A framework to test modified gravity using galaxy surveys

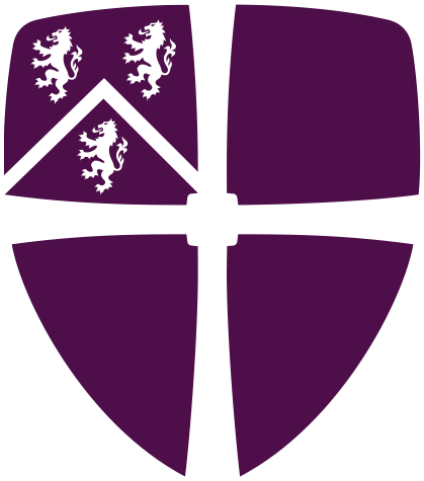
ARMIJO-TORRES, JOAQUIN,ANDRES

A framework to test modified gravity using galaxy surveys

Joaquín A. Armijo-Torres

A thesis presented for the degree of

Doctor of Philosophy



Institute for Computational Cosmology

Department of Physics

Durham University

United Kingdom

August 2022

A framework to test modified gravity using galaxy surveys

Joaquín A. Armijo-Torres

Abstract

We study the application of the marked correlation function, $\mathcal{M}(r_p)$, to probe gravity using the large scale structure of the Universe. We focus our efforts on testing the $f(R)$ modified gravity theory introduced by Hu et al. (2007). This model mimics the cosmic acceleration at late times through the introduction of new physics when the curvature terms are replaced by a function of the Ricci scalar R, rather than by invoking the cosmological constant. These modifications to the gravity equations lead to changes in the environments of large-scale structures that could, in principle, be used to distinguish this model from GR. We use data from the LOWZ and CMASS luminous red galaxy samples of SDSS-III BOSS to measure the marked statistic over a range of scales. To compare with the data, we create mock galaxy catalogues using the halo occupation distribution (HOD) prescription, to populate haloes from N-body simulations using GR and $f(R)$ gravity. Using the Monte Carlo Markov Chain algorithm, we extract the number density and two-point clustering from the mock catalogues and compare with the observational measurements to constrain the HOD model parameters. Weights for individual galaxies are based on the local density information, calculated using a Voronoi tessellation, and are used to mark galaxies when computing the marked correlation function. We find that when taking into account the 1-σ confidence interval for the best fitting HOD parameters, the marked correlation function only marginally distinguishes viable gravity models at the 1-σ level for separations $r_p < 1.7h^{-1}$ Mpc. As part of the process of evaluating the suitability of the N-body simulations to build mock catalogues, we address the question of the mass resolution of the halo catalogue, and introduce a simple scheme to allow the use of marginally resolved halos.

Supervisors: Carlton M. Baugh Peder Norberg and Nelson Padilla

Acknowledgements

This work could not being possible with the unconditional support of people around me, those who have contributed a lot in my life, specially during my time at Durham. First, I would like to thank Carlton, Peder, and Nelson. With your supervision and guidance I have reach really far in my academic path. I hope that I have learned a bit of all the wisdom you tried to provide me during these 4 years in my PhD. A special shout to Shufei, you made my life a lot easier, and I will be forever grateful for that. Your work in the ICC is of the highest quality.

Second, to all my friends and colleagues in Durham. You are a lot of people to be mentioned, but everyone I have meet in the last time has been really important, and I am happy for all the moments I have shared with you guys. Special thanks to my housemates and friends, Qiuhan and Sergio. It was a pleasure to live under the same roof, and knowing you more, guys. Specially the video-game nights (including factorio with Victor). To Arnau, Carol, Christoph, Aidan, and all the CDT members. I have learned so much from you, and I really thank you for that. Many thanks to the Astro-people, in special Jack, who disinterestedly offered his help to proofread this thesis and Ellen for being such a lovely friend. I will miss every coffe break, lunch, Friday-pub and dinners with you, wonderful people. To Ash and the football lads, you guys are awesome, and I really enjoyed my time being part of the Ustinov team. All the training and games, are memories I will hold back every time I enter the pitch.

To those who are far away right now: Thanks to all my friends in Chile. Regardless of the distance, you always managed to keep in touch and be there every time I needed. Pablo, Ignacio, Diego-Mono, Diego and Jorge, I miss you guys, but the Puma will rise again, stronger than ever. All the nights playing Age of Empires during lockdown were more than worth. To Maria Paz, your unconditional friendship and support makes me smile everyday, I am really grateful for that, and

I hope it goes forever. To Natalya, even though we are not close right now, thank you for being my friend, and I promise I will visit you soon.

To Ana María and Álvaro, my parents. Gracias por hacer de mi un hombre que está feliz por todo lo que ha logrado. Todo esto lo hago por ustedes, y espero que puedan estar orgullosos de su hijo. A pesar de la distancia, siempre están en mi mente y mi corazón. A mi hermano Álvaro, gracias por cuidar de mi Ema, y me hace muy feliz que ahora también tengamos a Canela. Espero verlos pronto.

It has been quite a long a eventful journey, and again, not the same without all of you on it.

Contents

Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, Department of Physics, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is the sole work of the author unless referenced to the contrary in the text.

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

Publications

- The content of Chapter 4 in this thesis has been published in:

  Joaquín Armijo et al. Monthly Notices of the Royal Astronomical Society: Letters, Volume 510, Issue 1, pp.29-33

- The content of Chapter 5 and 6 in this thesis is being prepared to be a paper to be submitted to a journal

  Joaquín Armijo et al. in prep.

List of Figures

List of Tables

Chapter 1

Using the large-scale structure of the Universe to constrain cosmology

1.1 The cosmological model

In this thesis we explore the large scale structure of the Universe, where galaxies are born, grow and evolve through different epochs, processes and environments. To understand the different phenomena that gave shape to the Universe we observe, first we need to study it as a whole in the context of a cosmological model. For example, on the largest scales, the Universe is governed by the laws of the theory of gravity, which can be described by Einstein's General Relativity (GR). Einstein proposed the fundamental field equation

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{1.1}$$

where $G_{\mu\nu}$ is the Einstein tensor, which encodes the curvature of space, and $T_{\mu\nu}$ is the energy-momentum tensor that describes the matter distribution for a perfect fluid. In this way the Universe can be described completely by considering its geometry and energy-density matter content, assuming certain properties encapsulated in the cosmological principle. This principle states that on large scales

the Universe is homogeneous and isotropic, which means that are no privileged directions or positions.

1.1.1 The Friedmann-Robertson-Walker metric

The geometrical properties of this homogeneous and isotropic space can be described by considering every element of the Universe as being modelled as a continuous fluid. Thus, any point in space-time has a set of "comoving coordinates" x^α, which is the position of the fluid element passing through the point and a proper time t measured by a clock moving with the fluid. This yields the Friedmann-Robertson-Walker metric, where the line element $d\tau^2$ has the form:

$$d\tau^2 = c^2dt^2 - a(t)^2\left[\frac{dr^2}{1-Kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right]. \tag{1.2}$$

Here $a(t)$ is a function to be determined, and has dimensions of length, the spherical polar coordinates, r (by convention dimensionless), θ and ϕ are the comoving coordinates, K is the curvature constant which is scaled to have a value of -1, 0 or 1, for a open, flat and closed universe, respectively. Then $d\tau$ represents the space-time interval between two points x^i and $x^i + dx^i$. For the case of an isotropic and homogeneous fluid, and an energy momentum tensor with a rest-mass energy density $\rho^2 c$ and pressure p, the Einstein Equation (Eqn.1.1) can be solved. These equations are called the Friedmann cosmological equations:

$$\ddot{a} = -\frac{4\pi G}{3}\left(\rho + 3\frac{p}{c^2}\right)a, \tag{1.3}$$

$$\dot{a}^2 = \frac{8\pi G}{3}a^2\rho - Kc^2, \tag{1.4}$$

where $\dot{a}$ and $\ddot{a}$ is the first and second derivatives, respectively, of the function $a(t)$, which is called the expansion parameter or "scale factor". In this way, the Friedmann equations tell us about the time evolution of the scale factor, if one knows the equation of state that relates the pressure p and the mass-energy density ρ for the different components of the universe. Also, from Eqn. 1.4 the curvature

can be written as:

$$\frac{K}{a^2} = \frac{1}{c^2}\left(\frac{\dot{a}}{a}\right)^2\left(\frac{\rho}{\rho_c} - 1\right), \tag{1.5}$$

where ρ_c is defined as the critical density. This quantity can be obtained by assuming a universe with $K = 0$:

$$\rho_c = \frac{3}{8\pi G}\left(\frac{\dot{a}}{a}\right)^2. \tag{1.6}$$

1.1.2 Cosmological distances

Understanding the geometrical properties of the Universe through the Friedmann equations permits us to define the concept of distance in cosmology. The proper distance, d_p, to a point, P, with coordinates r, θ and ϕ measured by an observer at the origin of the coordinate system is defined by a fixed ruler held by the observer to the position of P at given proper time t. Then for the FRW metric with $dt = 0$ this is

$$d_p = \int_0^r dr' \frac{a}{1 - Kr'^2} = a f(r), \tag{1.7}$$

where the function $f(r)$ depends on the value of K:

$$f(r) = \sin^{-1} r \quad (K = 1), \tag{1.8}$$

$$f(r) = r \quad (K = 0), \tag{1.9}$$

$$f(r) = \sinh^{-1} r \quad (K = -1). \tag{1.10}$$

As the proper distance changes with time because of the time dependence of a, we can define a comoving radial distance d_c by relating d_p at any time t with the value at the present time t_0 using 1.7

$$d_p(t_0) = a_0 f(r) = \frac{a_0}{a} d_p(t), \tag{1.11}$$

where a_0 is the value of a at t_0. Then we have a relation between d_c and d_p, which is just

$$d_c = \frac{a_0}{a} d_p. \tag{1.12}$$

As the proper distance of a source object at position P changes with time, there is an intrinsic radial velocity with respect to the observer, which is defined by Hubble's law

$$v_r = \dot{a} f(r) = \frac{\dot{a}}{a} d_p, \tag{1.13}$$

where the quantity $H(t) = \dot{a}/a$ is called the Hubble parameter. The value of the Hubble parameter today $H(t_0) = H_0$ is believed to be about $70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, though there is some tension between the results derived from different approaches (Solà et al., 2017). It is conventional to define $h = H_0/\left(100\mathrm{km\,s^{-1}\,Mpc^{-1}}\right)$, which is the dimensionless Hubble parameter.

1.1.2.1 Redshift

It is useful to define a variable related to the scale factor, $a(t)$, to explain the position of a distant object such as a galaxy or any extragalactic source in the expanding Universe. We call this variable the *redshift*:

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_0}, \tag{1.14}$$

which relates the wavelengths of the emitted (λ_e) and observed (λ_o) photons of the source moving with the expansion of the Universe. These photons travel from the source to the observer along the null geodesic with $d\tau^2 = 0$, therefore from the metric in Eqn. 1.2 we obtain:

$$\int_{t_e}^{t_0} \frac{cdt}{a} = \int_r^0 \frac{dr}{(1 - Kr^2)^{1/2}} = f(r), \tag{1.15}$$

where t_e is the time at which the photon was emitted at comoving distance r and t_0 is a later time when the photon is observed at the origin, $r = 0$. Then light from the source is emitted at $t'_e = t_e + \delta t_e$ and reaches the observer at $t'_0 = t_0 + \delta t_0$, at the same $f(r)$. Given that r is a comoving distance and both the observer and the

source are moving with the cosmological expansion, we can write

$$\begin{aligned}
\int_{t'_e}^{t'_0} \frac{cdt}{a} &= \int_{t_e}^{t_0} \frac{cdt}{a}, \qquad (1.16)\\
0 &= \int_{t_e}^{t_e+\delta t_e} \frac{cdt}{a} + \int_{t_e+\delta t_e}^{t_0} \frac{cdt}{a} - \int_{t_e+\delta t_e}^{t_0+\delta t_0} \frac{cdt}{a},\\
0 &= \int_{t_e}^{t_e+\delta t_e} \frac{cdt}{a} - \left(\int_{t_0}^{t_e+\delta t_e} \frac{cdt}{a} + \int_{t_e+\delta t_e}^{t_0+\delta t_0} \frac{cdt}{a} \right),\\
\int_{t_e}^{t_e+\delta t_e} \frac{cdt}{a} &= \int_{t_0}^{t_0+\delta t_0} \frac{cdt}{a}. \qquad (1.17)
\end{aligned}$$

If δt is small and, in particular, $\delta t_e = \lambda_e/c$ (and $\delta_{t_0} = \lambda_0/c$), which is equivalent to the frequencies of emitted and observed light respectively, the scale factor is essentially constant over t and $t + \delta t$, then we can integrate Eqn. 1.17 and obtain:

$$\frac{\lambda_e}{\lambda_0} = \frac{a_e}{a_0}. \qquad (1.18)$$

Then for any time $t = t_e$ and taking the definition of redshift en Eqn. 1.14 we can rewrite Eqn. 1.18 as

$$1 + z = \frac{a_0}{a}. \qquad (1.19)$$

1.1.2.2 The luminosity and angular-diameter distances

The distance to a source is calculated using the information we observe in the form of light, meaning there is a finite time interval and path the light has to travel to make a measurement of the distance. This means the proper distance is no longer valid. Therefore we need another definition for astronomically distant objects, such as the *luminosity distance*. To calculate the luminosity distance, d_L, we consider the flux obtained from a source at point P. The observed flux needs to take into account the expansion of the universe. Using Eqn. 1.18 we infer that the observed flux is $L_0 = L_e\,(a/a_0)^2$. The two factors of (a/a_0) come from the redshifting of the frequency for individual photons, and the small variation of time interval δt over

which the energy is emitted. The flux is given by

$$F = \frac{L_e}{4\pi a_0^2 r^2} \left(\frac{a}{a_0}\right)^2, \tag{1.20}$$

from the inverse-square law between flux and luminosity. From the definition of comoving distance (for the spatially flat Universe) and redshift in Eqns. 1.11 and 1.14 respectively, then the luminosity distance is

$$d_L = d_c(1+z). \tag{1.21}$$

Another useful distance to define is the angular diameter distance d_A, which relates the physical size of a distant object to its angular size for the observer. From the metric in Eqn. 1.2, the diameter D of a source with dr=0 and $d\phi \approx 0$, we obtain

$$d\theta = \frac{D}{ar}. \tag{1.22}$$

Then, considering that for a distant observer the angle $\Delta\theta = D/d_A$, then the angular diameter distance is simply:

$$d_A = ar. \tag{1.23}$$

Again, for the case of the Universe with K=0, we can express this distance in terms of the comoving distance, $d_A = d_c/(1+z)$.

1.1.3 The cosmological constant

The cosmological constant was first introduced by Albert Einstein to make the universe static, as scientists at the time were unaware of the cosmological expansion. By adding a constant factor Λ to the field equations, the Friedmann equations become

$$\ddot{a} = -\frac{4}{3}\pi G\left(\rho + 3\frac{p}{c^2}\right)a + \frac{\Lambda a}{3}, \tag{1.24}$$

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2 - Kc^2 + \frac{\Lambda a^2}{3}. \tag{1.25}$$

From these equations, the Λ term is equivalent to adding a new component to the energy-density of the universe on the right-hand size of the Einstein field equation, with density

$$\rho_\Lambda \equiv \frac{c^2}{8\pi G}\Lambda. \tag{1.26}$$

From the acceleration Eqn. 1.25, it can be seen that the cosmological constant has a different sign than the energy-matter density ρ, which means that Λ produces a negative pressure with a repulsive gravitational force. This conclusion can also be obtained by inferring the equation of state for this component, which relates the pressure to the density (ρ_Λ) through $p = w\rho c^2$, with $w = -1$. However, later on, observations showed that the Universe was expanding following Hubble's law, which prompted Einstein to discard the static universe model, and hence to declare Λ as the biggest blunder of his life. Currently, the cosmological constant is popular again following the observations that indicate that the expansion of the Universe is speeding up, and Λ has been adopted in the most accepted cosmological model, the ΛCDM universe (Riess et al., 1998; Perlmutter et al., 1999).

1.2 The Λ Cold Dark Matter universe

Now that we understand the context of the cosmological model, we try to understand the physical nature of the Universe, by measuring the evolution of the the different quantities defined in Section 1.1. The Universe is well described by the standard model of cosmology, Λ-cold-dark-matter (ΛCDM), which explains with great accuracy the different phenomena we observe. This model describes a homogeneous and isotropic Universe, which is currently undergoing an accelerated expansion, with various components which contribute to the overall energy-density today, $\Omega_0 \equiv \frac{\rho}{\rho_c}$. This means that the density of the Universe is measured by the contributions of matter and energy in comparison to the critical density ρ_c. The different components that contribute to Ω_0 are: The density for baryonic matter Ω_b, which is mainly in the form of hydrogen and helium, forms the stars and galax-

ies we observe in the sky, the one for cold dark matter Ω_c, and for dark energy Ω_Λ, which as of yet both their nature remain undiscovered. These are the cornerstones of the standard cosmological model.

The baryons we observe in the present time are in galaxies, stars and mostly in the hot intergalactic medium, and correspond to no more than 5% of the total energy-density composition of the ΛCDM universe. The rest is believed to be dark matter (24%) and dark energy (71%), as inferred by people, from the best fitting model to the observations by the Planck mission (Planck Collaboration et al., 2020a,b). The Universe then has $\Omega_0 = 1$, which means that we live in a universe with density equals to the critical value ρ_c (or almost), then the Universe is flat with $K = 0$. While dark matter is an active component in the formation of the large-scale structure of the Universe through its mass and gravity, it is thought to be "cold", hence the thermal velocity of one of the successful candidates must be non-relativistic (Chadha-Day et al., 2022), and does not interact with ordinary matter, apart from through gravity.

The most commonly invoked candidate for the dark energy is a cosmological constant, Λ, which can be interpreted as a negative pressure that makes the Universe to expand. This has a constant energy density which, in recent times, has become the dominant component of the Universe and hence results in an accelerating cosmic expansion today. The cosmological constant explanation of the accelerating cosmic expansion has unappealing aspects. The theoretically motivated value from vacuum energy considerations does not match that inferred from observations (Weinberg, 1989). This "vacuum catastrophy" is a hint that the nature of Λ is a mystery and that a different, more plausible explanation should be sought. Among the alternatives to the cosmological constant are exotic mechanisms which invoke more complex forms of dark energy, such as quintessence and dynamic scalar fields (Armendariz-Picon et al., 2000; Tsujikawa, 2013). Also, theories of modified gravity which add physical degrees of freedom (Carroll et al., 2004) to the gravity equations are being considered, switching the effect of dark energy to the curvature

side of Einstein's equation.

At the present time the Universe has evolved to become highly non-homogeneous, dominated by structures of several Megaparsecs (Mpc) in size, which we call the large-scale structure. The study of the evolution of the large-scale structures is one of the most active subjects in cosmology, which describes the hierarchical formation of structures, dark matter haloes and galaxies. Hierarchical growth can be understood as resulting from a primordial spectrum of fluctuations, with power on all scales. In particular, the shape of the small scale spectrum is such that the fluctuations on these scales become nonlinear first. This leads to a sequence of structure formation proceeding from small scales to larger scales, as the smaller structures merge or become larger by accreting more material.

1.3 Using the large-scale structure of the Universe to constrain cosmology

The cosmic web, which is the network pattern observed today formed by galaxies that trace the large-scale structures, grows from the primordial matter density field. This field is thought to be seeded during inflation, through the action of gravity. The small perturbations imprinted after inflation grow as the Universe is expanding. During the early phases of this growth structures can be modelled using perturbation theory, where the equations of motion can be approximated by linear and higher order equations to describe the dynamics of these structures.

The overdensities of baryonic and dark matter are defined by

$$\delta(\mathbf{r}) \equiv \frac{\rho(\mathbf{r}) - \langle\rho\rangle}{\langle\rho\rangle}, \tag{1.27}$$

where ρ is the density and $\langle\rho\rangle$ is the mean density of the field. The overdensity evolves first from a linear regime homogeneously, where the perturbations are much smaller than the horizon scale, and a linear analysis can be used in early times. As we approach late times, the overdensity grows due to the gravitational instability.

Here the perturbations become nonlinear, where $\delta \sim 1$, first on the small scales where the power is stronger and as time continues, larger scales transition to a strongly non-linear regime governed fully by gravity. Linear models that describe the gravitational instability process show how the initial density perturbations for baryons grow if they are larger than a characteristic length scale or mass, the Jeans mass, overcoming the pressure force and collapsing to form a gravitationally bound structures. For collisionless dark matter, the overdensity keeps growing over time without feeling the pressure force, which forms deep potential wells that structures follow after the recombination era.

Further on, the evolution of δ of structures becomes non-linear, where the growth of structures becomes inhomogeneous depending on the scale. Then, to describe δ, non-linear terms in the equations need to be added, as perturbation terms that depend on the initial density field (Carlson et al., 2009). Perturbation theories capture the significant information beyond linear theory, such as the mode-coupling dependencies, large-scale flows and free-streaming (Leclercq et al., 2013), allowing to explain the dynamical evolution of the density field. However, their accuracy is still limited in small scales. For instance, using the Zel'dovich approximation (Zel'dovich, 1970), we can understand how the initial non-linear collapse occurs in preferred directions forming sheet-like and filament-like structures, giving shape to the cosmic web. The Zel'dovich method is simply a linear approximation (first order) of the Lagrangian perturbation theory, which uses the position and the displacement of the massive particles.

To solve the small-scale regimes, accurate numerical simulations are required to study the evolution of structure. Structure formation occurs at nodes of the cosmic web, where matter collapses from every direction, moving through filaments and walls, where there is a preferred direction for the collapse, to enormous and almost empty cosmic voids, which form when most of the matter is evacuated from a specific region. We refer to hierarchical clustering as the process that connects the initial perturbations to the actual structures which form at late times. Later

on, using the Press-Schechter formalism (Press et al., 1974), we can predict the abundance of collapsed objects as a function of mass, which gives us a useful, yet approximate formalism to study gravitationally bound structures in the cosmic web. Nevertheless, various assumptions are made in the Press-Schechter analysis, which can be difficult to justify, but is still a practical predictive tool. The formalism can be refined when ellipsoidal dynamics are incorporated in the collapse to form dark matter haloes, which agrees better with the results from N-body simulations (Sheth et al., 1999).

The cosmic web serves as a rich laboratory with a variety of different environments nurturing the different types of galaxies during their lives. The Universe is thought to be dominated by the presence of dark matter over ordinary matter, which means that most of the interactions in the cosmic web occur due to gravity. However, we can only observe the galaxies lying at the density peaks of the matter density field, which means that we also need to understand the connection between these galaxies and the dark matter surrounding them. The relation between the matter density field and galaxies is called the halo-galaxy connection, which explains how galaxies populate density peaks of dark matter, known as dark matter haloes. Although halo models also include the effects of baryonic physics on the birth of galaxies, this process is still mainly driven by gravity (White et al., 1978, 1991). This results in galaxy formation as a two-stage theory: dark matter haloes form through the hierarchical clustering process first, and then the cooling mechanisms which operate within the hot gas inside the dark matter haloes lead to the formation of a disk of condensed cold gas. The halo model and the cosmic web in context of the ΛCDM universe have been tested with high precision over the years. These tests include probes of the early universe using observations of the cosmic microwave background (CMB) through a series of experiments culminating in the Wilkinson-Microwave-Anisotropy-Probe (WMAP) and Planck satellites (Hinshaw et al., 2013; Planck Collaboration et al., 2016, 2020a); and for the late Universe through large volume surveys of galaxies which measure the cosmic structure such

the 2-degree-Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky survey (SDSS) (Colless et al., 2003; Alam et al., 2015).

1.4 Probing the the Universe at Large-scales

The distribution of galaxies on the sky is not random. They cluster in groups in large numbers and, due to the connection between the formation of galaxies and the underlying dark matter distribution, galaxy clustering is an important cosmological probe (Benson et al., 2000). To study the clustering of galaxies a large volume needs to be sampled. The uncertainty on the two-point clustering decreases with the inverse square root of the surveyed volume. Also, the observed number of galaxies per unit volume (galaxy number density) used to trace the matter field has to be large enough to obtain a high signal-to-noise measurement of the clustering Feldman et al. (1994)*. A sufficiently high number density of galaxies and a large sample volume fix the amplitude of the clustering and reduces the uncertainties, which is crucial for constraining the parameters of the cosmological model. In Figure 1.1 we show what a large-scale survey looks like, and how they trace the cosmic web. The galaxy distribution shown corresponds to a large volume surveyed by SDSS-III, which is the largest published spectroscopic volume surveyed to date (covering a solid angle of $10,000$ deg^2 or almost $1/4$ of the sky).

1.4.1 Peculiar velocity and redshift space distortions

In reality the redshift we observe in galaxy surveys is a combination of the cosmological redshift we measure from expansion $z_{\cos}$ and the velocities of galaxies caused by the different dynamics that we can find within structures, for example inside cosmic nodes or other collapsing regions. This "peculiar" velocity v_p affects the line-of-sight distance measurement of the galaxy causing an additional Doppler

*In practice, this means that the shot noise arising due to the use of discrete tracers to sample the continuous density field has to be smaller than the clustering signal.

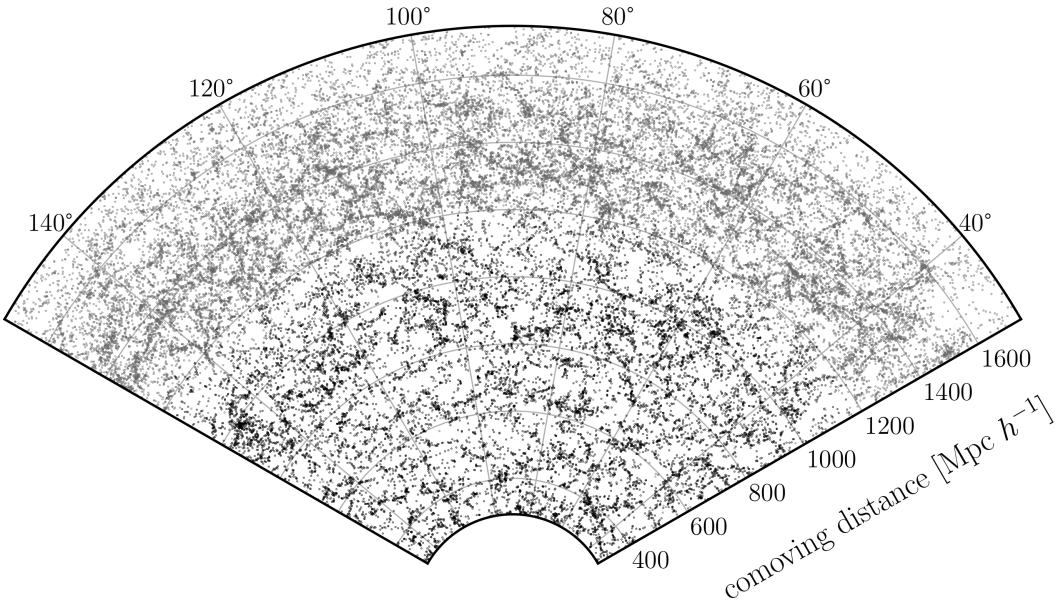


Figure 1.1: Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS). The plot shows the right ascension and comoving distance (obtained from the redshift) of the individual galaxies. Two samples, LOWZ (black dots), and CMASS (grey dots) are plotted over 140 deg. on the sky, in a slice of thickness 3 deg. in declination. These samples are described in more detail in Chapter 3.

effect in the emitted photons of the galaxy, which is indistinguishable to the one in Eqn. 1.14. Then the effect of peculiar velocities changes the observed redshift $z_{\rm obs}$:

$$z_{obs} = (1 + z_{\rm cos})(1 + \frac{v_p}{c}) - 1. \tag{1.28}$$

When ignoring peculiar velocities to obtain a correspondent comoving distance from the galaxy redshift $z_{\rm obs}$, then distortions in the line-of-sight distance will appear, for instance galaxy groups are elongated by peculiar velocities producing the so called "Finger-of-God" effect. Thus, the observed distance for a galaxy is now defined by

$$\mathbf{s} = \mathbf{r} + \frac{1+z}{H(z)}\mathbf{v} \cdot \hat{\boldsymbol{e}}_{\parallel}, \tag{1.29}$$

where $\mathbf{r}$ is the 3D position (in comoving coordinates) of the galaxy in real-space at redshift z, $\mathbf{v}$ is the 3D velocity of the galaxy, and $\hat{\boldsymbol{e}}_{\parallel}$ is the line-of-sight direction.

1.4.2 Power spectrum and galaxy-galaxy correlation function

To study the distribution of matter in the Universe one can use the matter power spectrum, which is defined as the mean square amplitude of the Fourier transforms of the density field, δ_k. This is appealing because in the linear regime of fluctuation growth, the Fourier modes evolve independently of one another. Then $P(k)$, the power spectrum, can be used to study the large-scale structures formed during the linear regime, which gives accurate predictions of the structure formation, particularly at very large scales. When the phases of the Fourier modes are independent, the density field $\delta(\mathbf{r})$ can be described by a normal probability distribution, and if the density field is homogeneous and isotropic as predicted by inflation, then all its statistical properties can be inferred from the power spectrum (the second moment of the field). On smaller scales the fluctuation growth rapidly becomes non-linear, hence higher order perturbation theory and or numerical N-body simulations are needed to resolve such regions. Additionally, as we cannot observe dark matter directly (and hence do not see the complete density field), we need to probe

the cosmic density field using individual galaxies, assuming there is an intrinsic connection between these and the density field we are modelling.

We can also study the two-point correlation function, which is the configuration-space version of the power spectrum (using the Fourier transform). This is defined by

$$\xi(\mathbf{r}) = \langle \delta_1 \delta_2 \rangle \,, \tag{1.30}$$

where $\mathbf{r} = |\mathbf{r}_1 - \mathbf{r}_2|$, which is defined by the positions of the objects measured. When using tracers of the density field (galaxies for example), the two-point correlation function can be understood as the excess of probability of finding a galaxy pair at separation r compared to a random distribution of N galaxies

$$\mathrm{d}N = n_\mathrm{g} \left[1 + \xi(r)\right] \mathrm{d}V, \tag{1.31}$$

where $\mathrm{d}N$ is the expected number of galaxies in the small volume $\mathrm{d}V$, and n_g is the mean number density of galaxies. Note that for a random distribution, $\xi(r) = 0$. There are advantages in using the two-point correlation function to measuring galaxy clustering over using the power spectrum. In general, for the scales we are interested to explore, the estimation of the clustering is more straight forward using the definition of ξ in Eqn. 1.31, even more so when we use the definition of "estimators" (see Section 3.5) with the information provided by observations. In this way we infer the density of a galaxy sample, given the limited sampled volume. Whereas for the power-spectrum we observe from a galaxy survey, this includes the convolution of the true spectrum with the window function of the survey. The shape of the correlation function for different scales gives a detailed description of how galaxies cluster in the large-scale structures. Nevertheless, higher-order moments are required to obtain a *complete* picture of the clustering, particularly on smaller scales which are in the nonlinear regime. As galaxies trace the underlying matter distribution in a non-local and stochastic way, there is a "bias" relation between the distribution of galaxies and the density field. The linear bias b is defined as:

$$\delta_g = b\delta_m, \tag{1.32}$$

where δ_g is the density field inferred from galaxies and δ_m is the density field for matter. This basic parameter b contains information about the content and dynamical evolution of the Universe, along with the shape of $\xi(r)$, which makes it a powerful tool to constrain the cosmological model. For example, assuming the cosmological model from the latest Planck observations, we have the matter correlation function, which when adding the form of the bias $b(M)$, we can infer the halo mass for the assumed cosmology.

1.5 Studying gravity with large-scale structures

By observing galaxies in wide-field surveys, to complement CMB measurements, the density of matter can be constrained to around 30% of the total energy-density content of the Universe. Given that these observations also support a universe with the critical density, mostly from CMB observations, this means that the rest of the energy density is made up of the mysterious dark component that we call dark energy; this component currently dominates the dynamics of the Universe, causing its accelerating expansion.

Although the cosmological constant Λ is part of the stationary universe solution erroneously introduced by Einstein in Eqn. 1.1, the nature of this component remains unknown and represents one of the biggest challenges in present day cosmology, and to the theory of gravity (Heymans et al., 2018; Baker et al., 2021). This means that the large-scale structure of the universe not only tests the cosmological model, but also confronts general relativity on cosmological scales. Moreover, many models that modify the theory of gravity from general relativity replicate the accelerated expansion without invoking a cosmological constant (Koyama, 2016). For such models, new degrees of freedom are introduced, which must be coupled to matter, modifying the formation of structure over time. Alternative models of gravity will inevitably modify structure formation in a manner that depends on the environment. To study such models, we need probes that gather information

related to these modifications, like the properties linked to the distribution of the galaxies in such environments. These impacts are seen in probes such as weak lensing (Kilbinger, 2015), redshift space distortions (Peacock et al., 2001), and the marked correlation function statistic (White, 2016). In this thesis we focus on the latter, which is a relatively new statistical method that contains information beyond the traditional galaxy-galaxy correlation function, whilst still being a second moment quantity. The marked correlation function has been used to study the connection between properties of galaxies and their environment, such as luminosity and environmental density, and halo mass (Sheth et al., 2004; Wechsler et al., 2006).

1.5.1 Modified gravity theories in large-scales structures

To better understand what kind of properties we can exploit to study general relativity and modified gravity models, we need to introduce the modified gravity models that are currently viable in terms of satisfying current constraints. In particular, viable models of modified gravity (MG) are those that include a screening mechanism on scales where gravity has been probed and is consistent with GR. Among such models one can find chameleon theories (Brax et al., 2013) and Vanshtein mechanism theories (Vainshtein, 1972). Although our interest lies in studying modified gravity models that change structure formation on cosmological scales, we need models that satisfy constraints on smaller scales, which to date are consistent with Einstein's general relativity.

Powerful constraints, such as the dynamics of the solar system, and the more recent detection of binary neutron star mergers, indicate that several classes of MG models that were recently under consideration have now been ruled out (Lombriser et al., 2016). Such theories of gravity modify the propagation velocity of gravitational waves detected in vacuum, which is not consistent with the current measurements, including measurements of the optical counterparts of such events. However, many models of modified gravity are still viable as they evade

local tests of gravity in the solar system and on galactic scales (Baker et al., 2017; Creminelli et al., 2017). Such models are continuously being tested and further constrained, and include chameleon theories, for example $f(R)$ gravity (De Felice et al., 2010) and Brans-Dicke type theories including the Dvali-Gabadadze-Porrati (DGP) model (Dvali et al., 2000). These MG models are important as they can be used to test GR and the equivalence principle on cosmological scales.

Our main aim in this thesis is to investigate if a marked correlation function, in which the mark depends on density, can distinguish between viable gravity models. To meet this aim, we have developed a pipeline to make realisations of mock galaxy catalogues from N-body simulations, using a simple halo model approach. A key feature of our analysis is an assessment of the resolution effects of the simulation and the uncertainty due to the range of halo models that give acceptable fits to the measured two-point correlation functions; this uncertainty is often ignored in the literature and could result in an overly optimistic view of the performance of any diagnostic that depends on clustering.

1.6 Outline of the thesis

The outline of this thesis is as follows: in § 2 we review the $f(R)$ theory of gravity, which is the model studied in this work. In § 3 the data from the Baryon Oscillation spectroscopic survey (BOSS) is presented, which is designed to collect large numbers of galaxy redshifts over a large volume to measure the large-scale structure of the Universe. The simulations used to understand the modelling of MG are presented in § 4, along with a discussion of the mass resolution of the halo catalogue and a simple scheme to make use of marginally resolved halos. The creation of mock galaxy catalogues to replicate the observations is described in § 5. The calculation of the marked correlation function for both data and mock catalogues is presented in § 6. Finally, we explain the direction in which this work could go in the future and draw our conclusions in § 7.

Chapter 2

The $f(R)$ theory of gravity

2.1 Overview

The $f(R)$ theory of gravity is a viable alternative to general relativity. In the standard ΛCDM cosmological model, the cosmological constant, Λ, drives the accelerated expansion of the universe at recent times. Instead of invoking Λ, $f(R)$ gravity models explain the quickening expansion by the invoking of new physics that arises from the additional degrees of freedom introduced in the equations of motion for gravity (Li et al., 2007). Such models can be understood as an extension to the standard GR model and can be tested by studying the effects of gravity on different physical scales.

