UAM
Universidad Autónoma
de Madrid

Universidad Autónoma de Madrid
Departamento de
Física Teórica

Redshift Error Modelling for Quasi-Stellar Objects

Exploring the Halo Occupation Distribution Model within DESI and AbacusSummit

Marco José Molina Pradillo

Directors

Dra. Violeta González-Pérez

Dr. Eusebio Sánchez Álvaro

Master's Thesis in Theoretical Physics
Astrophysics Speciality

Academic year 2022-2023

Acknowledgements

I would like to thank the professors of the master's degree in theoretical physics at the UAM for their enthusiasm and help in this year of intense learning, especially Violeta González-Pérez, the director of this thesis, for introducing me for the first time to the world of cosmological simulations in such depth. I am also very grateful to Eusebio Sánchez Álvaro, my thesis supervisor, for having accompanied me now that I am finishing my training as he did before I started. Thanks also to Sujatha Ramakrishnan, Bernhard Vos Ginés and Weiguang Cui for their willingness and offer to solve my doubts with the programming code whenever they could. Most of the work has been made possible by the resources of the DESI research collaboration at the National Energy Research Scientific Computing Center (NERSC). I would also like to thank them for their scientific work and willingness to share and explain to those of us who are learning, especially Antoine Rocher, Lehman Garrison and Sihan Yuan for their help and the whole community that got involved and contributed their bit to create something as exciting as science can be.

Abstract

English: This work provides a comprehensive analysis of an AbacusSummit suite cosmological simulation at redshift $z = 1.4$, being its main objective to model the redshift-space uncertainty of Quasars by adding Gaussian and Lorentzian noise to their peculiar velocities. The Halo Mass Function and Halo Bias are employed to characterize the distribution and clustering of dark matter halos. A Halo Occupation Distribution model based on DESI One Percent Survey data is then applied to generate a QSO mock catalogue. The redshift and real-space correlation functions of the QSO samples are examined, confirming their consistency with the Kaiser ratio on large scales. The resulting impact is assessed using the 2PCF, the multipoles and the projected correlation function w_{p}. The performed redshift smearing error model was validated implementing the redshift smearing distortion and studying the distribution of peculiar velocities, with noise expanding the range of expected values. The 2PCF analysis reveals fluctuations at small scales that are smoothed out when going to larger scales, but with a larger reach for higher intensity uncertainties. The quadrupole and hexadecapole show a homogenisation of their fluctuations around the range of distances affected by the distortion, more noticeable for larger modelled uncertainties. The projected correlation function w_{p} shows high agreement between perturbed and unperturbed QSO catalogues in terms of shape, while differences in clustering power indicate the weakening effect of velocity perturbations. This study emphasizes the importance of accounting for redshift uncertainties and provides insights into simulating noise in the observational redshift space.

Spanish: Este trabajo proporciona un amplio análisis de una simulación cosmológica de la suite AbacusSummit a desplazamiento al rojo $z = 1.4$, siendo su principal objetivo modelar la incertidumbre del espacio de desplazamiento al rojo de los Cuásares añadiendo ruido Gaussiano y Lorentziano a sus velocidades peculiares. La Función de Masa del Halo y el Sesgo del Halo se emplean para caracterizar la distribución y agrupación de los halos de materia oscura. A continuación, se aplica un modelo de distribución de la ocupación del halo basado en los datos del DESI One Percent Survey para generar un catálogo simulado de Cuásares. Se examinan las funciones de correlación entre el desplazamiento al rojo y el espacio real de las muestras de Cuásares, confirmando su coherencia con la relación de Kaiser a gran escala. El impacto resultante se evalúa utilizando la 2PCF, los multipolos y la función de correlación proyectada w_{p}. El modelo de error de corrimiento al rojo realizado se validó implementando la distorsión de corrimiento al rojo y estudiando la distribución de velocidades peculiares con ruido que amplía el rango de valores esperados. El análisis de la 2PCF revela fluctuaciones a escalas pequeñas que se suavizan al pasar a escalas mayores, pero con un alcance mayor para incertidumbres de intensidad más alta. El cuadrupolo y el hexadecapolo muestran una homogeneización de sus fluctuaciones en torno al rango de distancias afectadas por la distorsión, más notable para mayores incertidumbres modelizadas. La función de correlación proyectada w_{rmp} muestra una gran concordancia entre los catálogos de QSO perturbados y no perturbados en términos de forma, mientras que las diferencias en la potencia de agrupamiento indican el efecto debilitador de las perturbaciones de velocidad. Este estudio subraya la importancia de tener en cuenta las incertidumbres del corrimiento al rojo y proporciona ideas para simular el ruido en el espacio observacional de los corrimientos al rojo.

Contents

1 Introduction

In the realm of cosmology, these last decades have seen an apogee as seldom seen before. The possibility of conducting extensive observational surveys together with the capacity of following them side by side with computational simulations have opened the door to unravel some of the most enigmatic questions in science. Thanks to this, the cosmological parameters describing the flat ΛCDM model that explains the expansion history and structure formation of the Universe have an increasingly robust observational base. This is the case with Plank's comprehensive analysis of the cosmic microwave background temperature fluctuations (Aghanim et al., 2020) or the BOSS wide angle survey searching for galaxies to characterise the large-scale structure of the Universe (Dawson et al., 2016). This model is governed by the General theory of Relativity explaining gravity and includes collisionless cold dark matter (CDM) which can only be inferred gravitationally, and dark energy, which is attributed to the cosmological constant Λ, and is responsible of its accelerated expansion. Nevertheless, neither the nature of dark matter nor dark energy are well understood.

We now know that dark matter clumps together in a large-scale structure of halos and filaments that acts as a large skeleton for galaxies to inhabit. This structure must have had its seed in the first fluctuations of the matter distribution in the inflationary period of the universe and over time the accompanying baryonic matter dissipated and fell into the gravitational wells of the dark matter halos forming the first galaxies. From the exhaustive, statistical study of the galaxies we can see and their large-scale distribution we can infer properties of the underlying dark matter that is invisible to us. This is the key behind the halo-galaxy connection (Wechsler and Tinker, 2018). Furthermore, by studying the evolution and growth of this large-scale structure and characterising it at increasingly distant redshifts, some light can be shed about the nature of dark energy and its evolution.

