zero coefficients get before their contribution to the penalty is negligible. The cutoff value ν is set as $\nu = \epsilon$ for each run. The black dotted line shows the value of ϵ used in our standard configuration. When ϵ is large, all coefficients are above the cutoff value ν. For $\epsilon = 5 \times 10^{-4}$, all kept coefficients are significantly larger than 10^{-3}, indicating the adequacy of our nominal choice for ϵ.

5.4.3 Uncertainty on the models

Several of our coefficients C_i are associated with functions of the same form but with inputs from different redshifts (see Section 5.3.3). If the halo mass does not vary significantly between adjacent redshift slices, then the corresponding polynomial functions $f_i(x)$ are likely to show some correlation between them. In general different order combinations of correlated terms will also be correlated. Considering the above statements, it is possible that the parameter space of C_i coefficients has several local minima. This could be an issue for gradient descent algorithms, as by construction they will converge only toward the closest minimum. In practice we are satisfied with any reasonable minimum: for example, we do not have a preference between a feature being explained by the halo mass ratio at one specific redshift versus that of an adjacent redshift slice.

This, however, means that there might be slight variations in the surviving parameters of different models depending on the starting point of the minimisation and depending on the specific selection of the training set. We test for both aspects in turn.

To test how strong an effect the initial starting point is, we perform five different minimisations of the formation criterion model using 5 distinct starting points in the minimisation algorithm. We set $\lambda_{\rm min} = 932$, which is the optimised value found by running our methodology with our standard configuration. The initial point in the parameter space C_i is varied to random values between the five runs and is the only feature that is different between runs. Fig. 5.8 shows the best fit C_i coefficients obtained using 5 different sets of initial positions. All models have an equivalent accuracy with a RMSE within the range 0.249 ± 0.001. Three out of the five models use 19 parameters and the remaining two use 18. All resulting models have equivalent accuracy and simplicity and we can not select one as being significantly better than the rest.

We can tell that the most significant coefficients (i.e. those with a larger C_i) are kept constant amongst all runs, similarly there is a large subset of coefficients that are not necessary in any of the models. However, there is a subset of parameters that are interchangeable between different models. An example of this is shown by the green and blue circles, which correspond to the coefficients associated with $M_0 \times \mathbf{FC}_{30}$ and $M_0 \times \mathbf{FC}_{20}$ functions in runs 3 and 4 respectively. Both runs are very similar in almost every parameter, except that run 3 gives a very important role to the $M_0 \times \mathbf{FC}_{30}$ function and almost discards the $M_0 \times \mathbf{FC}_{20}$ function, while run 4 does the opposite. This indicates that both parameters are correlated with

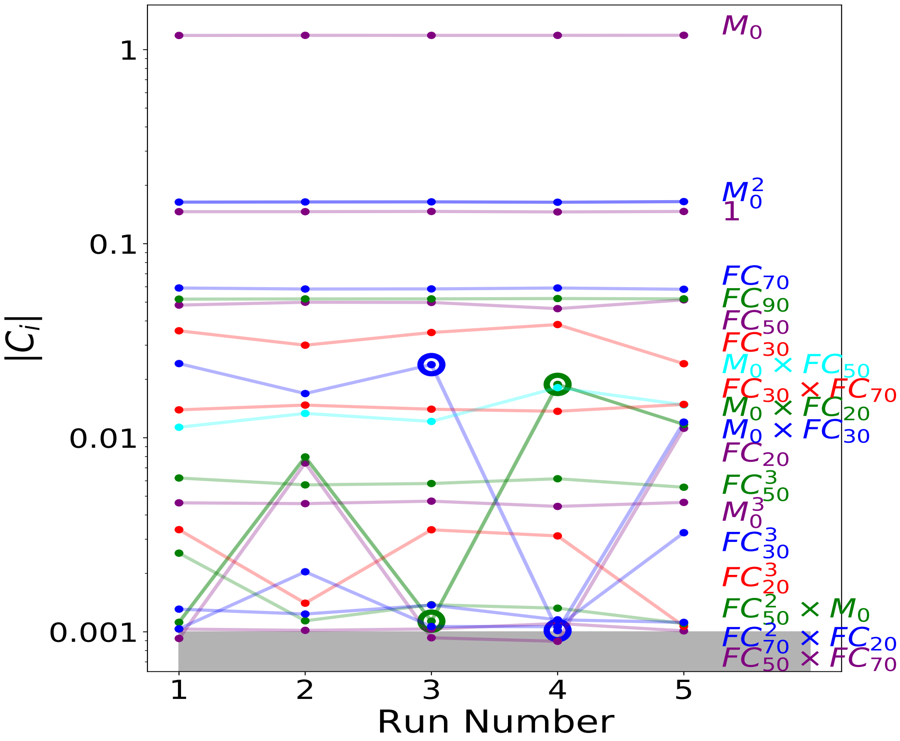


Figure 5.8: Best fit absolute values of coefficients C_i for the formation criterion model using 5 different initial positions. The lines connect coefficients that survived in at least one model, with the right hand key indicating which function they refer to. The colour coding of the lines is only there to help to differentiate between them. Blue and green circles correspond to the coefficients associated with $M_0 \times \mathbf{FC}_{30}$ and $M_0 \times \mathbf{FC}_{20}$ functions in runs 3 and 4 respectively. They are highlighted as an example of correlated variables associated with different local minima. The grey area represents the cutoff value ν.

each other and that our methodology can choose one or the other and still come up with equivalent solutions.

To test the variance of our methodology, we make six independent runs of the formation criterion model, varying only the holdout set, the data that is kept outside of the model fitting process. One of the holdout sets is our standard holdout, used throughout the paper. The other five correspond to five independent subsets of the training set with the same amount of points that the standard holdout set: the six independent holdouts considered have each 15% of the whole data set. The RMSE of the 6 resulting models are [0.167, 0.170, 0.169, 0.162, 0.162, 0.167] and they have [16, 14, 15, 15, 18, 17] surviving coefficients each respectively. Therefore all six models have similar accuracy and comparable simplicity. Fig. 5.9 shows the variations in the resulting C_i coefficients that survived in at least one of the six models. Solid line are used for the eleven coefficients that survived in all of six runs. This means that on average two thirds of all coefficients are the same irrespective of the specific holdout data set used. We note that the numerical values of those eleven coefficients are often of similar amplitude in all runs. Of the remaining coefficients, two are present in five of six models and a further two in four of six. Hence there are 15 coefficients present in nearly all six models, indicating how robust our algorithm is to changes in the holdout set used. We note that some of the other coefficients found in some runs are likely correlated with those ones and are sometimes present but discarded in at least half of the runs.

5.5 Results

We now present the results of our four models defined in Section 5.3.5, i.e.: (i) Mass ratio, (ii) Formation criterion, (iii) Mass ratio and $\vec{\mathbf{j}}$ and (iv) Formation criterion and $\vec{\mathbf{j}}$. The specific surviving coefficients C_i selected by each of the models are presented in Table 5.1, where coefficients are reported in standardised space. They can not be used directly to model the actual data, which needs to be transformed using Eq. 5.3. The standardised space is defined by the mean and the standard deviation of the logarithm of the stellar mass of galaxies and of the dependent variables $\vec{z}_i'$, which are shown in 5.2.

A striking feature of models (iii) and (iv), the two models with $\vec{\mathbf{j}}$, is that the algorithm does not select any specific angular momentum parameters in either of them. In fact the selected parameters of model (ii) and (iv) are almost identical. While there are some small differences between the coefficients chosen in models (i) and (iii), namely model (iii) selects two extra parameters, and the values of some of the common parameters are slightly different, these difference are consistent with the variance of the methodology reported in Section 5.4.3. This indicates that the

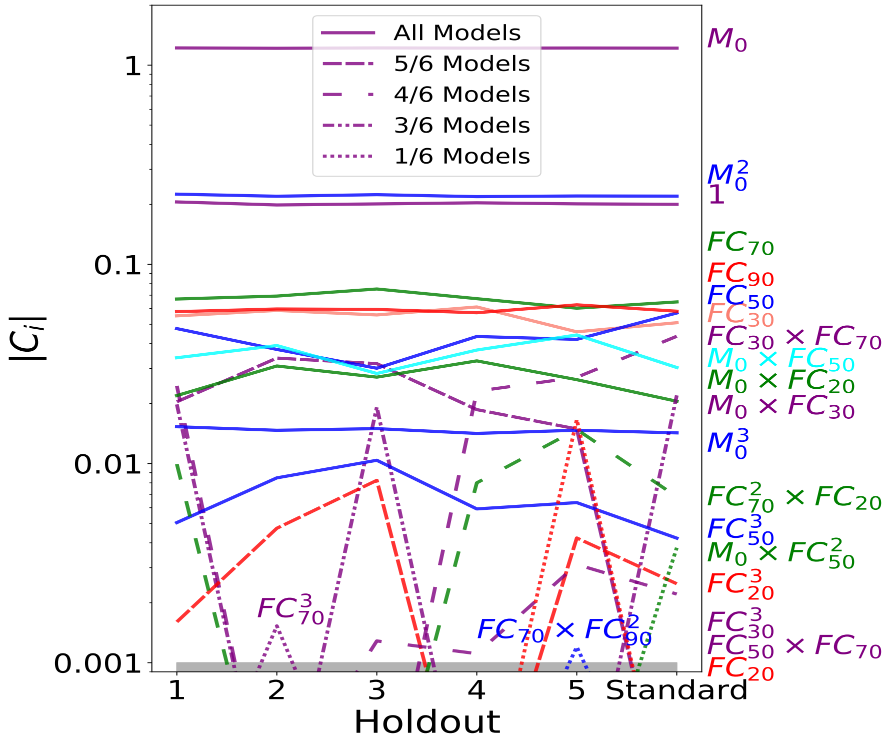


Figure 5.9: Best fit absolute values of the coefficients C_i for the formation criterion model using six different holdouts, with the right most one corresponding to the standard holdout set used throughout this paper. The lines connect coefficients that survived in at least one model, with the right hand key indicating which function they refer to. The line style indicates how often a given coefficient was kept by the best fit model (as indicated by the key). The colour coding of the lines is only there to help to differentiate between them. Each run uses %15 of the data as holdout set, each of which are disjoint from each other. The resulting models, which have similar accuracy (RMSE=0.166 ± 0.004), select a somewhat different subsets of surviving coefficients C_i, with the most important ones remaining the same and the less important ones often exchanged for comparable ones. See text for further discussion.

contribution to the accuracy of the model after including the angular momentum parameters is negligible: the sparse regression method found that no angular momentum parameters contributed additional information necessary to describe the SMHM relation that was not already provided using the rest of the parameters. This suggests that any correlation between the specific angular momentum history of a galaxy and that of its host halo should be the consequence of a correlation between the mass and specific angular momentum formation histories of host halos.

Fig. 5.10 shows the predicted values of the stellar mass for all galaxies in the holdout set for three models (omitting model (iv) as it is so similar to model (ii)) compared to their real values in the EAGLE simulation. The closer a point is to the one-to-one line (black dashed line), the better the model predicted its value. We also include the RMSE of each model, as given by Eq. 5.17.

A different estimate of the goodness of a fit is the R^2 statistic, which determines the amount of the variation in $\vec{y}$ that can be explained by a model[*]:

$$\mathbf{R}^2 = 1 - \frac{\mathbf{RMSE}^2}{\sigma_y^2}, \tag{5.26}$$

where σ_y is the standard deviation of $\vec{y}$. The usefulness of the R^2 comes from being intuitive to interpret: the closer to one the R^2 of a model is, the more accurate it is.

Both the RMSE and R^2 statistics show that the three models have very similar accuracy. The formation criterion model is slightly simpler than both mass ratio models, as the former has 17 free parameters, compared with 20 and 22 from the two mass ratio models. This suggests that the formation criteria parameters, $\mathbf{FC}_i$, are slightly more efficient at summarising the halo mass information than the mass ratio parameters, (M_z/M_0).

[*] R^2 estimators should be considered with caution as they are easily biased by inaccurate estimations of σ_{y_α} and can have deceivingly small (or large) values. They should be used as reference only. We also include RMSE errors as goodness of fit estimators, which are far more robust.

Models:	(i) mass ratio		(ii) formation criterion		(iii) mass ratio and $\vec{\mathbf{j}}$		(iv) formation criterion and $\vec{\mathbf{j}}$	
N	Coefficient	Value	Coefficient	Value	Coefficient	Value	Coefficient	Value
1	1	0.120	1	0.200	1	0.139	1	0.200
2	M_0	1.22	M_0	1.217	M_0	1.22	M_0	1.217
3	$M_{0.18}/M_0$	0.0335	$\mathbf{FC}_{30}$	0.0509	$M_{0.18}/M_0$	0.0334	$\mathbf{FC}_{30}$	0.0509
4	$M_{0.37}/M_0$	0.0469	$\mathbf{FC}_{50}$	0.0567	$M_{0.37}/M_0$	0.0465	$\mathbf{FC}_{50}$	0.0570
5	$M_{0.62}/M_0$	0.0662	$\mathbf{FC}_{70}$	0.0648	$M_{0.62}/M_0$	0.0671	$\mathbf{FC}_{70}$	0.0647
6	M_1/M_0	0.0410	$\mathbf{FC}_{90}$	0.0582	M_1/M_0	0.0412	$\mathbf{FC}_{90}$	0.0582
7	$M_{1.26}/M_0$	0.0312	M_0^2	-0.220	$M_{1.26}/M_0$	0.0284	M_0^2	-0.220
8	$M_{1.74}/M_0$	0.0578	$M_0 \times \mathbf{FC}_{20}$	-0.0205	$M_{1.74}/M_0$	0.0572	$M_0 \times \mathbf{FC}_{20}$	-0.0206
9	$M_{2.48}/M_0$	0.0545	$M_0 \times \mathbf{FC}_{30}$	-0.0220	$M_{2.48}/M_0$	0.0357	$M_0 \times \mathbf{FC}_{30}$	-0.0219
10	M_0^2	-0.172	$M_0 \times \mathbf{FC}_{50}$	-0.0304	$M_{3.02}/M_0$	0.0162	$M_0 \times \mathbf{FC}_{50}$	-0.0304
11	$(M_1/M_0)^2$	0.00936	$\mathbf{FC}_{30} \times \mathbf{FC}_{70}$	-0.0435	M_0^2	-0.215	$\mathbf{FC}_{30} \times \mathbf{FC}_{70}$	-0.0435
12	$(M_{1.74}/M_0)^2$	0.0136	M_0^3	0.0143	$(M_1/M_0)^2$	0.00928	M_0^3	0.0143
13	$(M_{2.48}/M_0)^2$	0.00747	$\mathbf{FC}_{20}^3$	0.00249	$(M_{1.74}/M_0)^2$	0.0130	$\mathbf{FC}_{20}^3$	0.00249
14	$(M_{3.02}/M_0)^2$	-0.00237	$\mathbf{FC}_{30}^3$	0.00221	$(M_{2.48}/M_0)^2$	0.00706	$\mathbf{FC}_{30}^3$	0.00221
15	$(M_{3.98}/M_0)^2$	0.00265	$M_0 \times \mathbf{FC}_{50}^2$	0.00378	$(M_{3.02}/M_0)^2$	-0.00326	$M_0 \times \mathbf{FC}_{50}^2$	0.00378
16	$M_0 \times (M_{0.62}/M_0)$	-0.0169	$\mathbf{FC}_{50}^3$	0.00422	$(M_{3.98}/M_0)^2$	0.00674	$\mathbf{FC}_{50}^3$	0.00422
17	$M_0 \times (M_{1.26}/M_0)$	-0.0114	$\mathbf{FC}_{70}^2 \times \mathbf{FC}_{20}$	0.00682	$M_0 \times (M_{0.62}/M_0)$	-0.0169	$\mathbf{FC}_{70}^2 \times \mathbf{FC}_{20}$	0.00682
18	$M_0 \times (M_{1.74}/M_0)$	-0.0139			$M_0 \times (M_{1.26}/M_0)$	-0.0117		
19	$M_0 \times (M_{3.02}/M_0)$	-0.0236			$M_0 \times (M_{1.74}/M_0)$	-0.0124		
20	$M_0^2 \times (M_{3.98}/M_0)$	-0.00415			$M_0 \times (M_{3.02}/M_0)$	-0.0348		
21					M_0^3	0.0142		
22					$M_0^2 \times (M_{3.98}/M_0)$	0.00643		

Table 5.1: Parameters and their values as selected by each of the four models of Section 5.3.5. Neither the mass ratio and $\vec{\mathbf{j}}$ model nor the formation criterion and $\vec{\mathbf{j}}$ model used any specific angular momentum parameters as part of their final coefficients. The formation criterion based models, i.e. models (ii) and (iv), are virtually identical with only very minor differences in some of the coefficient values. We note that the parameters in this table are all quoted in the standardised space, i.e. where all dependent variables have made used of Eq. 5.3. Parameters are shown to three significant figures, which we find are enough to make the RMSE accurate to the fourth significant figure.