The $f(R)$ model of gravity can be viewed as an extension of standard GR through the inclusion of a function f of the Ricci scalar, R, in the Einstein-Hilbert action

$$S = \int d^4x\sqrt{-g}\left(\frac{1}{2\kappa^2}\left[R + f(R)\right] + \mathcal{L}_m\right), \tag{2.1}$$

where $\kappa^2 = 8\pi G/c^4$, Einstein's constant, g, is the determinant of the metric $g_{\mu\nu}$ and $\mathcal{L}_m$ is the Lagrangian density of matter. The choice of this $f(R)$ function can be used to mimic the behaviour of the ΛCDM model which is well constrained by CMB observations, and to modify the cosmology at late times to replicate the accelerated expansion that we observe at low redshift. It is worth noting that

considering the specific shape of the $f(R)$ function, the ΛCDM model is recovered for $f(R) = 2\Lambda$, where Λ is a constant. The addition of higher order terms in 2.1 leads to the modifications of all the equations of GR, including the Einstein field equation in Eqn. 1.1:

$$G_{\mu\nu} + f_R R_{\mu\nu} - g_{\mu\nu}\left[\frac{1}{2}f - \nabla^2 f_R\right] - \nabla_\mu\nabla_\nu f = \kappa T_{\mu\nu}, \tag{2.2}$$

where ∇_μ is the covariant derivative of the metric tensor, $f_R \equiv \frac{df(R)}{dR}$ is the new, scalar and dynamical degree of freedom that arises from the introduction of the $f(R)$. To solve this new equation and obtain the equations of motion for massive particles, one can take the trace of Eqn. 2.2 and solve for the case of a perturbation around the standard Friedmann-Lemaître-Robertson-Walker metric in Eqn. 1.2. This leads to a modified set of the Friedman equations, which include several non-linear terms that combine $f(R)$, f_R and higher order. This description of the background evolution of the Universe gives two equations of motion. One is the modified Poisson equation for the gravitational potential:

$$\vec{\nabla}^2\Phi = \frac{16\pi G}{3}a^2[\rho_m - \bar{\rho}_m] + \frac{1}{6}a^2\left[R(f_R) - \bar{R}\right], \tag{2.3}$$

and for the new scalar field f_R

$$\vec{\nabla}^2 f_R = -\frac{1}{3}a^2\left[R(f_R) - \bar{R} + 8\pi G(\rho_m - \bar{\rho}_m)\right], \tag{2.4}$$

where ρ_m is the matter density field, and overbar indicate quantities ($\bar{\rho}_m$ and $\bar{R}$) defined as mean values for the background cosmology, which is solved in comoving coordinates. As we have now defined the Ricci scalar as a function of f_R in both Eqns 2.3 and 2.4, we can combine these to obtain

$$\vec{\nabla}^2\Phi = 4\pi G a^2\left[\rho_m - \bar{\rho}_m\right] - \frac{1}{2}\vec{\nabla}^2 f_R, \tag{2.5}$$

which is a new equation of motion for massive particles including a term which comes from the new scalar degree of freedom. We can understand this new term

as the potential $-\frac{1}{2}f_R$ of an extra force, the fifth force, mediated by the scalar field f_R, which is sometimes referred to as the scalaron (Gannouji et al., 2012).

2.2 The chameleon mechanism

The equations of motion of $f(R)$ gravity are different from the one in standard gravity, and different predictions may result. Nevertheless, local tests already constrain these predictions with great accuracy on certain scales, such as in the solar system (Guo, 2014), which means that modified gravity must include mechanisms to hide the new physics which arises from the extra degree of freedom in Eqn. 2.5. This feature is referred to as a screening mechanism (Khoury et al., 2004), and is a scale-dependent property of chameleon theories such as $f(R)$ gravity. In scales where the model is expected to behave as standard gravity, such as in the deep Newtonian potential of the Solar system, Eqn. 2.4 is dynamically driven to $|f_R| \to 0$. In this limit, Eqn. 2.5 reduces to the standard Poisson equation and GR is recovered, hence this theory is viable on these scales (Hu et al., 2007). On the other hand, on scales where the Newtonian potential becomes shallower, the term $R - \bar{R}$ in Eqn. 2.4 is negligible and Eqn. 2.5 is reduced to

$$\vec{\nabla}^2\Phi = \frac{16}{3}\pi G a^2[\rho_m - \bar{\rho}_m], \tag{2.6}$$

which is the same as the standard Poisson equation, but enhanced by a factor 4/3 when the amplitude of the fifth force is at its maximum and no screening is triggered. An interesting feature of this theory is that to obtain Eqn. 2.3 no assumption about the form of the $f(R)$ function is required in Eqn. 2.5, but some requirements are needed to avoid large deviations from GR on both cosmological and solar system scales. These constraints should include the predictions of ΛCDM as the limiting case on large scales, and can be achieved by ensuring the following:

$$\lim_{R\to\infty} f(R) = \text{const.}, \tag{2.7}$$

$$\lim_{R\to 0} f(R) = 0. \tag{2.8}$$

This class of model is referred as the Hu-Sawicki model (Hu et al., 2007).

2.3 The Hu & Sawicki model

The Hu-Sawicki $f(R)$ model gained traction over the past 15 years as a model that can reproduce almost the same expansion history as in the ΛCDM model (by construction) without the need for a cosmological constant. Also, the novel inclusion of a screening mechanism that allows the model to satisfy tests on solar system and galactic scales, makes this theory viable given the successful predictions of general relativity (Multamäki et al., 2006; Sotiriou, 2006). Additionally, the relative simple shape of the model, along with its dependence on only a few free-parameters, makes it possible to solve the equations of motion using N-body simulations (Oyaizu, 2008; Li et al., 2011). A popular choice for the functional form of $f(R)$ is proposed as following

$$f(R) = -m^2 \frac{c_1 \left(\frac{R}{m^2}\right)^n}{c_2 \left(\frac{R}{m^2}\right)^n + 1}, \tag{2.9}$$

where $m^2 \equiv 8\pi G\bar{\rho}_{m0}/3 = H_0^2\Omega_m$ is called the mass scale, and is a convenient way to express the regimes where the behaviour of the model shows or hides the effects of the scalaron f_R. $\bar{\rho}_{m0}$ is the value of the background matter density today, n, c_1 and c_2 are free parameters of the model. The form of this function is motivated by the aim of ensuring that for high curvature values compared to the mass scale, m^2, the term m^2/R goes to zero and $f(R)$ can be expanded as

$$f(R) \approx -\frac{c_1}{c_2} m^2 + \frac{c_1}{c_2^2} m^2 \left(\frac{m^2}{R}\right)^n. \tag{2.10}$$

In the limit $m^2/R \to 0$, the term c_1/c_2 acts as the cosmological constant of this model, and is independent of scale. As we have an explicit form for $f(R)$ we can set $c_1/c_2 = 6\Omega_{\Lambda,0}/\Omega_{m,0}$, where $\Omega_{m,0}$ is the matter density parameter today, and $\Omega_\Lambda = 1 - \Omega_m$. With this configuration the model follows the same expansion

history as the ΛCDM model by construction. Meanwhile, the scalaron field can also be approximated by

$$f_R \approx -n\frac{c_1}{c_2^2}\left(\frac{m^2}{R}\right)^{n+1}, \tag{2.11}$$

In the background cosmology the scalaron f_R sits in a minimum of the effective potential that governs the dynamics of massive particles, triggering the chameleon mechanism. When Eqn. 2.4 is solved, the effective potential $V_{\rm eff}$ for the scalaron has the form:

$$V_{\rm eff}(f_R) = \frac{1}{3}(R - f_R R + 2R + 8\pi G\rho_m). \tag{2.12}$$

This is a stable potential that requires $d^2V_{\rm eff}/df_R^2 > 0$, then for small oscillations the dependence of the scalaron in the Ricci scalar can be solved by using the background values $\bar{R}$ and $\bar{f}$ (Brax et al., 2012), then the Ricci scalar in the background cosmology can be written like

$$\bar{R} \approx 8\pi G\rho - 2\bar{f}(R) = 3m^2\left[a^{-3} + \frac{2}{3}\frac{c_1}{c_2}\right], \tag{2.13}$$

which, removes the dependence between $R(f_R)$ and the f_R. Then, The previous approximation can be used to set the term in Eqn. 2.11 once we evaluate with the values at present time:

$$\frac{c_1}{c_2^2} = -\frac{1}{n}\left[3\left(1 + 4\frac{\Omega_{\Lambda 0}}{\Omega_m 0}\right)\right]^{n+1} f_{R0}, \tag{2.14}$$

which is evaluated with the value of the scalaron today, f_{R0}. By fixing these values the model depends on only two free parameters, n and f_{R0}. To constrain these parameters, probing the large-scale structure at late times is required. One of the fundamental measurements to obtain these constraints is the power spectrum for a range of models with different values of the scalaron amplitude $|f_{R0}|$ when fixing $n = 1$.

In Figure 2.1 we show a comparison of the computation of $P(k)$ for F4, F5 and F6 models ($|f_{R0}| = 10^{-4}, 10^{-5}, 10^{-6}$ respectively), relative to the the calculation of $P(k)$ for GR-ΛCDM simulation. Here, a range of scales k is compared, first, for non-linear equations from Schmidt et al. (2009), and for the N-body simulations of Li

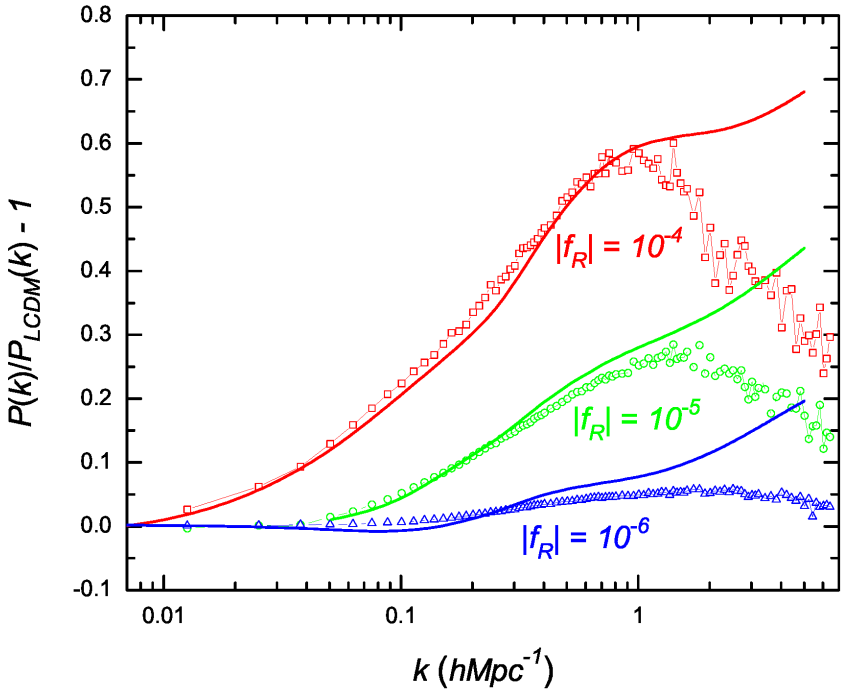


Figure 2.1: The non-linear power spectrum $P(k)$ ratio between $f(R)$ gravity models and ΛCDM cosmology as function of the scale k from Li et al. (2012). Open symbols are calculated using the simulations at $z = 0$ of F4 (red), F5 (green) and F6 (blue), whereas the solid lines correspond to the analytical fits from Schmidt et al. (2009).

et al. (2012). $P(k)$ at scales of $k = 0.1\ h$ Mpc^{-1} starts diverging for models like F4 and F5, which is expected when considering these models have a strong amplitude of the fifth force $|f_{R0}|$ (10^{-4}, 10^{-5} respectively) in comparison to F6, where this amplitude is much smaller. On smaller scales, for $k > 1\ h\,\mathrm{Mpc}^{-1}$, the non-linear power spectrum is no longer valid in comparison to the GR and modified gravity simulations. For example, the screening mechanism is not recovered from the non-linear equations of Schmidt et al. (2009), and N-body simulations are required. For the simulations in Li et al. (2012), the screening mechanism is triggered in high density regions at high k values, where some models with the higher fifth force amplitude, such as F4 and F5 differ from the $P(k)$ of GR more than 10%. Here modified gravity could still be distinguished due to its weaker screening, and N-body simulations play a strong role in studying this regime.

2.4 Large-scale N-body simulations in $f(R)$ modified gravity

N-body simulations have been used to investigate the impact of modified gravity on the large-scale structures. Interesting features arising from $f(R)$ gravity, such as the fifth force and the screening mechanism, can be studied by creating probes to exploit this new physics. As structures are expected to collapse at different rates, due to the additional enhancement in modified gravity, the growth of density fluctuations has a scale dependence different than the one coming from GR (Jennings et al., 2010). In Figure 2.2 the two dimensional power spectrum from Jennings et al. (2010), shows how for $f(R)$ modified gravity models the spherical symmetry seen in the real space power spectrum (left panel) is distorted in redshift space (right panel): the amplitude of the redshift-space power spectrum is larger and squashed to that in real-space at large scales, and more elongated along the line of sight in redshift-space compared to real-space. The effects are more pronounced in the F4 model, where the large scale boost appears larger than in GR. The redshift space $P(k)$ for $f(R)$ gravity looks far more distorted and asymmetrical than the redshift space $P(k)$ in GR. These results hint that the redshift distortion imprinted in the power spectrum and the subsequent two-point clustering is a relevant probe to study modify gravity at large-scales.

High resolution N-body simulations can shed light on how modified gravity differs from GR in large k values. In Figure 2.3 we show simulations from Arnold et al. (2019). When centering on some of the large mass haloes of simulations of GR and $f(R)$ we can see some of the extra features predicted by modified gravity. We highlight some of the regions in the top panel of Figure 2.3, where the different dynamics presented in the F5 simulation of Arnold et al. show how the structure formation is modified. Here, the most noticeable features are the enhancing of gravity in unscreened haloes, and the increased halo formation arising from the extra fifth force in MG models (Li et al., 2012). Another interesting feature

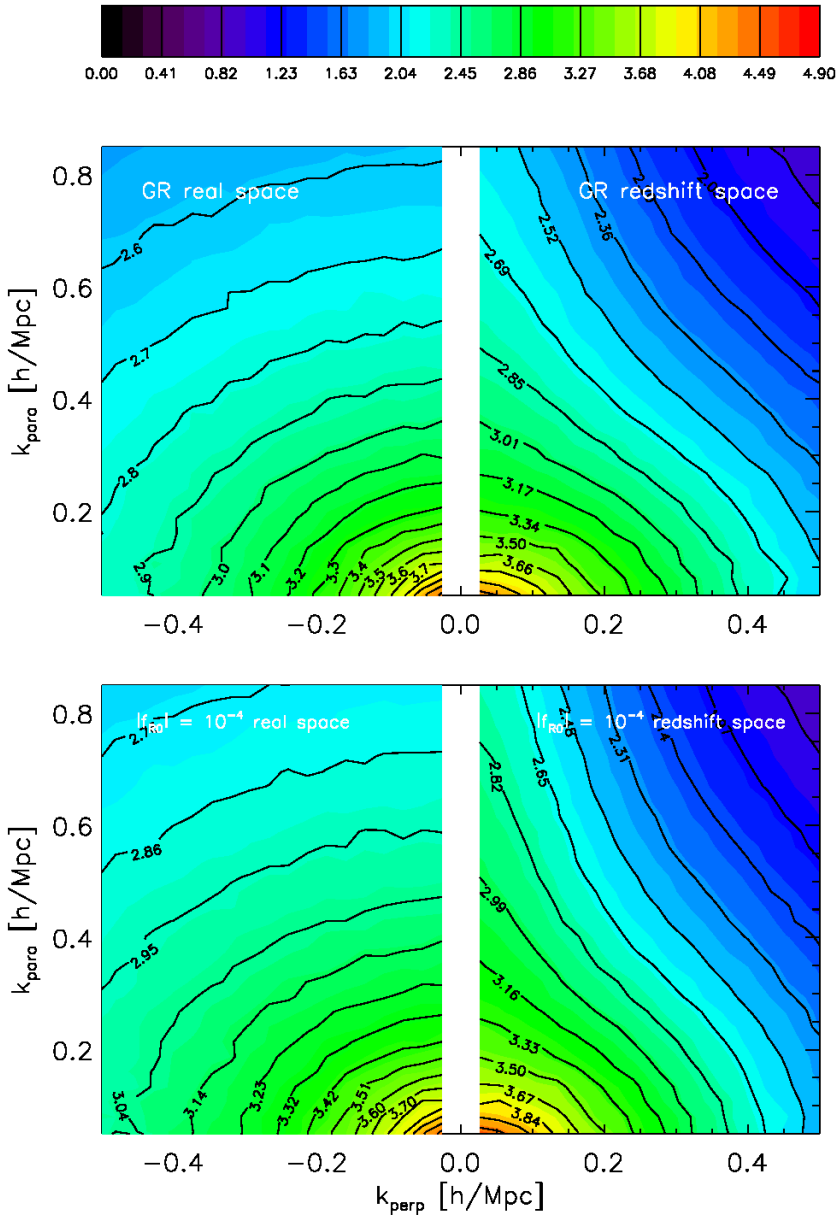


Figure 2.2: The two dimensional power spectrum in real space (left panels) and redshift space (right panels) for the GR (top panels) and F4 (bottom panels) simulations from Jennings et al. (2010). The colour gradient and the contour lines represent the amplitude of the power spectrum, $\log P$ as indicated by the line labels and color-bar at the top.

presented in the MG of Arnold et al. is the screening mechanisms predicted in $f(R)$. The bottom panel of Figure 2.3 shows the difference between the figures of the upper panel, where the differences between the maps are coloured using the density ρ calculated in each map. We coloured the map in the range of $-4/3 < \Delta\rho/\rho_{\rm GR} < 4/3$, which is the maximum gravity enhancement in units of the density ρ. As we define $\Delta\rho \equiv \rho_{\rm F5} - \rho_{\rm GR}$ positive values (red colour) indicates the enhancement in F5 in comparison to GR, whereas the negative values (blue colour) correspond to haloes with different dynamics and unsecreened haloes. The white inner structure of the large halo correspond to the regime where modified gravity is screened. As the individual densities are the same, hence $\Delta\rho \approx 0$, both GR and F5 models have the same density, and the dynamical equation of modified gravity (Equation 2.5) converges to the Poisson equation of GR. This is a remarkable visual result, as there are several screened haloes that can be identify in the top panel of Figure 2.3. These regions do not have to be confused with those in cosmic voids, here the individual densities are already quite low if not zero, which results in $\Delta\rho = 0$, henceforth these regions are also coloured white. These simulations provide a reasonable understanding on where $f(R)$ gravity has to be probed, focusing in the structure formation, and the observation of unscreened haloes, and avoiding screened regions.

In recent years, several tests have been proposed to constrain the amplitude of the fifth force in $f(R)$ gravity, including using weak lensing in cosmic voids (Cautun et al., 2018), the marked correlation function (Armijo et al., 2018; Hernández-Aguayo et al., 2018) and redshift space distortions (He et al., 2018; Ruan et al., 2022). All these probes exploit the extra information of $f(R)$ gravity caused by the additional fifth force triggered at different scales. In the case of cosmic voids, voids are predicted to be emptier of dark matter in modified gravity than in GR, due to the strongest fifth force in such region. This results in more matter being pushed to the boundaries of the defined voids in $f(R)$, having larger tangential shear than the voids in GR.

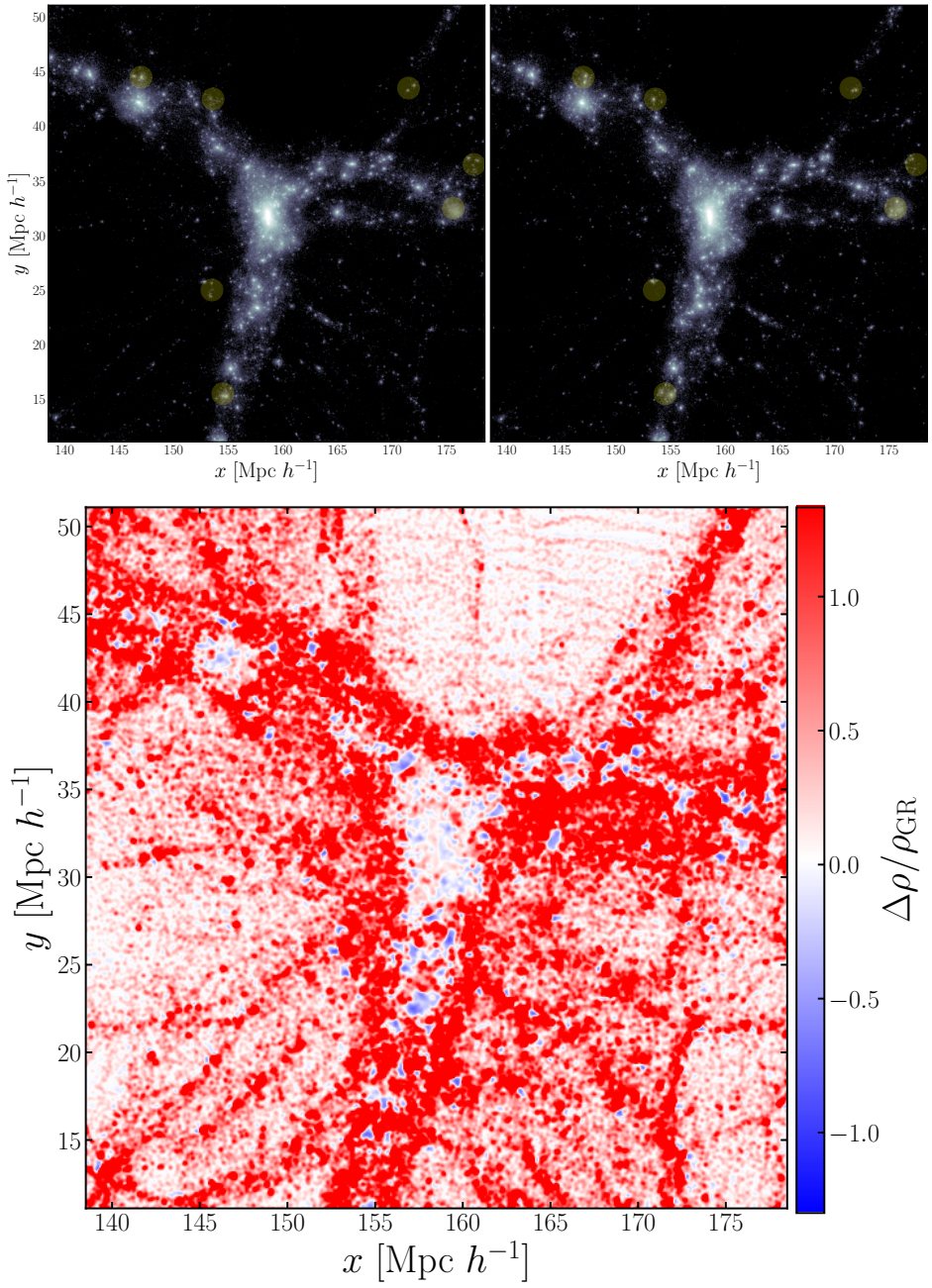


Figure 2.3: Top panel: The dark matter density field from the N-body simulations of GR (left) and MG (right) presented in Arnold et al. (2019). Both images show the density $\log \rho/\bar{\rho}$ of smooth particles in a box of $40 \times 40 h^{-1}$Mpc and a slice of $10 h^{-1}$ Mpc. Some regions have been highlighted to show the different formation of some haloes between the simulations of GR and MG. Bottom panel: The subtraction between the two images in the top panel coloured by $\Delta\rho/\rho_{GR}$, with $\Delta\rho = \rho_{\rm F5} - \rho_{\rm GR}$.

In case of the marked correlation function, this test can used to target properties arising from the fifth force when computing the two-point correlation function. Some options for marking galaxies go from local density and halo masses, both being properties which depend on the modified environment of unscreened regions. For redshift space distortions, improvements have been made in order to describe the effects of peculiar velocities in different cosmologies using more accurate models (Cuesta-Lazaro et al., 2020). These observational probes will take advantage of the upcoming observations from the new generation of surveys such as the Dark Energy Spectroscopic instrument (DESI) (Levi et al., 2013). Whilst, for weak lensing probes, surveys like the Large Synoptic Survey Camera (LSSTCam) of the Vera C. Rubin Observatory (VCO) (Blum et al., 2022), and the Euclid spacecraft mission (Laureijs et al., 2011), will shed new light to constrain or rule out these theories.

The current observational constraints on $f(R)$ gravity parameters correspond to $n = 1$ and $|f_{R0}| \leq 10^{-5}$ (Cataneo et al., 2016; Liu et al., 2016), using the abundance of massive clusters of galaxies and weak lensing peak statistic. More recent constraints using the modified velocity fields from $f(R)$ gravity can put tighter constraints on the fifth force $|f_{R0}| \leq 10^{-6}$ (He et al., 2018), who used redshift space information to compute the two-point clustering on several scales. In our case we decide not to compare directly with the results of He et al. (2018), as we are using the clustering information in the projected space.

Chapter 3

Luminous red galaxies in the Sloan Digital Sky Survey: The LOWZ and CMASS samples

3.1 Luminous red galaxies

The Sloan Digital Sky Survey (SDSS, York et al. (2000)) is a five (broad) band optical imaging and spectroscopic redshift survey. SDSS collected data for about 20 years covering more than 35% of the sky (Gunn et al., 2006), including more than four million of galaxy spectra, to shed light on different problems in astrophysics, including cosmology. Many galaxy samples have been created from spectroscopic SDSS observations. One of the first samples corresponds to a flux-limited sample with $r \sim 17.77$ (Strauss et al., 2002) and a median redshift of $z = 0.1$, called the Main galaxy sample (MGS). This galaxy sample has a mean surface density of 90 galaxies per deg^2 and is used to measure many independent modes of the density fluctuations on scales comparable to the peak of the galaxy power spectrum. An extension of MGS, correspond to fainter ($r \sim 19$) galaxies, which have intrinsically redder colours and higher redshift. These are early-type galaxies which meet a colour-magnitude selection and are called luminous red galaxies (LRG). The LRG

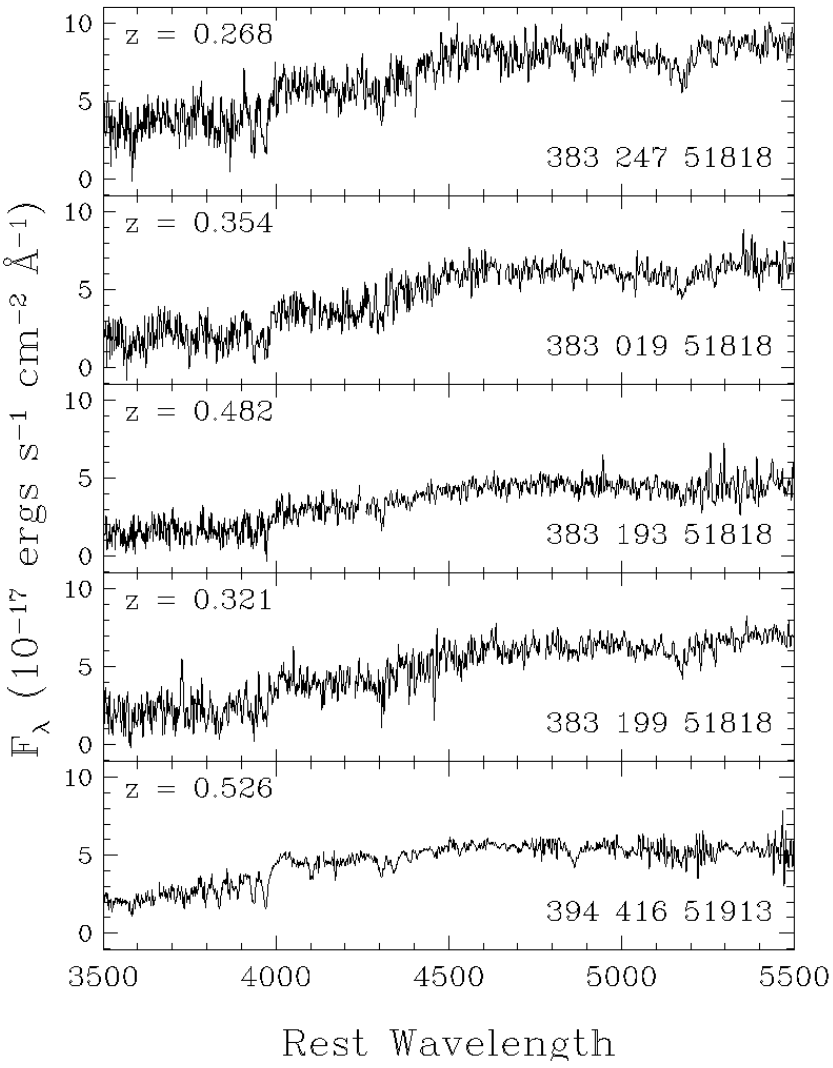


Figure 3.1: Rest-frame spectra of 5 LRGs from Eisenstein et al. (2001). The main feature of these spectra is the well defined 4000Å break.

spectra are relatively high signal-to-noise and clearly show a strong 4000Å break which is indicative of old stellar populations (Eisenstein et al., 2001). In Figure 3.1 we display some of the spectra selected in the LRG sample from Eisenstein et al. (2001). Each individual spectrum shows similar features present in LRG, where the different redshift values in each panel show the distinctive features in the LRG spectrum.

The luminous red galaxies (LRG) are intrinsically red, bright galaxies which are used to trace the large-scale structure efficiently over a large volume of the Universe (Eisenstein et al., 2001). LRGs are relatively passive-evolving early-type galaxies, selected up to redshift $z \approx 0.5$ which can be found in dense environments, such as large groups of galaxies and rich galaxy clusters (Postman et al., 1995). These samples also provide information about the evolution of elliptical galaxies in dense environments (Burke et al., 2000). The sample selection is based on the SDSS-I photometric colours. The 4000Å break provides a sharp feature that can be modelled by the galaxy spectral energy distribution (SED) templates to infer the redshift. The selection of LRG depends on color, for galaxies up to redshift $z = 0.38$, SDSS colours $g - r$ and $u - g$ are used to break the degeneracy between the position and strength of the 4000Å break, which can be used to obtain the selection redshift of the galaxy. For LRG samples at higher redshift, the $g - r$ and $r - i$ color space is used instead, because the $u - g$ colour is close to the redshift limits, leading to a noisy distribution of colours.

3.2 The baryon oscillation spectroscopic survey BOSS

After the SDSS-I LRG catalogue built by Eisenstein et al., new efforts were made to target fainter LRG samples at higher redshifts. In SDSS-III (Eisenstein et al. (2011)), spectra of 1.5 million galaxies spread over 10,000 $\deg^2$, with a magnitude limit of $i = 19.9$ were obtained up to redshift $z = 0.7$. As explained in Section 3.1, the selection of LRGs is colour dependent, which leads to the production of two

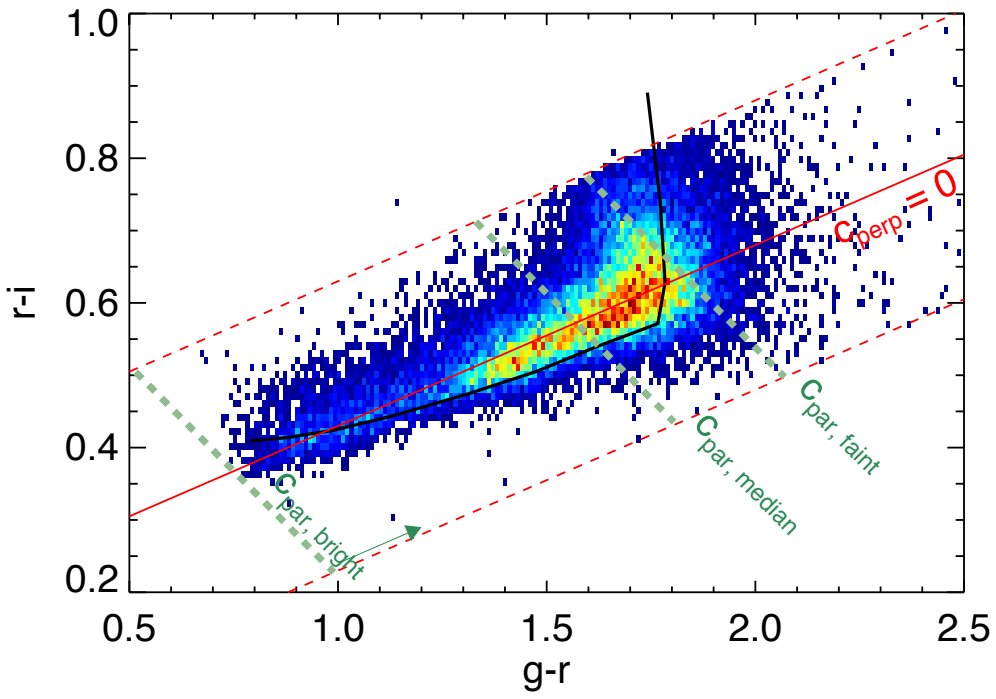


Figure 3.2: Colour-space density plot for LOWZ galaxies at low redshift from Reid et al. (2016). The diagram shows the color distribution in the $(g-r, r-i)$ plane detailing the selection with the definitions of Eqns. 3.1, 3.2 (red dashed lines). See the text in Reid et al. for more details.

independent samples: LOWZ for the lower redshift galaxies and CMASS for the higher redshift targets. The selection of LOWZ is defined by the following colour cuts:

$$c_{\parallel} = 0.7(g-r) + 1.2(r-i-0.18), \tag{3.1}$$

$$c_{\perp} = (r-i) - (g-r)/4.0 - 0.18, \tag{3.2}$$

where the colour bands are defined using the SDSS-III model magnitudes. The defined colours of Eqns. 3.1, 3.2 are simply rotations in the colour plane of the SDSS filters. This is pictured in Figure 3.2, where the density of galaxies in the colour plane shows a high surface density. Then, the selection is performed in terms of $c_{\parallel}$ and $c_{\perp}$. The peak in the distribution is located around $c_{\perp} = 0$ and the selection is extended up to $|c_{\perp}| < 0.2$, whereas $c_{\parallel}$ controls the selection given the galaxy brightness (green dashed lines). Thus fainter objects must be redder to pass the

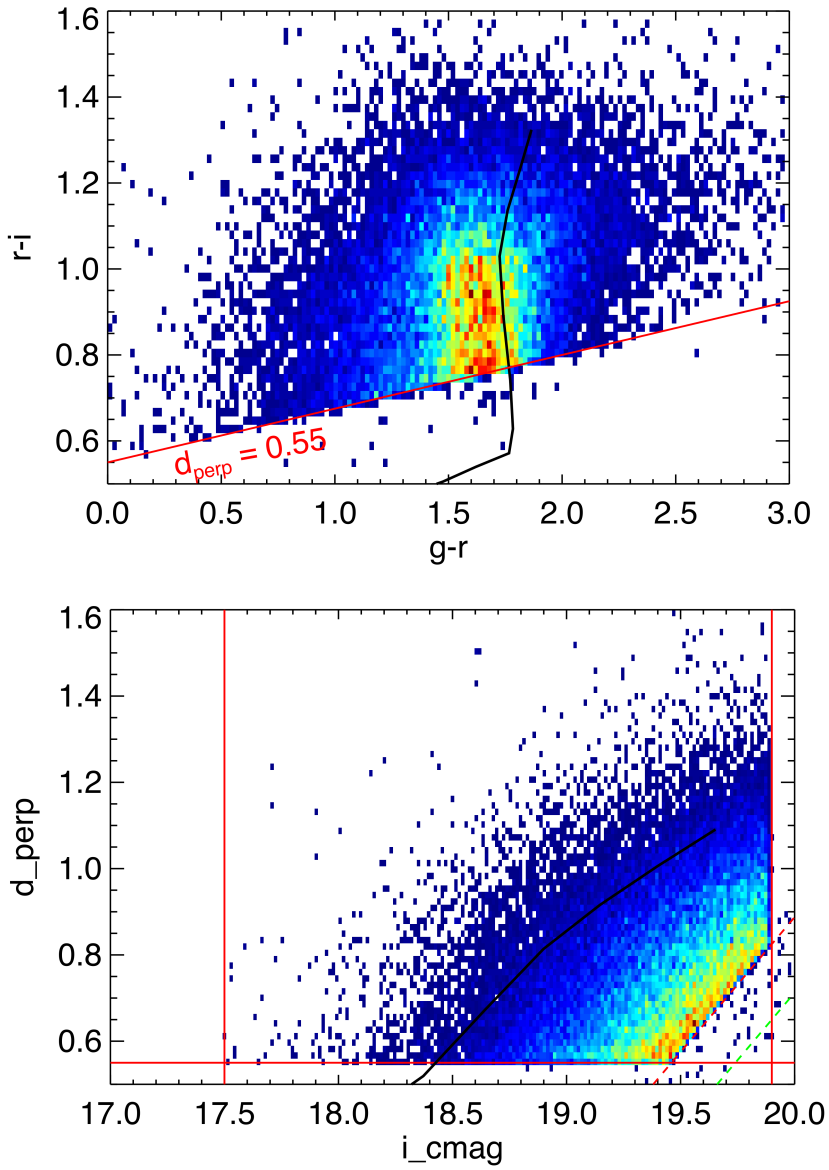


Figure 3.3: Colour-space density plot for SDSS-III CMASS galaxies from Reid et al. (2016). Top: The distribution in the colour-plane with a selection for the higher redshift sample with $z \gtrsim 0.4$. The additional colour selection (red line) is defined by the $d_\perp$ parameter. Bottom: The sliding cut in $d_\perp$ with the i-band magnitude. The color and magnitude cuts implemented for the samples are shown by the red lines in each axis See the text in Reid et al. for more details.

cut (Reid et al., 2016). The high signal-to-noise ratio in the clustering signal aimed for by SDSS produces a relatively high density sample ($\sim 3\times10^{-4}\,h^3\,\mathrm{Mpc}^{-3}$) which is ideal for studying the large-scale structure using biased tracers (Kaiser, 1986). For CMASS an additional colour-magnitude cut using the i-band is included to

select an approximately constant stellar mass limit over the CMASS redshift range $0.4 < z < 0.7$. In Figure 3.3 the selection of galaxies in the CMASS samples is displayed. The $d_\perp$ is defined in order to discard low-redshift galaxies from the color selection, by choosing

$$d_\perp > 0.55. \tag{3.3}$$

Then, an extra sliding color-magnitude in the i-band is introduced, according to the passively evolving model of Maraston et al. (2009):

$$i < \min\left(19.86 + 1.6(d_\perp - 0.8), 19.9\right). \tag{3.4}$$

The CMASS sample has an additional constant stellar threshold which includes bluer galaxies than those in LOWZ. This is derived from fitting stellar population models to the SDSS, and increases the number density of galaxies above $z = 0.4$ (see Tojeiro et al. (2012) for a more detailed description of the CMASS sample).