One way of inferring this crucial insights into the growth of cosmic structure and the nature of dark energy and dark matter is through the Redshift Space Distortions (RSD) analysis. When measuring the redshifts of faraway galaxies, the shifts are not only accounting for the cosmological expansion of the universe, but also for the proper velocities of these galaxies along the line of sight, induced by the gravitational interaction with the surrounding large-scale structure. This leads to anisotropic distortions in the spatial distribution of galaxies in redshift space, as initially described by Kaiser (1987). In this way, galaxies approaching to us will appear to be closer than they really are while those that move away appear to be further apart. By properly accounting for this redshift smear uncertainty, the growth rate of the cosmic structure, typically denoted as f, can be investigated. This is related to the matter content of the Universe in the ΛCDM model, and also provides valuable constraints on theories of modified gravity (Gong et al., 2008).

Surveys such as the one being performed by the DESI (Dark Energy Spectroscopic Instrument) are precisely designed to measure the redshifts of millions of galaxies using advanced spectroscopic techniques to create a comprehensive three-dimensional map of the large scale structure of the Universe and study the underlying dark matter structure and the dark energy effects and evolution (Levi et al., 2013). Objects like Quasars (QSO) are paradigmatic in this respect, as their brightness is unparalleled and they are visible at unsuspected distances, showing recognisable characteristics in their electromagnetic emission spectra that make them susceptible to being identified in redshift space, allowing distances and times far away to be studied. But they are also bound to many systematic errors such as their spectral line widths variability or the particular observing conditions (La Mura et al., 2017).

Accurately modeling and fitting the errors associated with this redshift measurements is thus of great importance. It is by comparing the different RSD estimation of the same observed objects that one can get an experimental profile distribution for the redshift smearing uncertainties. This was done in Lyke et al. (2020) for the Sloan Digital Sky Survey QSOs sample. There a fine tuned double Gaussian was proven to be a good model to mock the redshift uncertainties as used in Smith et al. (2020). This was also recently performed in Yu et al. (2023) with the very first Early Data Release from DESI using various galaxy tracers, including QSOs. Here the observational statistical distributions of the uncertainties where fitted to Lorentzian and Gaussian distributions for the different tracers and redshifts, finding that for QSOs, the Lorentzian profile is easily capable of mimicking the long tails observed in their redshift error profiles.

The aim of this work is to implement a redshift smearing error model based on the observational fit made in Yu et al. (2023) for the DESI One-Percent Survey QSOs at redshifts between $1.1 < z < 1.6$, using a Lorentzian distribution parameterised as the one described there to give the best fit to the experimental distribution of redshift uncertainties. Then a test will be performed on its efficiency when applied to a mock QSO catalogue generated according to the Halo Occupation Distribution model described in Yuan et al. (2023) using one of the fiducial dark matter simulations in the AbacusSummit suite and study its appreciable effects on the statistical observables derived from such QSOs exposed to the mocked uncertainty in the redshift space.

The structure of the thesis will be as follows. Section 2.1 will explain in detail the programming context of the dark matter simulation used, which will be then characterise in Section 2.2 by means of its bias and halo statistics. Section 3 will be devoted to describing the HOD and redshift models used to obtain the mock QSO catalogue and the redshift smear uncertainty, ending with the description of the statistical tools used to study them. The final results about the redshift uncertainty model and its effects on the QSOs clustering are presented in Section 4. The study is ended in Section 5 with a summary and conclusions.

2 Simulation

2.1 AbacusSummit

The present work has been performed entirely in the AbacusSummit environment, one of the largest set of high-accuracy N-body cosmological simulations produced to date, performed on the Summit supercomputer at Oak Ridge Leadership Computing Facility, involving around 60 trillion particles (Maksimova et al., 2021).

Exponential progress in computational techniques over the last decades has made it possible to render N-body simulations at ever better mass resolutions and increasing numbers of particles. AbacusSummit has been able to assemble simulations with up to 330 trillion particles across 97 different cosmologies in a single set thanks to the Abacus code and its improved Poisson equation solving algorithm and optimised software engineering based on GPU acceleration (Garrison et al., 2021).

In this way, AbacusSummit promises to supply the most extensive galaxy surveys that have been made to date, such as Euclid, LSST, or as it is the case with DESI, with the necessary comparison between observations and cosmological predictions, the tools for their systematic errors calibration and the examination of their data processing and analysis capabilities with mock catalogs. In this sense, N-body simulations are essential to analyze the nonlinear implications that the theory of large-scale gravitational structure cannot take into account in all its detail. These are reflected in AbacusSummit products such as halo catalogs, density maps or particle snapshots, not to mention subsamples of indexed particles with the potential to generate high fidelity merger trees as described in Bose et al. (2022).

The AbacusSummit suite placed special emphasis on the creation of a large cosmological simulation with the most up-to-date Lambda Cold Dark Matter (ΛCDM) model based on the