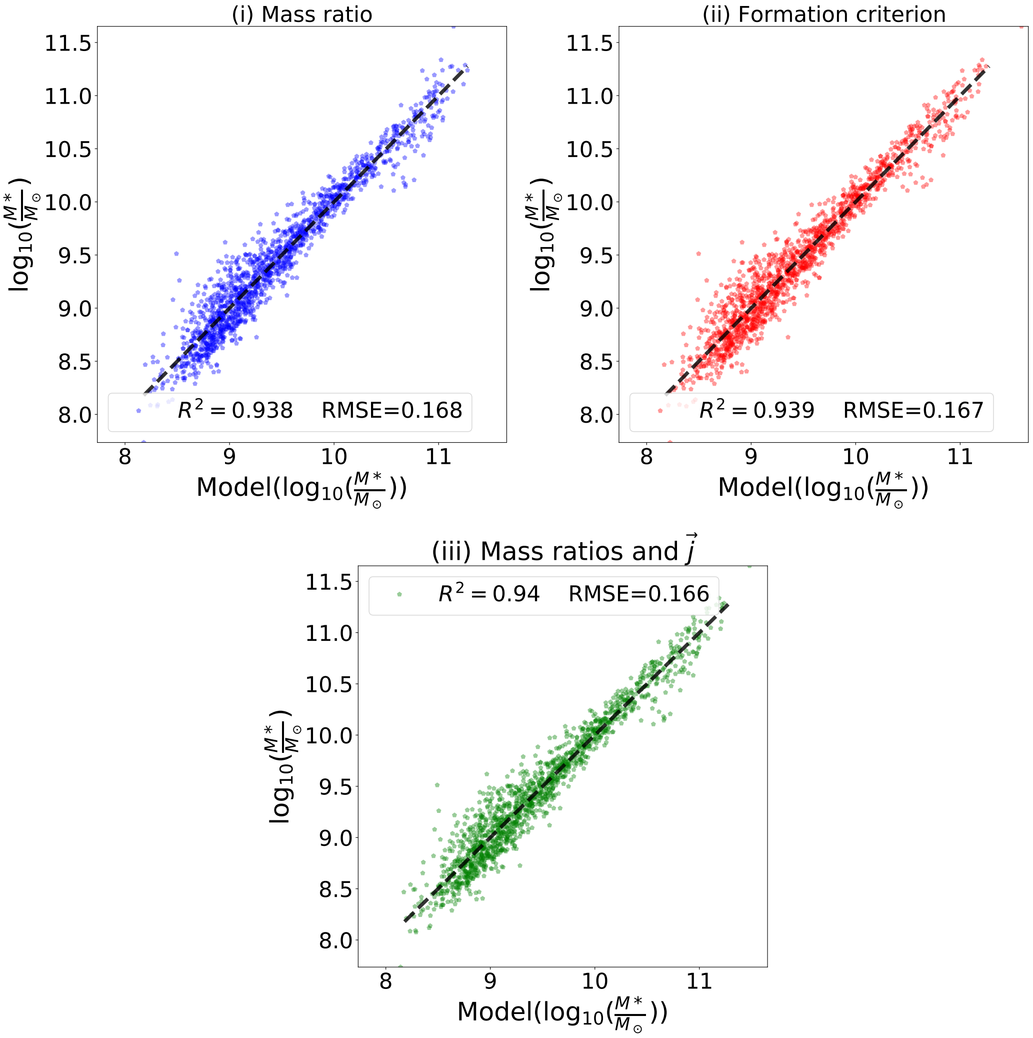


Figure 5.10: Comparison between the stellar mass predicted by the models and its actual value in the EAGLE simulation for all galaxies in the holdout set. The top left, top right and bottom panels correspond to the mass ratio, the formation criterion, and the mass ratio and $\vec{\mathbf{j}}$ models respectively (as indicated in the header of each panel). The closer each point is to the one-to-one relation (black dashed lines), the more accurate the model prediction is. The value of the RMSE and the R^2 statistic are included for each model. In general the three models have equivalent accuracy. As the formation criterion and $\vec{\mathbf{j}}$ model is virtually identical to the formation criterion model (see Table 5.1 for parameter values), we included only the latter one.

	$\log_{10} M^*/M_\odot$	M_0'	$\mathbf{FC}_{20}'$	$\mathbf{FC}_{30}'$	$\mathbf{FC}_{50}'$	$\mathbf{FC}_{70}'$	$\mathbf{FC}_{90}'$
μ	9.460	11.59	2.748	2.181	1.419	0.8794	0.3949
σ	0.6764	0.4723	0.7834	0.7389	0.5731	0.4125	0.2581

	$(M_{0.18}/M_0)'$	$(M_{0.37}/M_0)'$	$(M_{0.62}/M_0)'$	$(M_{1.0}/M_0)'$	$(M_{1.26}/M_0)'$
μ	-0.03263	-0.06793	-0.1233	-0.2254	-0.2973
σ	0.06933	0.09271	0.1147	0.1461	0.1660

	$(M_{1.74}/M_0)'$	$(M_{2.48}/M_0)'$	$(M_{3.02}/M_0)'$	$(M_{3.98}/M_0)'$
μ	-0.4352	-0.6548	-0.8086	-1.070
σ	0.2008	0.2468	0.2760	0.3254

Table 5.2: Normalisation parameters used for the stellar mass and the DM halo variables defined in section 5.3.3 and considered by our models. The μ and σ rows correspond to the mean and standard deviation of the variables respectively and are used in Eq. 5.3 to standardise the units of the variables considered.

5.5.1 Comparison with simpler models

While the LASSO approach uses only a fraction of the full set of available regression terms, the models it selects are still relatively complex and include non-linear combinations of terms characterizing the formation history. In this section, we compare our results to simpler models. Specifically, we compare the formation criterion model from the last section with the following two models:

- The first model is a third-order polynomial fit of the SMHM relation. This model includes the terms 1, M_0, M_0^2, and M_0^3. We label this model as M_0^3, with all four coefficients selected by our LASSO method*.
- Our second model is similar to the one presented in Equation (9) of Matthee et al. (2016). We include all terms of M_0 up to the third order and all linear terms of FC_{50}. More specifically, the eight possible terms are 1, M_0, M_0^2, M_0^3, $\mathbf{FC}_{50}$, $M_0 \times$ $\mathbf{FC}_{50}$, $M_0^2 \times$ $\mathbf{FC}_{50}$, and $M_0^3 \times$ $\mathbf{FC}_{50}$. We did not use the model presented in Matthee et al. (2016) directly because of small differences in the calibration redshift and in the methodology used for selecting and processing the EAGLE data sets. We label this model as (M_0^3 & $\mathbf{FC}_{50}$),

*The coefficients are $C(1) = 0.179$, $C(M_0) = 1.16$, $C(M_0^2) = -0.205$, $C(M_0^3) = 0.0152$, when quoted in the standardised space.

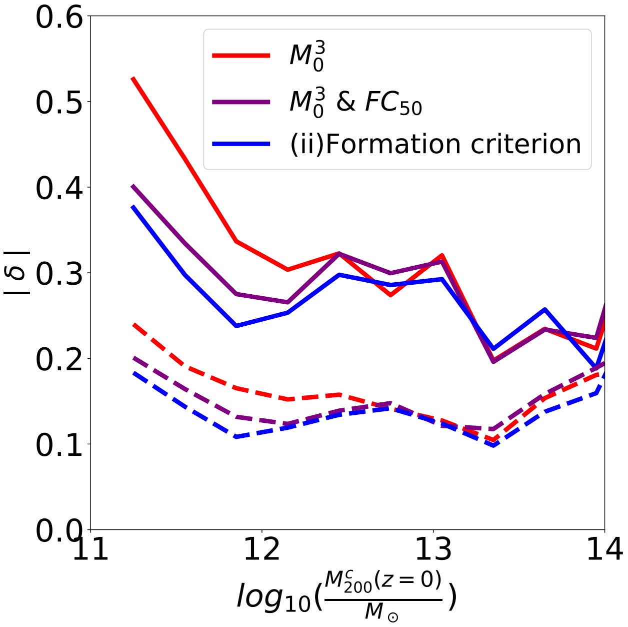


Figure 5.11: Comparison of the accuracy of the models discussed in Section 5.5.1, as traced by the δ error (also defined in §5.5.1) as a function of the present day halo mass. Dashed and solid lines correspond to the 68th and 95th percentiles of the absolute value of error distribution.

with six coefficients selected by our LASSO method*. We have tested the prediction of this model against the predictions of Matthee et al. (2016) † and find that the models are comparable.

As the models grow in complexity, their prediction of the stellar mass becomes more accurate, a way of quantifying this is by looking at the RMSE of our data set. The M_0^3 model has a RMSE of 0.225 when estimated with stellar mass units‡. We obtain very similar results looking both at the holdout set and the whole dataset.

*The coefficient are $C(1) = 0.156$, $C(M_0) = 1.22$, $C(\ \mathbf{FC}_{50}) = 0.199$, $C(M_0^2) = -0.169$, $C(M_0 \times\ \mathbf{FC}_{50}) = -0.274$, $C(M_0^3 \times\ \mathbf{FC}_{50}) = -0.00402$, in standardised space. The remaining terms were discarded by our LASSO methodology.

†Following a discussion with the authors, we identified an issue with the model in the way it was reported in the paper. The corrected model description is $\log_{10}(M^*) = \alpha - e^{\beta\, M_0^D + \gamma} - (a\, FC_{50} + b)$ where $M_0^D = M_0 - 12$, $a = 0.15048 + 0.21517\, M_0^D + 0.06412\, (M_0^D)^2 - 0.07217\, (M_0^D)^3$, $b = 0.20632 - 0.43077\, M_0^D + 0.25277\, (M_0^D)^2 + 0.34500\, (M_0^D)^3$ and α, β and γ are constants which values are given in Table 2 of Matthee et al. (2016)

‡In this sub-section, the RMSE is expressed in natural units, i.e. the logarithm of the stellar mass. This results in RMSE values which are more natural to understand, as here an RMSE of 0.2 implies that the mean error is $0.2 \log_{10}(M^*/M_\odot)$. We note that the RMSE depend on the parametrisation, which throughout the rest of this work is the one defined using standardised units (see Eq. 5.17).

For the M_0^3 & $\mathbf{FC}_{50}$ model, the stellar mass RMSE drops to 0.181, while it is 0.166 for the formation criterion model. Assuming that contributions from the different terms can be added in quadrature, this shows that 32% of the variance of the M_0^3 model is explained by including linear terms in $\mathbf{FC}_{50}$, while the more complex model selected by the LASSO process explains a further 10% of the variance, a modest but significant improvement. This suggests that the biggest improvement on the SMHM residuals (modelled by M_0^3) comes from the linear terms of $\mathbf{FC}_{50}$, the higher-order terms of $\mathbf{FC}_{50}$ and the terms corresponding to other formation criteria make a smaller but significant correction to the predicted stellar mass.

To explore the improvement of the model further, we define for each galaxy the error of a model as the difference between the actual stellar mass and the predicted one, or more precisely:

$$\delta = \log_{10}\left(\frac{M^*}{M_p^*(\vec{C},\ \mathbf{X'})}\right) \tag{5.27}$$

where M_p^* corresponds to the model predicted stellar mass of a galaxy of stellar mass M^*. Fig. 5.11 shows the 68th and 95th percentile ranges of $\mid \delta \mid$ as a function of halo mass for the reference formation criterion model (blue lines), and the M_0^3 & $\mathbf{FC}_{50}$ and M_0^3 models (purple and red lines respectively). The plot shows that the differences between the three models are most significant at small halo masses, while at halo masses larger than $\sim 10^{12.5} M_\odot$, all models are comparable. This suggests that galaxies in smaller halos are more readily explained by evolutionary effects (correlated with $\mathbf{FC}_p$ parameters), while the scatter in larger galaxies is perhaps more strongly influenced by stochastic baryonic processes, such as black hole accretion, that cannot be modelled using the halo mass history alone. This is in agreement with Matthee et al. (2016) that found no correlation between the scatter of the SMHM relation and formation time for halo masses larger than $\sim 10^{12.5} M_\odot$.

Rather than restricting, by hand, the choice of functions to terms that are linear in $\mathbf{FC}_{50}$, we can of course ask the LASSO methodology to simplify the formation criterion model, trading off an increase in variance for a reduction in complexity. It should be remembered, however, that this model will not provide optimal predictions for the stellar mass in a RMSE sense. We shift the balance to reduce complexity by increasing the penalty parameter λ of Eq. 5.6. As can be seen in Fig. 5.5, using a penalty λ three times larger than the one selected by the LASSO algorithm generates a model that is comparable to model M_0^3 & $\mathbf{FC}_{50}$ in terms of the RMSE and number of surviving terms. The terms retained by the model are: 1, M_0, $\mathbf{FC}_{50}$, $\mathbf{FC}_{70}$, $\mathbf{FC}_{90}$, $\mathbf{FC}_{20}^3$, $\mathbf{FC}_{30}^3$, $\mathbf{FC}_{50}^3$, with coefficients 0.0538, 1.13, 0.0315, 0.0534, 0.0242, 0.00590, 0.0104, 0.0108 respectively. Interestingly this model prefers to characterise the formation histories of the halos more precisely rather than to mix terms depending on halo mass and formation time.

5.5.2 Interpretation

The goal of this work is to make a model that is accurate and also explainable. With this in mind, we now try to give a physical interpretation to some of the terms kept in our model.

By looking at Table 5.1, we conclude that in general surviving parameters in all models can be divided into four different groups:

1. Terms forming a third order polynomial of M_0. Namely the terms 1, M_0, M_0^2, M_0^3.
2. Terms forming third order polynomials of the other dependent variables that are correlated with the mass at $z > 0$. Namely, terms of the form M_z/M_0, $(M_z/M_0)^2$ and $(M_z/M_0)^3$ for the mass ratio models with and without $\vec{\mathbf{j}}$, and terms of the form $\mathbf{FC}_p$, $\mathbf{FC}_p^2$ and $\mathbf{FC}_p^3$ for the formation criterion models with and without $\vec{\mathbf{j}}$.
3. Terms corresponding to the product of M_0 and either M_z/M_0 for the mass ratio models (i) and (iii) or $\mathbf{FC}_p$ for the formation criterion models (ii) and (iv).

4. Other terms corresponding to higher order combinations of crossed terms, which are more challenging to provide a physical interpretation of.

The terms in group (1) correspond to a direct modelling of the SMHM relation. Let us call $P^3(z=0)$ the polynomial built with the terms in group (1) and their associated coefficients, C_1^i:

$$P^3(z=0) = C_1^0 + C_1^1 M_0 + C_1^2 M_0^2 + C_1^3 M_0^3 \tag{5.28}$$

In order to compare our model stellar mass predictions with the EAGLE stellar masses, we transform our model from the standardised units to stellar mass units:

$$P'^3(z=0) = P^3(z=0)\,\sigma(\log_{10}(M^*)) + \mu(\log_{10}(M^*)) \tag{5.29}$$

where the stellar mass, M^*, is expressed in $M_\odot$, μ and σ are the mean and standard deviation operators considered in Eq. 5.3 already.

$P'^3(z=0)$ computed for the formation criterion model is shown as the black dashed curve of the left panel of Fig. 5.12. The figure shows that $P'^3(z=0)$ provides already a good model of the SMHM relation; however, there is some scatter around it that the model does not account for. We define the residual between each galaxy and the model prediction given by $P'^3(z=0)$ as δ':

$$\delta' = \log_{10}(M^*/M_\odot) - P'^3(z=0). \tag{5.30}$$

Galaxies in Fig. 5.12 are divided into four δ' bins. The yellow bin, which is the bin with the largest δ' values, correspond to galaxies for which their stellar masses are the most under-predicted by $P'^3(z=0)$, while the blue bin contains those with the most over predicted stellar masses. The right panel of in Fig. 5.12 shows the average mass of halos in each of the four bins as a function of redshift. On average galaxies in the yellow bin live inside host halos that attained their final mass early in their evolution when the characteristic density was higher. The deeper potential well of these halos allows the creation of massive galaxies. In contrast, the galaxies in the most over-predicted δ' bin (blue) live inside host halos that only achieved their final mass very recently and therefore had a lower characteristic density for a considerable period of time, compared to halos of the same mass in larger δ' bins. This implies that there is a correlation between δ' and the mass formation history and explains why coefficients in group (2) were selected by our model.

This conclusion is in agreement with Zentner et al. (2014), where formation time is used to model assembly bias, and with Matthee et al. (2016) where formation time is found to be the most correlated parameter with δ'. We emphasise that we arrive at this conclusion by using a completely different approach, that does not require any prior knowledge of the underlying physics correlating stellar mass with halo mass and formation time.

A novel result from our model is that all terms of $\mathbf{FC}_p$ with $p = [20, 30, 50, 70, 90]$ are needed in the final fit. This suggests that formation time alone is not enough for our model to remove the correlation with δ', but actually tracking the different formation times at which different percentages of the final halo mass were assembled leads to more accurate models.

Our model suggests that the assembly history dependence is itself a function of the final halo mass. In order to explore this, we write the polynomial fits to each of the $\mathbf{FC}_p$ terms from (2), and their associated coefficients, C_p^i:

$$P^3(p) = C_p^0 + C_p^1\mathbf{FC}_p + C_p^2\mathbf{FC}_p^2 + C_p^3\mathbf{FC}_p^3 \tag{5.31}$$

where $p = [20, 30, 50, 70, 90]$. We define the residual δ_p as the leftover residual once we have removed contributions from all terms from groups (1) and (2), i.e.:

$$\delta_p = y - P^3(z = 0) - \sum_p P^3(p) \tag{5.32}$$

where $p = [20, 30, 50, 70, 90]$. We note that δ_p is defined in standardised space, with positive δ_p corresponding to a model underprediction and negative δ_p a model overprediction.