The LRGs in SDSS-III have the same color selections as the LRGs in SDSS-I, but have a fainter magnitude cut, which increases the number density of galaxies at least by a factor of two (Ross et al., 2012). The sample is part of the Baryon Oscillation Spectroscopic Survey (BOSS) (Dawson et al., 2013), which was designed to improve the measurements of the baryon acoustic oscillation (BAO) scale that used the SDSS-I data from Eisenstein et al. (2001). The new samples from Eisenstein et al. (2011) obtain a more accurate measurement of the BAO (Anderson et al., 2012), than previous studies (Eisenstein et al., 2005; Percival et al., 2007) and at higher redshift, which can be used to probe cosmology by measuring the angular diameter distance $d_A(z)$ (Anderson et al., 2014), for a spherically averaged measurement.

Figure 3.4 shows the footprint area of the BOSS sample (Reid et al., 2016), focusing on the north galactic cap (NGC) of both samples which is the data used in this thesis. We decided to use only the NGC of both LOWZ and CMASS samples, instead of using NGC+SGC for practical convenience: as these correspond to different areas on the sky, we need to consider these as different surveys, with different

photometric properties and potentially different systematic errors. Furthermore, the NGC sample covers twice the solid angle of the SGC one.

3.3 Characteristics of the LOWZ and CMASS samples: number density and projected correlation function

The aim of the BOSS survey is to measure the BAO distance scale by efficiently mapping the large-scale structure of the Universe over a large volume (Anderson et al., 2012). As explained in Section 1.3, a quantitative measurement of the clustering of the large-scale structure can be made using either the power spectrum or the two-point correlation function, which contain the same information, but with a different emphasis. For example the use of $P(k)$ is focused on understanding the formation of structures at the largest scales, where the linear regime equations can be used. Additionally, the covariance matrix for such scales is almost diagonal due to the different modes k evolving independently. In the case of $\xi(r)$, the analysis is more focused on galaxies and how they cluster, where the smaller scales play a more important role.

We plan to make measurements of the marked correlation function by adding properties of individual galaxies to the clustering estimator. For the galaxy samples we use, the one- and two-point statistics are already well known, which means that the models of gravity we test need to replicate these measurements. There are two basic tests we can make to characterise the sample: 1) calculate the abundance of galaxies per unit volume, and 2) measure their two-point clustering. The first one is the galaxy number density, and as briefly described in Section 3.1, it is a function of redshift. In Figure 3.5 the number density $n(z)$ as a function of redshift z is shown for the two samples used in this study, LOWZ and CMASS. Here we can see how the distribution of galaxies varies as we go to higher redshift. The colour

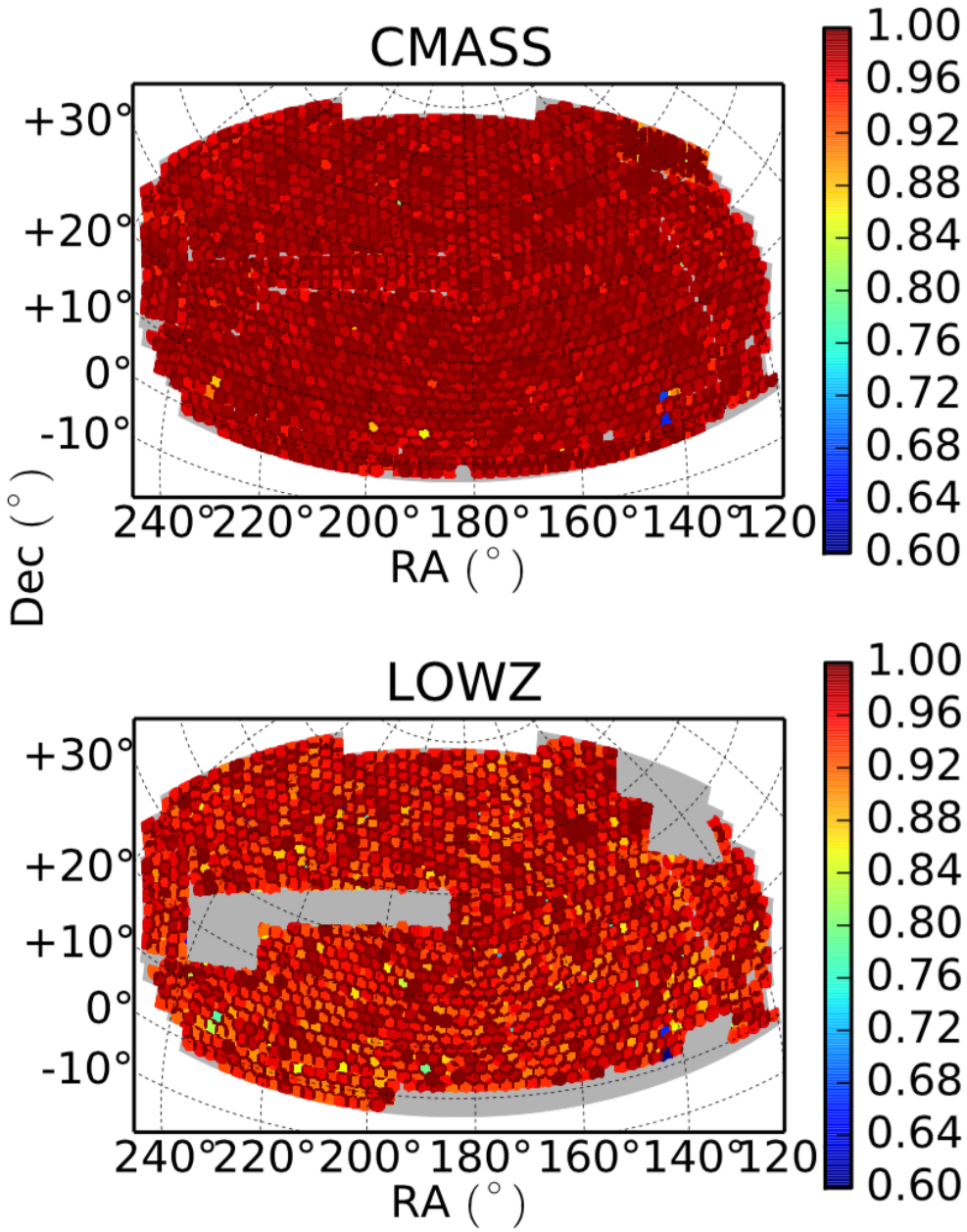


Figure 3.4: The angular coverage footprint of BOSS DR12, reproduced from Reid et al. (2016) catalogue showing the spectroscopic redshift completeness, which is the ratio of the number of galaxies with $z_{\text{spec.}}$ to the number of galaxies in the target catalogue. Individual patches corresponds to a plate with fibres measuring the redshift of target galaxies, coloured from blue to red by the overall completeness of that plate (the higher completeness goes to redder colours in the colour bar). The survey is divided into 2 samples with different areas and redshift ranges, with the LOWZ sample (bottom panel) at $0.10 < z < 0.43$ and CMASS (top panel) $0.43 < z < 0.7$.

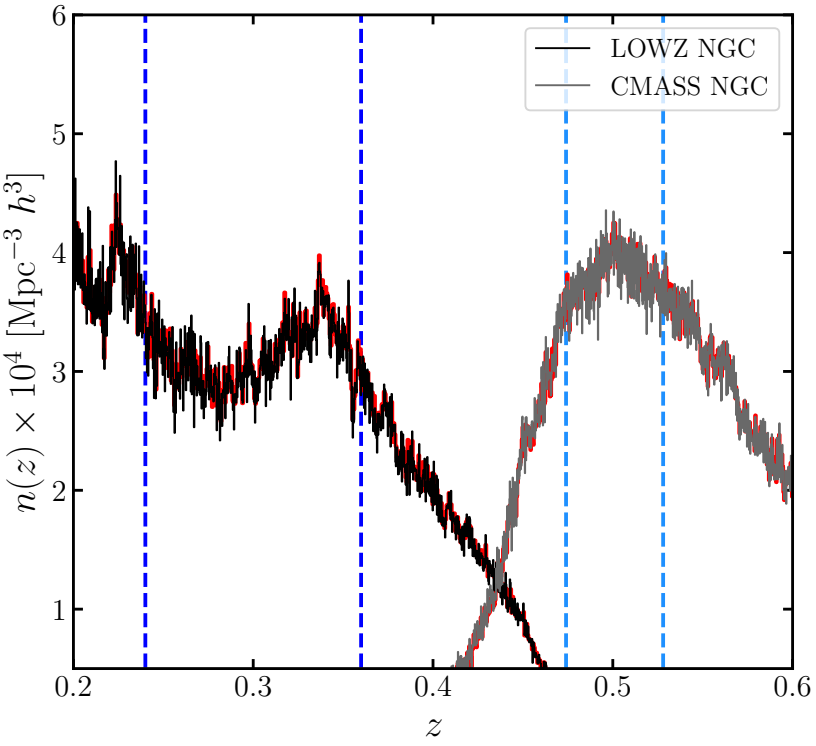


Figure 3.5: The galaxy number density $n(z)$ as function of redshift z for the BOSS DR12 NGC data. LOWZ (black) and CMASS (gray) samples have different selection functions which lead to different curves for $n(z)$. We also plot the distribution of the random galaxy catalogue (red) from Reid et al. (2016), used for clustering analysis, and the subsample selection for this study LOWZ $0.240 < z < 0.360$ (blue dashed line) and CMASS $0.474 < z < 0.528$ (light blue dashed line).

selection of the BOSS defines these samples at $0.1 < z < 0.43$ and $0.4 < z < 0.7$ for LOWZ and CMASS respectively. The NGC footprints of both surveys cover more than half of the survey (5836.21 deg^2 for LOWZ and 6851.42 deg^2 for CMASS). After $z > 0.35$, where the LRG distribution peaks, the LOWZ sample starts to decrease its number density systematically, due to the colour selection. In CMASS the selection changes, allowing LRGs to be detected at $z > 0.4$, with the additional stellar-mass threshold described in Eqn. 3.4, which leads to a number density even higher than LOWZ in average, with a peak at $z = 0.5$. After this, the number density of CMASS also starts decreasing up to redshift $z = 0.7$.

3.3.1 Galaxy number density

In the ideal case, given a cubic volume of galaxies, the galaxy number density is just a number represented by $n_{gal} = N_{gal}/V_{box}$, with N_{gal} the number of galaxies and V_{box} the volume of the box. The case of the observational data is different. In Figure 3.5 we see the dependence of n with redshift z, due to the selection function. This means that we are introducing new dependencies in the properties that depend on the number density when we compute the marked correlation function. To avoid this problem, we consider the total number density of the survey as the number of galaxies divided by the total volume $n_{obs} = N_{gal}/V_s$. Moreover, a more restricted volume is selected for both samples for which there is less variation in number density, which reduces the variations when computing the clustering and marked clustering. The dashed lines in Figure 3.5 show the redshift limits of these new subsamples, which define the new ranges $0.240 < z < 0.360$ for LOWZ and $0.474 < z < 0.528$ for CMASS. By using this selection, our aim is to achieve a sample with a roughly uniform number density within the full redshift range to test. We compare the samples with simulations of roughly the same volume when we create the mock catalogues.

3.3.2 Galaxy-galaxy two-point correlation function

Once we have selected the redshift range of the subsamples, the next step is to estimate the clustering of galaxies at different scales. As described in Section 1.3 the two-point correlation function can be computed as the excess of probability of finding a pair of galaxies at a given separation in comparison with a random distribution of points. Throughout this study, we measure the clustering using the projected correlation function w_p, which is the integral of the two-point correlation function $\xi(r_p, \pi)$, binned in distance along r_p, for the projected perpendicular distance, and π, the line-of-sight or parallel distance. The integral of $\xi(r_p, \pi)$ is over the line-of-sight parallel direction π, resulting in the clustering as a function

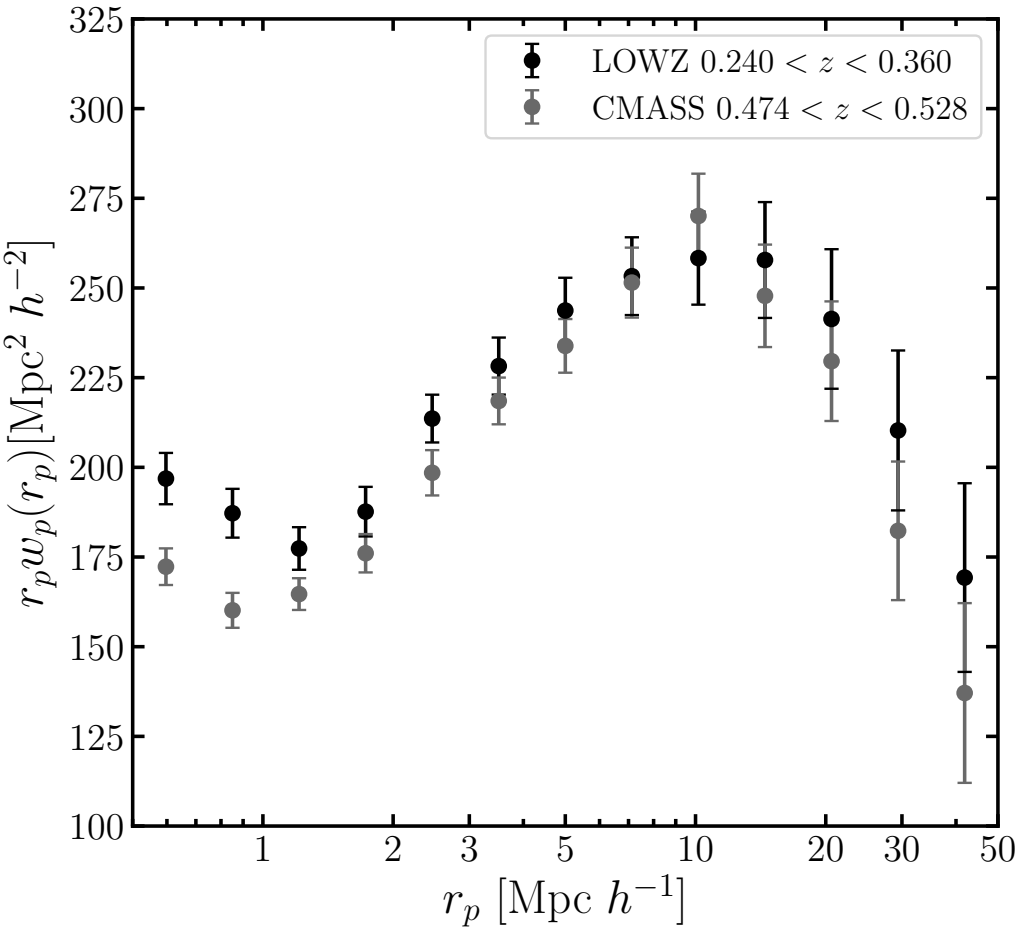


Figure 3.6: The projected two-point correlation function w_p scaled by r_p as a function of the projected perpendicular distance r_p for BOSS DR12 NGC. The clustering is calculated for the selected subsamples of LOWZ (black dots) and CMASS (gray dots) and scaled by r_p. Error bars are calculated using Jackknife resampling over 100 Jackknife regions (e.g Norberg et al. 2009).

of the perpendicular distance r_p only. In the distant-observer approximation, w_p is related to the correlation function in real space (Norberg et al., 2002), which simplifies the analysis. The correlation function binned in a two-dimensional grid is selected instead of the redshift space two-point correlation function $\xi(s)$ to avoid the influence of redshift space distortions in our results for the scales we are interested, which can complicate the prediction of the marked correlation function on such scales. We avoid measuring the correlation function and marked correlation function in redshift space to avoid the problems encountered by Satpathy et al. (2019), in which the marked correlation function of LOWZ is presented in redshift

space. These authors conclude that their results are driven by the limitations of modelling the clustering on such scales.

The correlation function can be computed using the Landy-Szalay estimator (Landy et al., 1993):

$$\xi(r_p, \pi) = \frac{DD - 2DR + RR}{RR}, \tag{3.5}$$

where DD, RR, DR are the normalised number of data-data, random-random and data-random pairs respectively for each separation bin. These terms are also normalized by the number of galaxies and randoms of the samples, and weights are added to address the systematic effects of the survey, including FKP weights (Feldman et al., 1994) which decrease the variance of the correlation function when taking into account the radial variation in number density, which can change strongly with redshift (he normalisation takes into account that the overall number of random points can be many times higher than the number of galaxies to reduce noise in the clustering estimation). For the case of the LOWZ and CMASS samples, these values are almost constant, as the number density $n(z)$ does not vary drastically as a function of redshift for our z range selection. To calculate the projected correlation function and obtain the clustering signal in real space we can integrate $\xi(r_p, \pi)$ in the π-direction:

$$\frac{w_p}{r_p} = \frac{2}{r_p} \int_0^\infty \xi(r_p, \pi) d\pi. \tag{3.6}$$

As we are not solving this integral analytically we bin $\xi(r_p, \pi)$ until π_{max}, which is a value chosen when the integral is converging to a stable value (Parejko et al., 2013). Considering the range of scales we are interested in, we choose $\pi_{max} = 80h^{-1}$ Mpc, after checking that the result of the w_p clustering converges around $\pi = 70h^{-1}$ Mpc. In Figure 3.6 we plot the results for the projected correlation function as a function of the perpendicular distance r_p on scales between $0.5 < r_p/(\text{Mpc } h^{-1}) < 50$ for both samples LOWZ and CMASS. Both correlation functions show similar features, with a small offset due to the different number

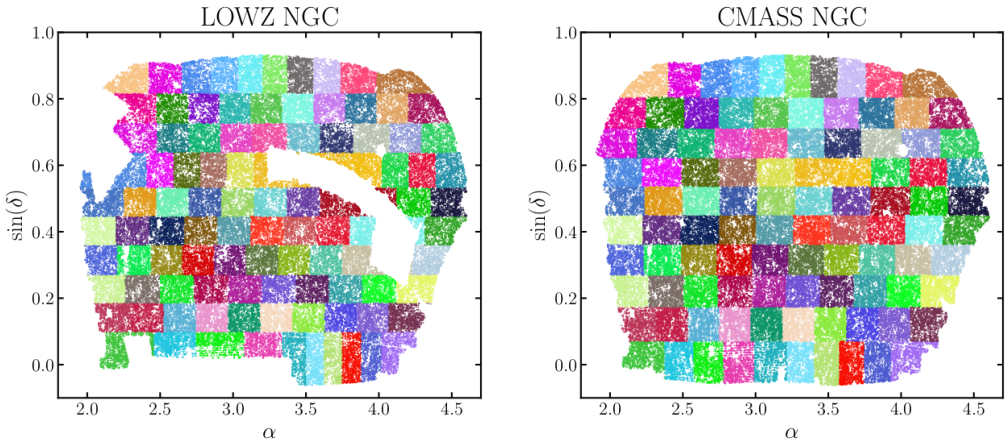


Figure 3.7: The footprint of LOWZ (left panel) and CMASS (right panel) samples including the Jackknife regions for the uncertainties in the clustering analysis. All colour regions have roughly the same area to create the resampling of the data.

densities that the subsamples have. We use the Jackknife re-sampling method to compute the uncertainties on the measurements of w_p. In Figure 3.7 we show the implementation of the resampling, showing the Jackknife areas over the BOSS NGC footprint. We used 100 jackknife areas, where each individual patch has an area of $A_{\rm JK} =\sim 60\,{\rm deg}^2$. Each different realization gives a similar result for w_p with a small variance, which is related to the Jackknife resampling uncertainties. To compute the error bars on w_p, the Jackknife method consists of calculating w_p 100 times omitting one of the regions each time. We end up with 100 calculations of w_p with a covariance matrix defined by

$$C_{ij} = \frac{N-1}{N}\sum_{k=1}^{N}\left(\xi_i^k - \bar{\xi}_i\right)\left(\xi_j^k - \bar{\xi}_j\right), \tag{3.7}$$

where ξ_i is the i-th measurement of the correlation function using the different Jackknife areas, and $\xi_i \equiv \sum_{k=1}^{N}\xi_i^k/N$. The covariance includes the factor $N-1$, which takes into account the dependency of the $N-1$ copies when doing the resampling, as only two areas change for each resampling. Jackknife errors are expected to represent the correct variance at large-scales $r_p > 10h^{-1}$ Mpc but to overestimate it on small scales $r < 2.5h^{-1}$ Mpc according to the tests carried out by Norberg et al. (2009). A better estimate for the errors on such scales can be

obtained by using the covariance matrix from an ensemble of a large number of N-body simulations. However these methods are still computationally expensive for the separation scales we want to resolve.

The projected correlation function in BOSS has been calculated before for earlier data releases (White et al., 2011; Parejko et al., 2013) including studies of weak lensing and intrinsic alignments (Singh et al., 2015) using LOWZ data. All these studies compute w_p for a similar redshift range and scales to the ones showed in Figure 3.6. White et al. divide the samples between low and high-redshift at $z = 0.55$ and a subsample at $0.4 < z < 0.7$, whereas Parejko et al., and Singh et al. use only LOWZ data in their analysis. Amplitudes and uncertainties agree between the different studies, but there are some differences between the error bars; however it is worth noting that all studies use different methods for these calculations. As White et al. analyse an early release of the BOSS-LRG data, commissioned for SDSS-II (Abazajian et al., 2009), they prefer to use Poisson realizations of their simulations, arguing that the area that this data covers is not enough to perform resampling methods. While Parejko et al. also use errors based on their mock catalogues from simulations, and use the standard deviation of 20 different galaxy correlation functions. Finally Singh et al. use Jackknife resampling for the uncertainties in their clustering data, which is the same method that we use. The amplitude of the error does not vary significantly between (<5%) the different studies, but it is important to mention that all papers use earlier data releases (DR9, DR11) than the one used on this thesis (DR12). The calculations of both the number density and the clustering of BOSS data will permit us to create mock catalogues, which will be used to better understand the systematic errors associated with the data and to predict the results of the marked correlation function for MG models.

Previous studies using the the projected-correlation function find robust results for the clustering and number density of galaxies, which tell how the large-scale structure can be traced by the LRG population. Such results, can be interpreted

as an observational constraint and need to be replicated when we create mock catalogues from different simulation models.

Chapter 4

N-body simulations of modified gravity: Making use of sub-resolution haloes

†

4.1 Introduction

The mass resolution limit of dark matter halo catalogues extracted from N-body simulations is often set to satisfy a range of requirements and, as a result, can appear unnecessarily conservative for some applications. The measurement of the internal properties of halos is challenging and requires that objects are resolved by several hundred particles. For example, Bett et al. (2007) demonstrated, using the Millennium simulation of Springel et al. (2005), that at least 300 particles are needed to measure halo spin robustly. On the other hand, many authors have used the same simulation to build semi-analytical galaxy formation models retaining halos down to 20 particles (e.g. Croton et al. (2006)), extending the mass resolution of the halo catalogue by more than an order of magnitude for this purpose,

†The text in this Chapter has been taken verbatim from Armijo et al. (2022)

compared with that used to measure halo spin.

Here we revisit how the mass resolution limit of a dark matter halo catalogue is set for use in a simple clustering study. The application in this case is to use the halos to build a galaxy catalogue, for example using a halo occupation distribution model (HOD) or a semi-analytical galaxy formation model (SAM) to populate the halos with galaxies. The resulting 'mock' galaxy catalogue will be compared to an observed sample, with the criteria for success being that the mock reproduces the abundance and clustering of the target sample to within some tolerance. Typical galaxy samples occupy a broad range of halo masses. If we impose an unduly restrictive mass limit on the halo catalogue that can be used from a simulation, this could result in the simulation not being suitable to probe a wide range of the parameter space in the HOD or SAM for a given galaxy selection. We judge the halo catalogue to be useful if it can be employed to reproduce the abundance and clustering of halos that would be measured in a higher resolution simulation; we show that this can be achieved for halos that are made up of a perhaps surprisingly low number of particles by employing a simple weighting scheme.

Here we address two issues relating to the use of simulated halos in clustering studies. The first is to devise a robust and reproducible way to determine the mass resolution limit of a halo catalogue extracted from an N-body simulation for a clustering study. The second is to see if we can still use the halos below this resolution limit in a clustering analysis, which, as we shall see, represent a fraction or subset of the true population of halos at these masses. As these halos are deemed to be below the mass resolution limit we have set, these 'sub-resolution' haloes will be treated in a different way to the resolved halos. We will show that considering the sub-resolution halos allows us to extend the useful dynamic range of the simulation by a factor of 10 below the formal resolution limit, so long as we are willing to tolerate some error in the clustering predictions. We describe our clustering analysis as simple since we do not consider secondary contributions to halo clustering besides mass; the halo resolution needed to use internal halo properties to build assembly

bias into mock catalogues has been discussed by Ramakrishnan et al. (2021).

4.2 The N-body simulations

We use three simulations of the standard cold dark matter cosmology with different mass resolutions. We mainly focus on two simulations from Arnold et al. (2019), but also consider the halo mass function from the P-Millennium Baugh et al. (2019). The simulations from Arnold et al. each use 2048^3 collisionless particles in cubic boxes of length $L_{\rm box} = 768h^{-1}\,{\rm Mpc}$ and $1536h^{-1}\,{\rm Mpc}$, resulting in particle masses of $M_{\rm p} = 4.9 \times 10^9$ and $3.6 \times 10^{10}h^{-1}{\rm M}_\odot$, respectively. Both simulations use the Planck cosmological parameters (Planck Collaboration et al., 2016): $h = 0.6774$, $\Omega_{\rm m} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.8159$, and $n_{\rm s} = 0.9667$. We use the simulation outputs at redshift $z = 0$. The P-Millennium run uses very similar but slightly different cosmological parameters to the above (e.g. $\Omega_{\rm M} = 0.307$; see Table 1 of Baugh et al. (2019)). The simulation box size in this case is $L_{\rm box} = 542.16h^{-1}{\rm Mpc}$ with the dark matter traced by 5040^3 particles, resulting in a particle mass of $1.08 \times 10^8h^{-1}{\rm M}_\odot$. The simulations were run with slightly different versions of the Gadget code (for the most recent description see Springel et al. (2020)). We henceforth refer to the Arnold et al. runs by their box lengths, as L1536, and L768. The L1536 and L768 runs form a sequence in mass resolution completed by the P-Millennium which has the best mass resolution.

Haloes are identified using SUBFIND Springel et al. (2001). The first step in this algorithm is to run the friends-of-friends (FoF) percolation scheme on the simulation particles. We set the minimum number of particles per group to be retained after the FoF step to be 20. SUBFIND then finds local density maxima in the FoF particle groups, and checks to see if these structures are gravitationally bound; these objects are called subhalos. Particles that are not gravitationally bound to the subhalo are removed from its membership list. The mass of the subhalo is obtained using the spherical overdensity (SO) method (Cole et al., 1996).

The SO method is applied to the gravitationally bound particles in the subhalo to find the radius within which the average density is 200 times the critical density of the universe. The halo mass, M_{200c}, is the sum of the particle masses within this radius. This results in some subhalos having masses with $M_{200c} < 20M_{\rm p}$, because small groups tend to be ellipsoidal in shape rather than spherical. We consider halo samples composed of main subhalos, i.e. the most massive subhalo within each FoF group.

4.3 The halo mass function and simulation resolution

We now look at the considerations that go into setting the mass resolution of the SUBFIND halo catalogues, by comparing the main subhalo mass functions measured in the different resolution simulations.

Fig. 4.1 compares the mass functions measured from the L1536 and L768 simulations from Arnold et al., with that obtained from the P-Millennium. To account for the very slightly different cosmology used in the P-Millennium, we generated analytic mass functions for the cosmologies used by Arnold et al. and Baugh et al. These analytic mass function are offset, and can be reconciled by applying a constant rescaling to the P-Millennium halo masses. After this correction, the differential mass functions measured from the three simulations agree with one another very well at high masses (i.e. for masses above a few times $10^{13}h^{-1}\mathrm{M}_{\odot}$), with some fluctuations at very high halo masses which arise due to sample variance. The lower panel of Fig. 4.1 shows the fractional difference of the mass functions with respect to that measured from the P-Millennium. The scheme we set out below depends on the comparison between the halo mass functions from the L1536 and L768 runs.

The green vertical dashed line in Fig. 4.1 shows a halo mass corresponding to 100 particles in the L1536 simulation, i.e. $3.6 \times 10^{12}h^{-1}\mathrm{M}_{\odot}$. At this mass, there is already a clear difference in the mass functions measured from the two simulation

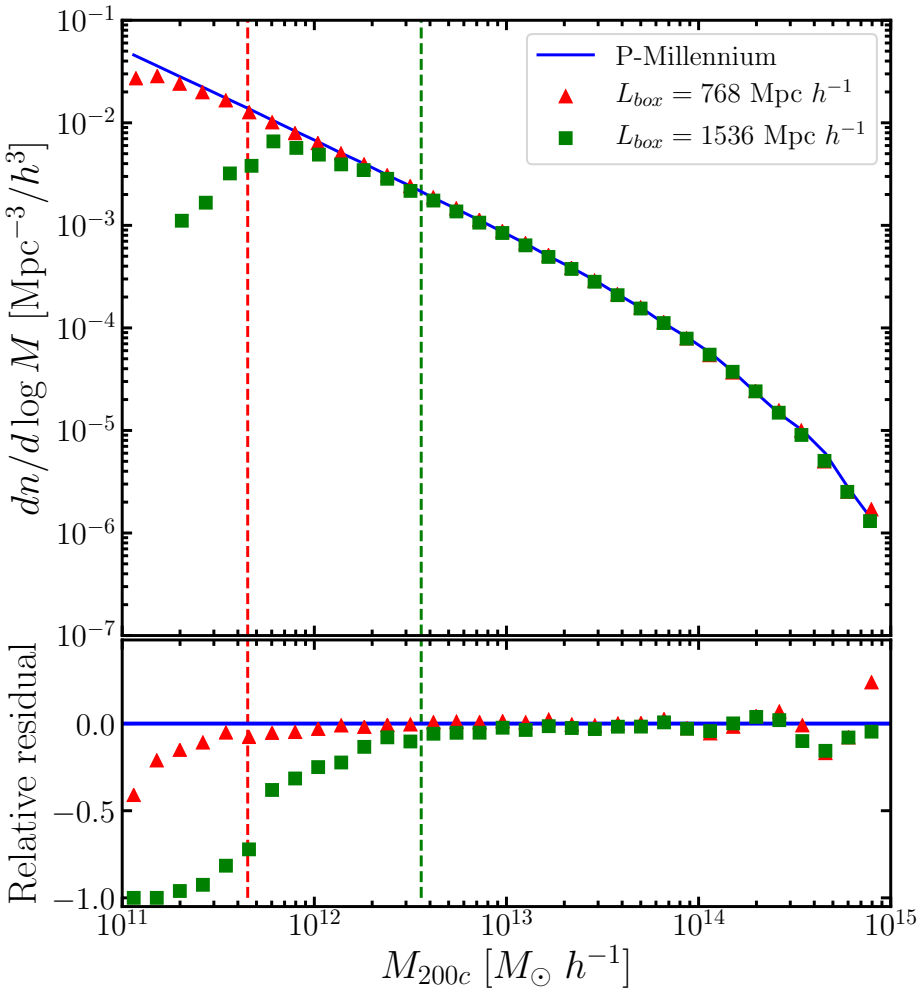


Figure 4.1: The differential halo mass function at $z = 0$. Top: results from the P-Millennium Baugh et al. (2019) (blue line), and the ΛCDM N-body simulations of Arnold et al. (2019) (points); red triangles show the mass function measured from the L768 simulation and the green squares show the L1536 run. The vertical dashed lines indicate a halo mass of 100 particles for the L768 (red) and L1536 (green) resolution runs. Bottom: fractional difference expressed relative to the P-Millennium halo mass function. A small correction has been applied to the masses in the P-Millennium mass function to account for the slightly different cosmological parameters used in this run and in Arnold et al. (see text for details).

boxes. To quantify these differences, above a mass threshold of $10^{13}h^{-1}\mathrm{M}_{\odot}$ there is already a 3 per cent deficit in the cumulative abundance of halos in the lower resolution L1536 run compared with the higher resolution L768 one; this rises to 12 per cent for a mass threshold of $10^{12}h^{-1}\mathrm{M}_{\odot}$ and 32 per cent for a mass limit of $4 \times 10^{11}h^{-1}\mathrm{M}_{\odot}$. We have checked that the difference in the slope of the mass function between the L1536 and L768 runs is due to the difference in mass resolution rather than sample variance in the smaller-volume/higher-resolution box by measuring the mass functions from the larger volume simulation after splitting it into eight smaller subvolumes, each equal in volume to that of the L768 run. We found that there is remarkably little variation in the slope of the mass function around $10^{13}h^{-1}\mathrm{M}_{\odot}$ due to sample variance.

Fig. 4.1 shows that moving to masses below 100 particles in L1536, there is a sudden drop in the number of halos recovered in the L1536 run compared to the L768 run around $10^{12}h^{-1}\mathrm{M}_{\odot}$. The red vertical dashed line is equivalent to 100 particles in the L768 run. The question of determining the halo mass resolution of the simulation can therefore be framed in terms of the tolerance for errors in the statistic of interest. If the halo mass function is of primary interest, then if we treat the L768 simulation as the reference or 'gold standard', we could choose the resolution limit of the L1536 run as being 100 particles, in the knowledge that this gives us a $\sim$5 per cent underestimate of the cumulative abundance of dark matter halos compared to the L768 run; if we require a better reproduction of the cumulative halo abundance, then we would need to apply a mass limit greater than 100 particles. If our interest in the halos is broader and extends to clustering then we also need to assess the errors made in statistics such as the two point correlation function. It is possible, however, as we demonstrate in the next section, to extend the useful resolution of the simulation by applying a simple weighting scheme to the halos when computing their abundance and clustering.

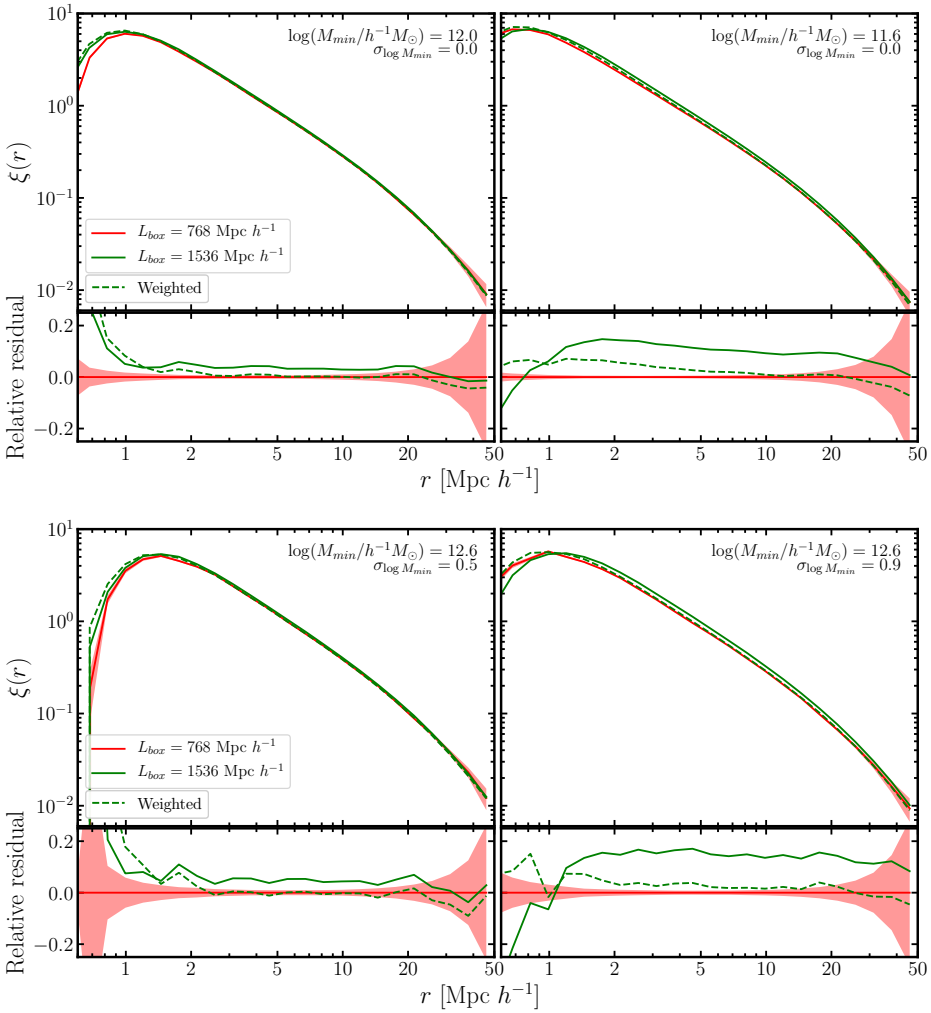


Figure 4.2: The correlation function measured in the HR (red) and LR (green) runs for subhalo samples defined by sharp lower mass cut (left and centre-left panels, corresponding to $\sigma_{\log M} = 0$) and by a HOD-style, more gradual mass cut (centre-right and right panels, defined by $\sigma_{\log M} > 0$; see Eqn. 1). For the correlation functions measured from the LR run, the solid lines shows the unweighted estimate and the dashed lines the weighted case. The lower panels show the fraction difference in the correlation function, relative to the HR measurement. The pink shading shows the error on the correlation function estimated by jackknife resampling.

4.4 Extending the resolution of the simulated halo catalogue

Typically, a clustering study involves using the number density of objects and the two-point correlation function estimated from an observational sample to constrain a model, such as setting the parameter values in an HOD model. Here, we present a simple weighting scheme that extends the resolution limit of a simulated halo catalogue down to lower halo masses than are generally considered for use in clustering analyses. The scheme returns the exact abundance of clustering tracers, by construction, and yields a more accurate prediction of the two-point clustering to within some tolerance. The procedure used to derive the resolution limit is transparent and reproducible.