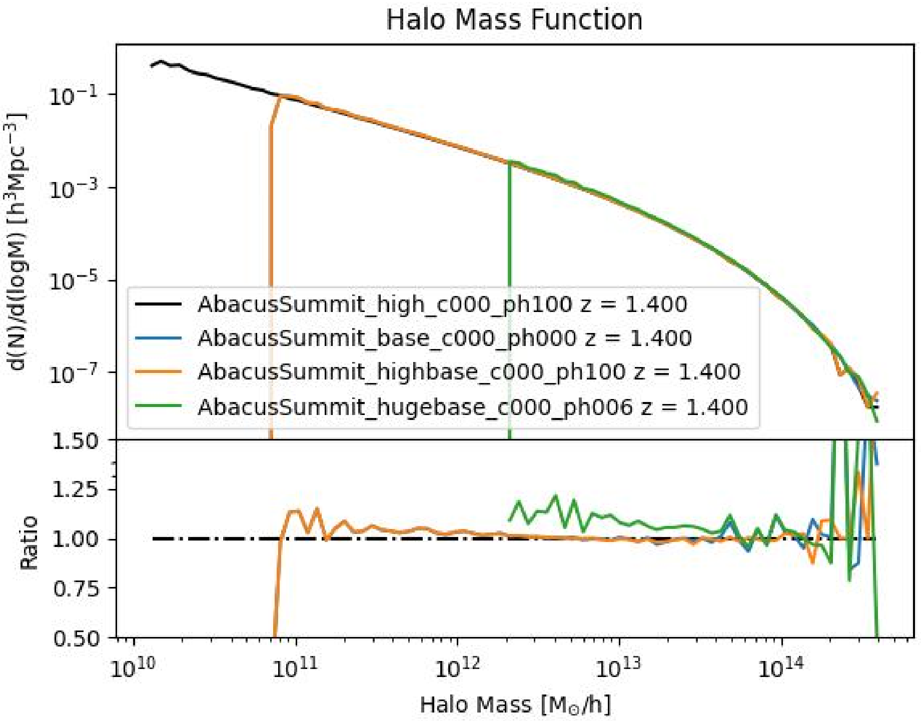

Figure 1: The halo mass function of the cleaned halos at z = 1.4 from four different simulations (`high`, `base`, `highbase` and `hugebase`) and three different phases (`ph100`, `ph000` `ph006`) in the fiducial Planck 2018 ΛCDM cosmology `c000`. The lower panel shows the ratio of these curves with respect to the high resolution box. The `base` and `highbase` boxes, with the same mass resolution, are in excellent agreement, but comparing a coarser simulation to a finer one gives an excess of haloes at sizes below a few hundred particles in general

latest Plank 2018 data release (Aghanim et al., 2020). To it is dedicated a huge volume of $400\,\mathrm{h}^{-3}\,\mathrm{Gpc}^3$ split into 25 base simulations, each $2\,\mathrm{h}^{-1}\,\mathrm{Gpc}$ box-sided and with approximately 330 billion particles with a $2\cdot10^9\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$ mass and softening length of $7.2\,\mathrm{h}^{-1}\,\mathrm{kpc}$ (Maksimova et al., 2021). It also features boxes of different sizes and resolutions for the creation of light cones or to study the statistical error, the covariance and boundary conditions of the box and the search for groups and their dependence on mass resolution.

Although the most representative of them is used here, AbacusSummit suite explores 96 other cosmologies with a similar box size and mass resolution in a parametric space of eight variables: the cold dark matter density $\omega_\mathrm{c} = \Omega_\mathrm{c}\cdot h^2$, the baryon density $\omega_\mathrm{b} = \Omega_\mathrm{b}\cdot h^2$, the spectral tilt n_s and its running α_s, the amplitude of structure of the CMB and baryons σ_8, the equation of state of dark energy $w(z) = w_0 + (1-a)w_a$, and the density of massless relics N_eff. All of them have an optical depth of $\tau = 0.054\,4$ and develop in a flat spatial curvature, where in most cases the Hubble parameter H_0 is selected so as to match with the CMB acoustic scale θ_* with $100\cdot\theta_* = 1.041\,533$ (Maksimova et al., 2021). The Boltzmann CLASS code solver takes these parameters as inputs to get the power spectrum from which the initial conditions are generated with the Abacus code first iterations while `HyRec` is used to model recombination (Lesgourgues, 2011; Ali-Haimoud and Hirata, 2011). These cosmologies are thought to analyse other relevant observational insights, the influence that different types of neutrinos might have, or to compare with the results of other large simulations.

The different cosmologies are classified with an alphanumeric code of type `cxxx` and the random seed initial conditions for the white noise of their Gaussian random field with a code of type `phxxx`. The various realisations with different simulation parameters are labelled with a key word. In this case, the simulation `AbacusSummit_base_c000_ph000` is used, where `c000` refers to the Planck 2018 ΛCDM cosmology and `base` indicates a configuration with $6\,912^3$

Table 1: Parameters of the AbacusSummit `base` simulation showing the final redshift to which it is run, the number of particles as the cubic particles per dimension, the length of the cubic simulation box and softening length together with the mass resolution for the cosmological parameters of the fiducial Planck 2018 ΛCDM cosmology used in the `c000` simulations.

`AbacusSummit_base_c000_ph000`

z_{final}	Particles	Box Size (h^{-1} Mpc)	Mass Resolution ($\mathrm{h}^{-1}\,\mathrm{M}_\odot$)	Softening Length (h^{-1} kpc)
0.1	$6\,912^3$	2 000	$2 \cdot 10^9$	7.2

Planck 2018 ΛCDM

ω_{b}	ω_{c}	h	$10^9\,A_{\mathrm{s}}$	n_{s}	α_{s}	N_{ur}	N_{ncdm}	$10^4\,\omega_{\mathrm{ncdm}}$	$w_{0,\mathrm{fld}}$	$w_{a,\mathrm{fld}}$	$\sigma_{8,\mathrm{m}}$	$\sigma_{8,\mathrm{cb}}$
0.022 37	0.120 0	0.673 6	2.083 0	0.964 9	0.0	2.032 8	1	6.442 0	-1.0	0.0	0.807 952	0.811 355

particles in a $2\,\mathrm{h}^{-1}$ Gpc -side box. This combination of cosmology and environment conditions gives a mass resolution of $2{\cdot}10^9\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$. All the information concerning the simulation employed here is summarised in Table 1. A fast comparison between simulations with different phases and configurations within this cosmology has been added in Figure 1 by studying their respective halo mass functions, which will be explained in more detail in Section 2.2.1.

About 1100 iterations are needed to get to $z = 0.1$ in `AbacusSummit_base_c000_ph000`. The Abacus code evolves all the particle dynamics with the same time step, generally constrained by the small-scale structure at each point in time. The scale at which the far-field gravitational interaction method is softened is 30% of the mean distance between particles, which at early times ($z > 11$) is around $7.2\,\mathrm{h}^{-1}$ kpc. Thus, the force law is modified to follow the $1/r^2$ behaviour when the two-body scattering scale comes into play.

The suite of simulations comprises 33 redshift snaps generating approximately 2 PB of data and results, carefully compressed and double-checked as they passed through the pipeline and were organised. Twelve of these snaps are classified as "primary redshifts" (z =0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 1.1, 1.4, 1.7, 2.0, 2.5, 3.0) and for them all information is displayed for a 10% subsample of the total particles including their position, velocity, ID number and kernel density estimate. For some of the boxes of some of these redshifts even the complete sample of all particles is given. The `abacusutils` package allows to decompress and read all this data in a direct way (abacusorg, 2023). In contrast, for the remaining 21 intermediate secondary redshifts, only the ID number of the particles belonging to the halos is given, as their main purpose is to enable the generation of merger trees for halos, being CompaSO, a halo search algorithm, one of the most attractive tools of Abacus.