Fig. 5.13 shows where galaxies are in the $\mathbf{FC}_{50}$ vs δ_p plane, where $\mathbf{FC}_{50}$ (in standardised units) corresponds to the redshift when 50% of the mass of a halo has

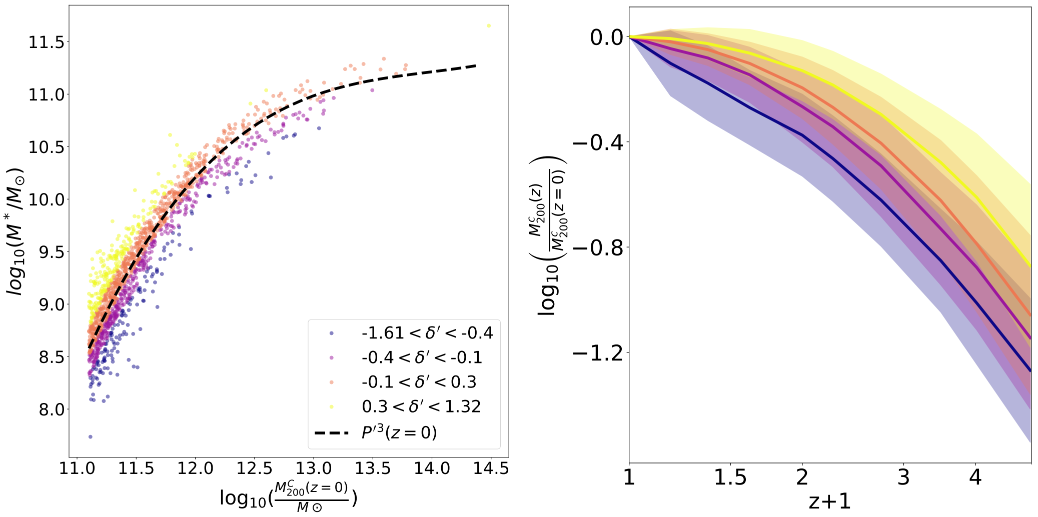
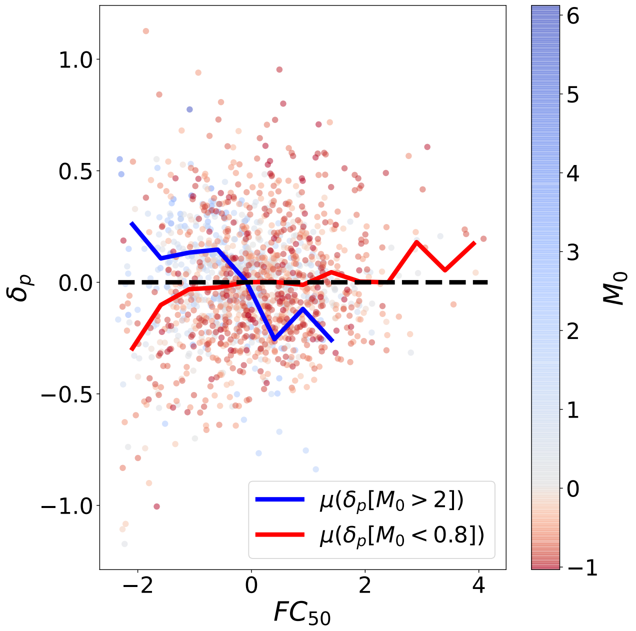


Figure 5.12: **Left panel:** The stellar mass – halo mass relation for galaxies in our holdout set. The black dashed line shows the polynomial $P'^3(z=0)$ from Eq. 5.29 for the formation criterion model. This line describes the trend well but there is some scatter around it. The colour coding split the galaxy sample by their residual δ' from Eq. 5.30 into four bins, with the δ' range indicated by the key. **Right panel:** Evolution of the halo mass for each residual δ' bin, as defined in the left panel, as function of redshift. The solid lines represent the mean of the logarithm of the halo mass ratios of Eq. 5.22 for all galaxies in each δ' bin, with the same colour scheme as in the left panel. The shaded contours indicate the corresponding standard deviation on the mean. Galaxies with the more negative δ' residuals reside in halos that recently assembled their final halo mass, while galaxies with the more positive δ' residuals reside in halos that primarily assembled their halo mass at an earlier stage of their evolution.

been formed. The blue and red solid lines show the average δ_p for very massive and very small halos respectively. When $\mathbf{FC}_{50}$ is negative (i.e. smaller redshifts than the average, i.e. at later times), galaxies living in massive host halos tend to be overpredicted by the model (as shown by the blue line being above zero) and galaxies living in small halos tend to be underpredicted (as shown by the red line being below zero). This shows why terms of the form $\mathbf{FC}_p \times M_0$, corresponding to coefficients in group (3) improve our model. The fact that the model selected terms of the form $\mathbf{FC}_p \times M_0$ suggests that it is not enough to model a linear relationship between stellar mass and formation time (or in our case formation criteria), but that this relation needs to be corrected by a factor that is dependent on the final halo mass. Assembly bias suggests that the stellar mass of galaxies



,

Figure 5.13: Relation between $\mathbf{FC}_{50}$ and the residual δ_p of Eq. 5.32 for all galaxies in our holdout set. $\mathbf{FC}_{50}$, which is in standardised units, maps to the redshift at which a halo acquired half of its mass. The galaxies are colour coded by M_0 (Eq. 5.21). The blue solid line shows the mean residual, $\mu(\delta_p)$, for galaxies in very massive halos, i.e. halos where $M_0 > 2$. Those halos are more that 2 standard deviations more massive than the mean. The red solid lines shows the mean residual, $\mu(\delta_p)$, for galaxies living in halos with very low mass ($M_0 < 0.8$). The blue and red lines have slopes of opposite sign, which is reflected in the presence of terms from group (3) in the solution (see Section 5.5.2). The plot shows that the strength of assembly bias is correlated with the final halo mass.

depends on formation history, our model also suggests that this dependency is in turn dependent on the final halo mass.

5.5.3 Stellar mass distribution and galaxy clustering of centrals

We have shown the models capability to reproduce the stellar mass of individual galaxies from the EAGLE simulation. We now discuss our models accuracy at reproducing other statistics from EAGLE such as the distribution of galaxy masses through the stellar mass function (SMF), and the clustering of the galaxies via 2-point correlation functions.

We consider the six realisations of the formation criterion model presented in

Fig. 5.9 as a way of providing some uncertainty on the best fit model predictions. Throughout this section any model comparison with EAGLE relates to comparisons with central galaxies in EAGLE, as our model only make predictions for such galaxies.

Furlong et al. (2015) shows that the SMF of the EAGLE hydrodynamical simulation at redshift zero agrees reasonably well with the one observed from SDSS (Li and White, 2009) and GAMA (Baldry et al., 2012). The red dashed line of Fig. 5.14 shows the central galaxy stellar mass function obtained from the stellar masses in our EAGLE data set. The red shaded region is an estimate of the error due to Poisson noise within the EAGLE sample and is computed with the bootstrap method (Efron, 1979).

The blue lines in Fig. 5.14 are the SMFs computed using the stellar masses predicted by each of our models. The predictions are so similar that it is difficult to differentiate between them, especially in the top panel. The bottom panel of Fig. 5.14 shows that the model SMFs are within 12% of the input EAGLE SMF over most of the mass range. At stellar masses above $\log_{10}(M^*/M_{\odot}) = 11.0$ the agreement of the models SMF worsens. This is likely due to the relatively small number of galaxies at this mass range in our sample (90 out of 9521). One of the many issues of including a comparatively small sample of galaxies is that the methodology has little incentive to fit them accurately as their contribution to the goodness of fit estimations is small. One possible way to improve this is to weigh their contribution more heavily than the one from galaxies in a lower mass ranges, this possibility will be explored in future iterations of this work. The scatter in the mass function between different models is smaller than the bootstrap error (shown as the shaded area), which suggests that the difference between the SMF of EAGLE and that of our model is not due to random sampling effects. There are notable deviations at $\log_{10}(M^*/M_{\odot}) = 9.0$ and $\log_{10}(M^*/M_{\odot}) = 10.5$. The disagreement at $\log_{10}(M^*/M_{\odot}) = 9.0$ is likely to be caused by selection effects, as we include a cut in halo mass which can have an effect in our model predictions at those lower stellar masses. At $\log_{10}(M^*/M_\odot) = 10.5$ the remaining residuals of the model are systematically larger and asymmetric, with the offset possibly correlated with other terms not included in our methodology. These parameters could be either other halo mass properties that we have not characterised, higher-order correlations of our input parameters, or the stochastic nature of baryonic processes. For example, feedback from super massive black holes has a highly non-linear effect on the stellar mass, either by affecting it directly, or through its influence on the baryon density inside halos (Bower et al., 2017; Martizzi et al., 2012). Whatever the cause, characterizing these asymmetric residuals remains a challenging but important problem.

As a result of the asymmetric scatter, we find that the SMFs predicted by the simpler models, M_0^3 and M_0^3 & $\mathbf{FC}_{50}$ from section 5.5.1, have only minor deviations from those predicted by the full model (the formation criterion model shown in Fig. 5.14). Although the more complex models predict more accurately the median stellar mass, all the models assume that the residuals are symmetric around this value: i.e., while the errors of a more complex model are smaller, they are not more symmetrical around their mean value. An improved treatment will have to characterise the spread of points as well as predicting a median of the relation.

The EAGLE hydrodynamical simulation has been shown to accurately reproduce the observed two point correlation function of galaxies from $1h^{-1}$Mpc and up to $6h^{-1}$Mpc (Artale et al., 2017). In order to test how well our model reproduces the correlation function of EAGLE galaxies, we divide the galaxies in each of our models into four stellar mass bins. We then compute the two point correlation function of galaxies in each mass bin. This is done by assigning to each model galaxy the same co-moving coordinates as that of the centre of its host halo.

Fig. 5.15 shows how the correlation functions of our models split by predicted model stellar mass compares with those obtained from the EAGLE simulation, split by the actual galaxy stellar mass. Each colour corresponds to a different mass bin, with each of our six models and for each stellar mass bin shown as solid faint lines. As with Fig. 5.14, the shaded areas show the bootstrap error estimate on the actual

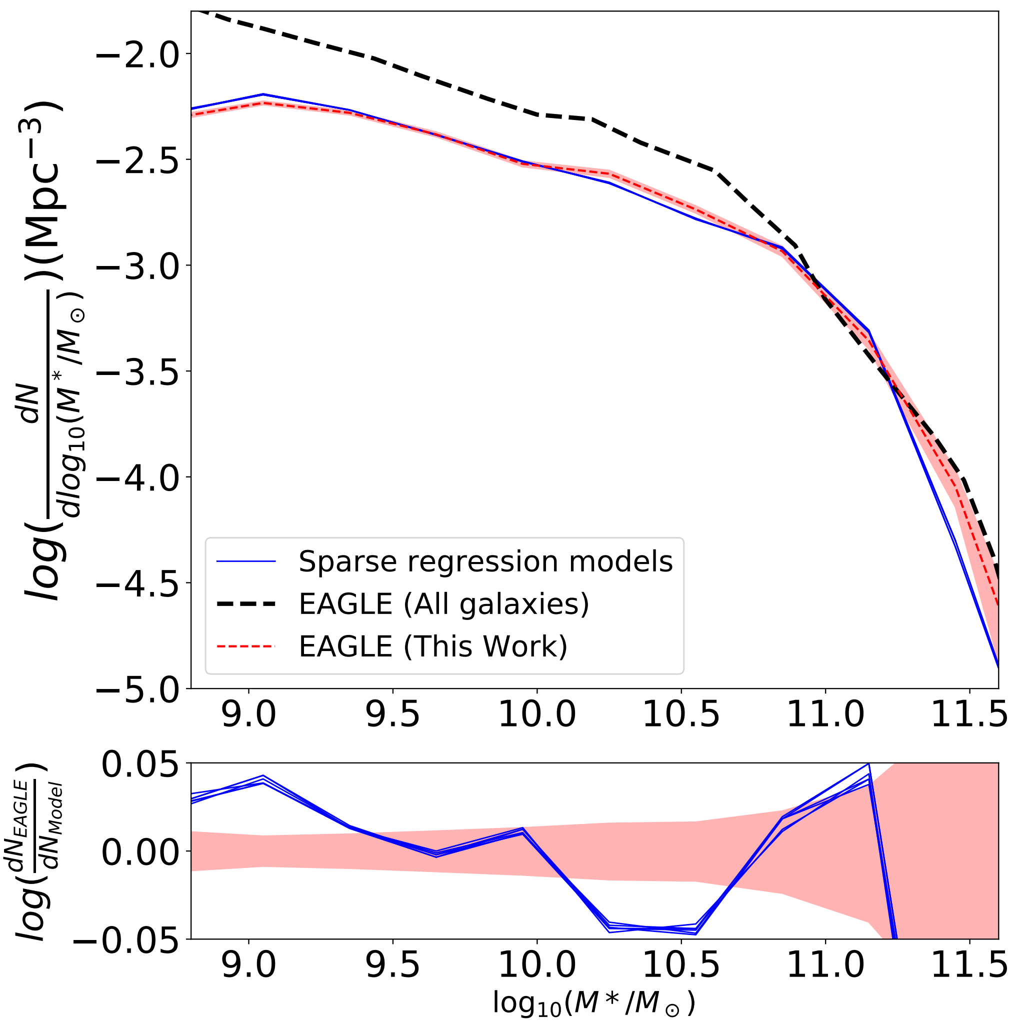


Figure 5.14: **Top panel:** the SMF of EAGLE galaxies used in this analysis (i.e. centrals) is shown in red. The SMF predicted by the six models built with our methodology is shown in blue. The red shaded area show the bootstrap errors on the SMF. For comparison, the SMF of all EAGLE galaxies is shown in black: this sample includes both centrals and satellites and does not include any halo mass cut. **Bottom panel:** the ratio of predicted to actual SMFs, indicating that our models result in SMF estimates which are within 12% of the input data on the stellar mass scales where the input data have good statistics.

EAGLE clustering. The bootstrap method is done on a galaxy basis, which is still adequate in this case as we are not trying to quantify the impact of sample (or cosmic) variance: the models use the same set of DM halos as the EAGLE data, with only the stellar mass of their host galaxies possibly differing. The correlations functions from each of our six models and for each stellar mass bin are shown as solid faint lines. Fig. 5.15 shows that our correlation functions agree within errors with the ones from EAGLE, which suggests that our models assign galaxy masses in a way that is sufficiently accurate to reproduce the stellar mass clustering of central galaxies up to $10h^{-1}$Mpc. It is also noticeable that the scatter on the correlation functions from our methodology is smaller that the one from bootstrap errors. Hence to be able to differentiate between the models a significantly larger simulation volume would be needed.

Hydrodynamical N-body simulations that model both the dark matter and the baryonic component of the universe are computationally expensive. This limits the volumes in which they can be computed to a few $(100\text{Mpc})^3$. Our models are informed by the physical processes relating the stellar mass of a galaxy and its host DM halo. Therefore, by populating DM-only simulations in larger volumes, our models could provide new tests of the hydrodynamical physics on larger scales than the ones permitted by direct comparisons with hydrodynamical simulations. The fact that we can reproduce accurately with our models both the stellar mass and the correlation functions of EAGLE, suggests that this approach is promising for populating DM only simulations.

5.6 Discussion and Conclusions

There is a well-known correlation between the stellar mass of a galaxy and the dark matter of its host halo (SMHM relation). However, this relation has significant scatter, which suggests that other properties are significant at determining the stellar mass of a halo. The halo mass evolution history and the specific angular

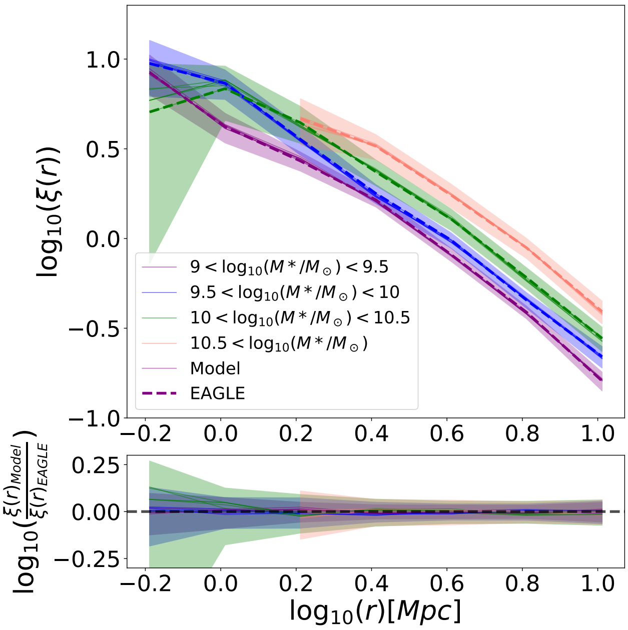


Figure 5.15: **Top panel:** correlation functions of EAGLE galaxies split into four stellar mass bins (coloured dashed lines as per key) compared to the clustering computed with our 6 models (i.e. 6 thin solid lines for each stellar mass bin). Bootstrap errors are shown on the EAGLE correlation functions. **Bottom panel:** the ratio of the predicted to the actual galaxy clustering for each stellar mass bin (same colour coding as in the upper panel). This indicates that our models result in galaxy clustering estimates split by stellar mass that agree well within the bootstrap errors with the actual clustering of EAGLE galaxies.

momentum have both been proposed to be correlated with this residuals.

We use a sparse regression methodology to model the governing equations relating the stellar mass of central galaxies to the properties of their host dark matter halos. This method builds accurate and explainable models without needing much physical knowledge of the processes that determine the stellar mass of a galaxy from the halo properties of its host. In sparse regression methods, the lack of physical knowledge is substituted by large numbers of free parameters, where each parameter models different behaviours of the dark matter halo properties. A LASSO algorithm is used to optimise solutions. This method heavily penalises the number of surviving parameters so that as few as possible are selected without losing accuracy. Here we have modified the form of the LASSO algorithm to be more efficient when combined with a gradient descent minimiser. This is achieved by including a regularisation term that smooths out discontinuities in the gradient that are present in standard LASSO when parameters are close to zero. This smoothing is characterised by a parameter ϵ that limits how close to zero coefficients need to get before being discarded by the algorithm. We also modify the method by which the minimiser decides which path to follow in such a way that we find performance gains in large dimensional spaces.

The size of the penalty is determined by the parameter λ, which is optimised using a k-fold methodology with $k = 10$. We use the one-standard-error rule to select a value of λ that is larger than the best-fit and therefore builds a slightly less accurate model with fewer free parameters and therefore with more explainability.