The weighting scheme is remarkably simple. A weight is defined in bins of halo mass such that the differential mass function of the L1536 simulation agrees with a reference mass function; here we use as the reference mass function the measurement from the L768 simulation. For mass bins in which the unweighted halo mass function in the L1536 run is below the target mass function, halos are assigned a weight greater than unity. By applying this weight to the halos in the L1536 run, the new, 'weighted' mass function agrees with the target mass function exactly by construction. In practice we set the weights to unity above some mass, e.g. $5 \times 10^{13} h^{-1}\mathrm{M}_{\odot}$, to avoid being affected by fluctuations at high masses in the mass function measured from the L768 run due to sample variance. The limiting factor which sets the new resolution limit of the weighted halo catalogue is the error that we are prepared to tolerate on the halo clustering.

With the weighting scheme, the halo correlation function is estimated by including the weight assigned to each halo in the pair count. As a first simple illustration we consider the clustering of samples of main subhalos defined by different lower mass thresholds in the top panels of Fig. 4.2. These samples are equivalent to

central galaxies in a simple HOD analysis with a sharp transition in the mean occupancy from 0 to 1 central per halo. In each case, the clustering of the halos in the L768 simulation is estimated without applying any weights, i.e. all halos in this case have the same weight of unity, whereas for the L1536 simulation the weights derived from forcing the halo mass function to match that in the L768 run are applied and included in the estimation of the correlation function. Due to the shape of the halo mass function, halos close to the minimum mass that defines each sample contribute importantly to the abundance of halos in the sample and to the clustering.

The top-left panel of Fig. 4.2 shows the clustering measured for a subhalo sample defined by a mass cut of $10^{12}h^{-1}$ $M_{\odot}$, close to which modest weights have been applied in the lower resolution run; for halos in which the weight is *not* unity the average weight applied in this case is 1.15. The clustering measured in the L1536 run for this halo sample, after applying the weights, agrees remarkably well with that measured in the L768 run, being within the estimated errors on the correlation function down to $\sim 3h^{-1}$Mpc. In the case without weights, the clustering measured for this halo sample in the lower resolution run is systematically shifted upwards by around 5 per cent compared to that measured in the higher resolution simulation.

The top-right panel of Fig. 4.2 shows the limit of the performance of our weighting scheme. This halo sample is again defined by a lower mass threshold than the one in the left-most panel. For the halos with a weight greater than unity, the average weight in this example is 1.8. Again, by construction, the weighted sample matches the abundance of halos in the L768 run to better than 1 per cent (the agreement could be further improved by using narrower bins to measure the halo mass function in the mass range where weights greater than unity are derived). The clustering in the weighted sample matches that in the L768 simulation over a reduced range of scales, compared to the other cases, being within the errors down to $10h^{-1}$Mpc. We note that the clustering of the unweighted halos for this sample is 10 to 15 per cent higher than the 'target' measurement from the higher resolution

simulation. If this is the error in the clustering that we are prepared to accept, agreement down to intermediate scales, rising to a 5 per cent excess approaching $\sim 1h^{-1}$Mpc, then the mass resolution of the halo catalogue has been extended down to halos with ~ 11 particles. To put this into context, the abundance of the halo samples starts to deviate between the L1536 and L768 simulations at a mass corresponding to around 550 particles.

As a second example we consider samples that are more comparable to those in HOD analyses, in which the occupation of halos by centrals moves from zero to one per halo more gradually than in the example above. The width of the transition is one of the HOD parameters; $\sigma_{\log M}$. Larger values of $\sigma_{\log M}$ mean that lower mass halos contribute central galaxies to the sample. (Note that we do not consider satellite galaxies in any of our examples; all galaxies within a halo would be assigned the weight of the halo to compute the abundance of galaxies and to estimate their clustering.)

In the popular five parameter HOD model the mean occupation of halos by centrals depends on the parameters $M_{\rm min}$ and $\sigma_{\log M}$ through (Eqn. 1 from Zheng et al. (2005)):

$$\langle N_{\rm cen}\rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right]. \tag{4.1}$$

The bottom panels of Fig. 4.2 show the correlation functions measured from the L768 and L1536 runs for a fixed value of $M_{\rm min} = 4 \times 10^{12}h^{-1}{\rm M}_{\odot}$, varying $\sigma_{\log M}$. The width of the transition from $\langle N_{\rm cen}\rangle = 0$ to $\langle N_{\rm cen}\rangle = 1$ gets broader in mass as the value of $\sigma_{\log M}$ increases. This means that lower mass subhalos are contributing to the correlation function shown in the bottom-right panel of Fig. 4.2 compared to the bottom-left panel. In the bottom-left panel of Fig. 4.2, the transition from all halos being empty to all containing a central is relatively narrow. As $\sigma_{\log M}$ increases, due to the shape of the halo mass function, the number of haloes in the samples increases. The bottom-left panel of Fig. 4.2 shows that applying the weighting scheme allows us to recover the correlation function down to $2.5h^{-1}$Mpc. Again, without applying any weights, the clustering measured in the L1536 box

would be systematically shifted upwards by 5 per cent. For the broadest transition considered, with $\sigma_{\log M} = 0.9$, the weighted correlation function is within a few per cent of the estimate from the higher resolution L768 simulation; without weights the estimate is too high by more than 15 per cent.

We have tested the performance of our method at $z = 1$. In this case, the marginally resolved halos have a clustering bias that is greater than unity and this poses a challenge to the method. In the simplest case, using a mass threshold of $\log M_{min} = 12.0\ h^{-1}\ \mathrm{M}_{\odot}$ to populate haloes with central galaxies, the initial disagreement in the measured clustering is around 6% on scales larger than $1\ h^{-1}$ Mpc. After applying the weighting scheme, this disagreement drops to 3%. The performance of the scheme is less good than at $z = 0$, but still represents an improvement over doing nothing. The situation is similar for the case with $\log M_{min} = 12.6\ h^{-1}\ \mathrm{M}_{\odot}$ and $\sigma_{\log M} = 0.5$. Here the difference in the correlation measured from the simulations without weighting differs by 8% on scales $1 < r/\mathrm{Mpc}\ h^{-1} < 20$. When we apply the weighting scheme, the discrepancy more than halves to a 3% of disagreement. In these two cases we are applying weights with values between 5 and 10 to haloes with around 20 particles.

We end by investigating the incomplete or 'partially' resolved halo population in the lower resolution L1536 simulation. What is special about the subhalos that are picked up by FoF and SUBFIND, at masses for which the subhalo samples in this run are incomplete? We address this by measuring the local environment around halos as a function of mass, by measuring the distribution of counts-in-cells centred on halos, and comparing the measurements between the L1536 and L768 runs. We use cubical cells of side $1.6h^{-1}$ Mpc which sample the density field defined by the dark matter particles. We find that the counts-in-cells distributions around subhalos that are well resolved in each simulation are essentially the same. The difference in shot noise (mean particle density) does not affect the count distributions because centring on a halo biases the counts to high densities. The top and middle panels of Fig. 4.3 contrast the cell count distributions measured in the two simulations

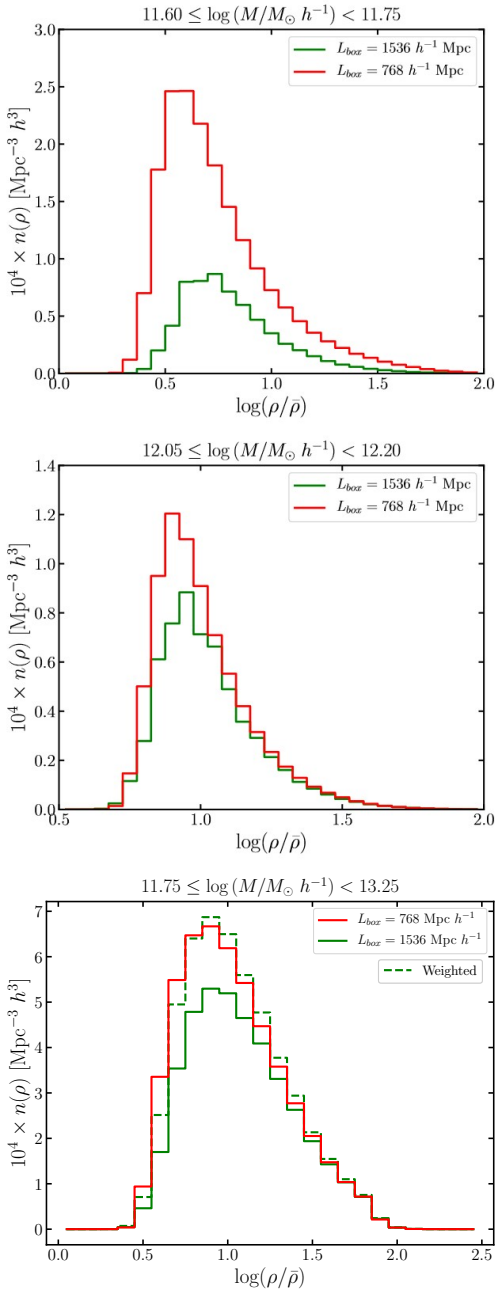


Figure 4.3: The distribution of matter density counts in cells of size 1.6 h^{-1} Mpc centred on halos in the stated mass range, measured from the L768 (red) and L1536 (green) simulations. The difference in volume of the L1536 and L768 runs has been taken into account in the normalisation. The left and central panels show the count in cells distributions for the bins used in the mass function (the bin limits are written at the top of each panel) for which the weights are greater than unity. The right panel shows the distribution of cells for a wider mass range covering all of the bins for which the weights in our scheme are greater than unity. Here the green dashed line shows the distribution of counts-in-cells in the L768 simulation after applying the weights.

in mass bins for which the mass functions are different. As the mass bin shifts to lower masses, the difference between the local environments of the subhalos that are identified changes, with marginally resolved subhalos from the L1536 run tending to be found in higher density environments than the true distribution, according to the measurement from the L768 run. The bottom panel of Fig. 4.3 shows the difference in the local density around subhalos for a sample with a sharp mass cut at $5.6 \times 10^{11} h^{-1} \mathrm{M}_{\odot}$. The mass range shown is that for which greater than unity weights are applied in our scheme. The unweighted cell-count distribution is shown by the solid green line; the weighted distribution, using the weights computed in 10 mass bins, is shown by the green dashed line and is remarkably close to the distribution found in the higher resolution L768 run.

4.5 Summary and Conclusions

Often the mass resolution limit of a simulated halo catalogue is presented as a suspiciously round number, 100+ particles, that may once have been checked but has long since passed into simulation folklore and has become an unquestioned rule of thumb. We have argued that for some studies, for example simple clustering analyses, such limits are overly conservative as we are not interested in quantities that are more difficult to calculate, such as the internal structure of the halo. We have gone a step further and presented a simple weighting scheme to compensate for 'missing' halos by upweighting those that are recovered by the halo finder. Our scheme is able, by construction, to reproduce a 'target' number density of halos, and returns improved estimates for the clustering of halo samples. Depending on one's error tolerance for the accuracy of the clustering predictions, we showed an example in which this scheme extended the mass resolution of a halo catalogue down to objects made of 11 particles.

As presented, our scheme requires at least two simulations. One is designated as the high resolution simulation and sets the target or benchmark for the halo

sample statistics. This simulation is used to provide the 'correct' answer for the halo mass function, and to provide some indication of the expected clustering for different halo samples. No weights are applied to the halos in the high resolution simulation. The second simulation is lower resolution, typically because it models the growth of structure in a much larger volume, with a similar or reduced number of particles than the high resolution simulation. The purpose of this simulation could be to access clustering predictions on larger scales than could be reached with the high resolution simulation, such as the scale of the baryonic acoustic oscillations. Also, many copies of the low-resolution simulation could be run using an approximate simulation method to generate many realisations of halo samples for error estimation. Examples of both these use cases can be found in Hernández-Aguayo et al. (2021). By extending the usable halo catalogue derived from the low-resolution run down to lower masses, significant computational resources can be saved.

The subhalo finding algorithm recovers a fraction of the expected halos in the mass range that is considered 'sub-resolution'. We showed that these objects have higher local overdensities than halos in the same mass range that are fully resolved in a higher resolution simulation. The details of which halos are found will no doubt depend somewhat on the subhalo finder algorithm used, and perhaps on the simulation code itself. Our scheme does not assign weights using any spatial information, and so cannot "correct" the clustering measured for halos in a single mass bin. Our approach works for samples defined by a mass threshold, for which there are several bins in the mass function from which the halos acquire different weights greater than unity.

The scheme that we have proposed allows the resolution of a halo catalogue to be extended down to small particle numbers by applying a correction to the halos that we do see to account for those that we do not find. Ultimately, the scheme breaks down at the halo mass for which the errors in the clustering prediction become unacceptable. This approach is therefore different to those that try to

account for assembly bias in marginally resolved halos (e.g. Ramakrishnan et al. (2021)). Assembly bias which arises when the clustering of halos in a given mass range also depends on an internal property, such as formation time, environment or concentration (Gao et al., 2007). Ramakrishnan et al. (2021) attempt to estimate internal halo properties from marginally resolved halos (e.g. with 30 particles) in order to build mock catalogues which include assembly bias. In principle it should be possible to combine the two approaches to build more accurate mock catalogues.

Chapter 5

The construction of accurate mock galaxy catalogues for the Baryon Oscillation Spectroscopic Survey galaxies.

5.1 Introduction

Here we create mock galaxy catalogues from the simulations presented in Section 4.2 to compare the theoretical models with observations, interpret the clustering measurements and to search for systematic errors associated with the measurements of the marked correlation functions from the observational data.

To create accurate mock galaxy catalogues we need simulations that are suitable for studying clustering and marked clustering. These simulations need to have sufficient resolution in both mass and length scales to allow the halos that are thought to host the target galaxy sample to be identified reliably. We also aim to model the inner structure of the haloes to obtain a realistic one halo contribution to the clustering. At the same time, we need a large volume computational box to simulate structures on scales in excess of $100h^{-1}\,\mathrm{Mpc}$ and to have the same

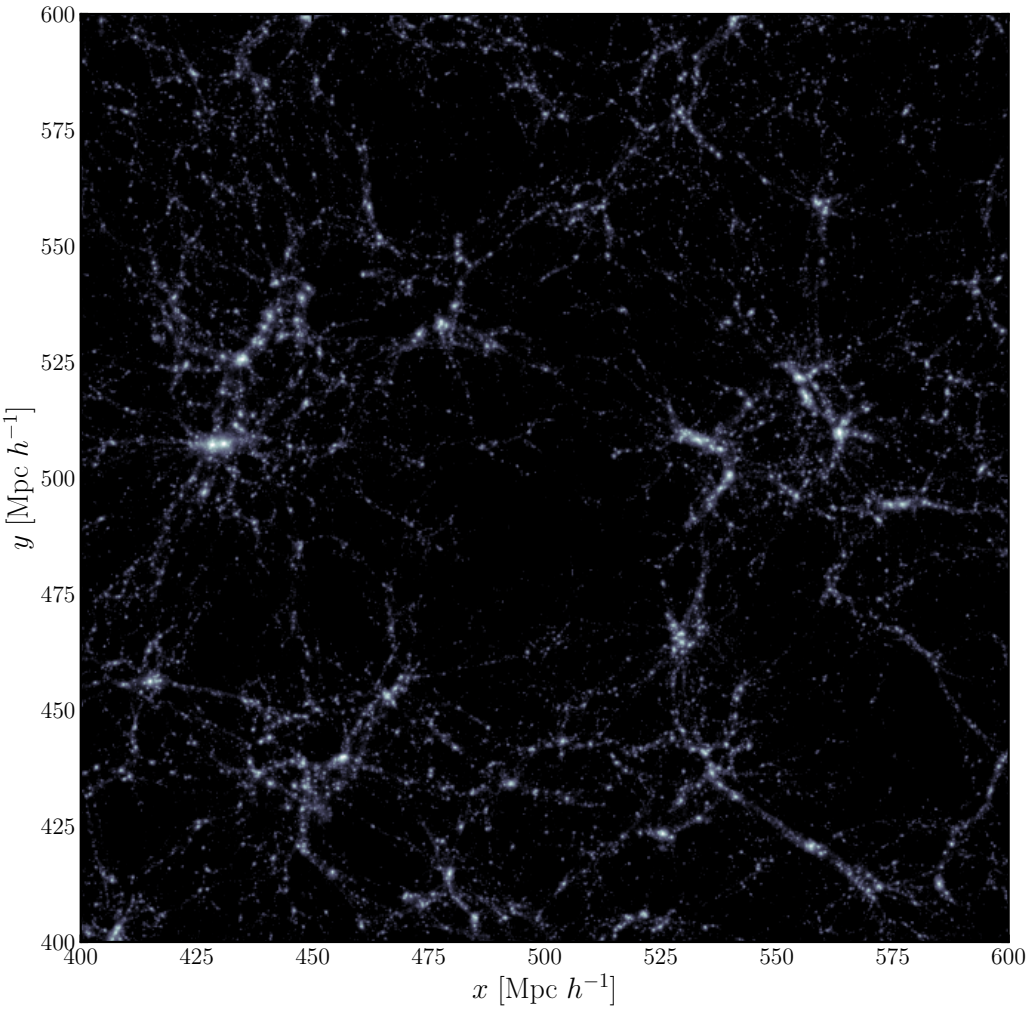


Figure 5.1: 2-D projection in the XY plane of the L768 N-body simulation for the snapshot at redshift $z = 0.3$ from Arnold et al. (2019). The distribution of smoothed dark matter particles is projected in a slice of $\Delta Z = 40\ h^{-1}$ Mpc. Highlighted regions are coloured using the density $\log \rho/\bar{\rho}$. The smoothing of the plotted particles was performed using the `swiftsimio` python-library (Borrow et al., 2020).

statistical errors in the correlation function estimation. We use the simulations studied in Chapter 4, where we test the resolution for the halo catalogues in two simulation box sizes and the impact on the clustering. This study needs to be extended for the effect of the resolution when using subhaloes in clustering studies, which we do in this chapter.

In order to model the Baryon Oscillation Spectroscopic Survey (BOSS) galaxies, we need to resolve scales in projected distance below $1h^{-1}$ Mpc, which means that a high-resolution simulation like the the L768 run introduced in Section 4.2 is needed. The box size of the L768 run ($L_{box} = 768\,h^{-1}$ Mpc) is comparable to the volume of the LOWZ and CMASS subsamples (see Chapter 3). This will be useful when the uncertainties in clustering measurement are considered. We are confident that L768 simulation provides both the resolution and volume needed to predict clustering accurately on the scales we want to study.

As the data covers a wide range of redshifts, we approximate the redshift range covered by the survey data, by using a single redshift snapshot from the simulations for our analysis. In this study, we do not create a lightcone mock catalogue with the geometry of the survey, as the data does not show a significant evolution for the used redshift range, neither does the number density of the samples vary strongly. The selection of the subsamples is made in a much narrower redshift range than the original samples, so the selection of individual redshift snapshots works well. As the size of the subsamples is small in redshift, the number density and clustering of the tracers is not expected to evolve. We decide to keep the snapshots at $z = 0.3$ to model the subsample of LOWZ at $0.24 < z < 0.36$, and the one at $z = 0.5$ to model CMASS in the $0.474 < z < 0.528$ redshift range.

In Figure 5.1 we show the distribution of dark matter particles in a thin slice of thickness $40h^{-1}$ Mpc for the L768 GR simulation. The particle distribution of a $200h^{-1}$ Mpc square and a depth of $40h^{-1}$ Mpc is projected into a 2-dimensional grid, with the density is plotted as $\log(\rho/\bar{\rho})$. Focusing on these smaller scales clearly shows the large-scale structure highlighting dark matter haloes as the brighter spots

in the projected density. We select simulation snapshots from the L768 run at $z = 0.3$ and $z = 0.5$ for model the LOWZ and CMASS samples respectively. For these boxes, the SUBFIND halo catalogues are extracted as explained in Chapter 4. We keep the subhalo catalogues as they will be used to model the one-halo term in the two-point correlation function. To populate haloes with galaxies, and to obtain the galaxy catalogue that matches the data from the samples, the halo occupation distribution method is used, which is described below.

5.2 The halo occupation distribution model

The HOD model (Peacock et al., 2000; Berlind et al., 2002) is an empirical relation that describes how the expected number of galaxies per halo varies as function of halo mass. By using the halo and subhalo catalogues we aim to recreate the galaxy populations of the BOSS LOWZ and CMASS LRG samples. We repeat the definition of the HOD prescription for centrals introduced in 4.4, but now add the term for the expected number of satellite galaxies, as defined in Zheng et al. (2007):

$$\langle N_{\rm cen} \rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right] \tag{5.1}$$

$$\langle N_{\rm sat} \rangle = \langle N_{\rm cen} \rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \text{ For } M > M_0 \tag{5.2}$$

where for centrals $N_{\rm cen}$ describes the number of central galaxies, and is a function of the mass of the halo M, with $M_{\rm min}$ and $\sigma_{\log M}$ being free parameters of the model. For satellites, Eqn. 5.2 is dependent on Eqn. 5.1 and M, because the satellite population of the halo is linked to whether or not there is a central galaxy. M_0, M_1, and α are free parameters in Eqn. 5.2. To better understand the model we need to give an interpretation of the meaning of the free parameters in Eqn. 5.1 and Eqn. 5.2: $M_{\rm min}$ is the minimum mass for a halo to host a central galaxy, $\sigma_{\log M}$ sets the probability for a halo with $M < M_{\rm min}$ to host a central galaxy (increasing this parameter increases the number density when the other parameters are held

fixed); M_0 is the initial mass where haloes start being populated with satellites, for $M < M_0$ the halo contains no satellites; M_1 is the typical mass where the expected number of satellites behaves like a power-law as function of halo mass; and α is the exponent of the power law for $M > M_1$, which controls the number of satellites in massive haloes. As this is an empirical relation, some restrictions on the allowed parameter values need to be considered. We require $M_0 > M_{\rm min}$ as satellite galaxies are not allowed if there is no central galaxy in the halo. Also $M_1 > M_0$ to ensure a smooth transition between haloes without/with satellite galaxies. We need $\alpha \sim 1$ to recreate a large number of satellites in the high-mass end of the HOD function, which is a result that most of the HOD models reproduce (Zehavi et al., 2005; Zheng et al., 2007; Manera et al., 2013). When $\alpha \ll 1$, then $\langle N \rangle$ is low and converges to unity, which is not realistic for observed galaxy samples, as this results in a model where all haloes with a central galaxy have either zero, one or a very small number of satellites. This does not agree with large haloes being modelled as clusters of galaxies, which have abundant red galaxies in the samples, as shown by cluster finder algorithms (Rykoff et al., 2014).

The inclusion of satellite galaxies allows us to resolve small scales in the distribution of the galaxies in the BOSS samples. Even though the fraction of satellites in the sample is small in comparison to fraction of central galaxies, contributing around ~ 10 per cent of the total number density of the simulated samples, they make a large contribution to the clustering of galaxies in the 1-halo term on scales of $r < 1\,h^{-1}\,$Mpc. These HOD mock galaxy catalogues are used to replicate the abundance and clustering of LOWZ and CMASS galaxies, which constrains the values of the other free parameters in the HOD model.

5.2.1 Modelling the one-halo term using HOD methods

The matter distribution inside a dark matter halo is not smooth, but smaller structures can be found which correspond to halos that have fallen into a larger structure at an earlier time and are still in the process of merging, having experienced some

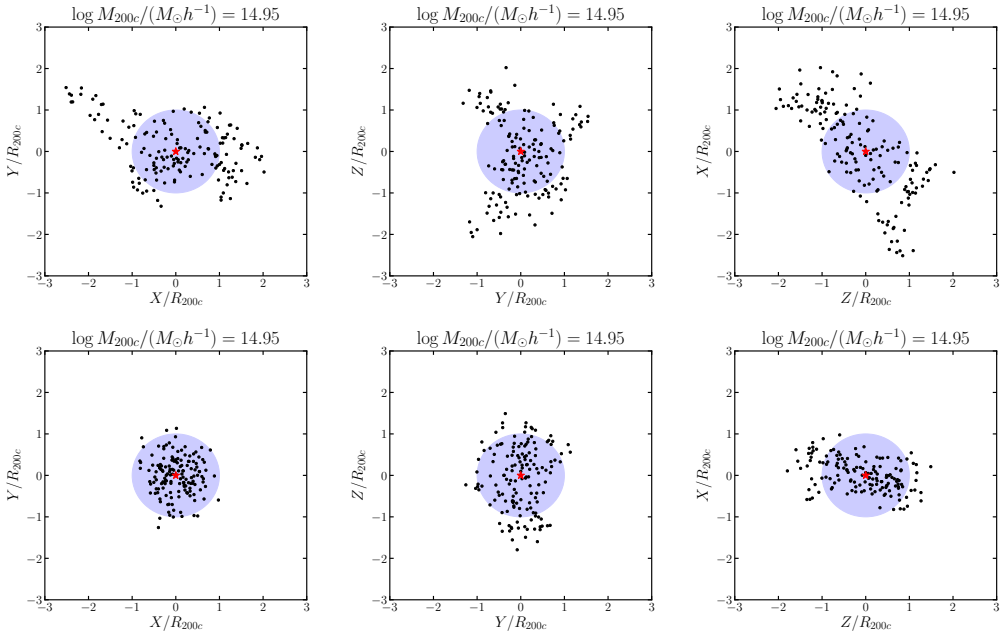


Figure 5.2: 2-D projection of the distribution of subhaloes in 2 arbitrary haloes with the same mass but different shapes. The coordinates are plotted in units of the respective value of R_{200c} radius (the blue circle marks unity in these units) and are centred on the main subhalo (red star). In each row we plot the same halo in XY, YZ and ZX projection for the distribution of subhaloes (black dots).

mass loss through tidal stripping. These clumps inside haloes or "subhaloes" can also be identified in the simulation by the halo finders such as SUBFIND and Rockstar (Springel et al., 2001; Behroozi et al., 2012). Following the application of the method described in Section 4.2 to identify the main haloes in the simulation using the FoF scheme, each FoF halo has its own set of bound particles within which substructures can be found.

Figure 5.2 shows two haloes and their respective subhalo distributions. Although both haloes have been chosen to have the same mass, they show completely different subhalo distributions. Subhaloes are local peaks in the matter distribution within a halo and can be populated with galaxies using the HOD prescription. The differences in these halo profiles provides an illustration of how we can describe the one-halo contribution to clustering of galaxies using subhaloes. In our scheme, satellite galaxies are assigned to subhalo positions rather than resorting to sampling spherically symmetric NFW profiles which end at the virial radius,

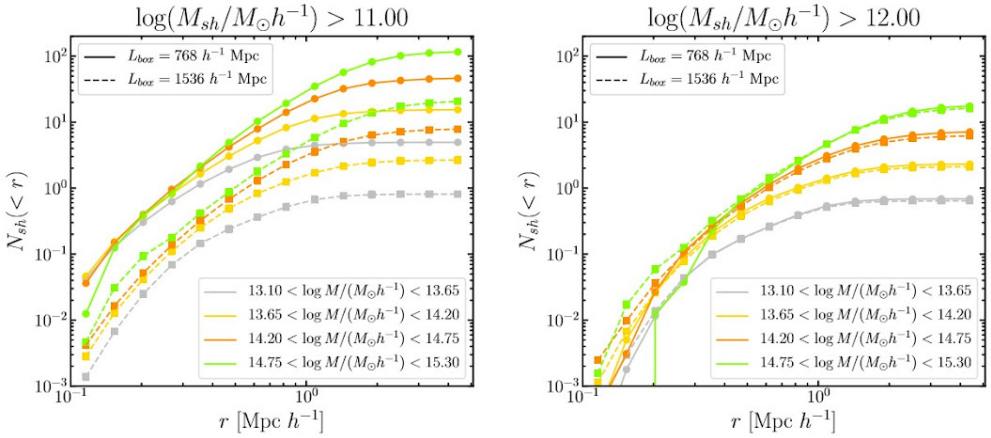


Figure 5.3: Cumulative number of subhaloes $N_{\rm sh}(< r)$ as a function of the radial distance from the halo centre using two different simulations, L768 (solid lines) and L1536 (dashed lines). We use halos with masses $M_{\rm sh} > 10^{11}\,h^{-1}\,M_\odot$ (left) and $M_{\rm sh} > 10^{12}\,h^{-1}\,M_\odot$ (right) to compute the subhalo density profiles in 3 different bins of subhalo mass as described in the legend.

as has been used in many previous studies (e.g. Cautun et al. 2018; Paillas et al. 2019; Armijo et al. 2018; Hernández-Aguayo et al. 2018). This choice was made in order to achieve better agreement between the mock catalogues and the observations: because we use the subhalo distribution to host satellites, we reduce the number of parameters needed to describe this population, which allows us to focus on constraining the other HOD parameters more tightly.

As we use the subhaloes found in the simulation, we need to verify that we have the resolution to probe the smaller scales in the clustering statistics we are trying to model, and that we have a sufficient number of subhaloes given the expected number of satellite galaxies. In Figure 5.3 we plot the cumulative subhalo profiles as function of distance r from the centre of the main subhalo. We choose two different subhalo mass thresholds to check how the simulation resolution affects the halo profiles for different halo masses. The subhalo mass is then defined by the number of bound particles only, instead of the classical M_{200c} used for the main halo masses. We find that when selecting subhaloes with masses $M_{\rm sh} > 10^{11}\,h^{-1}\,M_\odot$, a lower resolution simulation like the L1536 run yields almost an order of magnitude fewer subhalos than are found in the higher resolution L768 run. It is not until we select

subhaloes with masses $M_{\rm sh} > 10^{12}\,h^{-1}\,M_{\odot}$ that the profiles become insensitive to the simulation resolution; this is the minimum subhalo mass for which converged results can be obtained for the small scale clustering from the different resolution runs. Since the L1536 run can still be useful for some measurements we retain the threshold minimum subhalo mass as $M_{\rm sh} \gtrsim 10^{12}\,h^{-1}\,M_{\odot}$.

To determine whether or not this subhalo mass threshold is appropriate for our clustering analysis, we need to run some tests with example HOD parameters. A clear requirement is that the simulations agree with the expected values of the HOD model. If we attempt to use subhalo masses smaller than the subhalo mass threshold, the simulations will have a different number of satellites due to the resolution limit of L1536 as shown in Figure 5.3. If we increase the threshold for selecting subhaloes, we increase the minimum separation scale that can be probed by our clustering predictions, as it is expected that the main halo-subhalo distance also increases with subhalo mass (Angulo et al., 2009).

We test that our simulations recover the expected number of galaxies from the analytical HOD model for a reasonable range of parameters, using some of the HODs inferred from LRG surveys (Manera et al., 2013; Manera et al., 2014). We find that both simulations replicate the expected number of satellites and centrals when using each subhalo at most once when building the HOD catalogue. In cases when the number of satellites is higher than the subhaloes of an individual halo, then subhaloes can also be recycled. Nevertheless, this effect is sub-percent for the HOD parameters we use. Now that we know the limitations due to the resolution in the simulations for satellites for different threshold masses, and for central galaxies in low mass haloes as discussed in Chapter 4, we can search for HOD parameters to create mock galaxy catalogues that mimic the LRG catalogues of BOSS.

5.3 Inferring HOD parameters using the Monte Carlo Markov Chain method

The HOD framework provides a simple and accurate means of describing a galaxy population defined by a set of selection criteria, to allow a reproduction of the large-scale structure measured in a wide field survey. In the particular case of SDSS-LRGs, several studies have been performed to construct such mock galaxy catalogues (Parejko et al., 2013; Manera et al., 2013; Manera et al., 2014). These studies focused on developing methods to generate galaxy mock catalogues using N-body simulations for some of the early data collected by BOSS, and building mocks to allow covariance estimates. The galaxy mocks from these studies were used to test the clustering of LOWZ and CMASS, providing an estimate of the covariance matrix of galaxy clustering statistics over a wide range of scales. Whilst Manera et al. study the clustering of galaxies in redshift space on scales between $30 < s/h^{-1}\,\mathrm{Mpc} < 80$, Parejko et al. use the projected correlation function on scales between $0.4 < r_p/h^{-1}\ \mathrm{Mpc} < 40$; the latter is approximately a real-space clustering measurement (see Norberg et al. 2009).

The studies differ in some aspects in terms of how they extract the HOD parameters used to create mocks. Manera et al. minimize the χ^2 between clustering measured for the mock and observations, using a five-dimensional parameter search employing the simplex algorithm of Nelder et al. (1965). Parejko et al. use a Monte Carlo Markov Chain Method (MCMC) to fit the $w_p(r_p)$ function and overall galaxy number density. Nevertheless, neither of these studies consider that there is an intrinsic degeneracy between number density and clustering, while using HOD parameters, which needs to be addressed while inferring the best parameters for an observational sample. Parejko et al. assume a 15% error on the galaxy number density, although the number density of the samples used varies by more than 30% across the studied redshift range. The 15% error in the number density is justified to make sure that their MCMC chains converge to a solution for the clustering fit,

which is a valid reason for their methodology, but results in too wide a range of HOD parameters being compatible with the clustering measurements.

Here, we try to improve on the procedure of Parejko et al. in a number of ways: we restrict the redshift range of the samples as explained in Chapter 3, and we create a new scheme for fitting the number density along with the clustering. We describe our method in the next Section. Another method worth mentioning is that presented by Zhang et al. (2022), where HOD parameters are constrained by using high-order clustering statistics. Here, the combination of two, and three-point functions allows the exploration of the HOD five-dimensional space in a more accurate way. Although the study provides a precise estimation of some of the model parameters, such as the minimum mass of haloes which host a LRG, the computation of three-point functions is well known to be relatively time consuming (Guo et al., 2015). The authors approach this problem by using the tabulation methods described in (Zheng et al., 2016), which uses the combination of various weights to statistically emulate the 2,3-point functions directly from the HOD parameters, without explicitly producing a mock galaxy catalogue. Even though we could benefit from such an approach to compute the two-point clustering fast, as we are also computing the marked correlation function subsequently, we need the mock galaxy catalogues to compute the weights for creating the marked statistic of these samples.

5.3.1 The Metropolis-Hasting MCMC approach

We use the Metropolis-Hasting MCMC scheme (Metropolis et al., 1953; Hastings, 1970) to explore the 5-dimensional HOD parameter space, and obtain the best fitting parameters that replicate the number density and clustering of the LOWZ and CMASS samples. Here we describe some of the steps involved in the formalism and how it is applied to our particular problem:

1. A position in the chain is described by a set of parameters θ, which comes

from the proposal function $q(\theta^t|\theta^{t-1})$, which is normally modelled as a Normal distribution, computed for time t. In our case θ is the set of HOD parameters.

2. For the next potential move in the parameter at time t, we compute

$$\alpha = \frac{q(\theta^{t-1}|\theta^t)\pi(\theta^t)}{q(\theta^t|\theta^{t-1})\pi(\theta^{t-1})}, \tag{5.3}$$

where $q(\theta^{t-1}|\theta^t)$ is next move proposed, $q(\theta^t|\theta^{t-1})$ is the current position of the chain, and $\pi(\theta^t)$ is the probability of the proposed move $\left(\sim e^{-\frac{1}{2}\chi^2}\right)$, to be accepted. α is then the probability of accepting the move for the chain.

3. Draw a random number μ from a uniform $[0, 1)$ distribution. This is the Monte-Carlo step, where if $\mu < \alpha$, then θ^t is accepted.

4. If the proposed move is not accepted (with probability $1-\alpha$), then $\theta^t = \theta^{t-1}$, which means that the chain does not move to the next position. In general, $\alpha \sim 0$, when $q(\theta^t|\theta^{t-1}) \gg q(\theta^{t-1}|\theta^t)$, or the current state probability is much higher than the proposed one.

Normally it is advisable to work with logarithmic quantities for the density distributions rather than the density probability, so the ratio to determine the acceptance or rejection of the next step in the chain becomes the difference of the logged probabilities. In other words, the accept–reject step conditions can be written as $\ln q(\theta') - \ln q(\theta) > \ln \mu$. The algorithm is repeated a large number N of times to create a sample of size N which contains information about the true distribution that describes the best fitting parameters of the model. The main function that makes the MCMC algorithm work is the proposal function, $q(\theta^{t+1}|\theta^t)$, which in most scenarios can be described by a simple Gaussian probability distribution or any other symmetric distribution. Once the movement of the chain is decided by applying $q(\theta')$, the true distribution of the set of parameters θ, is given by a probability distribution function $p(\theta|X)$, which tells the probability of an event X, and an unnormalized distribution that comes from the product of the prior $\Pi(\theta)$ and the likelihood distribution $\mathcal{L}(X|\theta)$.

The priors are the initial distributions of every parameter from which the proposal function takes its values. Again, in the simplest case, the priors are bounded uniform distributions. If the proposed value is outside these bounds then the log-probability associated with the event (step in parameter space) is $-\infty$, or an event with zero probability. Then, the priors help to define the range within which the parameters are searched for in the log-likelihood distribution. The log-likelihood distribution describes the probability of a sample of events X to belong to the real $p(\theta)$ distribution. The log-likelihood distribution is proportional to $\Delta\chi^2$, which is defined by

$$\chi^2 = (\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{\Sigma^{-1}} (\mathbf{x} - \boldsymbol{\mu}), \tag{5.4}$$

where $\mathbf{x}$ is the realization value drawn from the set of parameters, θ, and $\boldsymbol{\mu}$ is the observable that we are trying to model. $\mathbf{\Sigma^{-1}}$ is the inverse of the covariance matrix, which includes the uncertainties in the observation of $\boldsymbol{\mu}$. In our case, $\mathbf{x}$ corresponds to the projected correlation function w_p and galaxy number density $n_{\rm gal}$ from every realization of the HOD parameters, and $\boldsymbol{\mu}$ represents the values of $w_{\rm p}$ and $n_{\rm gal}$ measured from the the observed data.