2.1.1 CompaSO

While cosmological N-body simulations use millions of particles identified as dark matter, they do not themselves represent any real observable. It is therefore necessary to find ways to group these particles into larger clusters that can be recognised as dark matter halos, as they are the ones hosting galaxies and revealing the cosmological structure on large scales. It is important to do this consistently as different classification criteria will result in different virialized particle structures in the simulation, partly determining the information one gets from its study.

The halo search algorithm CompaSO described in Hadzhiyska et al. (2021) first groups the particles into large disjoint sets called L0. To do so, it uses a slightly modified friend-of-friend (FoF) algorithm (Davis et al., 1985). Generally the FoF algorithm will assign two particles to the same group if they are less than a characteristic distance apart defined as the linking length $l_{\mathrm{FoF}} = 0.2 \cdot l_{\mathrm{mean}}$, where l_{mean} is the average inter-particle separation. L0 halos use this method only when applied to particles under a limit of their local density greater than what would correspond to them with a characteristic distance of $l_{\mathrm{FoF}} = 0.25 \cdot l_{\mathrm{mean}}$. In this way, a physical smoothing scale is imposed. Within these large groups are the L1 disjoint groups, which will be the actual halos defined with a modified spherical overdensity method. The particle with the highest local density among its neighbours in L0 is chosen as the centre of the halo and from it a spherical radius is increased until the virialization criterion is met. The process is repeated following a careful competitive guideline that takes into account the distance to other halos and their comparative density until all the particles in L0 have been assigned to a halo. Then this same process is carried out within each L1 halo with higher density criteria to generate the substructure in the form of subhalos corresponding to L2 groups.

All this is done on-the-fly, without the need to save the particle data for later, allowing for quick access to a wealth of statistical information about the haloes, such as an identification number, the number of particles contained in that halo, its parent halo L0 and its more massive subhalo, whose centre of mass acts as the centre of mass for the statistical calculations of the L1 halo, as is the case with the velocity dispersion of its particles. CompaSO works for all redshift snapshots, so the particle IDs allow for the creation of merger trees that can be used to clean the halo catalogues from low mass clusters that even if now appear as independent, have partially merge with others or lost most of its particles, improving performance over the problem of boundary particles and shared frontiers between merging halos along its history Bose et al. (2022).

2.2 Dark Matter Haloes

Several key methods are used in cosmological simulations to characterise its performance and to obtain a comprehensive understanding of the large-scale structure of the universe. Important tools in this respect are the halo mass function and the halo bias, based on the comparison of the clustering of dark matter and halos described by their respective correlation functions.

2.2.1 The Halo Mass Function

The Halo Mass Function (HMF) is a statistical insight that describes the abundance and mass distribution of dark matter halos in a given volume of the simulated universe. As a histogram, it provides the normalised number density per unit volume and mass range as a function of the halo mass. By analysing the HMF at different redshift snaps, the underlying patterns of structure formation and evolution can be discovered, the properties of dark matter particles investigated, and the cosmology used characterised. This can be tasted with Figure 1. From its study it can be clearly seen that different mass resolutions give rise to different halo distributions. In the `high` case, the highest resolution configuration, numerous less massive halos can form in the order of $10^{10}\,\mathrm{h}^{-1}\,\mathrm{M}_{\odot}$, while for coarser resolutions such as `hugebase` only halos from $10^{12}\,\mathrm{h}^{-1}\,\mathrm{M}_{\odot}$ onwards are visible. At intermediate mass resolutions such as the one used here, $2 \cdot 10^{9}\,\mathrm{h}^{-1}\,\mathrm{M}_{\odot}$, the least massive halos are found from $10^{11}\,\mathrm{h}^{-1}\,\mathrm{M}_{\odot}$ onwards. In addition, some combinations of phase and simulation configuration such as `base` and `highbase` can lead to the same mass resolution and therefore to similar HMF profiles. However, the distribution of haloes in the lower resolution simulations differs from that of the higher resolution ones for the more massive haloes, as the finite conditions of the simulation make these objects strange and few in number. The study of the HMF can thus help us to characterize the dependence of the simulation on its mass resolution.

In short, it can be obtained by classifying the different dark matter halos according to their mass into different boxes corresponding to different mass ranges and normalising. The CompaSO tool is used here to quickly build a catalogue of halos in the required cosmology and redshift from which L2 subhalos are discarded. This catalogue has a label `N` corresponding to the number of dark matter particles constituting each halo. Multiplying the number of particles by the particle mass resolution described in Table 1 gives the halo mass. After this,

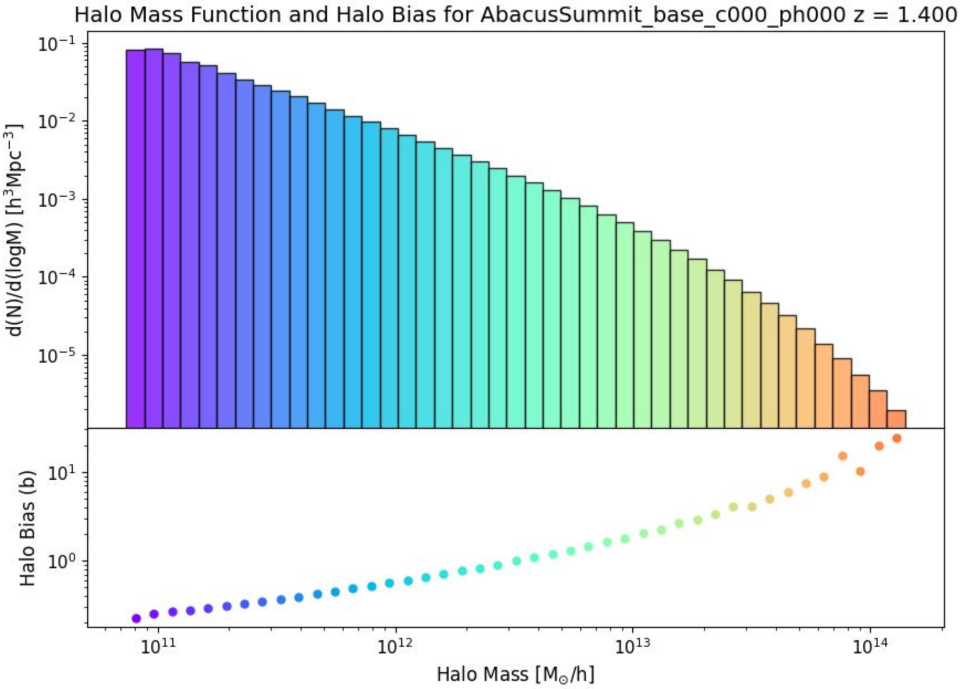

Figure 2: The halo mass function of the cleaned halos at z = 1.4 from the `base` simulation and phase `ph000` in the fiducial Planck 2018 ΛCDM cosmology `c000` with the normalised number density per unit volume and mass range as a function of the halo mass. The lower panel shows the dependence of the halo bias on the halo mass, being more relevant as masses grow. The higher mass ranges show a high uncertainty due to the lack of halos to correlate with in this finite volume.