The data that we use to build our models with comes from the EAGLE simulation. However, we emphasise that this method should be able to be calibrated against any simulation with similar results. We use a sample of 9521 central galaxies from the 100 cMpc box EAGLE suite of hydrodynamical simulations. The dark matter properties are read from a DM only simulation with the same initial conditions as our hydrodynamical simulation. The simulations are matched with each other in such a way that a pair is found for 99% of the DM halos.

We build four different models that differ by the independent parameters chosen to model the galaxy stellar mass. In the first instance, we consider two distinct model setups: (i) the mass ratio model uses the ratio between the mass of a halo at a redshift z and that at $z = 0$ to parametrise the mass history of the host halo; (ii) the formation criterion model uses the redshift at which a halo formed a specific percentage of its mass. For both models we include all linear, quadratic and cubic correlations of our independent variables as free parameters of the fits. Then we consider two additional models by extending the two previous models to include parameters related to the specific angular momentum ($\vec{\mathbf{j}}$) history of the halos. More specifically, we consider parameters that characterise both the magnitude and the direction of the specific angular momentum vector, and vary the radius of the DM halo over which to measure the magnitude of $\vec{\mathbf{j}}$. Due to computational restrictions, we include only linear terms of the free parameters related to $\vec{\mathbf{j}}$.

The computational time of our minimisation is correlated with the value of ϵ: a very large value would result in a very fast computational time, but it would be hard also to distinguish useful parameters from those that should be discarded. In Fig. 5.7 we show that a value of $\epsilon = 1 \times 10^{-3}$ selects the same coefficients as slower and more accurate runs without being too computationally expensive. Some input parameters are correlated with each other, for example, the mass ratio (M_z/M_0) at a given redshift and that at a neighbouring redshift slice. In principle, our answers could be susceptible to the starting point of the minimiser; however, we show in Figs. 5.8 that neither the explainability nor the accuracy of the model changes significantly between runs with different starting points. We show in Fig. 5.9 that models trained on different subsets of the same data arrive at equivalent models.

Our algorithm did not select any angular momentum parameters for either model that included specific angular momentum parameters. In fact, all the differences between these two models and their equivalent ones without angular momentum parameters are consistent with variations in our methodology. This suggests that any correlation between the linear terms of the angular momentum of a host halo and the residual of the SMHM relation is the consequence of correlations between the mass history of the halo and the history of its angular momentum. Given that model the formation criterion model is slightly simpler than the mass ratio model, we conclude that the formation criteria parameters, $\mathbf{FC}_p$, are slightly more efficient at summarizing the halo mass evolution information than the mass ratios (M_z/M_0).

The formation criterion model is more accurate, although more complex, than models that include only halo mass terms, or models that also include a linear dependence on a single formation time. The improvement is, however, modest. Including a single linear formation time explains 32% of the residual variance, while the full models improves this by a further 10%. If greater simplicity is required, this can be achieved (at the expense of accuracy) by increasing the penalty hyperparameter, λ. The resulting model prefers to select terms that more closely characterise the formation history of the halo rather than terms the mix formation time and halo mass, however.

A subset of our surviving terms can be combined into a polynomial of M_0 and is therefore a model of the SMHM relation. Other subsets of surviving terms can be combined into polynomials of either $\mathbf{FC}_p$ or M_z/M_0 (depending on the parametrisation of the halo mass evolution history) and therefore model the assembly bias. Terms of the shape $M_0 \times \mathbf{FC_p}$ (or $M_0 \times M_z/M_0$) add a significant correction to very small or very large halos. Our models suggest that a single formation time is not enough to model the variation in the SMHM relation, and that a better approach is to include the times at which different percentages of the mass have been formed. This is reflected in our model by the similar contribution of terms of the form $\mathbf{FC}_p$ for all p in $p = [20, 30, 50, 70, 90]$. Our model also suggests that the relation between the stellar mass and the formation times is not the same for all galaxies, but it depends on the halo mass at $z = 0$.

We have shown how the stellar mass function (SMF) of our model compares to that of EAGLE central galaxies. They agree well within the bootstrap errors at most stellar mass values, except around $\log_{10}(M^*/M_\odot) = 9.0$ and $\log_{10}(M^*/M_\odot) = 10.5$. The difference at lower stellar mass could be explained by selection effects given that our model includes a cut on halo mass that could affect the prediction of the lower stellar masses. At $\log_{10}(M^*/M_\odot) = 10.5$ on the other hand, the differences between the values predicted by our model and EAGLE are not symmetric around the mean. This suggests that the remaining residual of our model might be correlated with variables that have not been explored by our model. This could be either higher-order correlations of our current variables, DM variables that we have not considered yet, or the stochastic effects of the baryon physics affecting the stellar mass of the galaxy. These will be studied in further extensions of the model. We have also shown that the correlation function of EAGLE galaxies split by stellar mass is preserved in our models within the quoted bootstrap errors at all scales considered.

The fact that we can reproduce both the stellar mass and the correlation function of EAGLE accurately suggests that this method could be used to populate DM only simulations in larger volumes in a way that preserves these statistics. Our models are informed by the physical process that relates the stellar mass of a galaxy with the evolutionary and present properties of its host DM halo. Therefore DM only simulations that are populated using our methodology can provide tests of this physics on volumes where hydrodynamical simulations are prohibitively expensive to run. So far, however, our method has only been applied to central galaxies. Satellite galaxies in general have a weaker SMHM relation than halos. This is a consequence of satellites being subjected to processes like tidal stripping and heating. These processes modify the mass of subhalos and the galaxies they contain, meaning that the stellar mass of a satellite halo is different from what one would expect when comparing it with halos that were not stripped. By adapting our methodology to account for the more complex evolution of the satellite halo mass, for example by adding maximum progenitor masses to our list of variables, it should be possible to model the stellar mass of satellite galaxies as well. However, running our methodology with satellite galaxies would require to use a larger set of free parameters and a larger data set, as there are many more satellite galaxies than central galaxies. Therefore we should explore methods to optimise our minimisation without losing reliability. One approach could be to use methods like principal component analysis to transform free parameters into a parameter space where they are uncorrelated. However, this might transform free parameters into inputs that are harder to interpret and might reduce the explainability of our results. These ideas will be explored in the following chapter.

Chapter 6

A sparse regression approach for populating dark matter halos and subhalos with galaxies

6.1 Introduction

Within the Λ-CDM paradigm (e.g. Planck Collaboration et al., 2014), an expanding universe filled with particles that interact only through gravity can be accurately modeled using N-body simulations (e.g. Springel et al., 2005). Because of advances in computational methods, such simulations can track the formation of galaxy-scale dark matter halos within volumes approaching the size of the observable universe. However, these simulations do not include the baryonic component that leads to the formation of stars and galaxies. Hydrodynamical simulations that include baryons need to deal with complicated cooling and feedback processes and are strongly influenced by events happening at scales much smaller than the size of the simulation grid, this makes them significantly more expensive to run and limits their volume to about $1\,\mathrm{Gpc}^3$ (e.g. Springel et al., 2017). There is, therefore, an incentive for a hybrid approach, in which one uses hydrodynamic simulations to learn the relation between dark matter and baryonic tracers, and then uses these relations to populate N-body mock catalogs on larger volumes.

In chapter 5 we present a novel methodology that uses Sparse Regression Methods (SRM; Tibshirani, 1996; Hastie et al., 2015) to model the relations between the stellar mass of a galaxy and its host halo in the Evolution and Assembly of Galaxies and their Environments (EAGLE, Schaye et al., 2015; Crain et al., 2015; McAlpine et al., 2016) 100 Mpc hydrodynamical simulation. SRM are a set of machine learning algorithms designed to identify the parameters that better describe a dependent variable, then discard the remaining unnecessary ones. Recently they have been suggested as the appropriate framework to extract the equation of states of a physical system from collected data and with minimal knowledge of the physics of the system (Brunton et al., 2016).

In chapter 5 we were interested in developing and testing the methodology in a simple scenario without going into some of the more complicated challenges that populating a realistic N-body mock accurately would require. With that in mind, we tested our methodology on central galaxies (the main galaxy within each dark matter halo) only as they have monotonic growths with time which makes them easier to model. In this chapter, we test our methodology including satellite galaxies. Satellite galaxies (and their associated dark matter subhalos) are created when a smaller dark matter halo is accreted by a larger one. This is a common process in the Λ-CDM model. As they orbit within the larger halo, satellite galaxies (and their remnant dark matter subhalos) undergo a much more diverse range of physical processes than their central galaxy counterparts. Unlike the main dark matter halo, which undergoes monotonic mass growth, the remnants of smaller accreted halos may decay with time (e.g. Bower and Balogh, 2004; van den Bosch et al., 2018) as they lose mass due to processes such as tidal stripping and heating (Lynden-Bell, 1967; Merritt, 1983; Hayashi et al., 2003; Green and van den Bosch, 2019). Moreover, the satellite galaxies residing inside these remnant halos are subject to 'environmental' processes that remove cold gas and suppress the accretion of more material (Gunn and Gott, 1972; Vollmer et al., 2001; Larson et al., 1980; Bahé and McCarthy, 2015; Correa et al., 2019). As a result, star formation in satellite galaxies is significantly suppressed compared to central galaxies and we expect less stellar mass growth.

In EAGLE, the differentiation between central halos and subhalos is done by the SUBFIND algorithm (Springel et al., 2001). Within each sub-halo, the algorithm identifies the self-bound overdensities and classifies them as independent subhalos. The subhalo with the lowest potential energy is classified as the central halo and assigned any diffuse mass that has not already been associated with a sub-halo. This distinction is made separately at each output time and is not a fundamental differentiation, but dependent on the details of the algorithm. In some cases, this leads to anomalous behaviour, in particular inconsistent classifications of the same subhalo at different redshift slices (e.g. Behroozi et al., 2015). It is, therefore, desirable to use a methodology that does not make a fundamental distinction between central and satellite galaxies when modelling the stellar mass, but rather to use the same approach based on the overall halo mass history.

In this chapter, we also use a smaller cut in the halo mass of the galaxies host halos that we used last chapter in our central galaxy sample, reducing it from $M = 10^{11.1} M_{\odot}$ to $M = 10^{10.6} M_{\odot}$. This allows us to identify low mass halos which contain relatively large galaxies (with stellar masses greater than $10^{9} M_{\odot}$). This is a particularly important consideration for satellite galaxies, if we are to generate a stellar-mass complete catalog.

Other works have used machine learning algorithms to model the relationship between the halo and stellar properties inside a hydrodynamical simulation (e.g. Kamdar et al., 2016; Agarwal et al., 2018). Their models accurately reproduce several statistics of the original simulation. However, given that these types of models generate 'black box' answers it might be complicated to modify them to reproduce statistics from observations instead. Studies of this sort have also been done on the EAGLE simulation by Lovell et al. (2021) which uses random forests to learn the stellar mass properties inside both the EAGLE and C-EAGLE simulations and uses them to populate the P-Millennium N-body simulation (Baugh et al., 2019). Moster et al. (2021) uses a neural network approach that rewards the algorithm for reproducing observed statistics of a survey (like correlation functions and stellar mass functions) instead of properties of individual galaxies. This circumvents the problem of differences in statistics between the hydrodynamical simulation used to calibrate the model and those from an observational survey, at the cost of not requiring accuracy in the predictions of the individual values of galaxy properties. Given that our model is an equation of state with a set of input parameters fitted by the model, it is in principle possible to extract the best advantages of both approaches, extracting the important physical parameters by comparison to the simulation, but optimising the coefficients of these terms to reproduce the statistics of an observational data set.

This chapter is organised as follows. Section 6.2 introduces the data set that we use and any enhancements to the model that we have made to handle the more complex data-set. In particular, Section 6.2.1 explains the details of the bijective match between our hydrodynamical simulation and a EAGLE dark matter (DM) only simulation. Section 6.2.2 and Section 6.2.3 describe the methodology used to extract our training data set from the EAGLE DM only simulation as well as the new parameterisation of the model and the new weighting scheme adopted. The results from our different models are shown and analysed in Section 6.3. In particular, we show that a single model can be built to describe both central and satellite galaxies, avoiding the need to make a distinction beyond quantifying their halo mass histories. Our conclusions and thoughts on the potential of the current methodology are discussed in section 6.4.

6.2 Methodology

This chapter considers the same sparse regression method and uses the same simulations as in chapter 5: the EAGLE hydrodynamical and DM only simulations built in a 100 comoving Mpc box (section 5.3.1). This section summarizes the small differences in the method used to generate the models presented in this chapter compared to that of chapter 5. These differences are made with the objective of including satellite galaxies and applying a smaller mass cut to the data. Among these changes we include:

- a new re-run of our matching methodology (see section 6.2.1), which so far has only been done on central galaxies.
- a re-definition of our dependent variables, so that they can efficiently parametrize the growth of subhalos (see section 6.2.2).
- the addition of a new weighting scheme that incentivizes the method to fit the rare and massive galaxies equally well as the significantly more numerous and smaller galaxies (see section 6.2.3).

6.2.1 Matching

The goal of this work is to develop a fitting function that allows the mass of a galaxy to be estimated from knowledge of its DM halo's formation history only. Since DM halos in hydrodynamical simulations are affected by baryonic processes that might alter their density profile (Schaller et al., 2015; Martizzi et al., 2012; Navarro et al., 1996), or other properties like the shape of the halo (Katz and Gunn, 1991; Bryan et al., 2013), it is important that we match the halos in the hydrodynamical simulation with the same halos in a dark matter only simulation (with the identical cosmology and initial conditions). By making a one-to-one matching between the DM-only simulation and the hydrodynamical one, the properties of the DM-only simulations can be used as the input variables of the model (the column vectors $\vec{z'}_i$ of section 5.2.1) while the stellar mass is measured in the full-physics hydrodynamical simulation. The matching is done by following the procedure of Schaller et al. (2015). To summarise, we look at the 50 most bound DM particles of each halo or subhalo in the hydrodynamical simulation: if a halo or subhalo of the DM-only simulation contains at least half of these particles, then they are matched. The matching is done for all halos above $M_{\text{total}} > 2 \times 10^9$ and both halos need to be above this value to be matched, where M_{total} is the summed mass of all particles assigned to the halo or subhalo.

6.2.2 Halo Selection and Input Parameterisation

We begin our selection of halos by tracing the evolution of the halo mass at 19 redshift slices between $z = 0$ and $z = 4$. This initial selection is based on $M_{\text{total}}(z)$, the total mass of the particles associated to the halo or sub-halo by the SUBFIND algorithm. These trajectories summarise the evolution of the galaxy's host halo mass as a function of redshift and give us a relation between halo mass and time for each galaxy. In order to ensure that the trajectory is not overly affected by the algorithm used in the selection process, we use a Gaussian kernel with a σ of one redshift slice to smooth this evolution history. Since halo masses can increase as well as decrease (for satellite galaxies in particular), we base our halo selection on the maximum value of $M_{\text{total}}(z)$ in the smoothed trajectory. The success rate of the matching is dependent on the halo mass, with more massive halos being more likely to be matched. We found that $\mathbf{Max}(M_{\text{total}}(z)) = 10^{10.66} M_\odot$, corresponds to the threshold at which more than 95 per cent of halos are successfully matched, and we define this threshold as the cutout value of our sample. In order to avoid missing data, we discard all galaxies that do not have a well-defined main progenitor at all redshift slices up to $z = 4$. For $\mathbf{Max}(M_{\text{total}}(z)) > 10^{10.66} M_\odot$, this cut is unimportant and we keep 99.6 per cent of all galaxies. Our final sample consists of a total of 34,654 galaxies, of which 9,967 live inside subhalos, and 25,492 inside central halos[*].

[*] Note that satellite galaxies are generally smaller than centrals. The halo mass cut used in this work discards disproportionately more satellite halos than centrals and the final sample contains a larger amount of central galaxies.

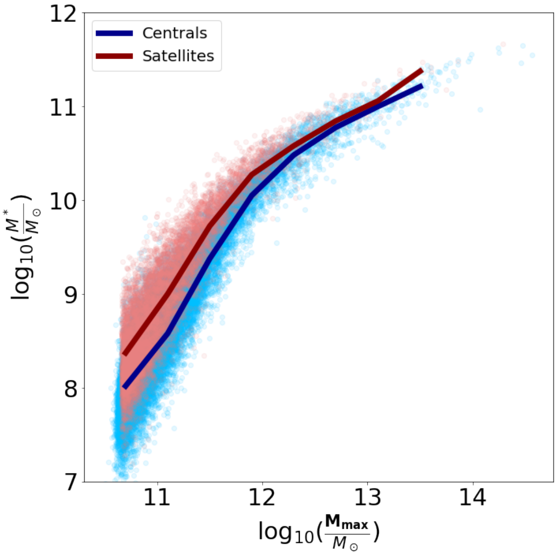


Figure 6.1: Distribution of the halos (blue dots) and subhalos (red dots) in our sample in the $\mathbf{M_{max}}$-M^* space, where $\mathbf{M_{max}}$ is the largest halo mass the halos main progenitors reached. The solid lines show the median value of the distributions. The plot shows that at a fixed $\mathbf{M_{max}}$ the median galaxy mass of a satellite galaxy is larger than that of a central galaxy.