5.3.2 Autocorrelation time and convergence

An essential point to consider when running the MCMC algorithm for any kind of problem is to determine when the search of parameter space has converged to a reliable solution. In other words: how long does the chain have to be run to represent the authentic posterior distribution of the parameters? The answer is not simple, as this is a fully stochastic method, where the longer the run the more consistent the answer. However, as the MCMC cannot be run for an infinite time, we need to define when a particular number of iterations is sufficient for a complete sample, which can be model dependent.

To do this we perform the autocorrelation time analysis of the model, which helps us to determine if the chains are sufficiently converged. The autocorrelation

time quantifies directly the Monte Carlo error in the sampler. This determines the robustness, in terms of the variance, of the sampling of a given distribution as function of the different steps. For an independent, finite sample of parameters, θ, the sampling variance estimator can be written as

$$\sigma^2 = \frac{1}{N}\mathrm{Var}_{p(\theta)}\left[f(\theta)\right], \tag{5.5}$$

where $f(\theta)$ is the function we are trying to sample over N realizations of the parameters θ, and this is drawn from a probability density $p(\theta)$. As the next move in the MCMC depends on its previous state, the samples of a particular chain are not independent and the error can be written as

$$\sigma^2 = \frac{\tau_f}{N}\mathrm{Var}_{p(\theta)}\left[f(\theta)\right], \tag{5.6}$$

where τ_f is the integrated autocorrelation time. Then, τ_f can be defined as the number of steps needed so the chain becomes independent of the starting point, and N/τ_f then is the effective number of samples for Eqn. 5.6 becoming the expected Monte Carlo error written in Eqn. 5.5. This means that by estimating τ_f, we can calculate the number of samples required to achieve convergence of the chain by calculating the number of samples required to obtain a constant τ_f as N increases.

To calculate τ_f we need to integrate over the time auto-correlation function of the process

$$\tau_f = \int_{\tau=-\infty}^{\infty} \rho_f(\tau)\mathrm{d}\tau, \tag{5.7}$$

where $\rho_f(\tau)$ is the normalized time autocorrelation function of the chain, representing a time series. Again, as we have a finite number N θ of samples, $\rho_f(\tau)$ can be calculated using an estimator which uses the finite chain of $f(\theta)$ for each set of parameters θ_n in the sample, $\{f_n\}_{n=1}^{N}$. The estimator $\hat{\rho}$ is defined as

$$\hat{\rho}(\tau) = \frac{1}{N-\tau}\sum_{n=1}^{N-\tau}(f_n - \mu_f)(f_{n+\tau} - \mu_f), \tag{5.8}$$

where $\mu_f = \frac{1}{N}\sum_{n=1}^{N} f_n$. An efficient way to compute Eqn. 5.8 for a large sample is by using a fast Fourier transform, which is computationally much cheaper than

summing Eqn. 5.8 directly. Then, the integrated autocorrelation time can be estimated from Eqn. 5.7

$$\tau_f = 1 + 2\sum_{\tau=1}^{N} \hat{\rho}(\tau). \tag{5.9}$$

As we sum Eqn. 5.9 over a large number of samples N, $\hat{\rho}$ adds more noise than signal, which can lead to variations in the value of τ_f and bias the estimation of the chain. It is recommended to monitor the dependence of τ_f as a function of the number of samples N, for a smaller subsample $M \leq N$. In this way, when the sample is large enough, τ_f(N) converges to a constant value.

An alternative to test the robustness of the posterior distribution sampled by the MCMC chains is the Gelman-Rubin diagnostic Gelman et al. (1992), R, defined as

$$R = \sqrt{\frac{\hat{V}}{W}}, \tag{5.10}$$

where R is the variance of one or all the parameters within a chain represented by the estimator $\hat{V}$, compared to the variance between all the chains W. The comparison tells us whether or not these two quantities are unbiased estimators of the true variance. Hence, as long as the chain grows and the values are stable, then $R \approx 1$. Whilst the calculation of the integrated autocorrelation time is an estimation of the error of the Monte Carlo integral and provides us with more understanding of how the chain converges, the Gelman-Rubin diagnostic is a more heuristic test which gives information about the variance of the individual parameters. The Gelman-Rubin test is related to the computation of the autocorrelation time in the sense that it gives positive results regarding the convergence of the chain, when these are long enough to reproduce $p(\theta)$ confidently. It is considered that ensuring the convergence of the chain is more a process of finding the number of samples that stabilize the values of either τ_f or R, instead of a simple value that tells if the chain is converged or not.

The convergence of the MCMC permit us to focus on our particular case, where we infer the parameters of the HOD model for the observed results provided by

the data through the calculation of the survey metrics. We show in the next Section how we set up the model in order to fit both the number density and the clustering of galaxies, and the analysis of convergence for the HOD parameters, when considering the production of the chain.

5.3.3 Studying the HOD parameter-space using the Markov Chain

To obtain the best fitting HOD parameters that most closely reproduce the galaxy number density and clustering measured from the observational samples, we need to test some aspects of the MCMC process. We focus on the behaviour of the chain evolution and how well it samples the parameter distribution. As this is a procedure that takes tens of thousands of iterations, it may become computationally expensive if some of the stages are not treated carefully. To run the MCMC we need to create an "ensemble", which consists of a set of individual samplers or "walkers" forming a chain that sample or "walk" the parameter space of the model to be fitted. In the Metropolis-Hasting scheme, the walking goes in the direction of the more probable positions in parameter space, according to the log-likelihood distribution. The final function of the ensemble is to sample the posterior distribution in the most complete way possible. In Figure 5.4 we show how the walkers move in the parameter space of the HOD model using 2-D projections. A walker starts from a random position in the parameter space, which is randomly selected from the prior. Then, the iterations begin accordingly to the Metropolis-Hasting algorithm explained in Section 5.3.1.

As the starting point may be far from the position of the final posterior distribution, the walker needs to wander its path for a sufficiently long time until it finds the location of the maximum likelihood. This is controlled by the definition of the χ^2 value used in the likelihood definition, which contains the probability information of the actual shape of $p(\theta)$. The stage in which the walkers of the ensemble have yet to reach the "best" regions of the log-likelihood that can be

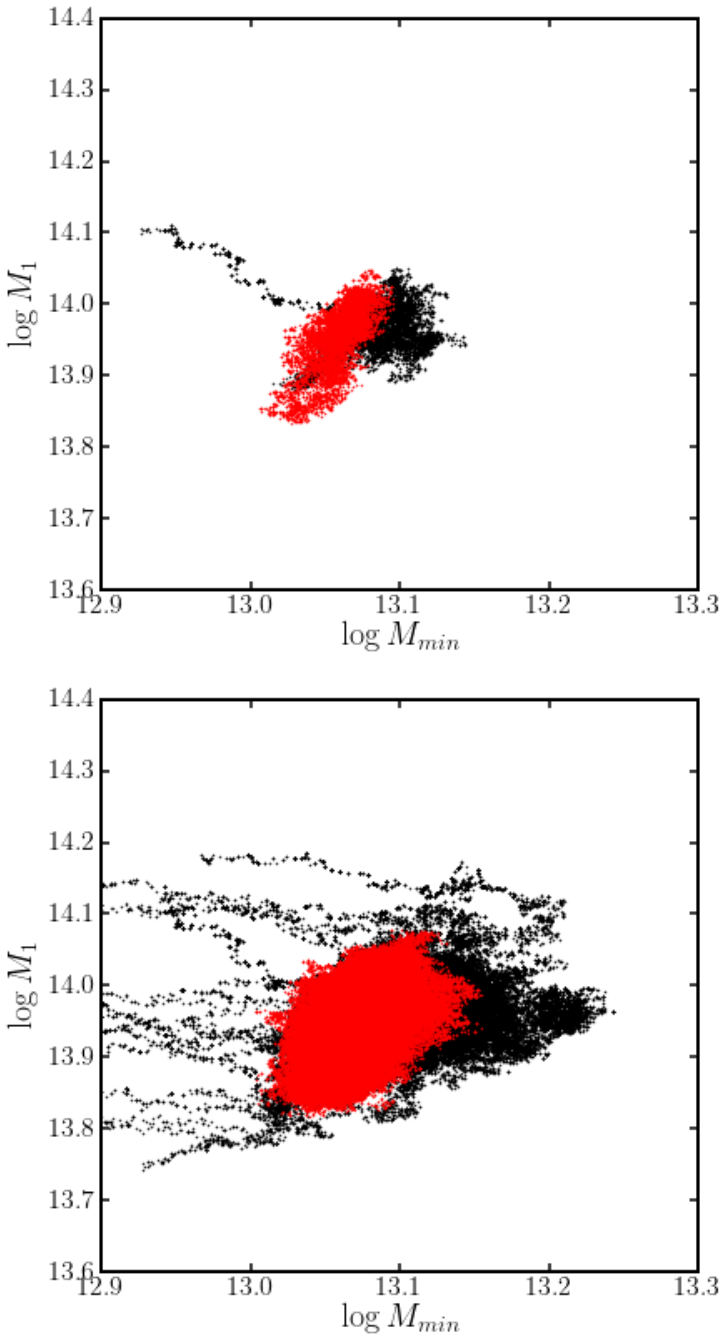


Figure 5.4: Top: the MH-MCMC sampling for a individual walker in the 2-D-projection space for the HOD parameters $\log M_{\rm min}$ and $\log M_1$. The sampling starts from a random position in the parameter space, with a "burn-in" stage (black dots), after which the "production" stage (red dots) starts. Bottom: Same as in the left panel, but for the complete MCMC ensemble composed of 28 independent walkers.

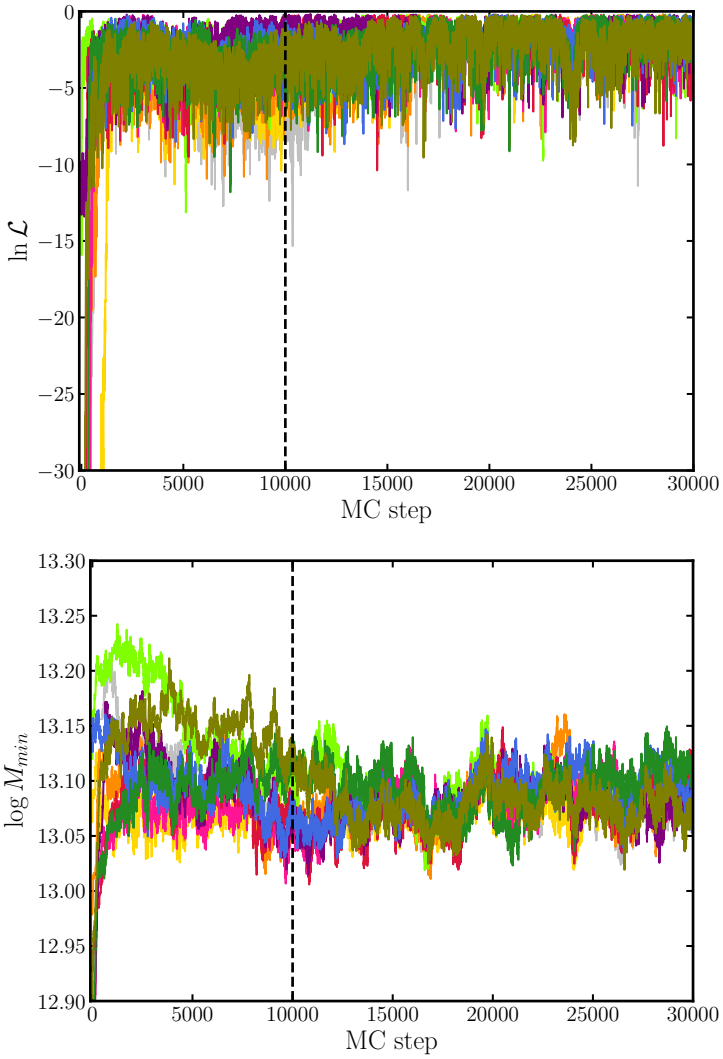


Figure 5.5: top: log-likelihood distribution, $\ln \mathcal{L}(\theta)$, as function of the Monte Carlo step for a realization of 30,000 samples. bottom: $\log M_{\rm min}$ distribution as function of MC step for the same realization as shown in the top panel. Only 10 individual walkers of a total of 28 walkers are plotted for clarity. The black dashed line indicates the burn-in stage, which is placed once the chain stabilize.

sampled is called the "burn-in" phase. This locus of points in parameter space is an unrepresentative sample that has to be cut away from the final sample. After the burn-in phase the chain becomes more representative in the sampling of the log-likelihood distribution, which is called the "production" stage. Here the chain samples for a longer time (larger number of steps) to recreate the final posterior distribution more accurately.

In the top panel of Figure 5.4 the burn-in phase can be identified as the path travelled by the walker in a 2-D projection during a certain time (black dots), until a stable region is found. In practice this correspond to the first 10,000 steps of the chain, after a visual inspection by looking at the evolution of the variance of the chain. This choosing is arbitrarily and it does not vary the final distribution of the parameters. After that, the region sampled by the walker, the likelihood, becomes more representative of the posterior distribution. The values samples by the chain during the burn-in stage can still be correct, as the walker moves relatively quick to the region of maximum likelihood. The large variance of the sample is what tell us we should discard this stage when calculating the posterior distribution. When the production stage begins (red dots), then we know the walker is actually sampling the log-likelihood region, and the variance is no longer an issue to be concerned about.

The procedure is repeated by all the walkers in the ensemble in parallel, as shown in the right panel of Figure 5.4. The selection of the number of iterations or Monte Carlo steps for burn-in and production stages is in principle arbitrary as it depends on how fast the model converges. A rule of thumb for dividing the MCMC run into these two stages is that the burn-in time is a half of the time used for production, $t_b = \frac{1}{2}t_p$, so the chain first runs until it is deemed to be stable and there is still enough steps remaining to record the sampling of the log-likelihood distribution. In Figure 5.5 we show the evolution of the chain over $30,000$ MC steps for 10 walkers. Although, the convergence to high likelihood values looks relatively quick, it is not until at least $10,000$ steps that the chain becomes more stable, as the

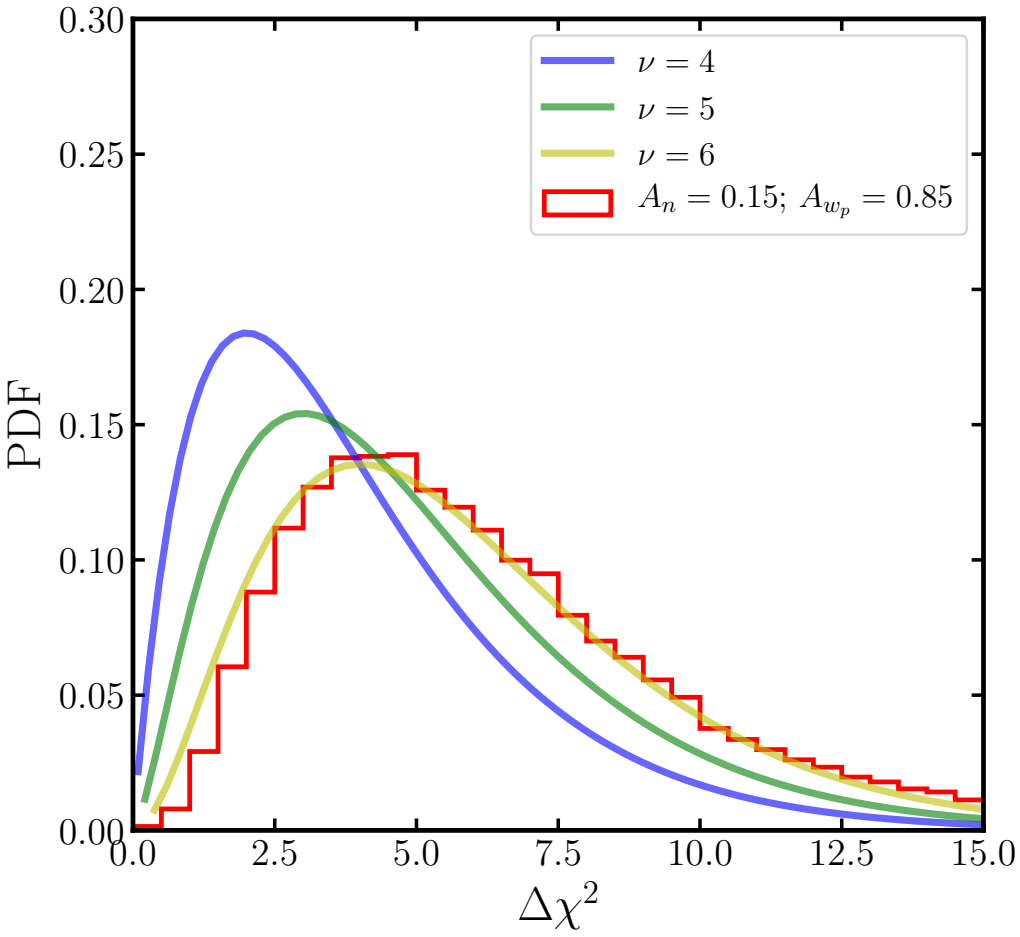


Figure 5.6: The $\Delta\chi^2$ probability density function for the MCMC run with $A_n = 0.15$ and $A_{w_p} = 0.85$ (red line). The shape of the distribution depends on the number of degrees of freedom (colour lines), the histogram is better fit by $\nu = 6$.

walkers start moving closer together. The same occurs for the evolution of some of the parameters, like the distribution of $\log M_{\rm min}$ (one of the parameters of the HOD model) for the same number of iterations, which suggests that $t_b \sim 10,000$.

5.3.4 Defining the χ^2 distribution in the MCMC

As the log-likelihood $\ln \mathcal{L}$ is a set of probability distribution functions (PDFs) for the parameters of our HOD model, there is not a unique answer for the problem we are trying to solve. Instead there are many sets of parameters, $\log M_{\mathbf{min}}$, $\log M_0$, $\log M_1$, σ, α, which yield acceptable fits to the observational measurements. To

obtain all values that can be considered valid, we need to perform an analysis on the χ^2 distribution, which provides a complete classification of the quality of our data given the χ^2 values. As we try to fit two quantities that are highly correlated, such as the clustering and number density of the galaxy sample, we need to consider this when defining the χ^2 that we want to measure. We define a new χ^2 by combining both measurements:

$$\chi^2 = A_n\chi_n^2 + A_{w_p}\chi_{w_p}^2, \tag{5.11}$$

where A_n and A_{w_p} are factors that weight the individual χ^2 for the number density n and the clustering w_p, respectively. By adding these quantities, we can fit models to the data using the given weights for these two metrics, which in turn can provide a better understanding of the correlation between the clustering and number density, and help us to determine if one is more important than the other when looking for the best fitting HOD parameters. To determine the optimal values for A_n and A_{w_p}, we need to test different values for the weights and assess the derived results.

First we need to understand the χ^2 distribution generated from our data. In general terms, we define the probability as the log-likelihood being proportional to the χ^2 defined in Eqn. 5.4, as we are reinterpreting the probability with an arbitrary normalization. To avoid the need to renormalize our data, we calculate $\Delta\chi^2 = \chi^2 - \chi^2_{\rm min}$, where $\chi^2_{\rm min}$ is for the value that maximizes the log-likelihood distribution. Then, the actual distribution is relative to the minimum χ^2 of the data, and the χ^2 distribution is simply

$$\Delta\chi^2 = -2\ln\mathcal{L}. \tag{5.12}$$

Figure 5.6 shows the distribution of $\Delta\chi^2$ for a MCMC run. Here, we define whether the log-likelihood distribution is well sampled by comparing the distribution with the original PDF for a χ^2 with ν degrees of freedom. For our case, the number of degrees of freedom is simply the number of free parameters in the HOD model, then $\nu = 5$. We also need to consider that in Eqn. 5.4 χ^2 depends on the number of bins of the vector $\mathbf{x}$ and the covariance matrix $\mathbf{\Sigma}$, which indicates

the level of correlation between the bins of $\mathbf{x}$. When combining these samples the calculation is more complicated, because of the correlations between the radial separation bins used when computing w_p. This indicates that the effective numbers of degrees of freedom in our model can vary as we modify the definition of the χ^2, by using the weights A_n and A_{w_p}. For example, the MCMC run displayed in Figure 5.6 with $A_n = 0.15$ and $A_{w_p} = 0.85$ is reproduced better by a χ^2 distribution with $\nu = 6$.

Now that we can describe the χ^2 distribution of the data we have the tools to obtain HOD parameters that provide a good fit to the data. Chapter 15.6 of Press et al. (1992) shows what values of $\Delta\chi^2$ we can consider appropriate to describe the data, by considering the numbers of degrees of freedom for our model. For example, when $\nu = 5$ all the sets of HOD parameters where $\Delta\chi^2 < 5.89$ correspond to 68% of the data we are trying to fit. To show a complete description of the χ^2 statistic, we plot the posterior distribution $p(\theta)$ of the parameters in our model in Figure 5.7. As we can only show the 2-D projection of the likelihood data (that has 5 dimensions in total), we need to consider $\nu = 2$ for drawing the contours in the off-diagonal subpanels of Figure 5.7. This plot shows what the parameter space likelihood looks like and how different parameters are correlated. Some of these correlations are expected, like the dependencies between $\log M_{\mathbf{min}}$ and σ that control the occupation rate of central galaxies in low mass haloes. Other correlations can be more interesting, like the one between σ and $\log M_0$, where the latter controls the haloes that produce satellite galaxies, once the low mass haloes with centrals have been fixed.

We can now test our HOD scheme by fitting measurements from the LOWZ and CMASS galaxy samples. We also need to determine the optimal weights A_n and A_{w_p} and the complete range of HOD parameters for each model.

For all the runs, we used `emcee` (Foreman-Mackey et al., 2013), which is a `Python` implementation of the MCMC algorithm that allows the Metropolis-Hasting ensemble sampler explained in Section 5.3.1. We build the ensemble using 28 walkers

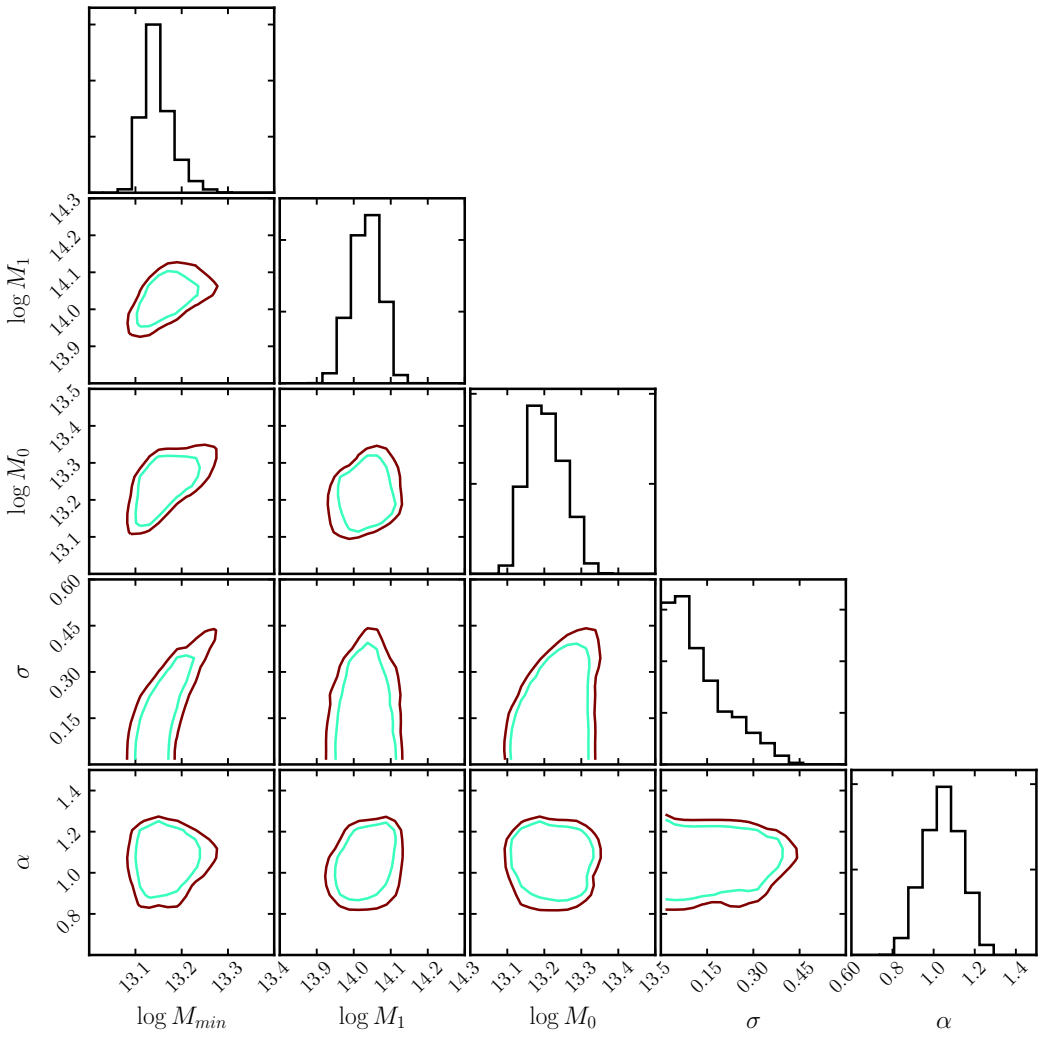


Figure 5.7: Corner plot of the MCMC posterior distribution for the HOD model parameters. The MCMC run fits the HOD model from a simulation (either GR or F5) over the data we want to replicate (either LOWZ or CMASS data). The diagonal subpanels show the 1-D distribution of the parameters (black lines) or posterior distribution $p(\theta)$ with θ being the HOD parameters. The off-diagonal subpanels show the 2-D projection of the parameters for all parameter combinations, where the contours are selected using the $\Delta\chi^2$, using 1-σ (cyan lines) and 2-σ (red lines).

HOD parameter-space limits for GR and F5 simulations	
θ	
$\log(M_{min}$ / $M_\odot\ h^{-1})$	$[12.7, 14.0]$
$\log(M_1$ / $M_\odot\ h^{-1})$	$[12.7, 14.8]$
$\log(M_0$ / $M_\odot\ h^{-1})$	$[12.7, 14.0]$
$\sigma_{\log M}$	$[0.0, 0.6]$
α	$[0.7, 1.6]$

Table 5.1: Uniform priors for the HOD parameters, θ. Extra conditions are applied to the prior distributions, like the fact that $\log M_0 > \log M_{\mathbf{min}}$ and that $\log M_1 > \log 5M_0$ for every set of HOD parameters.

each running for 30,000 iterations (10,000 for burn-in and 20,000 for production); these choices need to be corroborated using the autocorrelation time analysis (see Section 5.4 below) or the G-R diagnostic. We provide the parameter space limits applied to the priors, used for searching the HOD parameters, in Table 5.1. To investigate the optimal choice of weights we try three runs with different χ^2 definitions: $A_n = 0.15$, $A_{w_p} = 0.85$; $A_n = 0.85$, $A_{w_p} = 0.15$ and $A_n = 0.5$, $A_{w_p} = 0.5$. These cases are useful to study how we can adjust the metrics and check how the degeneracy works between these two parameters.

5.4 The HOD families that reproduce LOWZ and CMASS results

To search for the HOD parameters that give us mock galaxy samples that mimic the number density and clustering of the observational data, we need to understand how we adjust the model using these benchmarks. For instance, the number density, which is the mean number of galaxies per unit volume, is represented by one number for every HOD sample, ($n_{\rm gal} = N_{\rm gal}/V_{\rm box}$), where $V_{\rm box} = L^3_{\rm box}$. For the data, as explained in 3.3 we also consider $n_{\rm obs} = N_{\rm gal,obs}/V_{\rm s}$. Whilst for the clustering, a measurement of $w_{\rm p}$ is estimated in both the simulation box and the observational data, using 13 r_p bins in the projected-perpendicular distance range $0.5 < r_{\rm p}/(h^{-1}{\rm Mpc}) < 50$. For both observational metrics, the uncertainties are

estimated using Jackknife resampling to account for sample variance, using the full covariance matrix for $w_{\rm p}$. As we combine these measurements to fit the HOD model to the observational data, we need to make sure that this results in catalogues with accurate measurements of n_{gal} and w_p. For example by giving all the weight to the clustering by fitting $w_{\rm p}$ only, we will end up with the same two-point galaxy statistic but we will miss the number density of targets, by around 15-20% as shown by Parejko et al. (2013). Such a result will have a high influence on the calculation of the marked correlation function, which will impact on the utility of this test to probe modified gravity, by adding uncertainties in the ranges where we expect the models to differ. On the other hand, by giving more weight to the number density and less to the clustering, we will obtain poorer reproductions of the clustering. The range of "acceptable" HOD parameters will also be broader in the limit of giving increasing weight to the number density, as we are effectively trying to constrain the 5 HOD parameters from one measurement in the limit. Hence, a compromise is required in which both observational measurements are recovered without biases at an adequate statistical level of confidence.

As explained in Section 5.3.4 we test three scenarios for the weighting scheme. 1) $A_n = 0.15$, $A_{w_p} = 0.85$, where we give most of the weight to the clustering signal, by considering the number of bins used for w_p and n_{gal}. 2) $A_n = 0.5$ $A_{w_p} = 0.5$, where both clustering and number density have the same weight in the χ^2 definition. 3) $A_n = 0.85$, $A_{w_p} = 0.15$, where we give more weight to the number density. For these three cases we compare the best-fitting HOD parameters within their 1-σ confidence interval, and the corresponding results for the n_{gal} and w_p measurements.

In Figure 5.8 we show the resulting HOD functions and the associated clustering for the HOD galaxy catalogues. We plot a random subsample of 1000 of lines sampled from the acceptable HOD parameter space we find in one of the MCMC chains in the top panel of Figure 5.8. We choose to plot $A_n = 0.5$ $A_{w_p} = 0.5$ for clarity and because all three cases show the same features. For the three different

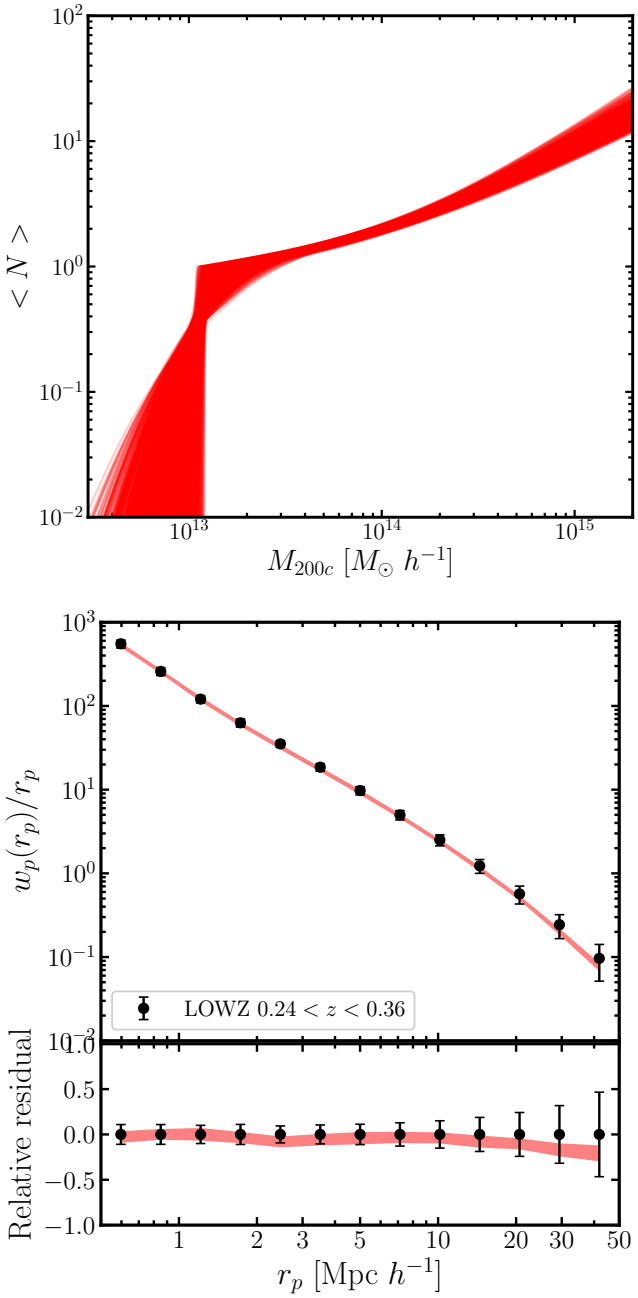


Figure 5.8: Top panel: The average number of galaxies in a halo, $\langle N \rangle$, as function of halo mass M_{200c} (red lines) for all the HOD parameter sets which lie within a $1\,\sigma$ confidence interval according to the χ^2 distribution. Bottom panel: The projected correlation function $w_{\rm p}(r_{\rm p})$ as function of the projected separation, $r_{\rm p}$, for galaxy catalogues created using the HOD samples shown in the top panel. The red region corresponds to that covered by all the w_p/r_p curves, and the black dots shows the measurement from LOWZ that we used to fit the model. Uncertainties for the observational measurements points have been calculated using the Jackknife as explained in Section 3.3. The bottom subpanel shows the residuals relative to the observational data.

weighting schemes we find similar results in terms of the range of values covered by the HOD parameters. An interesting feature of these HOD parameter is that all cases permit $\sigma_{\log M} = 0$, which corresponds to a sharp cutoff in the mass of low mass haloes that can host a central galaxy. We find that, in general, schemes where equal or higher weight is given to the clustering, $A_{w_p} = 0.5, 0.85$ cover the same parameter space. Whereas the model that gives more weight to number density (i.e. $A_n = 0.85$) expands the parameter range for those parameters that contribute less to the number density such as $\sigma_{\log M}$ and α, but constrains better those that contribute more such as $\log M_{\mathbf{min}}$. We also plot w_p for the same run. In this case we show the region covered by the individual w_p functions selected within the 1-σ region, which means that the shaded region represents the uncertainties due to the allowed values of the HOD parameters. Again, for the three cases considered we see the same features, as expected: the clustering is degenerate with the number density for the range of the HOD parameters we find, and the measurement of w_p is adequate for the different weighting schemes. These results indicate a good fit of the clustering overall, with a small deviation for the large-scales distances at $r_p > 20\,h^{-1}\,\mathrm{Mpc}$. Nevertheless this is consistent with the uncertainties from the Jackknife resampling. Additionally, our measurements of w_p are also consistent with those from Parejko et al. (2013), including the small deviation between mocks and data at these large scales.

Finally, we check the results for the number density in Figure 5.9, which shows the three different weight cases. The three panels show the distribution of the number density recovered using the HOD parameters sampled, denoted n_{sim}. When we compare the distributions to the value from the observational sample, n_{obs} we can test how good these fits are, paying attention to any significant systematic shifts. For the first case, $A_n = 0.15$, $A_{w_p} = 0.85$ (top panel), there is a clear mismatch between the mean of the n_{sim} distribution and n_{obs}. By comparing the distribution of n_{sim} with the Gaussian distribution with the same standard deviation, we find that the discrepancy is around 1-σ. In comparison, the other weight schemes

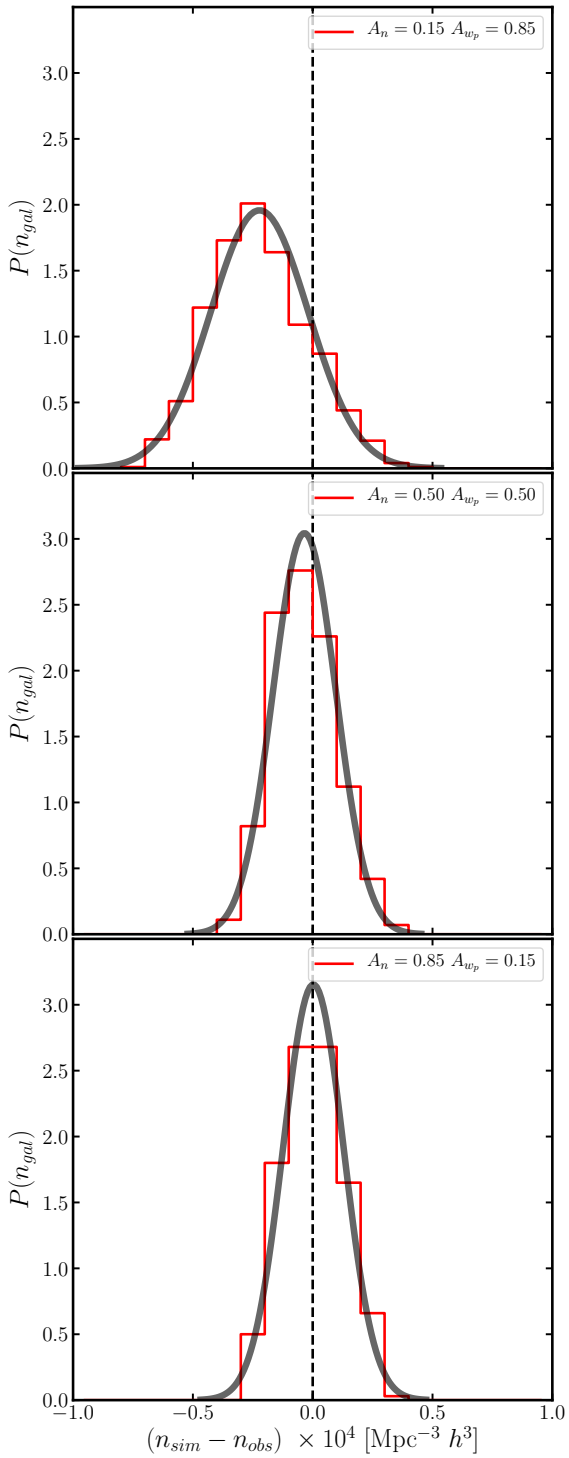


Figure 5.9: The distribution of the galaxy number density values $P(n_{\rm gal})$ recovered for the HOD samples in the different weighting schemes: $A_n = 0.15$, $A_{w_p} = 0.85$ (top panel); $A_n = 0.50$, $A_{w_p} = 0.50$ (middle panel) and $A_n = 0.85$, $A_{w_p} = 0.15$ (bottom panel). We draw over each $P(n_{\rm gal})$ a Gaussian with the same mean and standard deviation as the distributions. We have rescaled the x-axis to center each distribution on the target value we are fitting $n_{\rm obs}$, the number density of the LOWZ sample.