a few tens of bins are defined in a logarithmically equispaced mass range starting from the minimum mass among the haloes in the catalogue until the mass of the most massive one is reached. In this case, the minimum halo mass considered is $7 \cdot 10^{10}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$, corresponding to 35 particles, and 50 bins are used. Once the counts in each bin are obtained, the result is normalised dividing by the volume of the simulation $2^3\,\mathrm{h}^{-3}\,\mathrm{Gpc}^3$ and by the corresponding step in the mass range. The final result for the `AbacusSummit_base_c000_ph000` simulation at redshift $z = 1.4$ can be appreciated in Figure 2. As exposed, it can reveal key insights into the distribution and abundance of dark matter halos. The plot exhibits a characteristic shape, with a peak at the $10^{9.5}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$ halos that indicates the most common halo masses present in the simulated universe at the given redshift. The HMF also showcases a gradual decline in the number of halos towards the higher masses, demonstrating the hierarchical nature of structure formation. Due to the finite conditions of the simulation and the early epoch considered, the most massive halos are really strange and thus not suitable for statistical analysis, so is better to cut them. Using the values of the normalise density is easy to calculate the histogram bare counts to set a reasonable limit at $10^{-6}\,\mathrm{h}^3\,\mathrm{M}_\odot^{-3}$ under which no more than 1000 halos are found in each mass bin, as it is shown in this case.

2.2.2 The Two Point Correlation Function

The two-point correlation function (2PCF) measures the excess probability of finding a halo pair compared to a random distribution in a given volume. To calculate it, a simple estimator as the following can be used in simulations to account for the boundary periodic conditions of the box as explained in Garrison (2023):

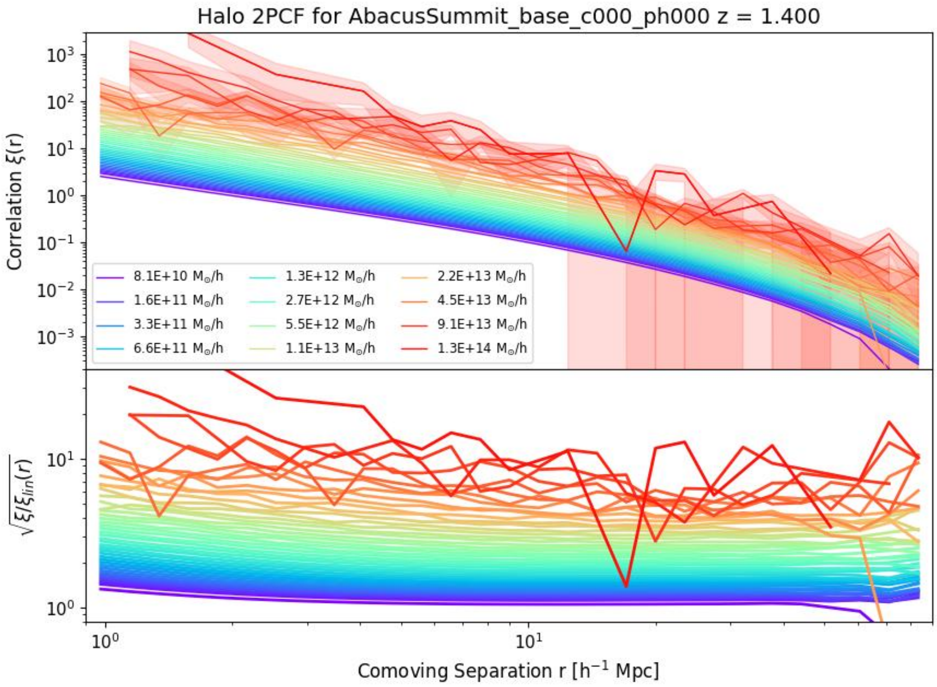

Figure 3: The two point correlation function for dark matter halos in different mass ranges with their corresponding Poisson uncertainty. Some vertical bars from the Poisson error appear due to intense shot noise, as in those cases only one halo pair could contribute to the correlation. Below is the ratio between the halo 2PCF and the real space correlation function of the underlying dark matter particles derived from the linear theory. This ratio is roughly constant for $7\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 50\,\mathrm{h}^{-1}\,\mathrm{Mpc}$.

$$\xi(r) = \frac{D_1 D_2(r)}{R_1 R_2(r)} - 1 \tag{1}$$

Here D_1D_2 is the number of halo-halo counts within r and r + dr separations measured in the simulation Cartesian coordinates. Allowing for periodic boundary conditions R_1R_2 can be calculated as:

$$RR = n^2\, V \Delta V \tag{2}$$

Where RR_i is the weighted expected number of random-random pairs within the separation bin r + dr, n is the mean number density of objects, V is the total simulated volume and ΔV is the volume of the spherical shell encircled by the r + dr separation bin (Gonzalez-Perez et al., 2011).

A simple estimation of the correlation uncertainty can be obtained with the Poisson error, used in what follows and defined with respect to the number of pairs $DD(r)$ at a given separation (e.g. Gonzalez-Perez et al., 2020):

$$\sigma_{\text{Poisson}} = (1 + \xi(r))/\sqrt{(}DD(r)) \tag{3}$$

In this work the `pycorr` python package is employed for this task as implemented in `Mohammad and Percival (2022)`. This package uses `Corrfunc` as its optimised, vectorised correlation engine for cosmological calculations (Sinha and Garrison, 2020; Sinha and Garrison, 2019).