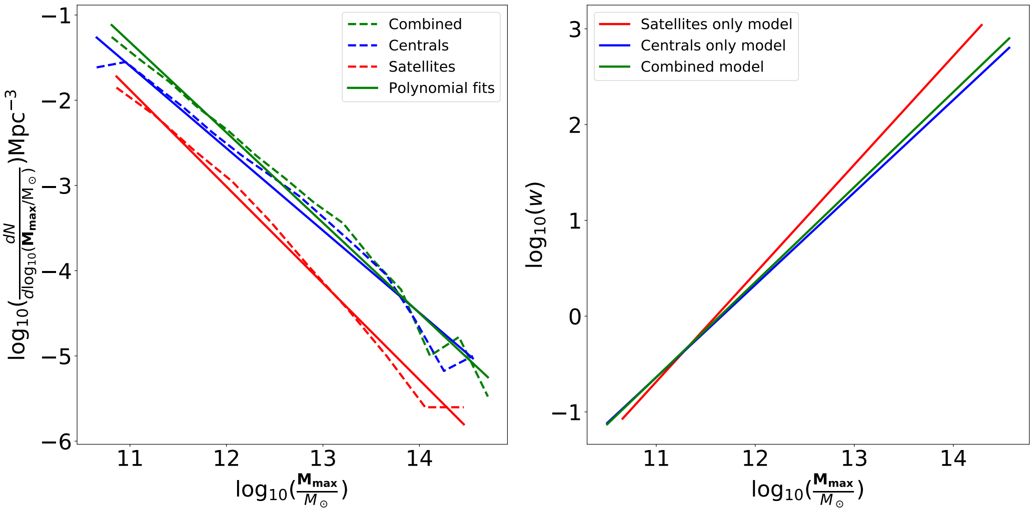


Figure 6.2: **Left**: halo mass function (HMF), as characterised by $\mathbf{M_{max}}$ (dashed lines) for each of our three samples: satellite galaxies (red), central galaxies (blue), and a combined sample with all centrals and satellites (green). The solid lines show the linear HMF fits used by the weighting scheme. **Right**: Values of the weights w as a function of $\mathbf{M_{max}}$.

We now describe the input parameters used in this work, which are the values of the column vectors $\vec{z'}_i$ of section 5.2.1. Although our selection of halos is based on $M_{\mathrm{total}}(z)$, the conventional way of defining the mass of a central halo is using M^c_{200}. For central galaxies, $M_{\mathrm{total}}(z)$ is derived from the FOF mass, which can deviate significantly from the intuitive notion of a halo. It is standard practice, therefore, to use the spherical M^c_{200} mass (i.e. the mass around the centre of a halo inside the radius at which the density is 200 times larger than the critical density of the universe). In practice we compared both mass definitions, and found that models based on M^c_{200} gave more accurate stellar mass predictions. At $z = 0$, M^c_{200} is defined for central halos only, therefore, the mass of satellite halos is computed using M_{total}. Of course, subhalo main progenitors are labeled as central halos before they merged with the main halo, therefore at some point in their history subhalo main progenitors should have a well-defined value of M^c_{200}. In practice, however, the redshift at which a subhalo reaches its maximum and the one at which it becomes a central galaxy is not always the same, and some subhalo main progenitors can oscillate between being considered central halos and subhalos throughout several redshift slices due to either complicated orbits or degeneracies in the classification scheme. With this in mind and to avoid any discontinuity due to changes in the mass definition, all halos classified as a subhalo at $z = 0$ have their evolution tracked with $M_{\mathrm{total}}(z)$ at all redshifts, in the same way all halos classified as centrals at $z = 0$ are tracked with $M^c_{200}(z)$. We use the interpolation scheme developed in 5 to ensure the halo masses of central galaxies are not affected by inconsistent classification between snapshots.

Since the satellite halo mass cannot be expected to grow monotonically with decreasing redshift, a more important parameter for each galaxy is instead its maximum halo mass. In the rest of the paper, we refer to this as $\mathbf{M_{max}}$:

$$\mathbf{M_{max}} = \begin{cases} \mathbf{Max}(M_{\mathrm{total}}(z)) & \text{for satellite galaxies} \\ \mathbf{Max}(M^c_{200}(z)) & \text{for central galaxies} \end{cases}$$

Central galaxies tend to grow monotonically with time, and $\mathbf{M_{max}}$ is correlated with the stellar mass through the $z = 0$ stellar mass - halo mass (SMHM) relation. In satellite galaxies, however, $\mathbf{M_{max}}$ corresponds to the redshift at which their host halo merges and becomes the subhalo of a larger system. Once a halo merges the mass of the halo declines due to tidal processes. We can expect, therefore, that the galaxy mass at $z = 0$ will be well correlated with the mass of the host halo before merging. Fig. 6.1 shows the distribution of galaxies in the $\mathbf{M_{max}}$-M^* space. Note that a small subset of central galaxies (blue dots) have values below $\mathbf{M_{max}}(z) = 10^{10.66} M_\odot$. This is due to galaxies in this subset satisfying $[\mathbf{M_{max}} = \mathbf{Max}(M^c_{200}) < 10^{10.66} M_\odot < \mathbf{Max}(M_{\rm total})]$. We also note that the median stellar mass of satellite galaxies is larger than that of centrals at fixed $\mathbf{M_{max}}$, i.e. for a fixed $\mathbf{M_{max}}$ satellite galaxies are more massive. The offset in the SMHM relation for satellites and centrals is driven by two competing processes. On the one hand, satellites may undergo a strong suppression of their star formation as they orbit within the main halo due to the combined effects of ram-pressure stripping (the removal of the interstellar medium of the galaxy by ram pressure) and 'strangulation' (the absence of gas infall onto the satellite). On the other hand, while the halo mass of the central continues to grow with cosmic time, the satellite reaches its peak mass and **Max** becomes frozen thereafter. The net offset is determined by whether the halo mass or the stellar mass grow fastest in the central galaxies, and by whether satellite galaxies are able to continue to grow in stellar mass after they are accreted (Behroozi et al., 2019). Because the effect on the stellar mass growth tends to be delayed compared to the effect on the halo, satellite galaxies tend to have larger stellar mass than their central counterparts.

In 5, we tested different parameterisations and concluded that parameters that measure the SMHM relation and the halo growth trajectory are the most useful for modelling the stellar mass at $z = 0$. We also found no improvement in our models when adding parameters correlated with the angular momentum evolution of the halo. The best model that we found used $\log_{10}(M_{200}(z = 0))$ as the input parameter that traced the SMHM relation, as well as a set of formation criteria parameters

	$\log_{10} M^*/M_\odot$	$\mathbf{lgM_{max}}$	$\mathbf{FC}'_{20}$	$\mathbf{FC}'_{30}$	$\mathbf{FC}'_{50}$	$\mathbf{FC}'_{70}$	$\mathbf{FC}'_{90}$
μ	8.760	11.12	3.096	2.547	1.743	1.151	0.6259
σ	0.8002	0.4586	0.7978	0.8427	0.7376	0.6053	0.4923

Table 6.1: Normalisation parameters used for the stellar mass and the DM halo variables. These parameters are for the model that mixes central and satellite galaxies. The μ and σ rows correspond to the mean and standard deviation of the variables respectively and are used in Eq. 5.3 to standardise the range of the variables considered.

$\mathbf{FC}_p$ that model the assembly history, where $\mathbf{FC}_p$ is the redshift by which a central galaxy has assembled $p = [20, 30, 50, 70, 90]$ per cent of its current mass. In order to accommodate satellite galaxies, we substitute the input parameter $\log_{10}(M_{200}(z = 0))$ with $\mathbf{M_{max}}$ and we define the dimensionless parameter

$$\mathbf{lgM_{max}} = \log_{10}(\mathbf{M_{max}}/M_\odot) \tag{6.1}$$

and redefine the formation criteria parameters $\mathbf{FC}_p$ as follows: We find the redshift z_i at which a halo or subhalo reaches M_{max}, we then look at the evolutionary history of the halo from $z = 4$ up until z_i, and find the redshift $(z_i < \mathbf{FC}_p < z = 4)$ at which the halo assembles the percentage p of M_{max}. Note that if z is such that $M(z) = \mathbf{M_{max}}$, then $z < \mathbf{FC}_{90} < \mathbf{FC}_{70}$. This parameterisation is almost equivalent to the one used in 5 when only considering central galaxies as in this case $M(z = 0) \sim \mathbf{M_{max}}$. As a check, we ran our methodology on the data set of 5 with the new parameterisation, the resulting model is comparable to the original one in accuracy and simplicity. In total we use six independent variables in our methodology $[\mathbf{lgM_{max}}, \mathbf{FC}_{20}, \mathbf{FC}_{30}, \mathbf{FC}_{50}, \mathbf{FC}_{70}, \mathbf{FC}_{90}]$. Each of these parameters is transformed to the standardised space defined by equation 5.3. Since we consider cubic combinations of these parameters this leads to a model with up to $D = 84$ parameters.

To test the differences between modelling satellite and central galaxies separately and modelling them together with a single model, we run three models independently of each other:

- A model that only contains central galaxies, with $N = 25,492$ data points.

- A model that only contains satellite galaxies, with $N = 9,967$ data points.
- A model that combines central and satellite galaxies and fits them all at the same time, with $N = 34,654$ data points.

6.2.3 Weighting the Cost Function

In 5, we used a simple χ^2 measure to assess the quality of the model's prediction of the data (i.e. χ^2 is the cost function). In the CDM paradigm, however, smaller halos are always much more numerous than massive ones. As a consequence, such methodology would have a stronger incentive to fit numerous smaller halos more accurately at the expense of a less accurate fit to less numerous massive ones. In 5, we concluded that our methodology became more inaccurate for galaxies larger than $\log_{10}(M^*/M_\odot) > 11.0$ (see discussion of Fig. 5.14) due to a relatively small fraction of galaxies above the threshold (90 out of ∼9,500). Given that in this iteration of the work we reduced the cutout value of galaxies even further, we now have a larger number of smaller galaxies making the issue even more problematic. A good solution to this problem is to assign a weight w_i' to each halo. This weight determines how much of an incentive the code will have to fit a particular halo mass correctly. If the weight w_i' is larger for galaxies in larger halos, then by modifying Eq. **??** to include a normalised weight w_i as below, we will give a larger importance to the rarer larger halos:

$$\chi_w^2 = \sum_{\alpha=0}^{N} \frac{w_\alpha (M_\alpha^* - M_{p\alpha}^*(C))^2}{N^2}. \tag{6.2}$$

To compute the weight of a halo we first look at the halo mass function (HMF) as a function of $\mathbf{lgM_{max}}$. These are shown as dashed lines in the left panel of Fig. 6.2. To avoid noisy weights from having a small number of objects in the more massive bins, we make use, in this plane, of a linear fit, shown as the solid lines, to the HMFs. Referring to the linear fits as $\mathrm{fl}(\mathbf{lgM_{max}})$, the weight of a halo is defined as:

$$w_\alpha' = \frac{10^{\mathrm{fl}(\mu)}}{10^{\mathrm{fl}(\mathbf{lgM_{max}}_\alpha)}} \tag{6.3}$$

Coefficient	Centrals	Satellites	Combined
Constant	0.132	-	0.118
$\mathbf{lgM_{max}}$	1.22	1.19	1.19
$\mathbf{lgM_{max}}^2$	-0.158	-0.169	-0.168
$\mathbf{lgM_{max}}^3$	0.00740	0.00833	0.0102
$\mathbf{FC}_{70}$	0.0959	-	0.128
$\mathbf{FC}_{90}$	-	-	0.0681
$\mathbf{FC}^2_{30}$	-	-	0.0227
$\mathbf{FC}^3_{20}$	0.0139	-	-
$\mathbf{FC}^3_{30}$	-	-	0.00394
$\mathbf{FC}^3_{50}$	0.0116	0.0467	-0.0150
$\mathbf{FC}^3_{70}$	-	0.00991	-
$\mathbf{lgM_{max}} \times \mathbf{FC}_{20}$	-	-	-0.0389
$\mathbf{lgM_{max}} \times \mathbf{FC}^2_{20}$	-	0.00196	-
$\mathbf{lgM_{max}} \times \mathbf{FC}^2_{30}$	-	0.00281	-
$\mathbf{lgM_{max}} \times \mathbf{FC}^2_{90}$	-	-	-0.00474
$\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{20}$	-0.00515	-	0.00306
$\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{50}$	-	-0.0267	-
$\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{70}$	-	0.0226	-
$\mathbf{FC}^2_{20} \times \mathbf{FC}_{50}$	-	-	0.0161
$\mathbf{FC}^2_{30} \times \mathbf{FC}_{20}$	0.00392	-	-
$\mathbf{FC}^2_{50} \times \mathbf{FC}_{20}$	-	-	0.0327

Table 6.2: Parameters and their respective values for the surviving coefficients of the three models. Note that the parameters presented here are in the standardised space defined by Eq. 5.3. Parameters are shown to three significant figures, which we find are enough to make the RMSE accurate to four significant figures.

where μ is the median value of $\mathbf{lgM_{max}}$. As a final step, we normalise the weights of a sample as follows

$$w_\alpha = \frac{N \times w'_\alpha}{\sum_{\alpha=1}^{N}(w'_\alpha)}. \tag{6.4}$$

The right panel of Fig. 6.2 shows the weights of a galaxy as a function of $\mathbf{lgM_{max}}$ for each of our three samples. Note that in the combined model, the weighting scheme does not distinguish between central and satellite galaxies.

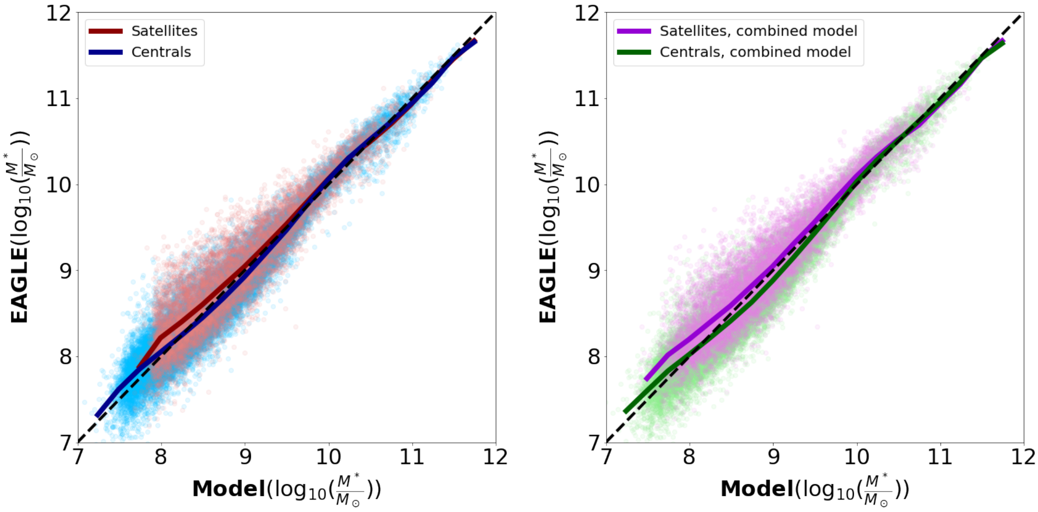


Figure 6.3: Comparison between the stellar mass of galaxies in EAGLE and the stellar masses predicted by the model. The coloured dots show the actual values of the individual galaxies and the solid lines are the mean values. The black dashed line corresponds to the one-to-one line. The left plot corresponds to the separate models where central (blue) and satellites (red) are run independently. The right plot is for the combined model, with the results separated into centrals (green) and satellites (purple).

6.3 Results

6.3.1 Comparing input and predicted stellar masses

We now present the results of each of our three models. The surviving coefficients and their respective values are shown in table 6.2. In order to extract a fitting function that can be applied directly to the input variables, one first needs to transform the input data using equation 5.3 (which requires the mean and standard deviation values of the dependent variables given in table 6.1[*].)

Fig. 6.3 shows a comparison between the stellar mass predicted by the models and its actual value in EAGLE. The left panel shows the results when running the central and satellite models independently, while the right panel shows the results when they are fitted simultaneously on the combined model. The plot shows that

[*] Note that the resulting stellar mass also needs to be converted from standardised units, and we have therefore included the stellar mass parameters in this table as well

the mean closely follows the one to one line (black dashed line) for all models above $\log_{10}(M^{\star}/M_{\odot}) = 9.5$, but is also evident that both the independent and the combined model tend to overpredict slightly the stellar mass of small satellites with $\log_{10}(M^{\star}/M_{\odot}) < 9.5$. This may be associated with the strongly asymmetric scatter in this region of the plot. Overall, however, the plot is very encouraging and shows that the properties of satellites, as well as centrals, can be accurately predicted by the SRM approach. This is an important pre-requisite for constructing accurate mock catalogs from dark matter simulations. We will explore the performance of the models in more detail below.

A subsidiary aim, however, is to determine whether it was necessary to explicitly distinguish between central and satellite galaxies in constructing the model. We test this by comparing the model in which central and satellite galaxies are fitted separately with one that combines all galaxies into one single model and relies on the methodology to distinguish between satellite and central galaxies only on the basis of their different formation histories. Note that in this model, there is no binary distinction between satellites and central and the model varies its predictions continuously. Removing this binary condition should result in an algorithm that is less dependent on the details of the SUBFIND algorithm, making results simpler to interpret.

In order to compare the accuracy of the models, we use the mean square error (RMSE) statistic defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\alpha=1}^{N}(M_{p\alpha} - M^{*}_{\alpha})^2}{N}}. \tag{6.5}$$

We find a RMSE of the central galaxies when modelled with our combined model of RMSE = 0.208, this is essentially indistinguishable from the error of the model ran with central galaxies only (RMSE = 0.210). We find similar results for satellite galaxies, where their individual model has a RMSE of 0.286, compared with a RMSE of 0.271 in the combined model. We can also look at all centrals and satellites of the individual models used together which have a RMSE of 0.232.

This is comparable with the combined model that has an RMSE of 0.229. This indicates that the individual models and the combined model have comparable accuracies. Note that the combined model ends up with 14 terms while modelling satellites and centrals individually requires 9 for each model (hence 18 terms in total).