HOD 1-σ confidence intervals for GR and F5 simulations		
θ	GR: $z = 0.3$, $z = 0.5$	F5: $z = 0.3$, $z = 0.5$
$\log(M_{\rm min}$ / $\mathrm{M}_\odot\ h^{-1})$	$[13.029, 13.205]$; $[12.928, 13.027]$	$[13.100, 13.282]$; $[12.996, 13.141]$
$\log(M_1$ / $\mathrm{M}_\odot\ h^{-1})$	$[13.853, 14.053]$; $[13.689, 13.869]$	$[13.953, 14.103]$; $[13.788, 13.927]$
$\log(M_0$ / $\mathrm{M}_\odot\ h^{-1})$	$[13.042, 13.262]$; $[12.937, 13.107]$	$[13.118, 13.341]$; $[13.007, 13.180]$
$\sigma_{\log M}$	$[0.003, 0.440]$; $[0.0, 0.327]$	$[0.0, 0.479]$; $[0.002, 0.431]$
α	$[0.802, 1.067]$; $[0.800, 1.045]$	$[0.801, 1.255]$; $[0.800, 1.302]$

Table 5.2: The 1-σ confidence intervals of the HOD parameters for the GR and F5 simulations at redshift $z = 0.3$ and $z = 0.5$, to match the clustering and abundance of galaxies in the LOWZ and CMASS samples.

(shown in the middle and bottom panels) seem to yield more accurate estimates of $n_{\rm obs}$. Comparing the different panels of Figure 5.9, we chose the weights $A_n \sim 0.5$ in order to obtain a correct, unbiased estimate of $n_{\rm obs}$.

We run the the autocorrelation time analysis and the G-R diagnostic to test the convergence of the three choices of weight scheme. In Figure 5.10 we plot the integrated autocorrelation time and the G-R diagnostic, for the chain $\tau_{\rm f}$ as a function of the number of samples N. The ensemble run using `emcee` is estimated to converge for $N > 50\tau_{\rm f}$, limited by the black dashed line in Figure 5.10. In fact, by looking at the value of τ_f where the curves start to flatten, we ensure that the chain has been running for enough time. For case 1) and 2), $\tau_{\rm f} \sim 400$, which is the number of samples needed for the chain to forget where it started, following the estimated number for the convergence suggested by `emcee`, these models need at least $20,000$ iterations. Case 3) seems to converge faster with $\tau_{\rm f} \sim 300$, which is expected, as this model allows a wider range of HOD parameters. Additionally, we compute the G-R diagnostic for the total samples in the different chains, obtaining $R = 1.149$, for case 1), $R = 1.087$ for case 2), $R = 1.071$ for case 3). Although, all cases seem to converge similarly, the higher R value for case 1) disfavours the scheme where $A_n = 0.15$ $A_{w_p} = 0.85$. This can be checked in the the bottom panel of Fig. 5.10, where the convergence of R for case 1) is slower and starts from a higher value. This indicates that the variance of the chain is higher than the one from the total sample in comparison to case 2) and 3) (and a slower convergence), possibly influenced by the smaller weight value A_n.

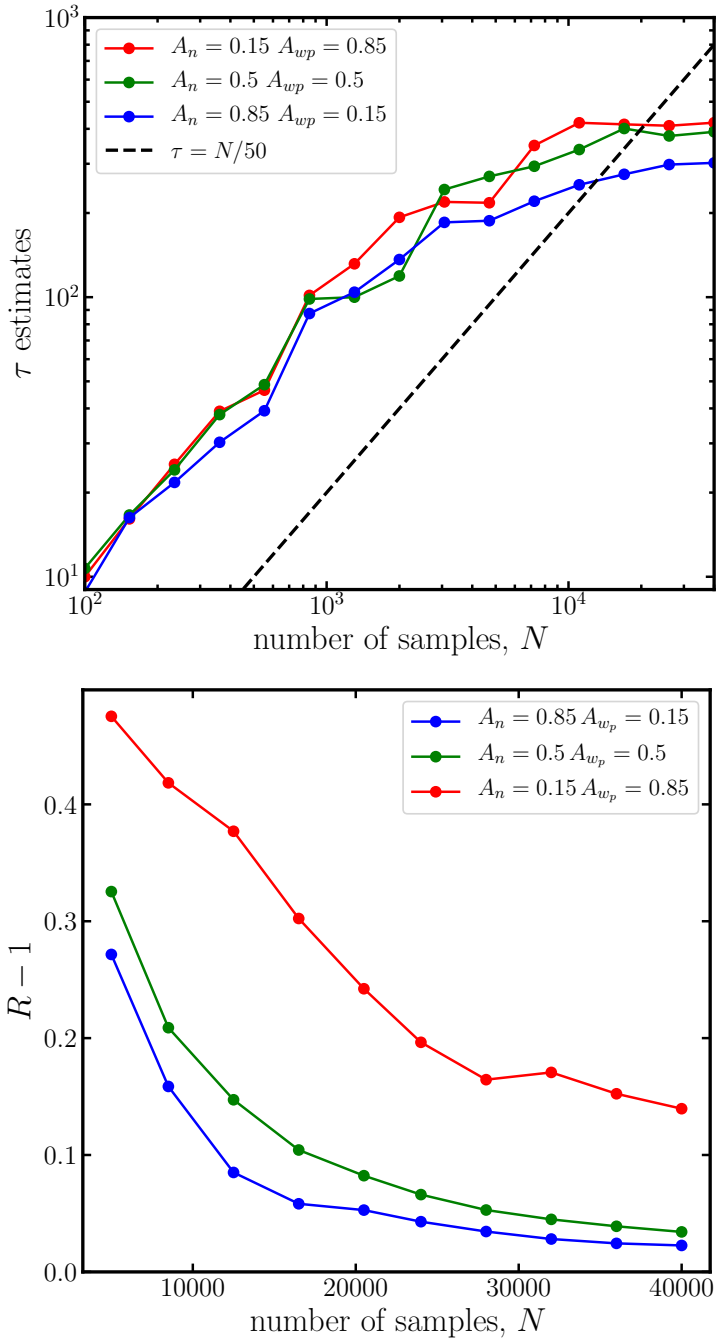


Figure 5.10: Top: The integrated autocorrelation time $\tau_{\rm f}$ as a function of the number of samples N. The curves show three different MCMC runs changing the weights that define the χ^2: $A_n = 0.15$ $A_{w_p} = 0.85$ (red), $A_n = 0.5$ $A_{w_p} = 0.5$ (green), and $A_n = 0.85$ $A_{w_p} = 0.15$ (blue). The $\tau_{\rm f} = N/50$ (black dashed line) is added to show how the models are predicted to start converging after crossing this value. Bottom: The G-R diagnostic showing the ration between the ratio $R - 1$ as function of the number of samples N for the same samples displayed in the top panel.

5.5 Discussion

By testing the three different cases for weighting the galaxy number density and clustering, we can choose the optimal scheme to obtain the HOD parameters that replicate the observations of LOWZ and CMASS samples. By examining the results of the fits in the number density and the G-R diagnostic values, we discard the scheme with weights $A_n = 0.15,\ A_{wp} = 0.85$. The level of accuracy of the fit to the clustering and number density for the other schemes looks adequate with a few caveats. The scheme with $A_n = 0.85\ A_{w_p} = 0.15$ also increases the HOD parameter range of the samples, which can be counterproductive, and a smaller effective number of degrees of freedom in the $\Delta\chi^2$ distribution with $\nu \sim 4$. We are giving more weight to the individual χ^2_n of n_{gal} and reducing the effective number of bins used to calculate $\chi^2_{w_p}$. This will have an impact on our conclusions regarding the HOD parameters within the 1-σ region and in future calculations of the marked correlation function. In summary, we choose to keep the scheme where $A_n = 0.5\ A_{w_p} = 0.5$, as we are giving equal relevance to both $n_{\rm gal}$ and w_p, this seems like the right choice to treat the degeneracy between these measurements and to extract the information from the observational data.

We run and compare the MCMC fitting for all models in Table 5.2. Here we add the information of the 1-σ confidence interval for the GR and F5 simulation snapshots at redshift $z = 0.3$ and $z = 0.5$. The F5 simulation has higher values for the characteristic masses, $\log M_{\rm min}$ being the one that contributes more to the number density and clustering amplitude. Also, the exponent of the power law for the number of satellite galaxies α goes to higher values in the F5 model. What we learn from the HOD parameters recovered for the GR and MG models is that the main differences are in the minimum mass needed to populate haloes with central galaxies and the rate of satellite galaxies per halo, both of which are higher in modified gravity. This is caused by enhanced halo formation in modified gravity models, where unscreened haloes increase the rate of central galaxies, which has to

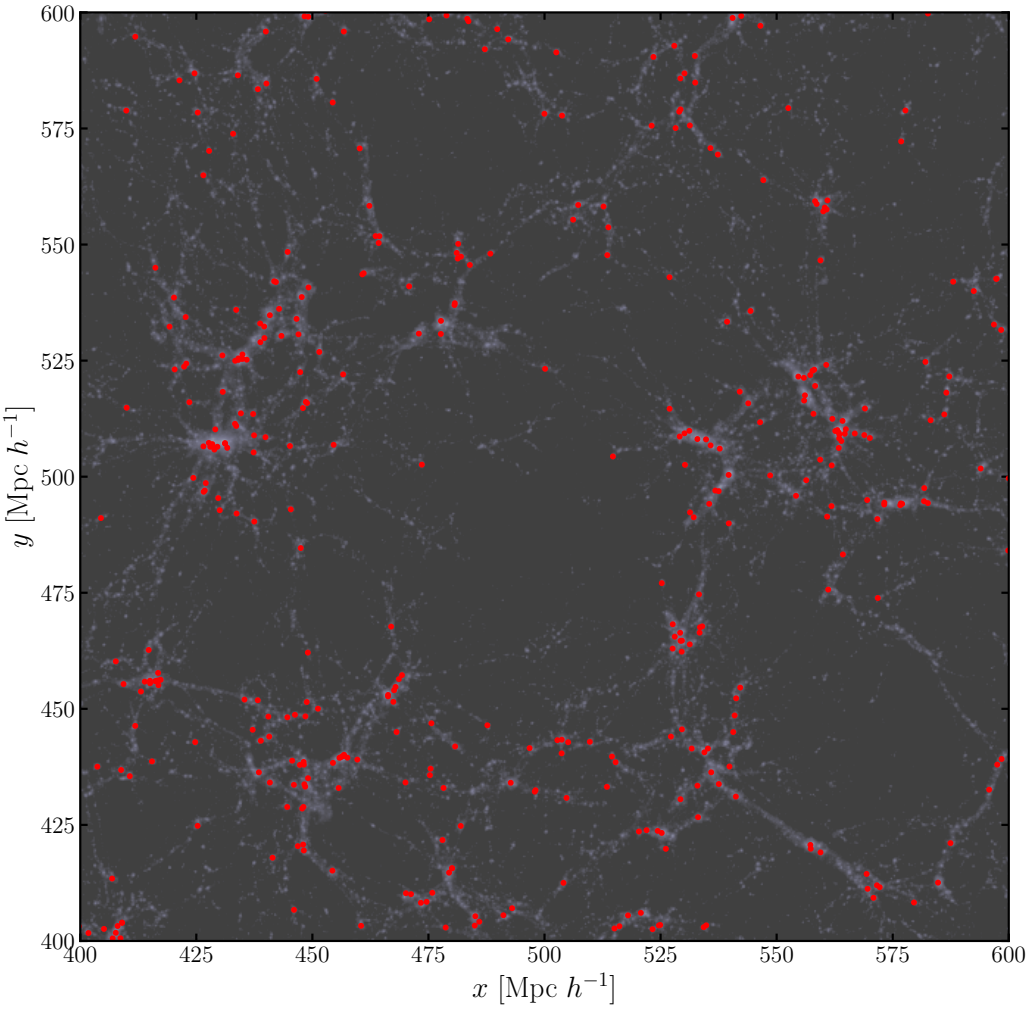


Figure 5.11: Same as figure 5.1, but adding the distribution of galaxies tracing the underlying dark matter. Galaxies are placed using the HOD method, with the parameters tuned to replicate the observed abundance and clustering of BOSS galaxies.

be compensated by higher values of $\log M_{\mathbf{min}}$, $\log M_1$, and $\log M_0$. This compensation also works for α, because fewer haloes will be populated with satellites, but more satellites will be added to obtain the same number density and clustering as in the catalogues produced from the GR simulations.

The catalogues created by this method have the same distributions of galaxies as the LOWZ and CMASS observational samples at the level of the one and two-point statistics, using measurements of the number density of galaxies and real-space clustering. In Figure 5.11 we show the HOD mock galaxy catalogues created using

this method, overlying the dark matter distribution shown in Figure 5.1. The HOD model shows good indication of reproducing the clustering of galaxies on various scales, in the simulations. It also appears to be a good tracer of the dark matter distribution in the simulations. These galaxy catalogues are used to calculate the marked correlation function, as we explain in the next chapter.

Chapter 6

The marked correlation function of LOWZ and CMASS galaxies as a test of modified gravity.

6.1 Introduction

The study of the clustering of galaxies at the level of the two-point correlation function is a robust test of the large-scale structure and has been used to study cosmology over the last decades. The accurate measurements provided by observations of galaxies that trace the density field offer tight constraints on how the clustering of galaxies looks like. This requires that when studying alternative theories that differ from the GR-ΛCDM universe, the number density and clustering of the mock catalogues from these simulations have to replicate the results of the real-projected-space correlation function of the most recent extragalactic surveys (Cautun et al., 2018). In other words, the two-point correlation function is limited as a means to distinguish GR from alternative gravity models.

We have shown that models such as the HOD prescription can be tuned to reproduce the number density and clustering as displayed by observational samples at the per cent level. The question then arises: What kind of test can we use to probe

modified gravity models? To answer this question we need to take into account the features of the modified gravity model we want to explore, such as the screening mechanism in $f(R)$ gravity that was explained in Chapter 2. In White (2016), the density-marked correlation function is proposed to study modified gravity models as an easy-to-compute test, as this uses a property that is expected to depend on modified gravity which is easy to obtain when observing galaxies in extragalactic surveys.

The idea of the marked correlation function has been tested using mock galaxy catalogues (Armijo et al., 2018; Hernández-Aguayo et al., 2018), motivated by the theoretical background presented in White (2016) using low order perturbation theory to explore the properties of the marked correlation function. In these studies, different definitions of weights applied to galaxies were investigated, including the local density for individual galaxies, the gravitational potential of different environments, and the host halo mass. All of these properties are expected to differ from the ΛCDM paradigm when calculated in modified gravity models, even when the 2-point clustering is matched. Moreover, some of these marks have already been tested in Satpathy et al. (2019), where the density-mark from White (2016) is applied to the LOWZ galaxy sample, using the marked correlation function defined in redshift-space. These authors concluded that their results are limited by the modelling of small scales in the simulations, where most of the differences between GR and MG models were found in previous studies. Nonetheless, no significant deviations from ΛCDM were found by Satpathy et al. on scales between $6 < s/(h^{-1}\,\mathrm{Mpc}) < 69$. The simulations used in Satpathy et al. (2019) have limited resolution when using the subhaloes to study their galaxy catalogues, which can drive the results on small scales, which motivates us to refine some aspects of their analysis. Furthermore, the analysis of Satpathy et al. is done testing the marked clustering in redshift-space, which is dominated by the pair-wise velocity distributions on small scales that require further modelling of differences between GR and modified gravity models.

As explained in Armijo et al. (2018), the same analysis in real space can be used to unveil modified gravity theories such as $f(R)$ in the scales not investigated by Satpathy et al. (2019). Another motivation for revisiting this study is to explore the option of using marks based on gravity (i.e. the gravitational potential) or mass, which can be viewed as providing more direct tests of modified gravity. For such marks, observations of weak lensing measurements become relevant, in addition to samples including other observational mass estimates, such as those of clusters of galaxies (Cataneo et al., 2018; Liu et al., 2021). For these reasons, we apply a real-space version of the marked correlation function based on the projected correlation function to the BOSS LOWZ and CMASS samples, which we explain in the next Section.

6.2 The projected marked correlation function

Following White (2016) we define the marked correlation function as

$$\mathcal{M}(r) = \frac{1 + W(r)}{1 + \xi(r)}, \tag{6.1}$$

where $\xi(r)$ is the two-point correlation function and $W(r)$ is the weighted or marked version of ξ. Rather than counting pairs of galaxies, the product of weights is counted for a given pair of galaxies. To implement the measurement of the mark correlation function we simply include the marks as additional weights in the correlation function estimator, adapting the example given in Eqn. 3.5, where the pair counts are replaced by the multiplication of the weights for each galaxy in the pair. We count pairs from the data and random catalogues, with the terms in the estimator

defined by Eqn.3.5 redefined to include the mark:

$$DD = \frac{1}{N_g(N_g-1)}\sum_{ij} w_{\mathrm{mark},i} w_{\mathrm{mark},j}, \tag{6.2}$$

$$DR = \frac{\bar{m}}{N_g N_r}\sum_{ij} w_{\mathrm{mark},i} w_{\mathrm{tot},j}, \tag{6.3}$$

$$RR = \frac{\bar{m}^2}{N_g N_r}\sum_{ij} w_{\mathrm{tot},i} w_{\mathrm{tot},j}, \tag{6.4}$$

where w_{mark} is the value of the mark for each galaxy, and w_{tot} includes the observational weights from the survey data. We note that randoms are marked by the mean mark $\bar{m}$.

The prescriptions for the construction of marks and weights is defined in Satpathy et al. (2019) to ensure that the weighted correlation functions depend on the local densities around galaxies. The definition of the total weight can be defined as

$$w_{\mathrm{mark},i} = m_i \times w_{\mathrm{tot},i}, \tag{6.5}$$

where m_i is the individual mark for each galaxy (see below). The $w_{\mathrm{tot},i}$ term is the total weight of a galaxy, including observational artefacts as explained in Reid et al. (2016), and the calculation of FKP weights (reviewed in 3.3), which gives an unbiased scheme for the estimation of the galaxy density field. For density-motivated definitions, the mark represents an estimation of the local density of an individual galaxy, ρ_i, which is defined as the inverse of the volume occupied by a galaxy in the density field, in terms of the mean density $\bar{\rho}$ of the field. Then we define marks of the form

$$m = \left(\frac{\rho}{\bar{\rho}}\right)^p, \tag{6.6}$$

where p is a free parameter we can vary, in order to up-weight different types of density environments. For example, a selection of $p < 0$ can be used to up-weight low density regions, where the additional gravity force in MG is triggered. On the other hand, $p > 0$ is equally useful as in this case high-density environments are favoured, and halos in unscreened regimes can be tested. Note that any normalization of ρ introduced in Eqn. 6.6 will be included in the value of $\bar{m}$ in the

estimators of Eqns. 6.2, 6.3 and 6.4. All these definitions produce similar results in distinguishing MG from GR than using the log-transform density field power spectrum and the clipped density field statistic (Valogiannis et al., 2018).

Here we focus on the real-space projected clustering, and some of the definitions change. Instead of measuring the correlation function in redshift-space, $\xi(s)$, as used in White 2016; Satpathy et al. 2019 we decide to utilize $w_p(r_p)/r_p$, the projected correlation function divided by the projected perpendicular distance r_p. The aim here is to avoid dealing with the modelling of redshift-space distortions, which add a layer of complication (see e.g. Cuesta-Lazaro et al. 2020) and can introduce noise in the final conclusions (Satpathy et al., 2019). Current RSD modelling performs best on intermediate to large scales, for which it is more challenging to distinguish modified gravity from GR (Paillas et al., 2019). Another reason for choosing to work in real-space is that the effects of RSD modify the local densities obtained from the Voronoi tessellation, as shown in Armijo et al. (2018), which reduces the signal of modified gravity in the amplitude of the marked correlation function. Finally, measuring RSD on these scales to test modified gravity is not in the scope of this study, which is already known to be difficult to model for $f(R)$ theories (Hernández-Aguayo et al., 2019). In the next section we explain more about the choice and calculation of density dependent galaxy marks.

6.3 Local density estimation: the Voronoi Tessellation

We decide to base the estimation of the galaxy local density on Voronoi tessellation (Voronoi, 1908) in 2D as we focus on projected-real space statistics. This is a computational method to tessellate the space according to a given geometrical criterion. The Voronoi tessellation is defined in general by a n-plane with N points, where each point generates a n-polytope* that contains all of the region closer to that point than to any other. The estimation of the local density for our galaxies

*The n-dimension generalization of a polyhedron

is performed in a 2D projection of the original XYZ 3D Cartesian coordinates. For the simulations, this is a straightforward procedure. In our case, a galaxy sample generates a set of Voronoi cells in two dimensions, each with an area, coming from a projected local volume (a thin 3D slice). With this area we define the volume V_i since the remaining dimension is provided by the thickness of the slice, and define the local projected density:

$$\rho_i = \frac{1}{V_i}. \tag{6.7}$$

Estimating the local density by using the Voronoi approach is a relatively inexpensive and intuitive method, where galaxies in overdense environments will have small volumes and hence high densities, and more isolated galaxies will have larger volumes and therefore smaller densities. Voronoi tessellations have been used in a wide range of problems in astrophysics and cosmology, such as the identification of cosmic voids (Platen et al., 2007; Neyrinck, 2008) and probing the primordial cosmology and galaxy formation (Paranjape et al., 2020). In Figure 6.1 we show the Voronoi diagram of the galaxy distribution drawn in Figure 5.11. In the top panel, we show the shape of the actual Voronoi cells in the 2D projection of the $40h^{-1}\,\mathrm{Mpc}$ thick slice, which comes from one of the HOD catalogues produced from the cubic box simulations. Here, the cells of different sizes are generated by tracers of the underlying matter field and are representative of the environment in which they reside. In the bottom panel, we relate these Voronoi cells to the actual marks m defined by Eqn. 6.6, with an arbitrary value for p, divided by the value of the mean mark $\bar{m}$. Then, we colour each Voronoi cell to show how different regions are up or down-weighted when the marked correlation function is computed. For example, small scales dominated by clusters and groups of galaxies are boosted when counting pairs, whereas pairs that include more isolated galaxies yield smaller marks.

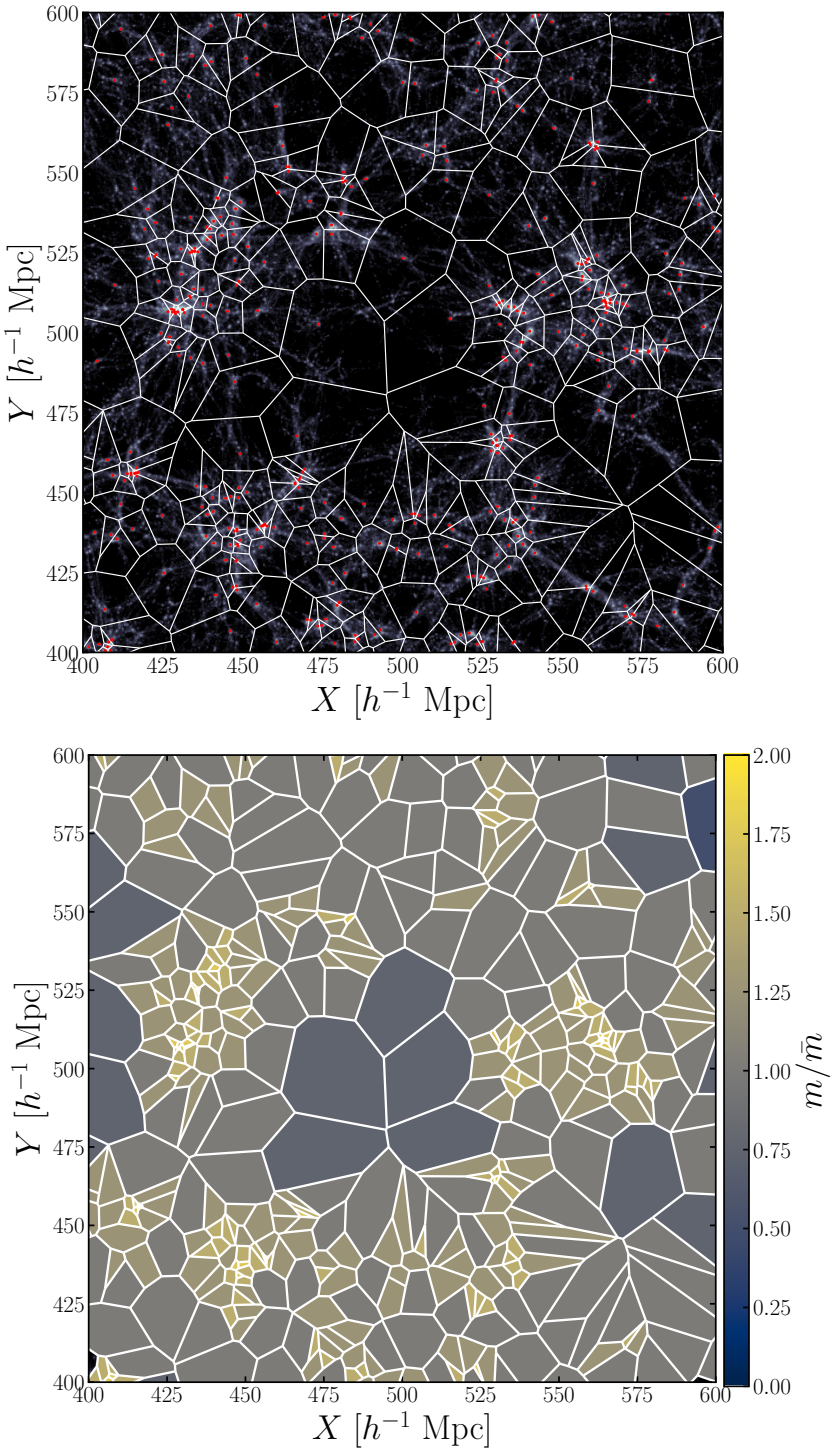


Figure 6.1: Top Panel: Two dimensional Voronoi diagram of the galaxy distribution shown in Figure 5.11. The polygons indicated by the white lines are calculated using the Voronoi tessellation for the projection of a slice of thickness $\Delta Z = 40h^{-1}\,\mathrm{Mpc}$ projected in the XY plane. Bottom panel: same as in the top panel but colouring individual Voronoi cells using the respective value of the mark of the galaxy in that cell, divided by the mean mark.

6.3.1 The shape of the local density distribution

A feature of Voronoi tessellation is that the distribution of the volumes has as its mean the inverse of the number density of tracers used to compute the tessellation, so for a set of N_g galaxies

$$\langle V \rangle = \frac{1}{N_{\rm g}} \sum_{i}^{N} V_i = \frac{V_{\rm tot}}{N_{\rm g}}. \tag{6.8}$$

Here $V_{\rm tot}$ is the total volume of the sample, and by definition $(V_{\rm tot}/N_{\rm g}) = 1/n_{\rm g}$. This means that the mean of the density distribution is equal to the galaxy number density of the tessellated sample, $n_{\rm g}$. Hence, when correctly normalized and if not dominated by shot noise, the distribution of densities of two samples tracing the matter density field in the same way are identical if they have the same number density.

We use the distant observer approximation, where the line-of-sight is taken to be parallel to a fixed, preferred axis of the simulation box. To project structures in their local environments only, we perform the Voronoi tessellation in 20 thin slices of $\Delta Z = 38.4 h^{-1}\,$Mpc for the total of the L768 simulation at redshift $z = 0.3$, and 25 slices with $\Delta Z = 30.72\, h^{-1}\,$Mpc for the snapshot $z = 0.5$. The choice of the number of slices at each redshift is made to preserve the mean volume $\bar{V}$ for the mocks with different number densities of tracers. This selection of slices allow us to avoid excessive projection effects when computing the Voronoi tessellation, which can decrease the amplitude of the marked correlation function on the scales in which we are interested. The selection of the number of slices is made to obtain an optimal projection, given that the slice has to be thick enough to avoid splitting structures between different slices. However, we also need to prevent projecting too many galaxies in one slice as this varies the projected number density of a single slice, and because of the projection more galaxies converge to this number, shrinking the distribution of volumes, as can be observed in Figure 6.2.

Here, we compare the distribution of Voronoi cell volumes for the same simulation changing the number of slices we divide the box when we create the tessellation.

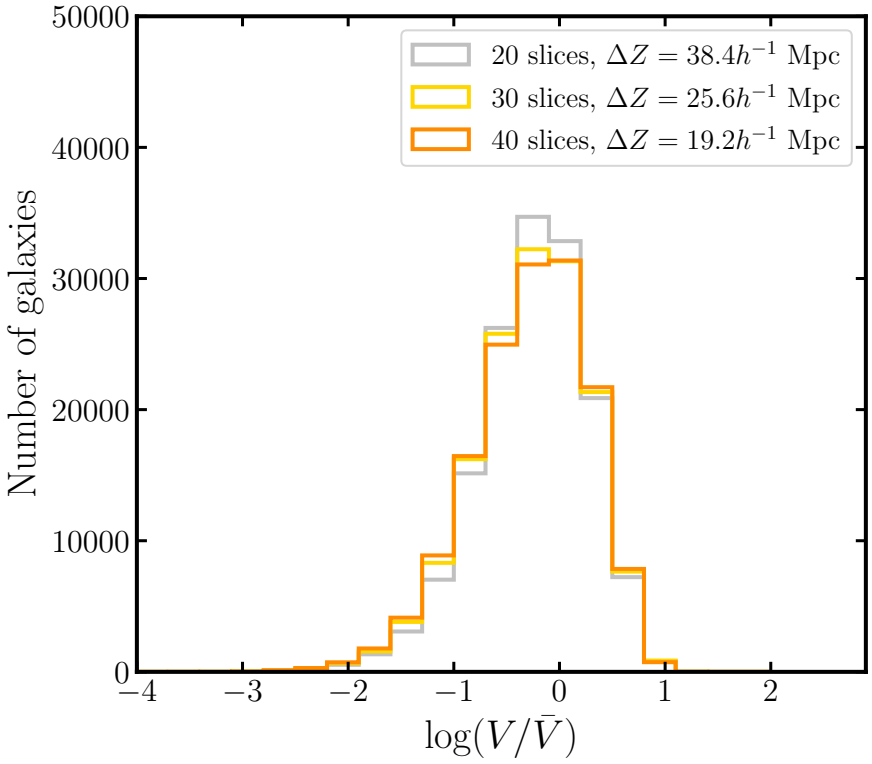


Figure 6.2: The distribution of the logarithm of Voronoi cell projected volumes, V, in units of the mean slice volume of the distribution, $\bar{V}$, for a HOD galaxy catalogue generated using the L768 simulation. The distributions are shown for different numbers of slices used to create the projection space before the 2D tessellation is performed: 20 (grey), 30 (yellow), 40 (orange).

Even though the different distributions have the same shape when divided by the mean volume $\bar{V}$ of each individual tessellation, we note that when using 20 slices there are more galaxies with values close to $\bar{V}$. Although one solution would be to create more slices, we also need to consider how many structures would be intersected or split by the slice boundaries. Nevertheless, we check that this effect is not a large systematic error in the marked correlation function, as the results do not vary strongly when using different number of slices.

6.3.2 Tessellation of the LOWZ and CMASS lightcones

The angular and radial distribution of galaxies in the real survey varies significantly in comparison to that in the idealised mock catalogue, where a cube box

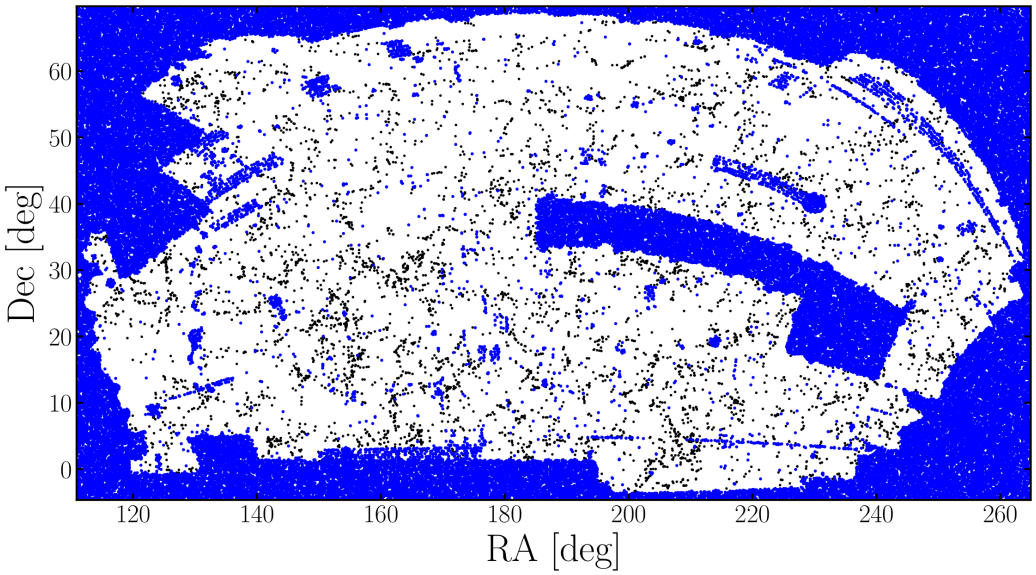


Figure 6.3: Right ascension (RA) and declination (Dec) for a set of galaxies in a thin redshift slice with $\Delta z = 0.008$ for the LOWZ sample. The black dots show galaxies within the survey in the redshift slice. Blue dots cover the survey mask.

with periodic boundary conditions is used to generate the galaxy sample. The window function imposed by surveying galaxies in the Universe is full of observational artefacts and limited by how much time and area were dedicated performing the observations. This means that a survey mask has to be applied if we want to perform an analysis that is as close as possible to that applied to the real observations.

For correlation functions, the estimators are designed to take into account irregular survey boundaries and varying radial number densities of galaxies. However, the calculation of the local density using Voronoi tessellation will be sensitive to the angular footprint of the survey and to any holes within the nominal survey boundary, hence the need to apply the angular mask to the mocks. In addition, for the observed galaxies, the sky position (RA,Dec) plus redshift z is measured, rather the 3D Cartesian position. Both frameworks need to be made in a consistent way, i.e. using the same positional information and applying the angular and radial selections, before tessellating the galaxies in the survey. For example, to deal with the edges of the surveyed sky region we generate a random sample of points that acts as a buffer to embed the survey within a rectangular patch of the sky, so that

we can stop the tessellation at the borders. This approach also applies for holes and regions within the surveyed area, due to the presence of artefacts or additional issues, that eliminate those pixels in the final samples.

In terms of the Voronoi tessellation, these random points act as a screen that prevents the tessellation going further than the edges of the survey and creating large cells for galaxies near the borders. In the same way, by filling the holes with the same random particles, galaxies near these regions will not be identified as being part of a void region, rather to have a more appropriate size for the Voronoi cell that the galaxy defines. In Figure 6.3 we plot both the distribution of galaxies and the random particles which "wrap around" the survey for a thin redshift slice. The number density of these random particles are denser than the $n(z)$, by an extra factor f_p to make sure that all the pixels in the mask holes are covered. This is an iterative process that we need to repeat until the measurements of the Voronoi cell volumes converge. We find that for $f_p = 10$ the distribution of volumes and the calculation of the marked correlation function becomes stable. By doing this the tessellation runs smoothly for the galaxies and random particles projected in the 2D space for a fixed redshift slice. In the case of our LOWZ subsample defined between $0.24 < z < 0.36$ we create 8 redshift slices with mean thickness of $\Delta Z = 38.42\,h^{-1}\,\mathrm{Mpc}$, whereas for CMASS, 4 samples are defined with a mean thickness of $\Delta Z = 30.72\,h^{-1}\,,\mathrm{Mpc}$. The slightly smaller slice thickness adopted for the CMASS slices are chosen to preserve $\bar{V}$ exactly as with the simulations, due to the higher galaxy number density of the CMASS sample in comparison to LOWZ.

The redshift slices for the LOWZ sample (and analogue for CMASS) are shown in Figure 6.4, where the dashed red lines mark the limits of the 2D projections used to perform the Voronoi tessellation. Here, it is possible to see the individual structures that we project in a single redshift slice. Once we tessellate both samples, we need to compare the tessellation of the data to that from the mock catalogues. In Figure 6.5 we compare both the tessellation of the data (the case for the LOWZ sample) and the HOD mock catalogues from the GR box simulation at redshift

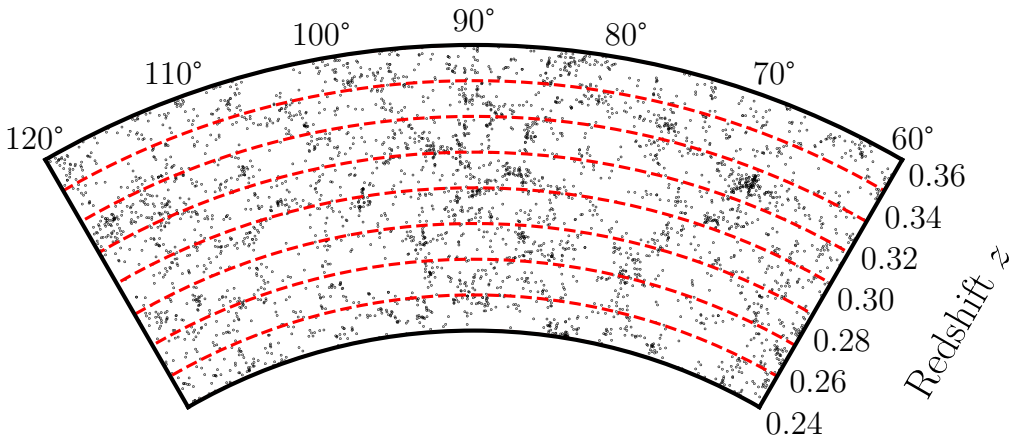


Figure 6.4: Angular distribution of galaxies in the LOWZ sample in a window of 60° in right ascension and a section of the radial coordinate, displaying the redshift, for a thin slice of ΔDec = 3.5 deg. in declination. We mark the 8 redshift slices (red dashed lines) with $\Delta z = 0.015$ used to perform the Voronoi tessellations in a 2D space.