The resulting 2PCF for the dark matter halos of this simulation according to their mass can be seen in Figure 3. The plot reveals the correlation strength as a function of halo separation, spanning from $0.1\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ to near $100\,\mathrm{h}^{-1}\,\mathrm{Mpc}$. At small separations, a pronounced upturn in the correlation function can be observed, indicating a higher probability of finding halos in close proximity to each other. This behavior reflects the presence of small-scale clustering and suggests the existence of halo groups or clusters. As the separation increases, the correlation function gradually diminish, illustrating a decrease in the clustering strength of halos on larger scales. This transition from strong clustering to weaker correlations at larger separations aligns with the hierarchical nature of structure formation, where smaller halos merge to form larger structures over cosmic time. Additionally, the plot highlights the dependence of the 2PCF on halo mass, with more massive halos exhibiting stronger clustering signals compared to their less massive counterparts, anticipating what the halo bias will come to confirm quantitatively in Section 2.2.3. It is noteworthy that the Poisson uncertainty becomes dominant where the number of correlated pairs decreases as dictated by Equation 3, i.e. for the most massive haloes in particular and in the large distance regime in general. There where only one halo pair was found, the Poisson uncertainty displays a vertical bar.

2.2.3 The Halo Bias

The halo bias refers to the statistical relationship between the distribution of dark matter halos and the underlying distribution of matter in the universe. It quantifies the statistical clustering of dark matter halos compared to the overall matter distribution. In its simplest description, the linear halo bias depends only on halo mass, with more massive halos being more strongly clustered than less massive halos. In other words, the halo bias describes how likely it is to find a dark matter halo of a given mass in regions with different densities of matter. A positive halo bias indicates that halos are more clustered than matter on average, while a negative halo bias means that halos are less clustered than matter. We study the halo bias through numerical simulations, observations of large-scale structure in the universe, and theoretical models. By comparing the observed or simulated clustering of galaxies or dark matter halos to the expected clustering based on the matter distribution, we can infer the properties of the underlying matter and the physical mechanisms influencing the formation and evolution of cosmic structures (Sato-Polito et al., 2019). Other secondary bias exists that correlate with the clustering capacity of halos, like its concentration or density profile. There has even been suggested a secondary bias specific to QSOs, the so-called merger bias, that argues that when modelling quasars in simulations it would be necessary to take into account the fact that the galaxy has undergone a recent merger as a bias for quasar formation (Bonoli et al., 2010).

In simulations, the halo bias can be quantified using various statistical measures, but a commonly employed one is the halo correlation function. The halo bias can be then calculated by comparing the clustering of halos to the clustering of dark matter on different scales by measuring the excess or deficit of halos in regions of the universe relative to what would be expected if halos were simply distributed randomly. The first necessary step is to calculate the 2PCF for the halos in each mass bin of the HMF, which is shown in Figure 3 for 30 equispaced separations in the $0.9\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 90\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ broad range.

The same thing could be done now with the dark matter particles, but the simulation peculiarities make it a hard task. As stated in Section 2.1, most of the simulation snaps do not save all the dark matter particles, and even if at $z = 1.4$ there should be an accessible 10% subsample, it is not evenly representative of the hole dark matter field as the halo particles are prioritised for the halo catalogue writing. Instead, one can access the primordial power spectra of the perturbations and cosmologies of the different theories predicted with Boltzmann codes like CLASS (Lesgourgues, 2011) using a package like `cosmoprimo` as build in Cosmodesi (2023), linearly evolve it to the desired redshift and go from Fourier space to the real space of the 2PCF to compare with that of the halos using the Hankel transformation (Avila et al., 2020):

$$\xi_{\text{lin}}(r) = \frac{1}{2\pi^2} \int_0^\infty P_{\text{lin}}(k)\, j_0(kr)\, k^2\, dk \tag{4}$$

Then one needs to find the separation regime where the ratio between the two clusterings is linear and the halo bias applicable, avoiding non-linearities affecting small scales and the BAO feature. As depicted at the bottom of Figure 3, the ratio between the correlation of the halos and that of the linear theory is very fluctuating at small and large scales, even more so for the most massive halos due to the scarcity of samples in the statistics. However, at an approximate scale of $7\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 50\,\mathrm{h}^{-1}\,\mathrm{Mpc}$, the ratio stabilises at a particular value visible as a horizontal line for each mass range on the graph, where one can consider a linear relationship between the two mediated by the value of a constant, i.e. the bias. The regime where this is applicable varies with the mass of the halos and so a conservative range of distances is chosen that applies to all on average.

For each mass bin, the value of the linear halo bias b_i minimizes χ^2 in the following expression where $\xi_{M_i}(r)$ is the 2PCF for each mass range i, $\sigma(\xi_{M_i})(r)$ is the corresponding standard deviation, $\overline{\xi}_{M_i}(r)$ is the mean 2PCF over that same range and $\xi_{\text{lin}}(r)$ is the real correlation from the linear theory, all evaluated along the separation linear regime r defined above (Avila et al., 2020):

$$\chi^2(b_i) = \sum_r \left(\frac{\xi_{\text{lin}}(r)\, b_i^2 - \overline{\xi}_{M_i}(r)}{\sigma(\xi_{M_i})(r)} \right)^2 \tag{5}$$

The resulting halo bias dependency with mass can be studied attending to Figure 2, which yields a strong correlation between the halo mass and clustering. As the halo bias increases with increasing halo mass, more massive halos tend to be more strongly clustered. The higher clustering strength of massive halos implies a preferential halo formation area in dense regions with enhanced gravitational potential.

3 Modelling Galaxies

The following is a description of the models used in this thesis, their implementation and the tools for their analysis. Specifically, a vanilla Halo Occupation Distribution model is described to create a mock catalogue of QSOs in the dark matter simulation environment described above. The noise implementation model for redshift smearing is then introduced followed by the statistical analysis tools shared by both in their observational characterization.

3.1 AbacusHOD

The Halo Occupation Distribution (HOD) models offers a valuable framework for describing the relationship between galaxies and dark matter halos. These models provide a probabilistic description of the distribution of galaxies within halos, considering factors such as halo mass and other relevant properties (Wechsler and Tinker, 2018). The HOD models are typically employed to quantify the probability distribution for the number of galaxies meeting certain criteria, e.g. a mass threshold, within a given halo. This distribution can be specifically characterized for both central galaxies, which reside at the center of halos (here L1 halos) and satellite galaxies, which orbit within the halos (here populating the L2 subhalos inside L1 halos). By considering the Mean Halo Occupation Distribution (MHOD), which encapsulates the average number of galaxies within halos of different masses, the standard HOD model its fully defined, offering insights into the clustering and abundance of galaxies in the cosmic landscape. This approach has not only proven particularly successful in galaxy redshift surveys, enabling a deeper understanding of galaxy evolution physics (Levi et al., 2013), but also has served as a valuable tool for the generation of mock catalogs that accurately reproduce these observed clustering patterns, allowing to test for different cosmological parameters (Yuan et al., 2023).