The accuracy of our models is comparable to what other methodologies have found. For example, Kamdar et al. (2016) find an RMSE of 0.35 when using extremely randomized trees to model the stellar mass in the Ilustrius simulation (Nelson et al., 2015). While our RMSE is smaller than theirs, they fit all galaxies in their simulation down to a stellar mass of $\sim 10^6 M_\odot$, and therefore larger errors are expected, as their galaxies are around two orders of magnitude smaller than those we consider. This makes it hard to compare the models directly but our RMSE is at least of similar amplitude to theirs. Given that it is difficult to know how much larger their errors should be when compared to ours, we can compare a statistic like the R^2 statistic defined in equation 5.26 that normalises for the standard deviation of the data set (although we should remember the warning in section 5.5 about considering R^2 estimates as they can be easily biased). Our combined model has a R^2 of 0.917 which is directly comparable with the values of 0.916 and 0.909 found by Kamdar et al. (2016) and Agarwal et al. (2018) [*] respectively.

A significant appeal of the SRM approach is that the surviving terms in Table 6.2 have a physical interpretation. Following the discussion in 5, we note that there are four types of surviving parameters:

- A constant, or normalisation, term.
- Terms that only include $\mathbf{lgM_{max}}$ and no formation criteria parameter: these terms model the underlying relation between $M_{\max}$ and M^*. For central galaxies they should correspond to a model of the SMHM relation.

[*] Agarwal et al. (2018) uses a random forest algorithm to reproduce galaxy properties of halos inside the MUFASA simulation (Davé et al., 2016). The code uses halo properties like halo mass, environment, spin, and recent growth history, and models the following baryonic properties: stellar mass, star formation rate, metallicity and neutral and molecular hydrogen mass.

- Terms that only include formation criteria parameters (e.g. $\mathbf{FC}_{50}$ and higher order combinations): these terms quantify the growth history of the halo, capturing scatter in the relation.
- Terms that are a product of halo mass, $\mathbf{lgM_{max}}$, and formation criteria parameters: these terms model the dependence of the assembly history on the final halo mass.

Comparing between the models, we see, firstly, that the $\mathbf{lgM_{max}}$ coefficients are similar between all three models. This reflects the similar underlying shape of the $\mathbf{M_{max}}$ and M^* relation. There is no constant term required for the satellite model, which implies that the mean halo mass corresponds to the mean satellite stellar mass[*], reflecting the offset between the central satellite relation seen in Fig. 6.1. In the combined model, central and satellite galaxies are treated on an equal footing and this offset is captured by the more complex dependence on formation time parameters. For example, the model depends strongly on the $\mathbf{FC}_{90}$ parameter: this quantifies the time at which the halo ceases to grow and becomes a satellite, so its presence in the model is expected. There are also more subtle interplays between terms in this model, with positive and negative coefficients appearing with similar magnitude. It is interesting that the model also has a strong cross dependence term between halo mass and formation time (i.e. $\mathbf{lgM_{max}} \times \mathbf{FC}_{20}$), suggesting that the satellite-central offset is strongly dependent on both satellite mass and its early formation history.

It is interesting to compare the central galaxy model with the one presented in 5. It is important to stress that we do not expect identical models, since we have broadened the range of masses considered and weighted the cost function to emphasise the importance of predicting stellar masses well over the full halo mass range. Interestingly, this change has resulted in a simpler model. The number of free parameters selected by the algorithm has been reduced from 17 to 9. Most

[*] Of course a constant would appear in this relation if we were to convert to the original, non-normalised, input space

noticeably, the only surviving linear term of any of the formation criteria parameters is $\mathbf{FC}_{70}$ (the model of 5 had 3 linear terms $\mathbf{FC}_{30}$, $\mathbf{FC}_{50}$, $\mathbf{FC}_{90}$). Similarly $\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{20}$ is the only term relating the dependence of the assembly history on the final halo mass, while previously we had five terms.

One difficulty becomes apparent when comparing the models in further detail, however. Because of the significant correlation between parameters: models of almost equivalent accuracy and complexity can vary in the final parameters chosen if these parameters are correlated. For example, the central model includes strong dependencies on terms in $\mathbf{FC}^3_{20}, \mathbf{FC}^3_{50}$ and $\mathbf{FC}_{70}$ while the satellite galaxies depend on $\mathbf{FC}^3_{50}$ and $\mathbf{FC}^3_{70}$. The satellite model also includes terms in $\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{50}$ and $\mathbf{lgM_{max}}^2 \times \mathbf{FC}_{70}$ with almost equal but opposite coefficients. It is difficult to decide on the significance of these differences because of the underlying correlations. Future investigations could consider methods like principal component analysis (Jolliffe, 2005) to transform our input functions into a parameter space where they are uncorrelated. However, this would lose the benefit of having a simple physical interpretation of the input parameters and the resulting model.

6.3.2 Predicting clustering and the stellar mass function

Figure 6.4 shows how the stellar mass function (SMF) of our models split by galaxy type (total in green, centrals in blue and satellites in red) compares to those from EAGLE. The plot shows the SMF of the combined model (solid lines), of the individual models (dotted lines) and of the EAGLE data (shaded area), the shading indicating a bootstrap error estimate to account for sampling effects (Efron, 1979). The different model SMFs are all comparable, as they seem to agree all similarly well with the EAGLE SMFs, with the agreement worsening somewhat for masses around $\log_{10}(M^\star/M_\odot) = 10.5$, as identified already in 5. As we suggested in that chapter, one possible reason behind this disagreement is the stochasticity of certain baryonic processes which might affect the stellar mass, for example the feedback from supermassive black holes (Bower et al., 2017; Martizzi et al., 2012). While this

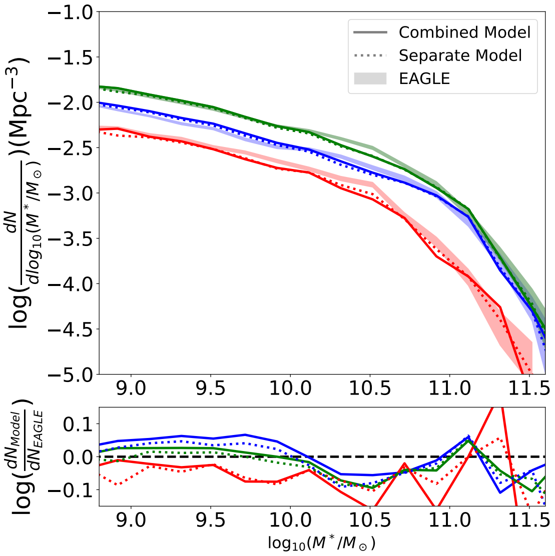


Figure 6.4: The galaxy SMF of EAGLE, represented as the shaded areas, compared to the galaxy SMF of our models, shown as solid (combined model) and dotted (individual models) lines. The green line corresponds to combined samples of all galaxies, and the red and blue to the satellite and central subsets respectively. The shaded region shows the bootstrap error on the EAGLE SMF estimate.

would be a challenging phenomenon to predict using input parameters from a DM only simulation, it should be possible to develop in a future work SRM models that estimate the stochastic scatter in predicted quantities as well as a central value.

As mentioned in section 4.3, Moster et al. (2021) uses a neural network approach called Galaxy-net that asks the algorithm to reproduce observed properties such as the SMF, the local clustering, the cosmic and specific SFRs, and the quenched fractions. Their model is fitted to the EMERGE data described in Moster et al. (2018). Their approach focuses on fitting these observed properties accurately instead of trying to fit the individual values of galaxy properties. Our predicted stellar mass function (SMF) agrees well with the one predicted by Galaxy-Net (SMF$_{GN}$) with $1 - \mathrm{SMF}_{GN}/SMF < 0.1$ for stellar masses bellow $10^{11} M_\odot$ and $1 - \mathrm{SMF}_{GN}/SMF \sim 0.2$ for larger stellar masses. Interestingly, Galaxy-Net also finds the same disagreement that we do between the stellar mass modeled and their target stellar mass from EMERGE at around $10^{10.5} M_\odot$ (see figure 6.4). The stellar mass function predicted by Galaxy-Net seems to have similar accuracy to ours with

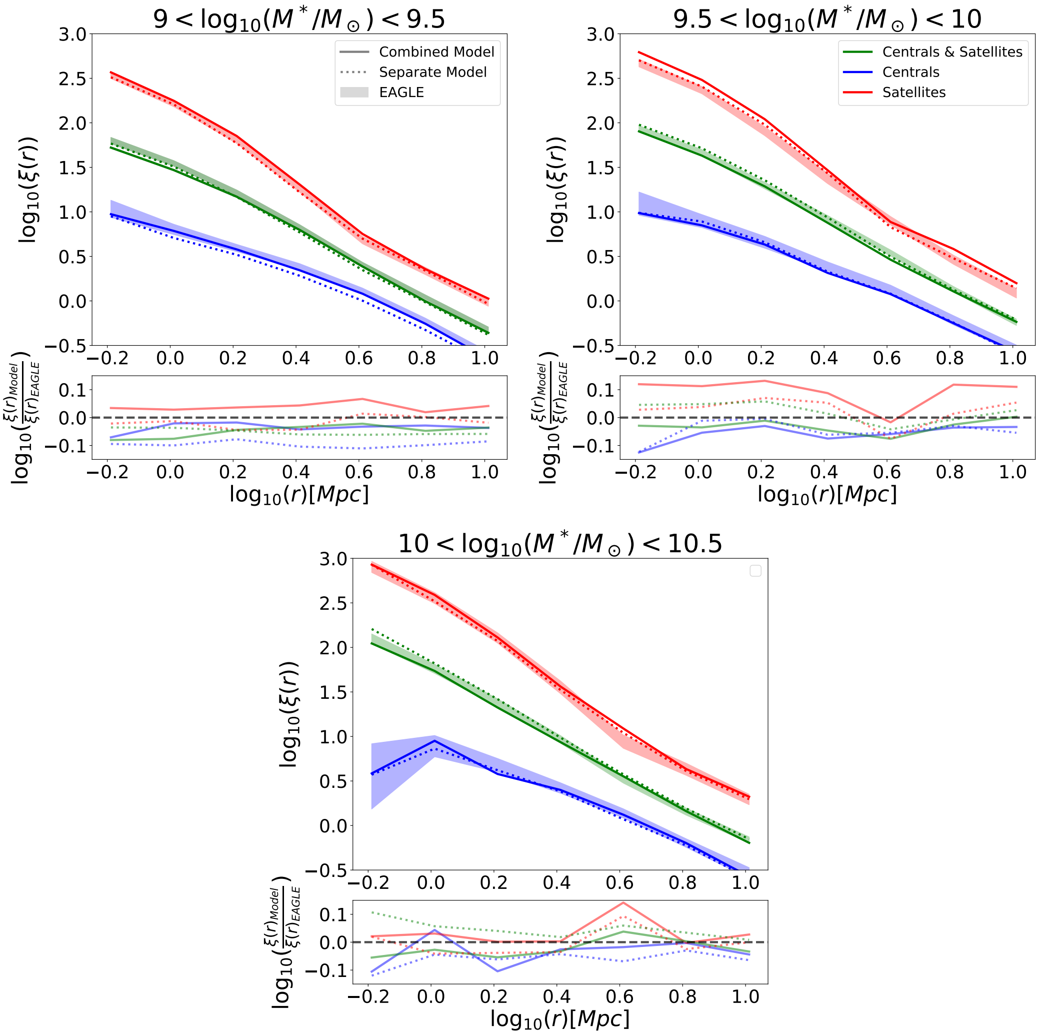


Figure 6.5: Correlation function of EAGLE galaxies split into different mass bins (as indicated in the title of the panels). The solid line shows the correlation function of all galaxies in our combined model. The dotted lines show the same but for our individual models. The shaded area corresponds to the correlation function of the corresponding EAGLE galaxies including bootstrap errors.

the differences between their predicted stellar mass and their target stellar mass being below 0.2 everywhere except around $10^{10.5} M_\odot$, this is comparable to what we found in the second panel of figure 6.4. At $10^{10.5} M_\odot$ the difference between our combined model rises to ~ 0.2 and the difference between Galaxy-Net and EMERGE to ~ 0.35.

Figure 6.5 shows the galaxy correlation functions of our models, split by the predicted galaxy stellar mass. The figure also includes the correlation function of galaxies when split by their stellar mass in the EAGLE simulation. As with figure 6.4 we have included an estimate of the error due to sampling effects using the bootstrap method. The correlation function of both models with central and satellites galaxies (green lines) agrees with the EAGLE correlation function within errors. For central galaxies (blue lines), the agreement with EAGLE is generally good, however, the separate model (dashed line) shows a slightly lower clustering amplitude for galaxies within $9 < \log_{10} M^*/M_\odot < 9.5$. Similarly, satellite galaxies (red colour) are slightly more strongly clustered within the $9.5 < \log_{10} M^*/M_\odot < 10$ bin in the combined model (solid line) compared to EAGLE. Interestingly, the discrepancies in the correlation functions are much less evident when satellites and central galaxies are modelled together. This is encouraging since the binary distinction between central and satellite galaxies is unnecessary in order to model the overall correlation function.

One of the advantages of our methodology over standard machine learning techniques is the fact that our solution is expressed as a simple equation of state with 14 free parameters fitted by the algorithm. This is important as the model can be modified to fit other data sets different from EAGLE, which is a requirement needed to use our method to populate DM-only simulations that would be used to analyze observational data set. This would not be achievable by a more complex 'black box' model.

6.4 Conclusions

In 5 we used a sparse regression methodology to fit the stellar mass of central galaxies as a function of properties of their host halos. In this chapter we expand our study to cover a wider halo mass range, and to model the properties of satellite galaxies. The distinction between central and satellite galaxies relies on identifying subhalos as self-bound substructures within larger halos, for example by using the SUBFIND algorithm. This classification is uncertain and may be inconsistent for the same subhalos in adjacent snapshots outputs. We therefore explored whether we need to make a fundamental distinction between halos and subhalos. With this in mind, we use the maximum mass that a halo has ever reached during its evolution, denoted $\mathbf{Max}(M_{\rm total}(z))$ and use this in place of the final (sub)halo mass at $z = 0$. Given that central galaxies grow monotonically then $\mathbf{Max}(M_{\rm total}(z)) \sim M(z = 0)$ and this results in little change. In subhalos, however, it corresponds to the mass of their main progenitor before merging with their central halo. In order to quantify the prior growth history of the halo, we define a set of formation criteria parameters, that measure the redshift at which a halo has formed a given percentage of its mass and before it reaches $\mathbf{Max}(M_{\rm total}(z))$.

Our data is taken from the EAGLE hydrodynamical simulation. In order to avoid selection biases when predicting stellar mass, we use a bijective matching between the EAGLE hydrodynamical simulation and a DM-only simulation with the same cosmology and initial conditions. We select all galaxies that have a halo mass larger than $\mathbf{Max}(M_{\rm total}(z)/M_{\odot}) > 10^{10.66}$, this value corresponds to the threshold at which our matching methodology successfully matches more than 95 percent of all galaxies. We use a total of 34,654 galaxies, 9,967 of them live inside subhalos and 25,492 inside central halos. Because our sample has significantly increased the fraction of low-mass galaxies considered compared to the previous work, we weight residuals according to stellar mass, giving a larger incentive to the model to accurately fit less represented galaxy masses.

The SMF of our models agrees well with that of EAGLE at all stellar masses except at $\log_{10}(M^{*}/M_{\odot}) = 10.5$ where our models tend to slightly under-predict the amount of galaxies when compared with the EAGLE simulation. This could be related to the stochasticity of baryonic processes that might alter the stellar mass of a galaxy, this would be hard to predict using parameters from a DM-only simulation. We also calculate the correlation function of our models when separated by their predicted stellar mass, and find this to also agree well with the EAGLE correlation functions. The model that combines central and satellite galaxies has comparable accuracy to the models in which centrals and satellites are treated independently, while using an overall smaller number of model parameters. This suggests that a binary classification is unnecessary and the stellar mass of both galaxy types can be predicted by suitable measurement of their halo mass history.

The SRM approach can be viewed as a machine learning algorithm. It can accurately model the stellar masses of EAGLE from the data itself and without requiring previous knowledge of physics behind the system. At the same time, the approach results in a prediction algorithm that is explicit and simple (compared with the solutions of other machine learning techniques), and the terms that are retained give physical insight into the important processes at work.

We have seen that the correlation function and the stellar mass function of our models agree well with the EAGLE data set. This is encouraging as both of these EAGLE statistics have been positively compared with observational data. For example, Furlong et al. (2015) has shown that the EAGLE SMF at $z = 0$ agrees reasonably well with the ones observed by the SDSS (Li and White, 2009) and GAMA (Baldry et al., 2012) surveys. Similarly (Artale et al., 2017) shows that the EAGLE correlation function reproduces observations accurately at $1h^{-1}$Mpc and up to $6h^{-1}$Mpc. This suggests that our methodology is a promising approach to populate N-body simulations with galaxies of the correct stellar mass and spatial distribution. However, the ultimate goal is to generate mock catalogs that provide an even better representation of the observed universe. An attractive idea is therefore to iterate on the coefficients of the terms selected by comparison to EAGLE, creating an even closer match to target observations. This would retain the same physical processes, but accept that their relative importance might differ between the true universe and the Eagle simulation. This is an interesting possibility that we will explore in more depth in future work.

Chapter 7

Summary and Future Work

In this thesis, I have presented two independent pieces of research that explore the nature of the dark sector of the universe. In the first research project, I performed a full shape analysis of the eBOSS LRG sample, where new constraints on specific RSD parameters were obtained. The second project proposes a novel approach to study the relation between galaxies and their host halos using sparse regression algorithms.

These two projects were the bulk of the work that I did during my PhD in Durham[†]. In this final chapter, I provide a summary of each of the two projects and briefly describe the next steps that I intend to explore in the future.