$z = 0.3$. This comparison shows that there are some differences in the shape of the the distribution of Voronoi volumes, or local densities, that could impact the possible measurements of the marked correlation function of the data and the models. For this reason, we think that we need to explore the impact of the geometry further,to study mark correlation function measurements in mock catalogues applying the footprint geometry of the survey.

As we are projecting the galaxy positions from a spherical shell to a 2D plane to perform the Voronoi tessellation, differences in the actual area for the local density estimation can arise. However, such differences are small, thanks to the small angles between galaxy pairs that define the Voronoi areas for objects in 2D,

which produces small Voronoi cells with similar areas if they are calculated from a flat or spherical plane. In fact, Na et al. (2002) shows that Voronoi tessellations in the sphere plane share the same properties as the same tessellation in a flat surface. We consider that the areas produced by the galaxies projected in a 2D surface are indeed small and the impact on the difference if we would consider a spherical surface negligible. In this scenario we transform points from the sphere surface to the 2D plane using the Stereographic projection. This projection defines XY points in the plane using the following transformation:

$$X = \frac{\sin\delta\cos\alpha}{1-\cos\delta} \tag{6.9}$$

$$Y = \frac{\sin\delta\sin\alpha}{1-\cos\delta} \tag{6.10}$$

For a sphere of radius $R = 1$, α the azimuthal angle, and δ the angle along the pole. The sphere has the origin in the "north pole". The impact on our results using this transformation is minimal, and is even smaller if we consider that the marked correlation function down-weights large areas with the definition of mark we adopt in Eq. 6.6.

6.3.3 Mock lightcones of the LOWZ and CMASS samples

As we mentioned earlier in Chapter 3 we make a comparison between mock catalogues built from N-body simulations and the survey data when we compute the clustering. However, as the calculation of the Voronoi tessellation is purely geometrical, we need to check if the angular footprint of the survey has an impact on the distribution of Voronoi cell sizes, the definition of the marks, and, in turn, the marked correlation function.

First, to compare the distribution of the Voronoi tessellation volumes between the simulation box and the survey, we need to move the latter to a 2D XY Cartesian coordinate system. Originally for the lightcone coordinates, we only have equatorial angles α and δ (right ascension and declination), so we need to apply the stereographical transformation in Eqns. 6.9 and 6.10. We know that in the distant

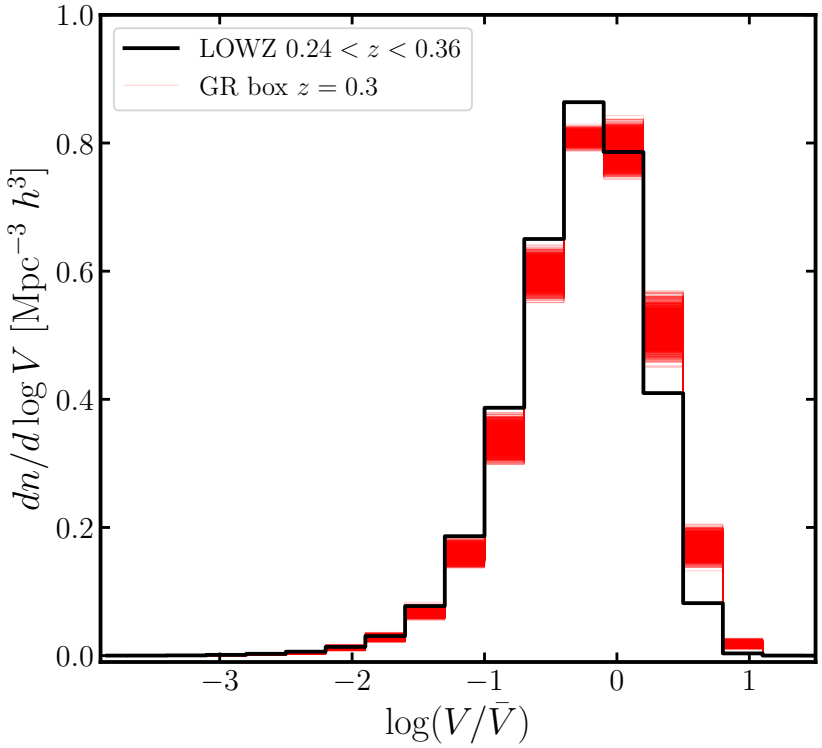


Figure 6.5: The distribution of Voronoi cell volumes for the projected slices for a comparison between the HOD mock catalogues from periodic simulation boxes (red lines) and the LOWZ $0.24 < z < 0.36$ data (black line). 1000 HOD catalogues selected from the random sampling explained in 5.4 are selected to represent the samples that match the galaxy number density and clustering.

observer approximation, two objects at the same redshift z have a comoving distance $d_C(z)$. If they are separated by an angle $\theta \sim 0$, then the transverse distance d is defined by

$$\tan\theta = \frac{d}{d_C(z)}, \tag{6.11}$$

$$d \approx \theta\, d_C(z). \tag{6.12}$$

For our coordinate system we can write $\theta = \alpha_1 - \alpha_0$, then the distance $d_x = \delta'\alpha\, d_C(z)$ if $\delta'\alpha$ is small, (with the analogue for the declination δ). These approximations are valid for the galaxies we are projecting into a 2D plane using the stereographical projection in a thin redshift slices, because we evaluate $d_C(z)$ at a fixed z_s, the mean redshift of the slice. Also, as the angular separation in α is small for all neighbouring galaxies, with $\alpha \ll 1\,\mathrm{deg.}$, when projected in the plane, which

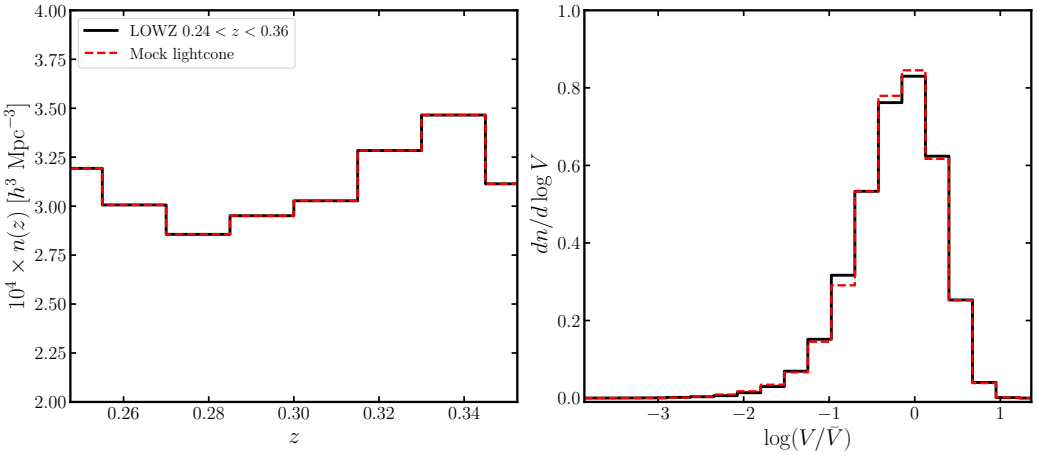


Figure 6.6: Left: The number density distribution $n(z)$ for the subsample of LOWZ and a mock lightcone which has been randomly sampled to have the same $n(z)$. Right: The distribution of Voronoi cell volumes $\mathrm{d}n/\mathrm{d}\log V$ for the mock lightcone and the LOWZ subsample.

means that the Voronoi cells are correctly defined, as these are only generated by the direct neighbours of each galaxy, which keeps the plane approximation locally. The definition of this Cartesian 2D plane allows us to compare the tessellation in the survey, and that in the mock lightcone, with the box simulation.

To create the mock lightcone we need to follow the following steps:

1. Move the origin to the centre of the box, where we place the observer.
2. Transform from Cartesian to spherical coordinates $(x, y, z) \rightarrow (r, \phi, \theta)$ for all galaxies in the mock catalogue.
3. Select the limits in comoving distance for the lightcone. The box is repeated periodically in case we reach any border.
4. Add the velocities of each galaxies projected along the line-of-sight direction.
5. Apply the sky mask of the survey to the lightcone.

We are not trying to re-create an accurate lightcone of the survey samples at this stage, rather our aim is to understand whether or not the geometry of the survey will affect the calculations of the marked correlation function. As the distribution

of the Voronoi cell volumes is related to the number density of tracers used in the tessellation, we need to mimic the number density distribution as a function of redshift $n(z)$. To do this, we use a mock galaxy catalogue with a number density similar to the one from the survey, which can be randomly sampled to obtain the same $n(z)$ distribution as the observations. We plot the results of this random sampling in the left panel of Figure 6.6, where the number of galaxies per redshift bin are virtually the same. Then, by comparing the distribution of Voronoi cell volumes $\mathrm{d}n/\mathrm{d}\log V$, in the right panel of Figure 6.6, we note that $\mathrm{d}n/\mathrm{d}\log V$ for both the mock lightcone and the survey are very close. Moreover, when comparing to Figure 6.5 the offsets are corrected and the smaller differences are more consistent with the uncertainties using different HOD catalogues, displayed by the different red lines in Figure 6.5 and the shaded area in Figure 6.7. Although, the differences showed by the histograms of the data and the mock, hint us that this effect is stronger than projecting the points of the sphere into the plane, which we already consider that affects large cells only.

We know that the random sampling of galaxies that matches $n(z)$ between mock and data does not impact the two-point correlation function calculation, which is already the same for mock and data, by design. Also, the distributions of Voronoi cell volumes are in good agreement after this procedure, meaning that the measurement of the Voronoi cell volumes is correct for both cases. The distribution of volumes for both samples has the same shape when we match the number density of the tracers in the complete volume, given the connection between the two shown in Eqn. 6.8. Considering this, we can now compare if by measuring the marked clustering using the volumes (local densities) as marks, we have comparable measurements between the marked correlation functions for mock catalogues and mock lightcones using Eqn. 6.1, applying the marked version of the estimators in Eqns. 6.2, 6.3, and 6.4. In Figure 6.7 we plot results for a marked correlation function of the same HOD catalogue from the GR box at $z = 0.3$, with the marks defined by the tessellations of the box and the lightcone. The comparison

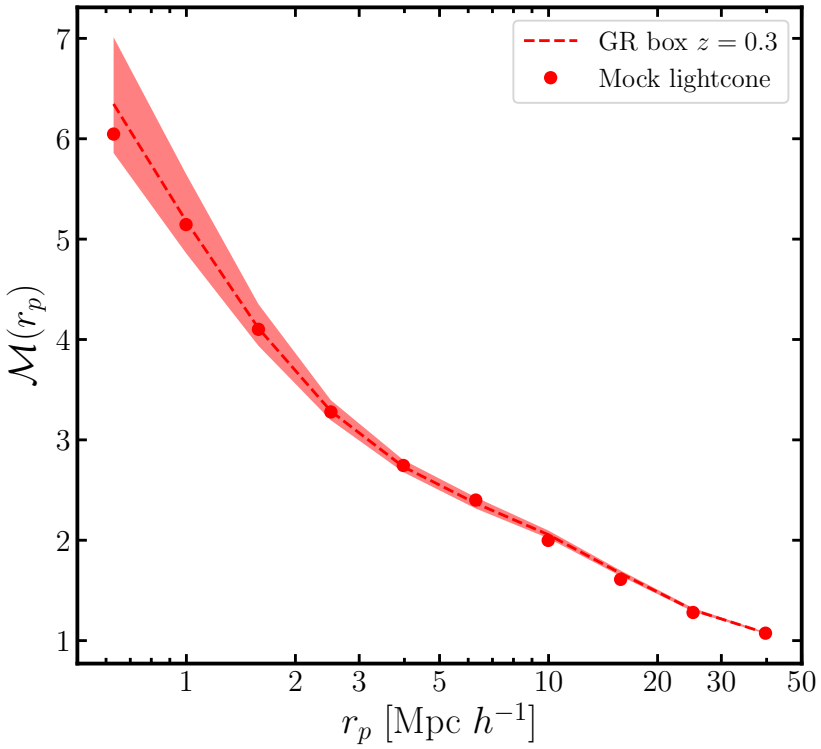


Figure 6.7: The marked correlation function $\mathcal{M}(r_p)$ as a function of the projected distance r_p using the same HOD mock catalogue from the original box (red dashed line) and the mock lightcone with the SDSS footprint geometry (red dots). The light-red shaded shows the uncertainties of the HOD model for the GR $z = 0.3$ simulations.

shows similar results for $\mathcal{M}(r_p)$ for both schemes, with deviations at small scales for r_p, but not higher than the uncertainties provided by the HOD modelling. This is expected when considering that the two-point clustering is the same for both the lightcone and the box regardless of the random sampling, and because if we examine the distribution of local densities for the different bins of separation r_p, the mean and the variance of the distribution of $\log(V/\bar{V})$ does not vary drastically. We conclude that the differences shown by the Voronoi volume distribution of Figure 6.5 can be conciliated when resampling the mock lightcone as shown in Figure 6.6. As this step does not strongly modify the results of the marked correlation function, we decide to make the comparison between the mock catalogues from the boxes and the survey data for our main study.

6.4 LOWZ and CMASS marked correlation functions

We calculate the marked correlation function of the LOWZ and CMASS samples using the marks derived from the local density measurements obtained using the Voronoi tessellation. To compute the terms in Eqn. 6.1 we use the Landy-Szalay estimator presented in Section 3.5 to calculate $\xi(r_p, \pi)$. When solving the integral in Eqn. 3.6 we consider separations in the line-of-sight direction, π, using logarithmic bins. By doing this, we obtain more accuracy to the integral calculation for the small π separations on which the correlation function changes rapidly. Then the differential term of the integral of Eqn. 3.6 can be written as $\mathrm{d}\pi = \pi \mathrm{d}(\log \pi)$. We use the publicly available TWOPCF* code to compute the $w_p(r_p)$ for the data and mock catalogues; this code supports logarithmic binning and estimators using weighted pairs. The code can also efficiently calculate jackknife errors in a single loop over the galaxy pairs. For the mock catalogues, we select a random sample of 1000 of the HOD parameters within the 1-σ confidence interval obtained in Section 5.4. To study the marked statistic of the HOD mock catalogues we select 68% of the total sample of values closer to the mean of $\mathcal{M}$ for each model. The argument for this is that from the whole range of the HOD catalogues selected, there is a probability to pick a HOD set that reproduce the correct marked correlation function for the data, as we are trying a different test than the one used to fit the data.

We plot the results of the marked correlation function $\mathcal{M}(r_p)$ for the LOWZ and CMASS subsamples in Figure 6.8. In the left panel of Figure 6.8, we compare $\mathcal{M}(r_p)$ for the GR and F5 models created from the snapshot at redshift $z = 0.3$, using the random sampling of the HOD parameters within the 1-σ confidence interval region. These results are compared with the measurement from the LOWZ sample in the redshift range $0.24 < z < 0.36$. Both models agree with the data at separations larger than $r_p > 0.8\,h^{-1}\,\mathrm{Mpc}$. It is only for $0.5 < r_p/(h^{-1}\,\mathrm{Mpc}) < 0.8$ that the GR model is a somewhat better match to the data than F5. These are

*https://github.com/lstothert/two_pcf

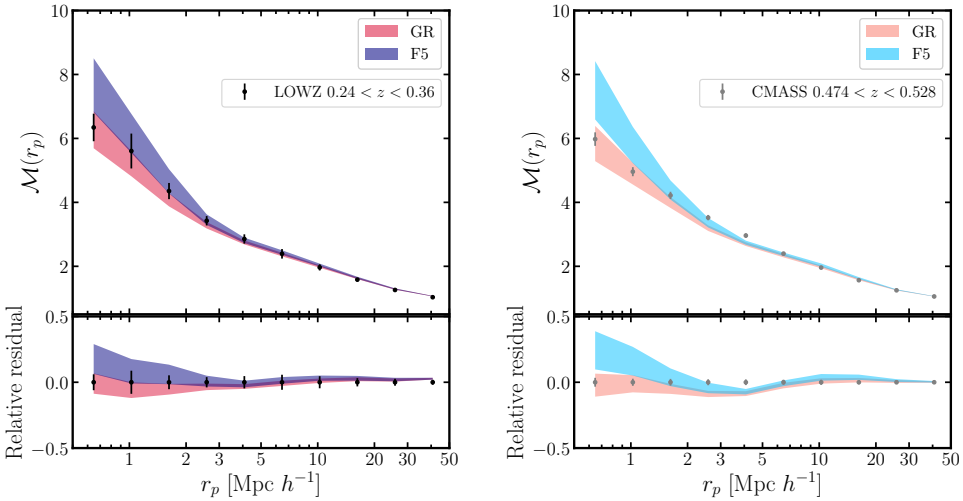


Figure 6.8: The marked correlation function $\mathcal{M}(r_p)$ as function of the projected distance r_p for the BOSS galaxy samples and the results from the respective HOD mock galaxy catalogues from the GR (red) and F5 (blue) simulations. Left panel: $\mathcal{M}(r_p)$ measured from LOWZ (black dots) at $0.24 < z < 0.36$ compared with the HOD mock catalogues within the 1-σ confidence interval from the MCMC fitting of the two-point clustering and number density. Right: same as left panel, but for the CMASS subsample (grey dots) at $0.474 < z < 0.528$. The shaded areas for the models come from selecting the 68% of all the family of HOD catalogues of each model, GR, F5 at redshift $z = 0.3$ (dark red and dark blue) and $z = 0.5$ (light red and light blue). The error bars on the data are estimated using Jackknife resampling, with 100 subvolumes of the data. In the bottom panels we show the relative residuals using the data measurements as a reference, meaning that we display $\mathcal{M}^{\rm mod}/\mathcal{M}^{\rm data} - 1$, with $\mathcal{M}^{\rm mod}$ the marked correlation function for each HOD set and $\mathcal{M}^{\rm data}$ is the marked correlation function of LOWZ and CMASS in left and right panels respectively.

the r_p separations where the range of acceptable HOD parameter values lead to a large spread in $\mathcal{M}(r_p)$ for both models. We compare the results of the GR and F5 simulations at redshift $z = 0.5$ with the CMASS measurement in the right panel of Figure 6.8. Although, similar conclusions are reached in this case as for the LOWZ data, there are other interesting features to note. The errors from CMASS are smaller, due to the higher number density in this sample compared with LOWZ, which makes it easier to see whether the data is likely to follow one model rather than the other. In fact, we note that below $r_p < 1.7h^{-1}\,$Mpc the CMASS data agrees with the GR model at the 1-σ level, whereas there is a clear tendency for the F5 model to disagree with the data. This result is derived considering only

the diagonal errors from the covariance matrix, which simplifies the analysis as we are not considering the covariance terms between different separation bins. To obtain a more detailed result on how significant this discrepancy is a full covariance matrix analysis should be adopted. This could be performed, for example, using an approach to generate large numbers of simulation volumes following the method proposed by Hernández-Aguayo et al. (2021). Both models fail to match the CMASS observational estimate on scales $2.0h^{-1}\,\mathrm{Mpc} < r_p < 5.0h^{-1}\,\mathrm{Mpc}$. We attribute this mismatch to the CMASS data being more difficult to replicate, when including extra information to define the samples as explained in Chapter 3.2. Finally, for separations $r_p > 10h^{-1}$ Mpc the data agrees better with both models, in particular the GR simulations. However the differences are smaller and there is a small overlap between the confidence regions of both models.

6.5 Discussion

We discuss the measurement of the marked correlation function for the LOWZ and CMASS samples and the comparison to HOD mock galaxy catalogues generated from the GR and modified gravity simulations. We find that for two models of gravity (GR and F5), we obtain different results for the marked correlation function. This is a limited result in terms of distinguishing modified gravity, as models differ by about 1-σ at most, in some scales, which is a marginal and not significant result. Nevertheless, the marked correlation function as a test, helps on breaking the degeneracy between modified gravity and the HOD modelling, which is expected for this kind of test (White et al., 2009). In regard to the latter, we check that HOD parameters within the 1-σ confidence interval reproduce the shape distribution of the local densities for the observational samples, with an interesting caveat. For the CMASS sample, the variations of $n(z)$ over the redshift range used proves to be difficult to model, which leads to disagreements between the data and the models over intermediate projected separation. The results for the distribution of

Voronoi volumes in Figure 6.5 show that even small differences between the number density of galaxies will generate a change in the shape of the volume histogram. We think that the fluctuations in the $n(z)$ of the CMASS subsample could drive the differences between the marked correlation function of CMASS galaxies displayed in right panel of Figure 6.8, where the differences between low and high number density bins as function of redshift could be larger than 20%. One solution here could be try a much thinner lightcone in redshift, to constraint the values of $n(Z)$ more close to the number density of galaxies in the mock catalogue. This is where the LOWZ observations become more relevant, where the more constant $n(z)$ is modelled correctly over the whole range of r_p we explore. At the same time, the higher number density of CMASS ($\sim$ 30% higher than in LOWZ) allows us to conclude that the data agrees slightly better with the GR model than F5, but only at the 1-σ level at small separations $r_p < 1.3h^{-1}$ Mpc. Such a difference is not enough to rule out a model like F5, but it suggests how we can explore further differences between GR and modified gravity.

The marked correlation function is a promising probe to distinguish between gravity models, as has been shown by previous studies such as (Valogiannis et al., 2018; Hernández-Aguayo et al., 2019; Armijo et al., 2018; Liu et al., 2021) and many others. However, the modelling of the number density seems to drive the results in terms of the amplitudes of $\mathcal{M}(r_p)$. Ongoing surveys such as DESI, which will yield larger LRG samples in terms of both volume and number density, may solve this issue (Zhou et al., 2020). The test with simulated data suggests that DESI will improve the number density by a factor of 2-3 in comparison to BOSS in a volume 9 times larger. Also, some modifications to this probe can be explored using different features of the MG models. Objects like galaxy clusters are promising observational probes to test these theories, where their masses can be incorporated into the mark if we cross-correlate them with the galaxy samples. As shown in Armijo et al. (2018) the mass of large haloes is in fact a better option for marking galaxies to distinguish between gravity models. Nevertheless, a large volume sample of galaxy

clusters including mass measurements is needed. A few candidates are available which could help with this issue, such as the COnstrain Dark Energy with X-ray (CODEX) clusters sample (Finoguenov et al., 2020), which provides measurements of clusters masses and their 3D space distribution. This sample is ideal to cross-correlate with BOSS galaxies as these surveys have the same footprint area. Also, probes motivated by studying the velocities of galaxies around clusters and weak lensing profiles will help us to obtain better constraints on modified gravity models.

Chapter 7

Summary and conclusions

We have introduced a new framework to test gravity on different scales using wide-field surveys. We use galaxies as tracers of the matter field to probe effects on the cosmic large scale structure introduced by modified gravity theories. Such models aim to provide an alternative to the cosmological constant to explain the accelerating cosmic expansion. These alternative theories of gravity are constrained by the successful predictions of the general relativity in a range of scales that goes from the solar system to the propagation of gravitational waves (Lombriser et al., 2016). The viable models we study present two interesting features: the screening mechanism to hide the modifications where GR is shown to be accurate, and the additional fifth force arising from the new degrees of freedom in modified gravity. Then, this fifth force can be detected in regions where the fifth force is unscreened, where GR still needs to be proven (Li et al., 2007).

To trace the large-scale structures we use observations of luminous red galaxies from the SDSS-III survey. We apply marked clustering statistics to samples of bright red galaxies from the Baryon Oscillation Spectroscopic Survey (BOSS, the LOWZ and CMASS samples) to detect the existence of a fifth force in the $f(R)$ theory of gravity. To achieve this, we select volume-limited subsamples of the catalogues at two redsfhit ranges, and compare with N-body simulations to search for the impact of modified gravity in these measurements.

We produce accurate mock catalogues that match the number density and unmarked two-point clustering of the observational samples. We find the HOD parameters that best fit these observational measurements using the MCMC algorithm, which leads to a set of mock catalogues that we use to predict the form of the marked correlation function. We also review possible systematic effects in the calculation of marks when projecting slices in an arbitrary direction in both the simulations and the survey.

We define a density-dependent marked correlation function using an estimation of the local galaxy density based on Voronoi tessellation. We provide conclusions for the main chapters of this thesis, highlighting the results of the marked correlation function test.

- Several tests of modified gravity (Cataneo et al., 2015; Liu et al., 2016; Armijo et al., 2018; Hernández-Aguayo et al., 2018; Valogiannis et al., 2018; Liu et al., 2021; Ruan et al., 2022) have been proposed recently to constrain or even rule out models. We present a new methodology to add to this canon of tests which applies the marked correlation function from White (2016) to probe the impact of modified gravity on cosmic large scale structure, taking advantage of the projected real-space information of the clustering of galaxies over a wide range of separations. These observations cover a substantial portion of the sky and sample a large volume over which our test can be applied.

- From the theoretical side, and to predict the behaviour of the marked correlation function, we prepare mock galaxy catalogues using simulations of a ΛCDM-GR universe, and compare these with mocks from a simulation which uses $f(R)$ theory of gravity with fifth force amplitude of $|f_{R0}| = 10^{-5}$ (in the parameterisation of Hu et al. 2007). We use the HOD prescription to populate haloes and subhaloes with central and satellite galaxies, from which we extract the best fitting parameters in terms of the reproduction of the projected correlation function $w_p(r_p)$ and galaxy number density $n_{\rm gal}$.

- We have presented a simple weighting scheme to compensate for 'missing' halos by upweighting those that are recovered by the halo finder. Our scheme is able, by construction, to reproduce a 'target' number density of halos, and returns improved estimates for the clustering of halo samples. As presented, our scheme requires at least two simulations. One is designated as the high resolution simulation and sets the target or benchmark for the halo sample statistics. By extending the usable halo catalogue derived from the low resolution run down to lower masses, significant computational resources can be saved. The scheme that we have proposed allows the resolution of a halo catalogue to be extended down to small particle numbers by applying a correction to the halos that we do see to account for those that we do not find. Ultimately, the scheme breaks down at the halo mass for which the errors in the clustering prediction become unacceptable. We design this approach to build more accurate mock catalogues, when using the halo information from the low resolution simulations.

- In order to obtain the HOD parameters that reproduce the observational metrics computed for the LOWZ and CMASS subsamples, we perform an MCMC search to find the best fitting parameters. We combine the measures of the galaxy number density, $n_{\rm gal}$, and projected two-point correlation function, w_p, in the least squares search for its parameters. This definition of the χ^2 allows us to test the dependency of these two measurements on the different HOD parameters. We test different weighting schemes of number density and projected correlation function in the definition of the overall χ^2. We find that better results are obtained if equal weights are given to both measurements. This decision is based on the definitions of convergence and the individual chains, in addition to the precision with which the measurements can be recovered. In the case of the number density, if too little weight is assigned to its contribution to the overall χ^2, the target value is not recovered with the uncertainties included, which favours models of the χ^2 where equal weight is

given to both the number density and clustering. Using the χ^2 distribution, we choose a range of HOD parameters within the 1- σ confidence interval to create mocks for both the GR and F5 simulations.

- The marked correlation function of the LOWZ and CMASS galaxies is calculated for subsamples selected in narrow ranges of redshift, $0.24 < z < 0.36$ and $0.474 < z < 0.528$ respectively. We create estimations of the 2D projected local density based on the Voronoi tessellation, for galaxies in thin redshift slices with $\Delta z = 0.015$ for LOWZ and $\Delta z = 0.0135$ for CMASS, which are compared with the tessellation of the simulation boxes with the same slice thickness in comoving coordinates, assuming the cosmology of the simulation. We also compare the tessellation of the observational and mock lightcone data adding the geometry imposed by the survey mask, which can be used to assess the robustness of the local density measurements. We find that to match the distribution of densities, the mock lightcone needs to reproduce the same galaxy number density as the observational data, which can be done by using a simple random sampling of the $n(z)$ distribution. This is an additional requirement imposed after searching for the best HOD parameters, to check the impact of the number density distributions on the marked correlation function. We check that this new sampling does not affect the two-point clustering and that it has a negligible effect on the shape of the marked correlation function, allowing the comparison of data to the HOD mock catalogues from periodic simulated boxes. When comparing the observations with the HOD mock catalogues, we find that for the LOWZ sample the data agrees with both mocks from both gravity models for separations larger than $r_p > 1.3\,h^{-1}\,\mathrm{Mpc}$. Our conclusions are probably limited by the sample variance of the data and the uncertainties introduced by the HOD modelling. The situation for the CMASS samples is better, however, with the GR model giving a better reproduction of the data for separations below $r_p < 1.3\,h^{-1}\,\mathrm{Mpc}$. There are also differences in the models that are appar-

ent at separations larger than $r_p > 10\,h^{-1}\,\mathrm{Mpc}$. These smaller differences between simulations suggest that a different analysis, such the one provided by adding RSD effects, could help to test such scales. The results provided by our method allow us to constrain $f(R)$ gravity using only spatial information from a galaxy sample. Nevertheless the uncertainties introduced by the HOD modelling restrict the analysis to a marginal 1-σ discrepancy between GR and $f(R)$ which is not enough to rule out a model as F5 with $|f_{R0}| = 10^{-5}$. Results from He et al. (2018) claim that a model like F6 ($|f_{R0}| = 10^{-6}$) can already be ruled out by using the redshift-space galaxy-galaxy correlation function, but using a different method for populating haloes with galaxies. Although, both methods are similar in terms of testing gravity using the distribution of galaxies at the two-point level, we introduce the uncertainties of the HOD modelling, which have not been considered before, whereas He et al. (2018) uses a subhalo abundance matching (SHAM) technique with a different estimation of the uncertainties for its model parameters. An extension of this work using the same SHAM method than He et al. (2018) to create the mock catalogue could improve our results.

- The modelling and results of the marked correlation function are highly dependent on the galaxy number density of the samples. For instance, the larger number density of CMASS galaxies improves the tendency of data to agree with the GR model, where it also deviates more from F5. Nevertheless, there is a regime in which the models fail to reproduce the data. We believe that the disagreement arises due to the CMASS sample being less adequate to compare to the HOD catalogues in periodic boxes, because of the additional selection of the sample that makes it more challenging to model. We think that this matter can be solved by future observations of LRG samples, with a higher and more uniform distribution of $n(z)$, or by making the comparison directly to mock lightcones with varying number density with redshift. We also propose to try new definitions of marks for galaxies or other probes to

include in the clustering estimations, such as mass estimations for clusters of galaxies, which are already proven to provide more information about modified gravity (Armijo et al., 2018). We will explore some of these ideas in the future work.

7.1 Future work

Another probe of the large-scale structure we can use to test modified gravity (MG) are clusters of galaxies. These objects correspond to massive perturbations of the initial power spectrum, and their abundance, for a fixed mass, depends on the growth rate of cosmic structures and the expansion history of the Universe. Various studies have used galaxy clusters to test both cosmology and deviations from the general relativity (Mak et al., 2012; Cataneo et al., 2015; Mitchell et al., 2018). For instance, Galaxy clusters are useful probes to test MG models, even if they correspond to high density regions and hence are expected to be screened from any fifth force, as their abundance is expected to be different than in GR (Cataneo et al., 2018). This is because in $f(R)$ gravity, unscreened haloes are formed more efficiently compared to standard gravity (Cai et al., 2015), and also, because galaxy cluster structures are changed by the emergence of the fifth force on several scales (Khoury et al., 2004). Another possibility is to test the effects of MG in the environments for galaxy clusters. Armijo et al. (2018) shows how the mass of haloes can be used to mark galaxies to measure the marked correlation function. They find that large mass haloes, such as clusters of galaxies, allow more differences between the clustering of GR and MG to be found in simulations. In future work we aim to understand the enhancement of gravity environments related to galaxy clusters, and how their structure is modified by the action of the fifth force. For this several tests can be tried: from testing the abundance of clusters in unscreened regimes, weak lensing peak statistic, measuring the velocities of galaxies in clusters, the cross-correlation between clusters and galaxies, and using the marked correlation

function for testing cluster properties. We provide a description of the latter, which can be a direct extension of the work of this thesis.

The marked correlation function defined in this thesis is a valid test to apply to clusters, but its definition has to be extended. We want to compute the cross-correlation between clusters and galaxies, where the two samples come from different survey experiments and with different properties, which can be used to also expand the definition of how we mark the tracers. In this scenario, several cluster properties that should be different between GR and MG models can be used to mark clusters. In Mitchell (2021) several of these properties are modelled in the framework of modified gravity theories such as $f(R)$, the mass being the main property that we can test. Then, the extension of the marked correlation function we can apply for clusters and galaxies, includes the mass for the clusters and measurements of the local density for galaxies as possible marks. Further more, mass measurements for galaxies could also be considered as ideal marks, but an accurate mass estimation or proxy for its measurement is limited for current data.

We can select large mass haloes and define samples from the simulations, to then compute the cross-correlation between galaxy clusters (as examples of high mass halos) and other tracers of the large-scale structure, such as LRGs. The selection can be treated in a similar way than the one for the HOD mock catalogues. In principle, catalogues of galaxies that reproduce the same number density and clustering than the LRG samples from BOSS, will also reproduce the cross-correlation when including the clusters as long as the selection is done based on the mass of the cluster. As we do not observe the mass directly, an extra step to infer the mass from the observation is needed. Currently, different methods to estimate the mass of galaxy clusters can be used, such as weak lensing, the Sunyaev-Zeldovich effect, the X-ray luminosity, and the optical richness (Baxter et al., 2018; Nagai et al., 2007; Fabjan et al., 2011; Capasso et al., 2019). There are several samples of galaxy clusters that include mass estimates made with different methods, which can be used together with a large sample of galaxies to compute the marked cross-

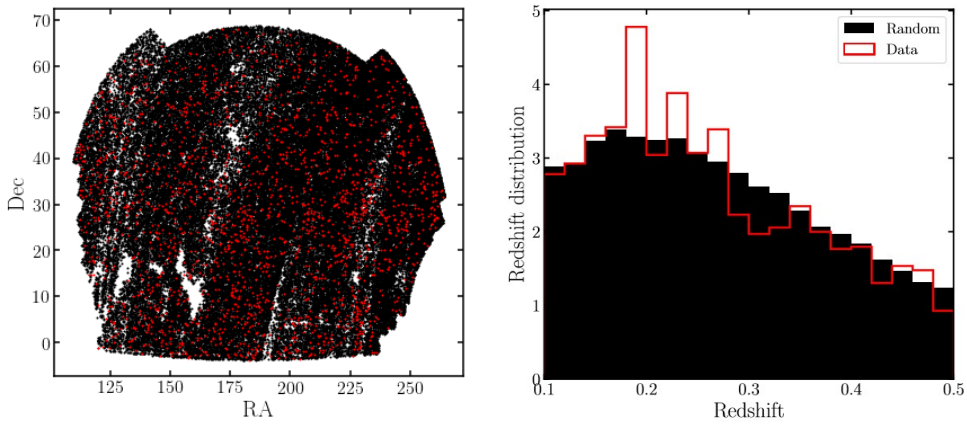


Figure 7.1: Left: the footprint coverage of the CODEX cluster catalogue from Lindholm et al. (2021). We plot the distribution of randoms (black) and the cluster (red) samples, following the area of the SDSS-Legacy survey with the X-ray mask from Clerc et al. (2020). Right: The redshift distribution for the random and cluster samples from Lindholm et al. (2021)

correlation function.

7.2 The Constrain Dark Energy with X-ray clusters sample (CODEX)

A suitable choice for the cluster sample to use in the cluster-galaxy cross correlation is the COnstrain Dark Energy with X-ray (CODEX) clusters sample, which covers the same area as the SDSS-DR12 footprint with an average cluster candidate density of 0.8 deg^{-2} up to a maximum redshift of $z \sim 0.6$ (Finoguenov et al., 2020). The selection of CODEX clusters is based on extended X-ray sources over the complete SDSS legacy-footprint ($\sim 7,500$ deg^2). First, the galaxy clusters are identified through their X-ray emission using measurements from satellites such as ROSAT (Wang et al., 2011) and XMM-Newton (Fassbender et al., 2011). Then, the red-sequence Matched-filter Probabilistic Percolation (redMaPPer) algorithm (Rykoff et al., 2014) is used to find red-sequence galaxies at the same redshift in the optical counterparts of these objects. RedMaPPer finds the red sequence galaxy members in the BOSS survey (LOWZ and CMASS), with a given probab-

ility that these galaxies belong to the galaxy cluster. The cluster galaxies have SDSS spectroscopic information, as they belong to the samples of BOSS, and can be used to estimate the redshift of the center of mass of the galaxy cluster. A final validation is performed using the SDSS imaging to obtain a final cluster sample with optical properties. Furthermore, the CODEX cluster members are targets for the SPIDERS survey of BOSS which completes the member identification of these clusters Clerc et al. (2020). The complete catalogue includes the centre of the object and the spectroscopic redshift, their richness, which can be used to infer the cluster mass. The final sample corresponds to 5,424 cluster candidates with masses between $10^{14} < M/\mathrm{M}_{\odot} < 10^{15}$ in a redshift slice of $0.031 < z < 0.658$. Lindholm et al. (2021) provides a final clean sample of CODEX with the selection method of Finoguenov et al. (2020), which is shown in Figure 7.1. This cluster sample is used to study cosmological parameters performing a two-point correlation function analysis of clusters in projected-real space. In this work, the samples are divided in 2 redshift bins at $0.1 < z < 0.3$ and $0.3 < z < 0.5$, to calculate two-point correlation functions, which can be used to estimate the bias as a function of the halo mass $b(M)$. The calculation of $b(M)$ is depends on the selection of the cosmology, which means that it can be used to constrain cosmological parameters such as σ_8 and Ω_m. The estimation of the mass measurements for the galaxy clusters is obtained using the richness-mass relation calibration provided in (Capasso et al., 2019).