More precisely, the HOD can be understood as a probability distribution $P(n_\mathrm{g}|X_\mathrm{h})$ that assigns a number of galaxies n_g to a given halo according to its properties X_h. As with the halo bias, in the base HOD model it is assumed that the halo mass M_h is the only and most relevant property to take into account, although as described in section 2.2.3 for halo clustering, there is certainly a dependence on other properties such as their assembly, known as secondary biases, which will not be taken into account here. This baseline model assumes a Bernoulli occupation distribution for the central galaxies and a Poisson distribution for the satellites, which are properly modeled through some key parameters (Smith et al., 2020).

The AbacusHOD code is used for this purpose to efficiently implement a physically informed QSO HOD based on the DESI One Percent Survey data as described by Yuan et al. (2023). AbacusHOD can be used to find best-fit HODs from data and to further reproduce them with mock galaxy catalogues for analysis as detailed in Yuan et al. (2021). The tool is part of the `abacusutils` package (abacusorg, 2023) and works in a very direct way. The HOD parameters and the target simulation details are collected in a `.yaml` configuration file. In a first step, the necessary simulation dark matter and halo data are loaded and prepared for its future fast access and manipulation. Next and last one can build as many mock galaxy catalogues as desired with the given HOD parameters or new others to perform statistical analysis with the build-in tools from AbacusHOD such as $\xi(r_\mathrm{p}, r_\pi)$ or w_p, as described in detail in section 3.3.

3.1.1 Vanilla Model

The HOD vanilla model for QSOs mimicked here follows the probabilistic scheme first described for luminous red galaxies in Zheng et al. (2007) as modified and implemented in Yuan et al. (2023). The mean number of central galaxies $\langle N_\mathrm{cen}\rangle$ within a halo of mass M_h is given by:

$$\langle N_\mathrm{cen}\rangle = \frac{f_\mathrm{ic}}{2}\,\mathrm{erfc}\left[\frac{\log_{10}(M_\mathrm{cut}/M_\mathrm{h})}{\sqrt{2}\sigma}\right], \tag{6}$$

Where erfc(x) is the error function, M_cut is the minimum halo mass for hosting a central galaxy, σ is the width of the transition region from 0 to 1 in the number of central galaxies, and f_ic is an incompleteness parameter between $0 < f_\mathrm{ic} \leq 1$ that acts as a downsampling factor to control the number density of the mock galaxies.

The mean number of satellite galaxies $\langle N_\mathrm{sat}\rangle$ within a halo of mass M_h is given by:

$$\langle N_\mathrm{sat}\rangle = \left[\frac{M_\mathrm{h} - \kappa M_\mathrm{cut}}{M_1}\right]^\alpha, \tag{7}$$

where M_1 is the characteristic mass at which, on average, one satellite galaxy resides in the halo, α is the power-law slope of the satellite HOD, and κM_cut is the minimum halo mass to host a satellite galaxy.

The vanilla model is then determined by 6 HOD parameters ($M_\mathrm{cut}, M_1, \sigma, \alpha, \kappa$ and f_ic) that will enter in the `.yaml` configuration file to build the mock QSO catalogues. Its values have been derived in Yuan et al. (2023) performing HOD best-fits analysis with respect to the DESI One Percent Survey using the statistical observables explained in 3.3, i.e. constructing and minimizing likelihood functions with different HOD models statistical analysis with respect to the observations. DESI survey is willing to obtain spectroscopic data from 40 million galaxies to perform the most precise measurement of the expansion history of the universe to date. The

Table 2: Main parameters of the AbacusHOD vanilla model describing the central and satellite occupation profiles of QSOs. Mass units are given by $\mathrm{h^{-1}\,M_\odot}$

HOD Vanilla Model Parameters					
$\log M_{\mathrm{cut}}$	$\log M_1$	σ	α	κ	f_{ic}
12.67	15.00	0.58	1.09	0.74	0.041

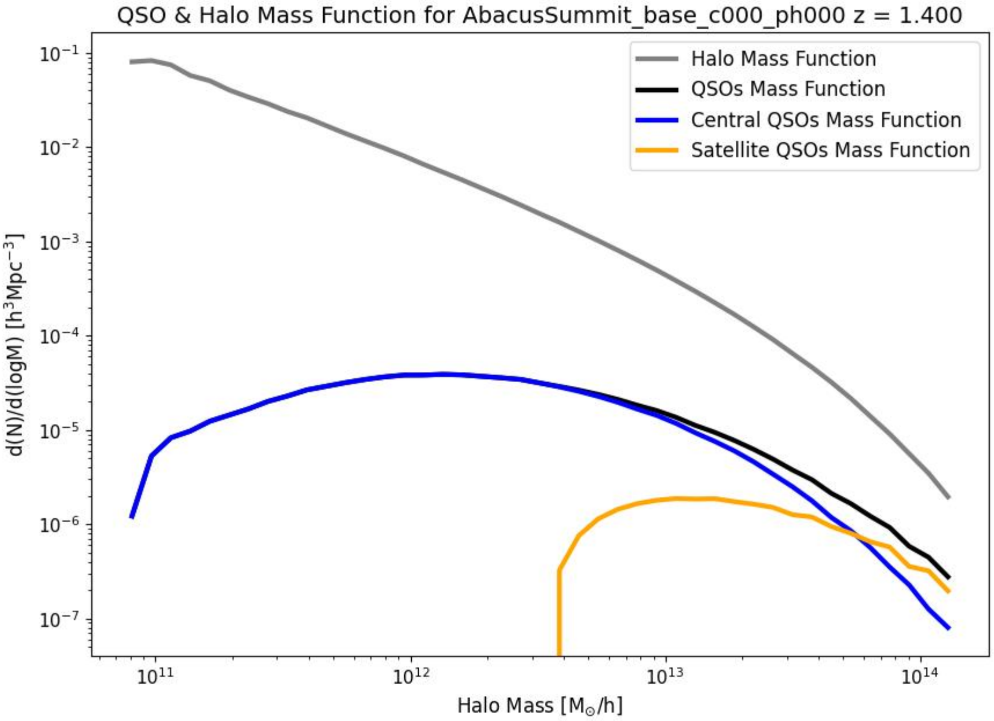

Figure 4: The QSO and halo mass function at $z = 1.4$ with the normalised number density per unit volume and mass range as a function of the halo mass. The blue curve corresponds to the mass function of the halos inhabited by central QSOs, while the orange one depicts that of the halos where a satellite QSO inhabits a subhalo.