7.1 Summary of RSD modelling

Measurements of standard rulers (like the BAO scale) and standard candles (like SN-Ia analysis) are not enough to constrain a cosmological model, as an appropriate

[†] As part of the Centre for Doctoral Training in Data Intensive Science (CDT-DIS) program, I worked on two projects outside of academia that were not related to astronomy. The first of these projects was an internship of five months in total that I did in the consumer goods company Procter & Gamble (P&G). During this time we developed a sparse regression methodology that predicts physical properties of laundry powder as a function of the configuration of the manufacturing machine, which then led on to our astronomy focused SRM work presented in chapters 5 and 6. In a second independent project, I worked as part of Durham University's JUNE project (Bullock et al., 2020). We modeled the covid-19 pandemic in the UK using detailed census data to accurately reproduce the population distribution across the country and their interactions. My work whitin this collaboration mostly centred on modelling the different infection progression paths that an individual would follow once infected depending on their age and gender.

selection of the equation of state can result in the same expansion history for different models. Therefore, there is a need for a new independent parameter that can break this degeneracy. One parameter that is commonly used is the growth rate f which in turn is related through the continuity equation to the peculiar velocity of galaxies. One can study these peculiar velocities through their contribution to the redshift that one measures in large-scale structure surveys. When observed in redshift space, this effect causes an anisotropy in the distribution of galaxies which can be studied using two-point correlation function statistics.

A RSD analysis uses an estimate of the 2-point correlation function of a given survey and compares it to the prediction of a theoretical model that predicts the evolution of $\delta(\vec{r})$. On linear scales, the peculiar velocity of the galaxy should be a function of the growth rate f, and therefore theoretical models include it as a free parameter. One can also include parameters that give the model the freedom to explore if the fiducial cosmology assumed in the analysis could be different from the true cosmology of the universe. This can be done using the Alcock-Paczynski parametrisation.

In chapter 3 we present the first full-shape analysis (e.g. Alcock-Paczynski and RSD) done with the eBOSS LRGs sample. The sample includes galaxies from the 14^{th} data release of the survey corresponding to all observations done in the survey's first two years.

We compute the monopole, quadrupole, and hexadecapole of the correlation function of our sample. We use a model that combines Convolution Lagrangian Perturbation Theory with Gaussian Streaming theory to model these multipoles within the ΛCDM cosmology model and assuming Planck 2018 parameters. The model considers four RSD parameters (f, F', F'', σ_{FOG}) and two Alcock-Paczynski parameters (α,ϵ). Using a MCMC algorithm we explore the parameter space to find the regions of high likelihood. The covariance matrix used for the likelihood estimations is computed from a set of 1000 QPM mocks.

When fitting our model to the mean of the N-series high precision mocks we concluded that the best-fit models do not match the amplitude of the hexadecapole (see discussion around figure 3.7). We considered that this needed further exploration. Given that this were the only high-precision mocks with the properties of eBOSS LRGs available, we decided to follow a conservative approach and report the fits done with ξ_0 and ξ_2 as our final result, but include the result from models with ξ_4 for consistency.

The data sample correlation had a very high amplitude in its quadrupole at large scales, which is well outside the 1σ variances of the mocks. At the moment of finishing this study, this feature was not properly understood and it was speculated that it could be a consequence of either a large statistical fluctuation or a systematic effect that was not properly understood. Since then, the final data release of eBOSS has been published, and it did not present this feature (e.g. Bautista et al., 2021). Therefore, a statistical fluctuation is now the preferred explanation.

Our final constraints for our parameters of interest are $f(0.72)\sigma_8(0.72) = 0.454 \pm 0.134$, $D_A(0.72) = (1466.5 \pm 133.2)(r_s/r_s^{fid})$ and $H(0.72) = (105.8 \pm 15.7)(r_s^{fid}/r_s)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. These results are consistent with previous measurements and with a ΛCDM model using the Planck 2018 cosmology.

7.2 Summary of SRM

Hydrodynamical simulations are arguably the best tool to study the relation between galaxies and the underlying DM distribution theoretically. However, hydrodynamical simulations with good resolution are limited in volume to around $\sim (300[\mathrm{Mpc}])^3$.

Surveys designed to study galaxy evolution require mock catalogs of galaxies with a volume of around $\sim (1[\mathrm{Gpc}])^3$. There is therefore an incentive to better understand the relation between galaxies and their host halos inside hydrodynamical simulations and use them to populate halo catalogs of the required size.

The models they produce can be very accurate and reproduce statistics like the stellar mass function or the correlation function of the original hydrodynamical simulation with remarkable accuracy. One of their shortcomings is that the resulting models are complex 'black box' answers that are difficult to modify. This becomes an issue for certain statistics for which the volumes of hydrodynamical simulations are too small to predict accurately, e.g. the large-scale correlation function.

We propose a novel approach where we use sparse regression methods, which are a type of machine learning algorithm, to model an equation that relates the stellar mass with properties of its host DM halo. The resulting equations are linear combinations of a set of functional forms and are therefore explainable and easy to modify.

To build a model without knowledge of the underlying physics, one uses a large number of free parameters, each of which models a different behavior of the halo properties. The optimised value of the parameters is found with a LASSO algorithm, which is designed to penalise the number of relevant coefficients used in such a way that it discards all but the most relevant. This is achieved by minimizing the LASSO function, which combines a penalty term, which incentivises the minimisation to discard unnecessary free parameters, with a goodness of fit estimator. The relative magnitude of the penalty term with respect to the goodness of the fit is regulated by the hyperparameter λ, which is optimised with the k-fold methodology.

The models presented here are built using the EAGLE hydrodynamical simulation 100 Mpc box. To avoid any bias due to alterations of the halo properties, which could be generated by the presence of baryons, we collect the halo data from a DM-only simulation with the same initial conditions and with the same volume as the full hydrodynamical simulation. Then we perform a matching between both simulations and read the baryonic properties from the hydrodynamical simulation.

In the first iteration of our work, we restrict our data set to central galaxies with a relatively large halo mass cut of $M(z = 0) > 10^{11.1}M_{\odot}$. The parameters of the host halos selected for this first interaction included the halo mass at $M(z = 0)$ (so that our model could predict the SMHM relation), the formation criteria parameters FC_p, which measures the redshift at which a halo has assembled a percentage p of its final mass, and parameters that quantify the angular momentum evolution of the host halo.

Our resulting models discarded all angular momentum parameters, which suggests that the observed correlation between the residuals of the SMHM relation and the angular momentum of the host halo is a consequence of a correlation between angular momentum and the mass history of the halo.

We find that most models include terms that relate the halo mass of the galaxy with our formation criteria parameters at different orders. This suggests that the relation between the stellar mass and the history of the halo mass is not the same for all galaxies, but it depends on the halo mass at $z = 0$.

In the second iteration of our work, we included satellite galaxies, which were originally left out due to their host halos having a complicated evolution after merging with their central halo. This is in contrast with central halos that grow monotonically during their evolution. We also reduced our halo mass cut to $M = 10^{10.66}M_{\odot}$, which is the threshold at which our matching methodology has a success rate of 95%.

In this second iteration, we were interested in exploring whether or not we were required to make a fundamental distinction between satellites and central galaxies in our models. With this in mind, we modified our parametrisation: we select the largest mass the halo reached during its evolution ($\mathbf{Max}(M(z))$) as a substitute parameter for $M(z = 0)$, and at the same time the formation criteria parameters $\mathbf{FC}_p$ are modified so that they represent the redshift before the halo reached its maximum where a percetnage p of $\mathbf{Max}(M(z))$ has been assembled. We note that given that central galaxies grow monotonically, we have typically for central galaxies $\mathbf{Max}(M(z)) \sim M(z = 0)$.

Also given that our halo mass cut is lower in this second iteration the fraction of low-mass galaxies has increased considerably. We needed to incentivise the model to accurately fit the larger galaxies, which is achieved with the introduction of stellar mass-dependent weights.

We find that the accuracy of modelling both central and satellites galaxies separately is comparable to when we combine them into one sample. This suggests that this binary classification is unnecessary for our methodology.

Our resulting models can accurately reproduce the stellar mass function of EAGLE. This is true for all stellar-mass values, except for a bump at $\log_{10}(M*/M_{\odot}) = 10.5$. We suspect that this might be related to the stochasticity of baryonic processes altering the stellar mass. We also find good agreement between the correlation function of EAGLE and the one predicted by our model when these are split into stellar mass bins.

7.3 Future work

We have mentioned that one of the current issues with machine learning methods is that they can generate somewhat inflexible models. While some methods can model the properties of galaxies very accurately, they are fitted to a specific hydrodynamical simulation. The statistics of the data sets generated by these models, e.g the correlation function or the stellar mass function, will be similar to those of the original hydrodynamical simulation. This becomes an issue when the original simulations does not have a large enough volume to reproduce these statistics accurately. For example, it has been shown that simulations underestimate the correlation function, even at scales much smaller than the simulation length Gelb and Bertschinger (1994); Bagla and Prasad (2006); Bagla and Ray (2005) and therefore the relatively small volumes of hydrodynamical simulations result in underpredicted correlation function at large scales.

For some machine learning models it is hard to modify them to fit the galaxy properties in such a way that they would reproduce statistics of other data sets that are different from those in the original hydrodynamical simulation. However, our sparse regression models are polynomial equations, and the algorithm decides the subset of polynomial terms it needs and the value of the coefficients associated with each polynomial. We suggest that, while there is physics determining which parameters are kept, the exact values of the coefficients are tailored to EAGLE. The shape and the amplitude of statistics like the SMF and the 2-pt correlation function depend on the values of these coefficients.

We propose to run, e.g., a MCMC chain over the parameters, intending to optimise the models of a set of target statistics like the SMF and the 2-pt correlation function of a given survey. This algorithm should be limited by a set of priors that preserves the general shape of the original model. Building this methodology should allow us to use our sparse regression methods to populate galaxy mocks useful for analysing data of upcoming galaxy surveys by ensuring the mock reproduces some key properties of the sample considered. An application to DESI BGS might be of particular interest and rather timely, given that DESI started survey operations in May 2021.

One possible application of our method is to build mocks to be used in galaxy evolution surveys. In general, these mocks can benefit from other galaxy information besides the stellar mass. We are interested in expanding our methodology to include the prediction of other free parameters like the star formation rate, the metallicity, or the luminosity of a galaxy. Other machine learning methods (e.g. Agarwal et al., 2018) have been able to model these properties fairly accurately using halo properties as their only input, so there are encouraging precedents for our project.

One of the current shortcomings of our sparse regression method is the lack of reproducibility of the resulting models: given that there is a strong correlation between several of our functional forms, the method tends to select different subsets of the parameters depending on, e.g., the initial position of the minimiser (see discussion about figure 5.8) or the way one divides the data set into a training and a holdout set (see discussion about figure 5.9). The resulting models are of comparable accuracy and complexity so these differences do not affect the quality of the model. The study of chapter 5 showed that 2/3 of the parameters seem to be chosen consistently, so the reproducibility issue in that analysis was only present in the remaining 1/3 of the parameters.

One approach to deal with this reproducibility issue is to use a method that transforms the parameter space into one where free parameters are not correlated, e.g. using principal component analysis. This method will transform the parameter space into a new one with no correlations between parameters, but the new parameters can be generated with complicated functions of the old ones. This will make it harder to give a physical interpretation to the surviving coefficients. However, the resulting model will still be dependent on a small set of free parameters and therefore it will still be possible to fit it to a desired set of statistics.

Bibliography

Abadi M.G., Moore B. and Bower R.G. 1999, MNRAS, 308(4), 947–954.

Abolfathi B. et al. 2017, preprint (arXiv:1707.09322).

Abolfathi B. et al. 2018, ApJS, 235(2):42.

Agarwal S., Davé R. and Bassett B.A. 2018, MNRAS, 478(3), 3410–3422.

Alam S. et al. 2015, MNRAS, 453(2), 1754–1767.

Alam S. et al. 2017, MNRAS, 470, 2617–2652.

Alcock C. and Paczynski B. 1979, Nature, 281, 358.

Alcock C. et al. 2000, ApJ, 542(1), 281–307.

Allen S.W., Evrard A.E. and Mantz A.B. 2011, ARA&A, 49(1), 409–470.

Alonso D. 2012, preprint (arXiv:1210.1833).

Anderson L. et al. 2014, MNRAS, 441, 24–62.

Arfken G., 1985. Mathematical Methods for Physicists. San Diego, third edition.

Artale M.C. et al. 2017, MNRAS, 470(2), 1771–1787.

Audren B. et al. 2013, JCAP, 1302, 001.

Bagla J.S. and Prasad J. 2006, MNRAS, 370(2), 993–1002.

Bagla J.S. and Ray S. 2005, MNRAS, 358(3), 1076–1082.

Bahé Y.M. and McCarthy I.G. 2015, MNRAS, 447(1), 969–992.

Baker J.G. et al. 2006, ApJ, 653(2), L93–L96.

Baldry I.K. et al. 2012, MNRAS, 421(1), 621–634.

Barnes J. and Hut P. 1986, Nature, 324(6096), 446–449.

Baugh C.M. et al. 2019, MNRAS, 483(4), 4922–4937.

Bautista J.E. et al. 2018, ApJ, 863(1):110.

Bautista J.E. et al. 2021, MNRAS, 500(1), 736–762.

Behroozi P. et al. 2015, MNRAS, 454(3), 3020–3029.

Behroozi P. et al. 2019, MNRAS, 488(3), 3143–3194.

Benson A.J. 2001, MNRAS, 325(3), 1039–1044.

Benson A.J. et al. 2001, MNRAS, 320(2), 261–280.

Berger P. and Stein G. 2019, MNRAS, 482(3), 2861–2871.

Berlind A.A. et al. 2003, ApJ, 593(1), 1–25.

Bernardeau F. et al. 2002, Physics Reports, 367(1-3), 1–248.

Bertone G. and Hooper D. 2018, Reviews of Modern Physics, 90(4).

Bessel F.W. 1844, MNRAS, 6, 136–141.

Bharadwaj S. 1994, ApJ, 428, 419.

Blake C. and Glazebrook K. 2003, ApJ, 594(2), 665–673.

Blanton M.R. and Moustakas J. 2009, ARA&A, 47(1), 159–210.

Blanton M.R. et al. 2017, AJ, 154:28.

Blumenthal G.R. et al. 1984, Nature, 311, 517–525.

Bond J.R. et al 1984, In Audouze J. and Tran Thanh Van J., editors, *Formation and Evolution of Galaxies and Large Structures in the Universe*. page 87.

Booth C.M. and Schaye J. 2009, MNRAS, 398(1), 53–74.

Boselli A. and Gavazzi G. 2006, PASP, 118(842), 517–559.

Bower R.G. and Balogh M.L. 2004, In Mulchaey J.S., Dressler A. and Oemler A., editors, *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution*. page 325 (`arXiv:astro-ph/0306342`).

Bower R.G., McCarthy I.G. and Benson A.J. 2008, MNRAS, 390(4), 1399–1410.

Bower R.G. et al. 2017, MNRAS, 465(1), 32–44.

Breiman L. 2001, Machine Learning, 45, 5–32.

Brunton S.L., Proctor J.L. and Kutz J.N. 2016, Proceedings of the National Academy of Sciences, 113(15), 3932–3937.

Bruzual A. G. 1983, ApJ, 273, 105–127.

Bryan S.E. et al. 2013, MNRAS, 429(4), 3316–3329.

Bullock J. et al. 2020, medRxiv.

Cabré A. and Gaztañaga E. 2009, MNRAS, 393(4), 1183–1208.

Carlson J., Reid B. and White M. 2013, MNRAS, 429, 1674–1685.

Carroll S.M. 1998, Physical Review Letters, 81(15), 3067–3070.

Carroll S.M. 2001, Living Reviews in Relativity, 4(1).

Cen R. and Ostriker J.P. 1992, ApJ, 399, L113.

Chandrasekhar S. 1943, ApJ, 97, 255.

Chaves-Montero J. et al. 2016, MNRAS, 460(3), 3100–3118.

Chuang C.H. et al. 2015, MNRAS, 446(3), 2621–2628.

Chuang C.H. et al. 2015, MNRAS, 452(1), 686–700.

Coc A. and Vangioni E. 2017, International Journal of Modern Physics E, 26(08), 1741002.

Cole S. et al. 1994, MNRAS, 271(4), 781–806.

Cole S., Fisher K.B. and Weinberg D.H. 1995, MNRAS, 275(2), 515–526.

Cole S. et al 2000, In Mazure A., Le Fèvre O. and Le Brun V., editors, *Clustering at High Redshift*. page 109 (`arXiv:astro-ph/9910233`).

Cole S. et al. 2005, MNRAS, 362(2), 505–534.

Coles P. 1993, MNRAS, 262(4), 1065–1075.

Coles P. and Jones B. 1991, MNRAS, 248(1), 1–13.

Colless M. et al. 2001, MNRAS, 328(4), 1039–1063.

Correa C.A., Schaye J. and Trayford J.W. 2019, MNRAS, 484(4), 4401–4412.

Crain R.A. et al. 2015, MNRAS, 450(2), 1937–1961.

Croom S.M. et al. 2001, MNRAS, 322(4), L29–L36.

Dalla Vecchia C. and Schaye J. 2012, MNRAS, 426(1), 140–158.

Dalton G.B. et al. 1997, MNRAS, 289(2), 263–284.

Danovich M. et al. 2015, MNRAS, 449(2), 2087–2111.

Davis M. et al. 1985, ApJ, 292, 371–394.

Davé R., Thompson R. and Hopkins P.F. 2016, Monthly Notices of the Royal Astronomical Society, 462(3), 3265–3284.

Dawson K.S. et al. 2012, AJ, 145(1), 10.

Dawson K.S. et al. 2016, AJ, 151:44.

de Bernardis P. et al. 2000, Nature, 404(6781), 955–959.

de Mattia A. et al. 2021, MNRAS, 501(4), 5616–5645.

DESI Collaboration et al. 2016, preprint (arXiv:1611.00036).

Drinkwater M.J. et al. 2010, MNRAS, 401(3), 1429–1452.

Driver S.P. et al. 2019, The Messenger, 175, 46–49.

du Mas des Bourboux H. et al. 2017, A&A, 608:A130.