7.2.1 Richness-Mass relation

As the CODEX sample is obtained using the optical information measured by the redMaPPer algorithm, the selection depends on the optical definition of the cluster richness, λ, provided in Rykoff et al. (2014). This parameter is a proxy of the mass estimation, which can be calibrated to create a scaling relation to obtain the mass of the CODEX clusters (Capasso et al., 2019). We use these results to obtain richness estimations from the masses of our simulated clusters, to apply the same selection as in the CODEX clusters provided in Lindholm et al. (2021). In Figure 7.2 we

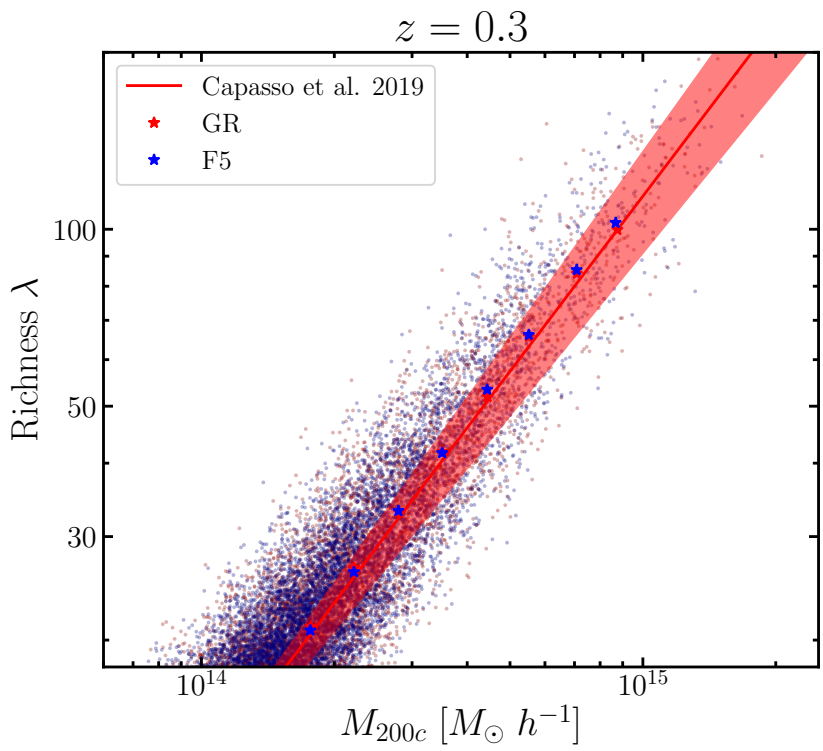


Figure 7.2: The richness-mass relation, $\lambda(M)$, for galaxy clusters from Capasso et al. (2019) (red line) with the respective model uncertainty (red shaded area). We add the richness λ and its variance estimation σ^2_λ for mock cluster catalogues from simulations of GR and F5 at redshift $z = 0.3$.

use the $\lambda(M)$ relation of Capasso et al. (2019) to estimate λ for the clusters in the simulation. Once we calculate the richness λ, we add an uncertainty to the estimation using the richness variance σ^2_λ, which is obtained from the redMaPPer estimations, including observational effects of the survey. The aim of these process is to create a mock cluster catalogue that follows the CODEX selection.

7.2.2 Future plans

Once we understand the impact of the richness selection of the CODEX clusters, we will use the mock catalogues to predict the marked correlation function for the cross-correlation with BOSS-LRGs, using mass as weight for clusters and density as weight for galaxies. In principle, when changing the selection to richness instead of mass, more low-mass clusters are included in the selection when we include the

uncertainties given by σ_λ. These extra objects can improve the differences between models if we are adding unscreened haloes with different formation history than GR predicts. We also need to check if this selection leads to the matching of the projected-cross-correlation function of CODEX-LOWZ and CODEX-CMASS samples, when compared to the mock catalogues of galaxies and clusters from the simulations of GR and F5. Again, maybe the richness cluster selection impacts in the results from the models, and this can be detected by using two-point statistics. These aspects could also help to predict what level of differences we can expect for the marked correlation function to constrain modified gravity models and to tell if GR is enough to describe the mark correlation function of the observations.

7.2.3 Final remarks

Although the physics of galaxy clusters indicates that their mass estimations occur in a screened regime where modified gravity is hidden, it is the additional information coming from the dynamics and the environment of these objects, that can be used to test gravity (Clampitt et al., 2012; Lam et al., 2012). Here, is where marked correlation function becomes relevant as an independent test in addition to cluster abundance tests to find possible signatures of modified gravity.

Bibliography

ABAZAJIAN K N, ADELMAN-MCCARTHY J K, AGüEROS M A, et al. 2009. THE SEVENTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY [J/OL]. The Astrophysical Journal Supplement Series, 182(2): 543-558. https://doi.org/10.1088/0067-0049/182/2/543.

Alam S, Albareti F D, Allende Prieto C, et al. 2015. The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III[J/OL]. , 219(1): 12. DOI: 10.1088/0067-0049/219/1/12.

Anderson L, Aubourg E, Bailey S, et al. 2012. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample[J/OL]. , 427(4): 3435-3467. DOI: 10.1111/j.1365-2966.2012.22066.x.

Anderson L, Aubourg E, Bailey S, et al. 2014. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring D_A and H at z = 0.57 from the baryon acoustic peak in the Data Release 9 spectroscopic Galaxy sample[J/OL]. , 439(1): 83-101. DOI: 10.1093/mnras/stt2206.

Angulo R E, Lacey C G, Baugh C M, et al. 2009. The fate of substructures in cold dark matter haloes[J/OL]. , 399(2): 983-995. DOI: 10.1111/j.1365-2966.2009.15333.x.

Armendariz-Picon C, Mukhanov V, Steinhardt P J. 2000. Dynamical Solution to the Problem of a Small Cosmological Constant and Late-Time Cosmic Acceleration[J/OL]. , 85(21): 4438-4441. DOI: 10.1103/PhysRevLett.85.4438.

Armijo J, Cai Y C, Padilla N, et al. 2018. Testing modified gravity using a marked correlation function[J/OL]. , 478(3): 3627-3632. DOI: 10.1093/mnras/sty1335.

Armijo J, Baugh C M, Padilla N D, et al. 2022. Making use of sub-resolution haloes in N-body simulations[J/OL]. , 510(1): 29-33. DOI: 10.1093/mnrasl/slab122.

Arnold C, Fosalba P, Springel V, et al. 2019. The modified gravity light-cone simulation project - I. Statistics of matter and halo distributions[J/OL]. , 483 (1): 790-805. DOI: 10.1093/mnras/sty3044.

BAKER T, BELLINI E, FERREIRA P G, et al. 2017. Strong constraints on cosmological gravity from gw170817 and grb 170817a[J/OL]. Phys. Rev. Lett., 119: 251301. https://link.aps.org/doi/10.1103/PhysRevLett.119.251301.

Baker T, Barreira A, Desmond H, et al. 2021. Novel Probes Project: Tests of gravity on astrophysical scales[J/OL]. Reviews of Modern Physics, 93(1): 015003. DOI: 10.1103/RevModPhys.93.015003.

Baugh C M, Gonzalez-Perez V, Lagos C D P, et al. 2019. Galaxy formation in the Planck Millennium: the atomic hydrogen content of dark matter halos[J/OL]. , 483(4): 4922-4937. DOI: 10.1093/mnras/sty3427.

BAXTER E J, RAGHUNATHAN S, CRAWFORD T M, et al. 2018. A measurement of CMB cluster lensing with SPT and DES year 1 data[J/OL]. Monthly Notices of the Royal Astronomical Society, 476(2): 2674-2688. https://doi.org/10.1093/mnras/sty305.

BEHROOZI P S, WECHSLER R H, WU H Y. 2012. THE ROCKSTAR PHASE-SPACE TEMPORAL HALO FINDER AND THE VELOCITY OFFSETS OF CLUSTER CORES[J/OL]. The Astrophysical Journal, 762(2): 109. https://doi.org/10.1088/0004-637x/762/2/109.

BENSON A J, COLE S, FRENK C S, et al. 2000. The nature of galaxy bias and clustering[J/OL]. Monthly Notices of the Royal Astronomical Society, 311(4): 793-808. https://doi.org/10.1046/j.1365-8711.2000.03101.x.

Berlind A A, Weinberg D H. 2002. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass[J/OL]. , 575(2): 587-616. DOI: 10.1086/341469.

Bett P, Eke V, Frenk C S, et al. 2007. The spin and shape of dark matter haloes in the Millennium simulation of a Λ cold dark matter universe[J/OL]. , 376(1): 215-232. DOI: 10.1111/j.1365-2966.2007.11432.x.

Blum B, Digel S W, Drlica-Wagner A, et al. 2022. Snowmass2021 Cosmic Frontier White Paper: Rubin Observatory after LSST[A]. arXiv:2203.07220. arXiv: 2203.07220.

BORROW J, BORRISOV A. 2020. swiftsimio: A python library for reading swift data[J/OL]. Journal of Open Source Software, 5(52): 2430. https://doi.org/10.21105/joss.02430.

Brax P, Davis A C, Li B, et al. 2012. Unified description of screened modified gravity[J/OL]. , 86(4): 044015. DOI: 10.1103/PhysRevD.86.044015.

BRAX P, DAVIS A C, LI B, et al. 2013. Systematic simulations of modified gravity: chameleon models[J/OL]. Journal of Cosmology and Astroparticle Physics, 2013 (04): 029-029. https://doi.org/10.1088/1475-7516/2013/04/029.

Burke D J, Collins C A, Mann R G. 2000. Cluster Selection and the Evolution of Brightest Cluster Galaxies[J/OL]. , 532(2): L105-L108. DOI: 10.1086/312579.

Cai Y C, Padilla N, Li B. 2015. Testing gravity using cosmic voids[J/OL]. , 451 (1): 1036-1055. DOI: 10.1093/mnras/stv777.

Capasso R, Mohr J J, Saro A, et al. 2019. Mass calibration of the CODEX cluster sample using SPIDERS spectroscopy - I. The richness-mass relation[J/OL]. , 486 (2): 1594-1607. DOI: 10.1093/mnras/stz931.

Carlson J, White M, Padmanabhan N. 2009. Critical look at cosmological perturbation theory techniques[J/OL]. , 80(4): 043531. DOI: 10.1103/PhysRevD.80.043531.

Carroll S M, Duvvuri V, Trodden M, et al. 2004. Is cosmic speed-up due to new gravitational physics?[J/OL]. , 70(4): 043528. DOI: 10.1103/PhysRevD.70.043528.

Cataneo M, Rapetti D. 2018. Tests of gravity with galaxy clusters[J/OL]. International Journal of Modern Physics D, 27(15): 1848006-936. DOI: 10.1142/S0218271818480061.

Cataneo M, Rapetti D, Schmidt F, et al. 2015. New constraints on f(R) gravity from clusters of galaxies[J/OL]. , 92(4): 044009. DOI: 10.1103/PhysRevD.92.044009.

Cataneo M, Rapetti D, Lombriser L, et al. 2016. Cluster abundance in chameleon f(R) gravity I: toward an accurate halo mass function prediction[J/OL]. , 2016 (12): 024. DOI: 10.1088/1475-7516/2016/12/024.

CAUTUN M, PAILLAS E, CAI Y C, et al. 2018. The Santiago–Harvard–Edinburgh–Durham void comparison – I. SHEDding light on chameleon gravity tests[J/OL]. Monthly Notices of the Royal Astronomical Society, 476(3): 3195-3217. https://doi.org/10.1093/mnras/sty463.

Chadha-Day F, Ellis J, Marsh D J E. 2022. Axion dark matter: What is it and why now?[J/OL]. Science Advances, 8(8): eabj3618. DOI: 10.1126/sciadv.abj3618.

CLAMPITT J, JAIN B, KHOURY J. 2012. Halo scale predictions of symmetron modified gravity[J/OL]. Journal of Cosmology and Astroparticle Physics, 2012 (01): 030-030. https://doi.org/10.1088/1475-7516/2012/01/030.

Clerc N, Kirkpatrick C C, Finoguenov A, et al. 2020. SPIDERS: overview of the X-ray galaxy cluster follow-up and the final spectroscopic data release[J/OL]. , 497(3): 3976-3992. DOI: 10.1093/mnras/staa2066.

COLE S, LACEY C. 1996. The structure of dark matter haloes in hierarchical clustering models[J/OL]. Monthly Notices of the Royal Astronomical Society, 281(2): 716-736. https://doi.org/10.1093/mnras/281.2.716.

Colless M, Peterson B A, Jackson C, et al. 2003. The 2dF Galaxy Redshift Survey: Final Data Release[A]. astro-ph/0306581. arXiv: astro-ph/0306581.

CREMINELLI P, VERNIZZI F. 2017. Dark energy after gw170817 and grb170817a [J/OL]. Phys. Rev. Lett., 119: 251302. https://link.aps.org/doi/10.1103/PhysRevLett.119.251302.

Croton D J, Springel V, White S D M, et al. 2006. The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies[J/OL]. , 365(1): 11-28. DOI: 10.1111/j.1365-2966.2005.09675.x.

CUESTA-LAZARO C, LI B, EGGEMEIER A, et al. 2020. Towards a non-Gaussian model of redshift space distortions[J/OL]. Monthly Notices of the Royal Astronomical Society, 498(1): 1175-1193. https://doi.org/10.1093/mnras/staa2249.

Dawson K S, Schlegel D J, Ahn C P, et al. 2013. The Baryon Oscillation Spectroscopic Survey of SDSS-III[J/OL]. , 145(1): 10. DOI: 10.1088/0004-6256/145/1/10.

De Felice A, Tsujikawa S. 2010. f(R) Theories[J/OL]. Living Reviews in Relativity, 13(1): 3. DOI: 10.12942/lrr-2010-3.

DVALI G, GABADADZE G, PORRATI M. 2000. 4d gravity on a brane in 5d minkowski space[J/OL]. Physics Letters B, 485(1): 208-214. https://www.sciencedirect.com/science/article/pii/S0370269300006699. DOI: https://doi.org/10.1016/S0370-2693(00)00669-9.

Eisenstein D J, Annis J, Gunn J E, et al. 2001. Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample[J/OL]. , 122 (5): 2267-2280. DOI: 10.1086/323717.

Eisenstein D J, Zehavi I, Hogg D W, et al. 2005. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies [J/OL]. , 633(2): 560-574. DOI: 10.1086/466512.

Eisenstein D J, Weinberg D H, Agol E, et al. 2011. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems[J/OL]. , 142(3): 72. DOI: 10.1088/0004-6256/142/3/72.

FABJAN D, BORGANI S, RASIA E, et al. 2011. X-ray mass proxies from hydrodynamic simulations of galaxy clusters – I[J/OL]. Monthly Notices of the Royal Astronomical Society, 416(2): 801-816. https://doi.org/10.1111/j.1365-2966.2011.18497.x.

FASSBENDER R, BöHRINGER H, NASTASI A, et al. 2011. The x-ray luminous galaxy cluster population at 0.9 &ltiz/i$\lesssim$ 1.6 as revealed by the XMM-inewton/idistant cluster project[J/OL]. New Journal of Physics, 13(12): 125014. https://doi.org/10.1088/1367-2630/13/12/125014.

Feldman H A, Kaiser N, Peacock J A. 1994. Power-Spectrum Analysis of Three-dimensional Redshift Surveys[J/OL]. , 426: 23. DOI: 10.1086/174036.

Finoguenov A, Rykoff E, Clerc N, et al. 2020. CODEX clusters. Survey, catalog, and cosmology of the X-ray luminosity function[J/OL]. , 638: A114. DOI: 10.1051/0004-6361/201937283.

Foreman-Mackey D, Hogg D W, Lang D, et al. 2013. emcee: The mcmc hammer [J/OL]. PASP, 125: 306-312. DOI: 10.1086/670067.

GANNOUJI R, SAMI M, THONGKOOL I. 2012. Generic f(r) theories and classicality of their scalarons[J/OL]. Physics Letters B, 716(2): 255-259. https:

//www.sciencedirect.com/science/article/pii/S0370269312008520. DOI: https://doi.org/10.1016/j.physletb.2012.08.015.

Gao L, White S D M. 2007. Assembly bias in the clustering of dark matter haloes [J/OL]. , 377(1): L5-L9. DOI: 10.1111/j.1745-3933.2007.00292.x.

GELMAN A, RUBIN D B. 1992. Inference from Iterative Simulation Using Multiple Sequences[J/OL]. Statistical Science, 7(4): 457 - 472. https://doi.org/10.1214/ss/1177011136.

GUNN J E, SIEGMUND W A, MANNERY E J, et al. 2006. The 2.5 m telescope of the sloan digital sky survey[J/OL]. The Astronomical Journal, 131(4): 2332-2359. https://doi.org/10.1086/500975.

GUO H, ZHENG Z, JING Y P, et al. 2015. Modelling the redshift-space three-point correlation function in SDSS-III[J/OL]. Monthly Notices of the Royal Astronomical Society: Letters, 449(1): L95-L99. https://doi.org/10.1093/mnrasl/slv020.

Guo J Q. 2014. Solar System Tests of f(R) Gravity[J/OL]. International Journal of Modern Physics D, 23(4): 1450036. DOI: 10.1142/S0218271814500369.

HASTINGS W K. 1970. Monte Carlo sampling methods using Markov chains and their applications[J/OL]. Biometrika, 57(1): 97-109. https://doi.org/10.1093/biomet/57.1.97.

He J H, Guzzo L, Li B, et al. 2018. No evidence for modifications of gravity from galaxy motions on cosmological scales[J/OL]. Nature Astronomy, 2: 967-972. DOI: 10.1038/s41550-018-0573-2.

Hernández-Aguayo C, Baugh C M, Li B. 2018. Marked clustering statistics in f(R) gravity cosmologies[J/OL]. , 479(4): 4824-4835. DOI: 10.1093/mnras/sty1822.

Hernández-Aguayo C, Hou J, Li B, et al. 2019. Large-scale redshift space distortions in modified gravity theories[J/OL]. , 485(2): 2194-2213. DOI: 10.1093/mnras/stz516.

Hernández-Aguayo C, Prada F, Baugh C M, et al. 2021. Building a digital twin of a luminous red galaxy spectroscopic survey: galaxy properties and clustering covariance[J/OL]. , 503(2): 2318-2339. DOI: 10.1093/mnras/stab434.

Heymans C, Zhao G B. 2018. Large-scale structure probes of modified gravity [J/OL]. International Journal of Modern Physics D, 27(15): 1848005. DOI: 10.1142/S021827181848005X.

Hinshaw G, Larson D, Komatsu E, et al. 2013. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results[J/OL]. , 208(2): 19. DOI: 10.1088/0067-0049/208/2/19.

Hu W, Sawicki I. 2007. Models of f(R) cosmic acceleration that evade solar system tests[J/OL]. , 76(6): 064004. DOI: 10.1103/PhysRevD.76.064004.

JENNINGS E, BAUGH C M, PASCOLI S. 2010. TESTING GRAVITY USING THE GROWTH OF LARGE-SCALE STRUCTURE IN THE UNIVERSE [J/OL]. The Astrophysical Journal, 727(1): L9. https://doi.org/10.1088/2041-8205/727/1/l9.

Kaiser N. 1986. Evolution and clustering of rich clusters.[J/OL]. , 222: 323-345. DOI: 10.1093/mnras/222.2.323.

KHOURY J, WELTMAN A. 2004. Chameleon cosmology[J/OL]. Phys. Rev. D, 69: 044026. https://link.aps.org/doi/10.1103/PhysRevD.69.044026.

Kilbinger M. 2015. Cosmology with cosmic shear observations: a review[J/OL]. Reports on Progress in Physics, 78(8): 086901. DOI: 10.1088/0034-4885/78/8/086901.

Koyama K. 2016. Cosmological tests of modified gravity[J/OL]. Reports on Progress in Physics, 79(4): 046902. DOI: 10.1088/0034-4885/79/4/046902.

LAM T Y, NISHIMICHI T, SCHMIDT F, et al. 2012. Testing gravity with the stacked phase space around galaxy clusters[J/OL]. Phys. Rev. Lett., 109: 051301. https://link.aps.org/doi/10.1103/PhysRevLett.109.051301.

Landy S D, Szalay A S. 1993. Bias and Variance of Angular Correlation Functions [J/OL]. , 412: 64. DOI: 10.1086/172900.

Laureijs R, Amiaux J, Arduini S, et al. 2011. Euclid Definition Study Report[A]. arXiv:1110.3193. arXiv: 1110.3193.

Leclercq F, Jasche J, Gil-Marín H, et al. 2013. One-point remapping of Lagrangian perturbation theory in the mildly non-linear regime of cosmic structure formation [J/OL]. , 2013(11): 048. DOI: 10.1088/1475-7516/2013/11/048.

Levi M, Bebek C, Beers T, et al. 2013. The DESI Experiment, a whitepaper for Snowmass 2013[A]. arXiv:1308.0847. arXiv: 1308.0847.

LI B, BARROW J D, MOTA D F. 2007. Cosmology of ricci-tensor-squared gravity in the palatini variational approach[J/OL]. Phys. Rev. D, 76: 104047. https://link.aps.org/doi/10.1103/PhysRevD.76.104047.

Li B, Zhao G B, Teyssier R, et al. 2012. ECOSMOG: an Efficient COde for Simulating MOdified Gravity[J/OL]. , 2012(1): 051. DOI: 10.1088/1475-7516/2012/01/051.

LI B, ZHAO G, KOYAMA K. 2012. Haloes and voids in f(R) gravity[J/OL]. Monthly Notices of the Royal Astronomical Society, 421(4): 3481-3487. https://doi.org/10.1111/j.1365-2966.2012.20573.x.

Li Y, Hu W. 2011. Chameleon halo modeling in f(R) gravity[J/OL]. , 84(8): 084033. DOI: 10.1103/PhysRevD.84.084033.

Lindholm V, Finoguenov A, Comparat J, et al. 2021. Clustering of CODEX clusters [J/OL]. , 646: A8. DOI: 10.1051/0004-6361/202038807.

Liu R, Valogiannis G, Battaglia N, et al. 2021. Constraints on f (R) and normal-branch Dvali-Gabadadze-Porrati modified gravity model parameters with cluster abundances and galaxy clustering[J/OL]. , 104(10): 103519. DOI: 10.1103/PhysRevD.104.103519.

LIU X, LI B, ZHAO G B, et al. 2016. Constraining $f(r)$ gravity theory using weak lensing peak statistics from the canada-france-hawaii-telescope lensing survey[J/OL]. Phys. Rev. Lett., 117: 051101. https://link.aps.org/doi/10.1103/PhysRevLett.117.051101.

Lombriser L, Taylor A. 2016. Breaking a dark degeneracy with gravitational waves [J/OL]. , 2016(3): 031. DOI: 10.1088/1475-7516/2016/03/031.

Mak D S Y, Pierpaoli E, Schmidt F, et al. 2012. Constraints on modified gravity from Sunyaev-Zeldovich cluster surveys[J/OL]. , 85(12): 123513. DOI: 10.1103/PhysRevD.85.123513.

Manera M, Scoccimarro R, Percival W J, et al. 2013. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues[J/OL]. , 428(2): 1036-1054. DOI: 10.1093/mnras/sts084.

MANERA M, SAMUSHIA L, TOJEIRO R, et al. 2014. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the low-redshift sample[J/OL]. Monthly Notices of the Royal Astronomical Society, 447(1): 437-445. https://doi.org/10.1093/mnras/stu2465.

Maraston C, Strömbäck G, Thomas D, et al. 2009. Modelling the colour evolution of luminous red galaxies - improvements with empirical stellar spectra[J/OL]. , 394(1): L107-L111. DOI: 10.1111/j.1745-3933.2009.00621.x.

METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. 1953. Equation of state calculations by fast computing machines[J/OL]. The Journal of Chemical Physics, 21(6): 1087-1092. https://doi.org/10.1063/1.1699114.

Mitchell M A. 2021. A general framework for unbiased tests of gravity using galaxy clusters[A]. arXiv:2110.14564. arXiv: 2110.14564.

Mitchell M A, He J H, Arnold C, et al. 2018. A general framework to test gravity using galaxy clusters - I. Modelling the dynamical mass of haloes in f(R) gravity [J/OL]. , 477(1): 1133-1152. DOI: 10.1093/mnras/sty636.

MULTAMÄKI T, VILJA I. 2006. Cosmological expansion and the uniqueness of the gravitational action[J/OL]. Phys. Rev. D, 73: 024018. https://link.aps.org/doi/10.1103/PhysRevD.73.024018.

NA H S, LEE C N, CHEONG O. 2002. Voronoi diagrams on the sphere [J/OL]. Computational Geometry, 23(2): 183-194. https://www.sciencedirect.com/science/article/pii/S0925772102000779. DOI: https://doi.org/10.1016/S0925-7721(02)00077-9.

Nagai D, Vikhlinin A, Kravtsov A V. 2007. Testing X-Ray Measurements of Galaxy Clusters with Cosmological Simulations[J/OL]. , 655(1): 98-108. DOI: 10.1086/509868.

NELDER J A, MEAD R. 1965. A Simplex Method for Function Minimization [J/OL]. The Computer Journal, 7(4): 308-313. https://doi.org/10.1093/comjnl/7.4.308.

Neyrinck M C. 2008. ZOBOV: a parameter-free void-finding algorithm[J/OL]. , 386(4): 2101-2109. DOI: 10.1111/j.1365-2966.2008.13180.x.

Norberg P, Baugh C M, Gaztañaga E, et al. 2009. Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics[J/OL]. , 396(1): 19-38. DOI: 10.1111/j.1365-2966.2009.14389.x.

NORBERG P, BAUGH C M, HAWKINS E, et al. 2002. The 2dF Galaxy Redshift Survey: the dependence of galaxy clustering on luminosity and spectral type [J/OL]. Monthly Notices of the Royal Astronomical Society, 332(4): 827-838. https://doi.org/10.1046/j.1365-8711.2002.05348.x.

OYAIZU H. 2008. Nonlinear evolution of $f(r)$ cosmologies. i. methodology[J/OL]. Phys. Rev. D, 78: 123523. https://link.aps.org/doi/10.1103/PhysRevD.78.123523.

Paillas E, Cautun M, Li B, et al. 2019. The Santiago-Harvard-Edinburgh-Durham void comparison II: unveiling the Vainshtein screening using weak lensing[J/OL]. , 484(1): 1149-1165. DOI: 10.1093/mnras/stz022.

Paranjape A, Alam S. 2020. Voronoi volume function: a new probe of cosmology and galaxy evolution[J/OL]. , 495(3): 3233-3251. DOI: 10.1093/mnras/staa1379.

Parejko J K, Sunayama T, Padmanabhan N, et al. 2013. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: the low-redshift sample [J/OL]. , 429(1): 98-112. DOI: 10.1093/mnras/sts314.

Peacock J A, Smith R E. 2000. Halo occupation numbers and galaxy bias[J/OL]. , 318(4): 1144-1156. DOI: 10.1046/j.1365-8711.2000.03779.x.

Peacock J A, Cole S, Norberg P, et al. 2001. A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey[J]. , 410(6825): 169-173.

PERCIVAL W J, COLE S, EISENSTEIN D J, et al. 2007. Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and 2dF Galaxy Redshift Survey[J/OL]. Monthly Notices of the Royal Astronomical Society, 381 (3): 1053-1066. https://doi.org/10.1111/j.1365-2966.2007.12268.x.

PERLMUTTER S, ALDERING G, GOLDHABER G, et al. 1999. Measurements of and from 42 high-redshift supernovae[J/OL]. The Astrophysical Journal, 517 (2): 565. https://dx.doi.org/10.1086/307221.

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016. Planck 2015 results - xiii. cosmological parameters[J/OL]. A&A, 594: A13. https://doi.org/10.1051/0004-6361/201525830.

Planck Collaboration, Aghanim N, Akrami Y, et al. 2020a. Planck 2018 results. VI. Cosmological parameters[J/OL]. , 641: A6. DOI: 10.1051/0004-6361/201833910.

Planck Collaboration, Akrami Y, Arroja F, et al. 2020b. Planck 2018 results. X. Constraints on inflation[J/OL]. , 641: A10. DOI: 10.1051/0004-6361/201833887.

Platen E, van de Weygaert R, Jones B J T. 2007. A cosmic watershed: the WVF void detection technique[J/OL]. , 380(2): 551-570. DOI: 10.1111/j.1365-2966.2007.12125.x.

Postman M, Lauer T R. 1995. Brightest Cluster Galaxies as Standard Candles [J/OL]. , 440: 28. DOI: 10.1086/175245.

Press W H, Schechter P. 1974. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation[J/OL]. , 187: 425-438. DOI: 10.1086/152650.

PRESS W H, TEUKOLSKY S A, VETTERLING W T, et al. 1992. Numerical recipes in c[M]. Second ed. Cambridge, USA: Cambridge University Press.

Ramakrishnan S, Paranjape A, Sheth R K. 2021. Mock halo catalogues: assigning unresolved halo properties using correlations with local halo environment[J/OL]. , 503(2): 2053-2064. DOI: 10.1093/mnras/stab541.

Reid B, Ho S, Padmanabhan N, et al. 2016. SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues[J/OL]. , 455(2): 1553-1573. DOI: 10.1093/mnras/stv2382.

Riess A G, Filippenko A V, Challis P, et al. 1998. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant[J/OL]. , 116(3): 1009-1038. DOI: 10.1086/300499.

Ross A J, Percival W J, Sánchez A G, et al. 2012. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics[J/OL]. , 424(1): 564-590. DOI: 10.1111/j.1365-2966.2012.21235.x.

Ruan C Z, Cuesta-Lazaro C, Eggemeier A, et al. 2022. Towards an accurate model of small-scale redshift-space distortions in modified gravity[J/OL]. , 514(1): 440-459. DOI: 10.1093/mnras/stac1345.

Rykoff E S, Rozo E, Busha M T, et al. 2014. redMaPPer. I. Algorithm and SDSS DR8 Catalog[J/OL]. , 785(2): 104. DOI: 10.1088/0004-637X/785/2/104.

RYKOFF E S, ROZO E, BUSHA M T, et al. 2014. redMaPPer. i. ALGORITHM AND SDSS DR8 CATALOG[J/OL]. The Astrophysical Journal, 785(2): 104. https://doi.org/10.1088/0004-637x/785/2/104.

SATPATHY S, A C CROFT R, HO S, et al. 2019. Measurement of marked correlation functions in SDSS-III Baryon Oscillation Spectroscopic Survey using LOWZ galaxies in Data Release 12[J/OL]. Monthly Notices of the Royal Astronomical Society, 484(2): 2148-2165. https://doi.org/10.1093/mnras/stz009.

Schmidt F, Lima M, Oyaizu H, et al. 2009. Nonlinear evolution of f(R) cosmologies. III. Halo statistics[J/OL]. , 79(8): 083518. DOI: 10.1103/PhysRevD.79.083518.

Sheth R K, Tormen G. 1999. Large-scale bias and the peak background split[J/OL]. , 308(1): 119-126. DOI: 10.1046/j.1365-8711.1999.02692.x.

Sheth R K, Tormen G. 2004. On the environmental dependence of halo formation [J/OL]. , 350(4): 1385-1390. DOI: 10.1111/j.1365-2966.2004.07733.x.

SINGH S, MANDELBAUM R, MORE S. 2015. Intrinsic alignments of SDSS-III BOSS LOWZ sample galaxies[J/OL]. Monthly Notices of the Royal Astronomical Society, 450(2): 2195-2216. https://doi.org/10.1093/mnras/stv778.

SOLà J, GóMEZ-VALENT A, de Cruz Pérez J. 2017. The h0 tension in light of vacuum dynamics in the universe[J/OL]. Physics Letters B, 774: 317-324. https://www.sciencedirect.com/science/article/pii/S0370269317307852. DOI: https://doi.org/10.1016/j.physletb.2017.09.073.

Sotiriou T P. 2006. The nearly Newtonian regime in non-linear theories of gravity [J/OL]. General Relativity and Gravitation, 38(9): 1407-1417. DOI: 10.1007/s10714-006-0328-8.

Springel V, White S D M, Tormen G, et al. 2001. Populating a cluster of galaxies - I. Results at [formmu2]z=0[J/OL]. , 328(3): 726-750. DOI: 10.1046/j.1365-8711.2001.04912.x.

Springel V, White S D M, Jenkins A, et al. 2005. Simulations of the formation, evolution and clustering of galaxies and quasars[J/OL]. , 435(7042): 629-636. DOI: 10.1038/nature03597.

Springel V, Pakmor R, Zier O, et al. 2020. Simulating cosmic structure formation with the GADGET-4 code[A]. arXiv:2010.03567. arXiv: 2010.03567.

Strauss M A, Weinberg D H, Lupton R H, et al. 2002. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample[J/OL]. , 124(3): 1810-1824. DOI: 10.1086/342343.

TOJEIRO R, PERCIVAL W J, BRINKMANN J, et al. 2012. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring structure growth using passive galaxies[J/OL]. Monthly Notices of the Royal Astronomical Society, 424(3): 2339-2344. https://doi.org/10.1111/j.1365-2966.2012.21404.x.

Tsujikawa S. 2013. Quintessence: a review[J/OL]. Classical and Quantum Gravity, 30(21): 214003. DOI: 10.1088/0264-9381/30/21/214003.

VAINSHTEIN A. 1972. To the problem of nonvanishing gravitation mass[J/OL]. Physics Letters B, 39(3): 393-394. https://www.sciencedirect.com/science/article/pii/0370269372901475. DOI: https://doi.org/10.1016/0370-2693(72)90147-5.

Valogiannis G, Bean R. 2018. Beyond δ : Tailoring marked statistics to reveal modified gravity[J/OL]. , 97(2): 023535. DOI: 10.1103/PhysRevD.97.023535.

VORONOI G. 1908. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites.[J/OL]. Journal für die reine und angewandte Mathematik (Crelles Journal), 1908(133): 97-102. https://doi.org/10.1515/crll.1908.133.97. DOI: doi:10.1515/crll.1908.133.97.

Wang L, Yang X, Luo W, et al. 2011. Cross identification between X-ray and Optical Clusters of Galaxies in the SDSS DR7 Field[A]. arXiv:1110.1987. arXiv:1110.1987.

Wechsler R H, Zentner A R, Bullock J S, et al. 2006. The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation[J/OL]. , 652(1): 71-84. DOI: 10.1086/507120.

WEINBERG S. 1989. The cosmological constant problem[J/OL]. Rev. Mod. Phys., 61: 1-23. https://link.aps.org/doi/10.1103/RevModPhys.61.1.

WHITE M. 2016. A marked correlation function for constraining modified gravity models[J/OL]. Journal of Cosmology and Astroparticle Physics, 2016(11): 057-057. https://doi.org/10.1088/1475-7516/2016/11/057.

White M, Padmanabhan N. 2009. Breaking halo occupation degeneracies with marked statistics[J/OL]. , 395(4): 2381-2384. DOI: 10.1111/j.1365-2966.2009.14732.x.

White M, Blanton M, Bolton A, et al. 2011. The Clustering of Massive Galaxies at z $\sim$0.5 from the First Semester of BOSS Data[J/OL]. , 728(2): 126. DOI: 10.1088/0004-637X/728/2/126.

White S D M, Rees M J. 1978. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering.[J/OL]. , 183: 341-358. DOI: 10.1093/mnras/183.3.341.

White S D M, Frenk C S. 1991. Galaxy Formation through Hierarchical Clustering [J/OL]. , 379: 52. DOI: 10.1086/170483.

York D G, Adelman J, Anderson J, John E., et al. 2000. The Sloan Digital Sky Survey: Technical Summary[J/OL]. , 120(3): 1579-1587. DOI: 10.1086/301513.

ZEHAVI I, ZHENG Z, WEINBERG D H, et al. 2005. The luminosity and color dependence of the galaxy correlation function[J/OL]. The Astrophysical Journal, 630(1): 1-27. https://doi.org/10.1086/431891.

Zel'dovich Y B. 1970. Gravitational instability: An approximate theory for large density perturbations.[J]. , 5: 84-89.

Zhang H, Samushia L, Brooks D, et al. 2022. Constraining galaxy-halo connection with high-order statistics[A]. arXiv:2203.17214. arXiv: 2203.17214.

ZHENG Z, GUO H. 2016. Accurate and efficient halo-based galaxy clustering modelling with simulations[J/OL]. Monthly Notices of the Royal Astronomical Society, 458(4): 4015-4024. https://doi.org/10.1093/mnras/stw523.

Zheng Z, Berlind A A, Weinberg D H, et al. 2005. Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies[J/OL]. , 633 (2): 791-809. DOI: 10.1086/466510.

Zheng Z, Coil A L, Zehavi I. 2007. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering[J/OL]. , 667(2): 760-779. DOI: 10.1086/521074.

Zhou R, Newman J A, Dawson K S, et al. 2020. Preliminary Target Selection for the DESI Luminous Red Galaxy (LRG) Sample[J/OL]. Research Notes of the American Astronomical Society, 4(10): 181. DOI: 10.3847/2515-5172/abc0f4.

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2_ε. 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.