Dunkley J. et al. 2005, MNRAS, 356(3), 925–936.

Efron B. 1979, Ann. Statist., 7(1), 1–26.

Efron B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. URL `https://epubs.siam.org/doi/abs/10.1137/1.9781611970319`.

Einasto J. et al. 1984, MNRAS, 206, 529–558.

Eisenstein D.J. et al. 2001, AJ, 122(5), 2267–2280.

Eisenstein D.J. et al. 2005, ApJ, 633(2), 560–574.

Eisenstein D.J. et al. 2011, AJ, 142(3):72.

Erb D.K. 2008, ApJ, 674(1), 151–156.

Fall S.M. and Romanowsky A.J. 2013, ApJ, 769(2), L26.

Fang W. et al. 2008, Phys. Rev. D, 78(10):103509.

Feldman H.A., Kaiser N. and Peacock J.A. 1994, ApJ, 426, 23–37.

Fisher K.B. 1995, ApJ, 448, 494.

Fry J.N. 1984, ApJ, 279, 499–510.

Fukugita M. et al. 1996, AJ, 111, 1748.

Furlong M. et al. 2015, MNRAS, 450(4), 4486–4504.

Gao L. and White S.D.M. 2007, MNRAS: Letters, 377(1), L5–L9.

Gao L., Springel V. and White S.D.M. 2005, MNRAS, 363(1), L66–L70.

Garrison L. et al. 2021, doi:10.13139/OLCF/1811689.

Gelb J.M. and Bertschinger E. 1994, ApJ, 436, 467.

Gil-Marín H. et al. 2018, MNRAS, 477(2), 1604–1638.

Gil-Marín H. et al. 2020, MNRAS, 498(2), 2492–2531.

Gilman D. et al. 2019, MNRAS, 491(4), 6077–6101.

Giocoli C., Tormen G. and Van Den Bosch F.C. 2008, MNRAS, 386(4), 2135–2144.

Golub G.H. 1970, Reinsch, C..

Green S.B. and van den Bosch F.C. 2019, MNRAS, 490(2), 2091–2101.

Gunn J.E. and Gott, J. Richard I. 1972, ApJ, 176, 1.

Gunn J.E. et al. 1998, AJ, 116, 3040–3081.

Gunn J.E. et al. 2006, AJ, 131, 2332–2359.

Hamilton A.J.S. 1992, ApJ, 385, L5.

Hamilton A.J.S. 1998. Linear Redshift Distortions: a Review. In *The Evolving Universe*, pages 185–275.

Hamilton A.J.S., 1998. The Evolving Universe.

Harrison E.R. 1974, ApJ, 191, L51.

Hartlap J., Simon P. and Schneider P. 2007, A&A, 464, 399–404.

Harvey A. 2012, preprint (`arXiv:1211.6338`).

Hastie T., Tibshirani R. and Wainwright M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations.

Hayashi E. et al. 2003, ApJ, 584(2), 541–558.

Helly J.C. et al. 2003, MNRAS, 338(4), 913–925.

Higuchi Y. et al. 2020, MNRAS, 497(1), 52–66.

Høg E. et al. 2000, A&A, 355, L27–L30.

Hopkins A.M., McClure-Griffiths N.M. and Gaensler B.M. 2008, ApJ, 682(1), L13.

Hou J. et al. 2018, MNRAS, 480(2), 2521–2534.

Howlett C. et al. 2015, MNRAS, 449(1), 848–866.

Huchra J.P. and Geller M.J. 1982, ApJ, 257, 423–437.

Hui L. and Bertschinger E. 1996, ApJ, 471(1), 1–12.

Huterer D. et al. 2015, Astroparticle Physics, 63, 23–41.

Icaza-Lizaola M. et al. 2020, MNRAS, 492(3), 4189–4215.

Ivezić Ž. et al. 2019, ApJ, 873(2):111.

Jolliffe I. 2005, Principal Component Analysis. American Cancer Society (https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013192.bsa501). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa501`.

Joyce A. et al. 2015, Phys. Rep., 568, 1–98.

Kaiser N. 1987, MNRAS, 227, 1–21.

Kamdar H.M., Turk M.J. and Brunner R.J. 2016, MNRAS, 457(2), 1162–1179.

Kampakoglou M. and Benson A.J. 2007, MNRAS, 374(3), 775–786.

Katz N. and Gunn J.E. 1991, ApJ, 377, 365.

Kauffmann G., White S.D.M. and Guiderdoni B. 1993, MNRAS, 264(1), 201–218.

Kauffmann G. et al. 1999, MNRAS, 303(1), 188–206.

Kauffmann G. et al. 2004, MNRAS, 353(3), 713–731.

Kembhavi A.K. and Narlikar J.V., 1999. Quasars and Active Galactic Nuclei: An Introduction.

Khandai N. et al. 2015, MNRAS, 450(2), 1349–1374.

Kitaura F.S. and Heß S. 2013, MNRAS: Letters, 435(1), L78–L82.

Klypin A. and Holtzman J. 1997, arXiv e-prints, art. astro-ph/9712217.

Klypin A. et al 2013, Halo Abundance Matching: accuracy and conditions for numerical convergence (arXiv:1310.3740).

Klypin A. et al. 2016, MNRAS, 457(4), 4340–4359.

Knobel C. 2013, An Introduction into the Theory of Cosmological Structure Formation (arXiv:1208.5931).

Knox L., Christensen N. and Skordis C. 2001, ApJ, 563(2), L95–L98.

Lacey C. and Cole S. 1994, MNRAS, 271, 676.

Lacey C.G. et al. 2016, MNRAS, 462(4), 3854–3911.

Lamoreaux S.K. 1997, Phys. Rev. Lett., 78, 5–8.

Landy S.D. and Szalay A.S. 1993, ApJ, 412, 64–71.

Larson R.B., Tinsley B.M. and Caldwell C.N. 1980, ApJ, 237, 692–707.

Laureijs R. et al. 2011, preprint (arXiv:1110.3193).

Lecun Y., Bengio Y. and Hinton G. 2015, Nature, 521(7553), 436–444.

Lewis A., Challinor A. and Lasenby A. 2000, ApJ, 538, 473–476.

Li C. and White S.D.M. 2009, MNRAS, 398(4), 2177–2187.

Lilly S.J. et al. 2007, ApJS, 172(1), 70–85.

Lin Y.T. et al. 2016, ApJ, 819(2), 119.

Linder E.V. 2005, Phys. Rev. D, 72(4):043529.

Lovell C.C. et al 2021, A machine learning approach to mapping baryons onto dark matter halos using the EAGLE and C-EAGLE simulations (`arXiv:2106.04980`).

Lucie-Smith L. et al. 2018, MNRAS, 479(3), 3405–3414.

Lupton R., 1993. Statistics in theory and practice.

Lynden-Bell D. 1967, MNRAS, 136, 101.

Madau P. and Dickinson M. 2014, ARA&A, 52(1), 415–486.

Manera M. et al. 2012, MNRAS, 428(2), 1036–1054.

Manera M. et al. 2015, MNRAS, 447(1), 437–445.

Marri S. and White S.D.M. 2003, MNRAS, 345(2), 561–574.

Martizzi D. et al. 2012, MNRAS, 422(4), 3081–3091.

Mastropietro C. et al. 2005, MNRAS, 364(2), 607–619.

Matsubara T. 2008, Phys. Rev. D, 77(6):063530.

Matsubara T. 2015, Phys. Rev. D, 92(2).

Matthee J. et al. 2016, MNRAS, 465(2), 2381–2396.

McAlpine S. et al. 2016, Astronomy and Computing, 15, 72–89.

Merritt D. 1983, ApJ, 264, 24–48.

Merson A. et al. 2018, MNRAS, 474(1), 177–196.

Metropolis N. et al. 1953, J. Chem. Phys., 21(6), 1087–1092.

Michell J. 1784, Philosophical Transactions of the Royal Society of London Series I, 74, 35–57.

Milgrom M. 1983, ApJ, 270, 365–370.

Milonni P.W. and Eberlein C. 1994, American Journal of Physics, 62(12), 1154–1154.

Mo H., van den Bosch F.C. and White S., 2010. Galaxy Formation and Evolution.

Monaco P., Theuns T. and Taffoni G. 2002, MNRAS, 331(3), 587–608.

Moore B. et al. 1996, Nature, 379(6566), 613–616.

More S. et al. 2010, MNRAS, 410(1), 210–226.

Moster B.P., Naab T. and White S.D.M. 2018, Monthly Notices of the Royal Astronomical Society, 477(2), 1822–1852.

Moster B.P. et al. 2021, MNRAS.

Mueller E.M. et al. 2014, preprint (`arXiv:1408.6248`).

Myers A.D. et al. 2015, ApJS, 221(2):27.

Navarro J.F., Eke V.R. and Frenk C.S. 1996, MNRAS, 283(3), L72–L78.

Nelson D. et al. 2015, Astronomy and Computing, 13, 12–37.

Nelson D. et al. 2015, Astronomy and Computing, 13, 12–37.

Neveux R. et al. 2020, MNRAS, 499(1), 210–229.

Newman J.A. et al. 2013, ApJS, 208(1), 5.

Norberg P. et al. 2009, MNRAS, 396(1), 19–38.

Oliver S.J. et al. 1996, MNRAS, 280(3), 673–688.

Oppenheimer B.D. and Davé R. 2008, MNRAS, 387(2), 577–600.

Peacock J.A. 1997, MNRAS, 284(4), 885–898.

Peacock J.A., 1998. Cosmological Physics.

Peacock J.A., 1999. Cosmological Physics.

Peacock J.A. and Dodds S.J. 1994, MNRAS, 267(4), 1020–1034.

Peacock J.A. et al. 2001, Nature, 410(6825), 169–173.

Peacock J.A. et al 2002, In Metcalfe N. and Shanks T., editors, *A New Era in Cosmology*. page 19 (`arXiv:astro-ph/0204239`).

Pearce F.R. et al. 1999, ApJ, 521(2), L99–L102.

Peebles P.J.E., 1980. The large-scale structure of the universe.

Peng Y., Maiolino R. and Cochrane R. 2015, Nature, 521(7551), 192–195.

Percival W.J. and White M. 2009, MNRAS, 393(1), 297–308.

Percival W.J. et al. 2014, MNRAS, 439(3), 2531–2541.

Perivolaropoulos L. and Skara F. 2021, Challenges for ΛCDM: An update (`arXiv:2105.05208`).

Perlmutter S. et al. 1999, ApJ, 517, 565–586.

Phillips M.M. 1993, ApJ, 413, L105.

Planck Collaboration et al. 2014, A&A, 571:A1.

Planck Collaboration et al. 2016a, A&A, 594:A16.

Planck Collaboration et al. 2016b, A&A, 594:A13.

Planck Collaboration et al. 2020, A&A, 641, A6.

Poincare H. 1906, Popular Astronomy, 14, 475–488.

Prakash A. et al. 2016, ApJS, 224:34.

Press W.H. et al, 2002. Numerical recipes in C++ : the art of scientific computing.

Qu Y. et al. 2017, MNRAS, 464(2), 1659–1675.

Raichoor A. et al. 2017, MNRAS, 471(4), 3955–3973.

Raichoor A. et al. 2020, Research Notes of the AAS, 4(10), 180.

Ramakrishnan S. et al. 2019, MNRAS, 489(3), 2977–2996.

Rampf C. and Buchert T. 2012, J. Cosmology Astropart. Phys., 2012(6):021.

Reid B.A. and White M. 2011, MNRAS, 417, 1913–1927.

Reid B.A. et al. 2012, MNRAS, 426, 2719–2737.

Riess A.G. et al. 1998, AJ, 116, 1009–1038.

Roberts S. and Everson R., 2001. Independent Component Analysis: Principles and Practice.

Rodríguez-Torres S.A. et al. 2016, MNRAS, 460(2), 1173–1187.

Rosas-Guevara Y.M. et al. 2015, MNRAS, 454(1), 1038–1057.

Ross A.J., Percival W.J. and Manera M. 2015, MNRAS, 451(2), 1331–1340.

Ross A.J. et al. 2017, MNRAS, 464, 1168–1191.

Ross A.J. et al. 2020, MNRAS, 498(2), 2354–2371.

Ross N.P. et al. 2012, ApJS, 199(1):3.

Rubin V.C., Ford, W. K. J. and Thonnard N. 1980, ApJ, 238, 471–487.

Rykoff E.S. et al. 2014, ApJ, 785:104.

Safonova S., Norberg P. and Cole S. 2021, MNRAS, 505(1), 325–338.

Salcedo A.N. et al. 2020, preprint (arXiv:2010.04176).

Samushia L. et al. 2014, MNRAS, 439(4), 3504–3519.

Satpathy S. et al. 2017, MNRAS, 469(2), 1369–1382.

Schaller M. et al. 2015, MNRAS, 454(3), 2277–2291.

Schaller M. et al. 2015, MNRAS, 451(2), 1247–1267.

Schaller M. et al. 2015, MNRAS, 452(1), 343–355.

Schaye J. and Dalla Vecchia C. 2008, MNRAS, 383(3), 1210–1222.

Schaye J. et al. 2010, MNRAS, 402(3), 1536–1560.

Schaye J. et al. 2015, MNRAS, 446(1), 521–554.

Scherrer R.J. and Weinberg D.H. 1998, ApJ, 504(2), 607–611.

Schlafly E.F. and Finkbeiner D.P. 2011, ApJ, 737:103.

Schmidt M., Schneider D.P. and Gunn J.E. 1995, AJ, 110, 68.

Scoccimarro R. 2004, Phys. Rev. D, 70, 083007.

Scodeggio M. et al. 2018, A&A, 609, A84.

Secchi A., 1877. L'astronomia in Roma nel pontificato DI Pio IX.

Sheldon E.S. et al. 2004, AJ, 127, 2544–2564.

Sheth R.K. and Tormen G. 2004, MNRAS, 350(4), 1385–1390.

Smee S.A. et al. 2013, AJ, 146:32.

Smith P. and Lewin J. 1990, Physics Reports, 187(5), 203–280.

Sotiriou T.P. and Faraoni V. 2010, Reviews of Modern Physics, 82, 451–497.

Spergel D.N. et al. 2003, ApJS, 148(1), 175–194.

Springel V. 2005, MNRAS, 364, 1105–1134.

Springel V. and Hernquist L. 2002, MNRAS, 333(3), 649–664.

Springel V. et al. 2001, MNRAS, 328(3), 726–750.

Springel V., Yoshida N. and White S.D. 2001, New Astronomy, 6(2), 79–117.

Springel V. et al. 2005, Nature, 435(7042), 629–636.

Springel V. et al. 2017, MNRAS, 475(1), 676–698.

Tassev S., Zaldarriaga M. and Eisenstein D.J. 2013, J. Cosmology Astropart. Phys., 2013(6):036.

Taylor A.N. et al. 1998, ApJ, 501(2), 539–553.

Thacker R.J. and Couchman H.M.P. 2000, ApJ, 545(2), 728–752.

Thomson and Kelvin 1904, LECTURE I. Cambridge University Press, page 5–21.

Tibshirani R. 1996, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Tibshirani R. and Friedman J. 2017, preprint (`arXiv:1712.00484`).

Tinker J.L. et al. 2012, ApJ, 745:16.

Tisserand P. et al. 2007, A&A, 469(2), 387–404.

Tojeiro R. et al. 2017, MNRAS, 470(3), 3720–3741.

Trayford J.W. et al. 2016, MNRAS, 460(4), 3925–3939.

van den Bosch F.C. et al. 2018, MNRAS, 474(3), 3043–3066.

Vargas-Magaña M. et al. 2014, MNRAS, 445, 2–28.

Vollmer B. et al. 2001, ApJ, 561(2), 708–726.

Wang L., Reid B. and White M. 2014, MNRAS, 437, 588–599.

White M. 2002, ApJS, 143(2), 241–255.

White M., Hernquist L. and Springel V. 2001, ApJ, 550(2), L129–L132.

White M., Tinker J.L. and McBride C.K. 2013, MNRAS, 437(3), 2594–2606.

White S.D.M. and Frenk C.S. 1991, ApJ, 379, 52.

White S.D.M. and Rees M.J. 1978, MNRAS, 183(3), 341–358.

Wiersma R.P.C., Schaye J. and Smith B.D. 2009a, MNRAS, 393(1), 99–107.

Wiersma R.P.C. et al. 2009b, MNRAS, 399(2), 574–600.

Wright E.L. et al. 2010, AJ, 140:1868-1881.

Xu X. et al. 2012, MNRAS, 427, 2146–2167.

York D.G. et al. 2000, AJ, 120(3), 1579–1587.

Yoshikawa K. et al. 2001, ApJ, 558(2), 520–534.

Yèche C. et al. 2020, Research Notes of the AAS, 4(10), 179.

Zaninetti L. 2018, International Journal of Astronomy and Astrophysics, 8(3), 258–266.

Zarrouk P. 2021, In *American Astronomical Society Meeting Abstracts.* page 303.03.

Zarrouk P. et al. 2018, MNRAS, 477(2), 1639–1663.

Zavala J. et al. 2016, MNRAS, 460(4), 4466–4482.

Zel'Dovich Y.B. 1970, A&A, 500, 13–18.

Zentner A.R., Hearin A.P. and van den Bosch F.C. 2014, MNRAS, 443(4), 3044–3067.

Zhai Z. et al. 2017, ApJ, 848:76.

Zhou R. et al. 2020, Research Notes of the AAS, 4(10), 181.

Zu Y. and Mandelbaum R. 2015, MNRAS, 454(2), 1161–1191.

Zwicky F. 1933, Helvetica Physica Acta, 6, 110–127.

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2$_{\varepsilon}$. It was created using the *memoir* package, maintained by Lars Madsen, with the *madsen* chapter style. The font used is Latin Modern, derived from fonts designed by Donald E. Kunith.