Early Data Release (EDR) used in this HOD tuning was obtained in the survey validation campaign just before the actual survey starts, and covers roughly 1% of the total $14\,000 deg^2$ aimed volume, hence its name. DESI targets bright galaxies, luminous red galaxies, emission line galaxies and Quasars, on which this HOD model focuses. Being some of the brightest luminous extragalactic sources, QSOs are perfect tracers to explore high redshifts, and the EDR provides 24 182 of them in the $0.8 < z < 2.1$ redshift range with more or less constant number density. That is why $z = 1.4$ was selected as the intermediate snap to validate the HOD model in Yuan et al. (2023) and thus chosen in this work. The concrete values of the HOD parameters are shown in Table 2.

3.1.2 Mock QSOs

Figure 4 depicts together the total halo and QSO mass functions derived from the simulation at redshift $z = 1.4$ and the HOD vanilla model described in Section 3.1.1, offering valuable information about the population and distribution of QSOs within the dark matter halos. The HOD distinguishes between the contributions from central galaxies and satellite galaxies. The central galaxy contribution represents the QSOs hosted by central galaxies inhabiting L1 halos.

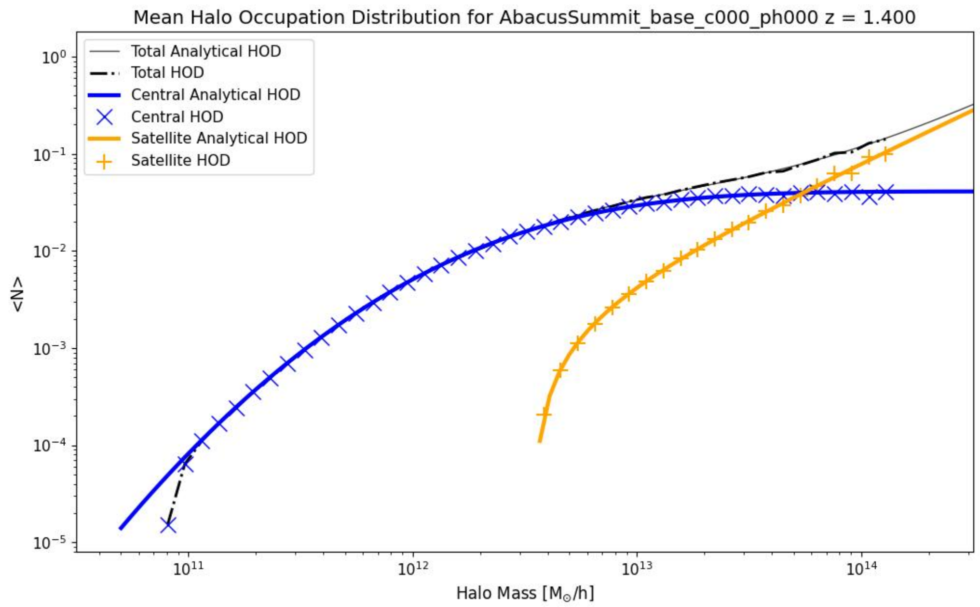

Figure 5: The vanilla model Mean Halo Occupation Distribution for central and satellite QSOs at $z = 1.4$ with the mean number occupation as a function of the halo mass. The blue curve corresponds to the mass function of the halos inhabited by central QSOs, while the orange one depicts that of the halos where a satellite QSO inhabits a subhalo. The analytical profile is that described in Equations 6 and 7

However, the satellite QSO mass function, despite its name, represents the QSOs in terms of the central L1 halo mass they inhabit, rather than the satellite L2 halos where they are indeed evolving.

Upon analyzing the plot, notable features are observable for both the halo and QSO mass functions. The halo mass function here is just the contour of the one depicted in Figure 2. The halo population is far from being saturated with this kind of galaxies, especially for the more common low mass ranges, as the QSOs number density is always some orders of magnitude under the HMF curve. This is not the case for the few top mass halos, which are much more occupied and may even have satellite QSOs together with a central one. For its part, the central galaxy contribution dominates the low and medium halo mass range among the total galaxy mass function, with an occupation peak at $10^{12.5}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$ halos, slowly decreasing with the HMF towards larger halos due to fewer candidates to accommodate them and also towards less massive halos as the HOD dictates. The satellite galaxy contribution makes its entrance at approximately $5 \cdot 10^{12}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$ as the $\log M_{\mathrm{cut}}$ HOD parameter said in 7. It becomes more significant at higher masses and peaks at $10^{13}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$, slowly decreasing with growing mass but surpassing the central QSO occupation at $10^{13.6}\,\mathrm{h}^{-1}\,\mathrm{M}_\odot$, indicating that for this huge halos it is more probable to find satellite QSOs rather than central.

The Mean Halo Occupation Distribution (MHOD) of the given simulation and mock QSO catalogue is shown in Figure 5 together with the analytical expressions of the HOD exposed in 3.1. The MHOD quantifies the average number of QSOs hosted by dark matter halos of different masses and can be obtained by dividing the QSO mass function by the total HMF in Figure 4. Analyzing the plot one can observe the agreement or discrepancy between the actual QSO distribution obtained from the simulation and the analytical HOD guidance. As expected, the simulated QSO distribution matches the analytical orientation of the HOD from which it comes, although it disagrees slightly at high and low halo masses due to the shortage in the number of QSOs. The MHOD also gives a better perspective from which to understand what was discussed about QSO occupation in Figure 4.

3.2 Redshift Smear Model

While in the framework of a cosmological simulation the observer occupies a privileged position with access to the velocities and positions of halos and galaxies with respect to the origin of coordinates, when conducting surveys such as DESI, such relative distances are only available in the plane perpendicular to the line of sight, while along it the distance measurements are made in redshift space. The redshift of the electromagnetic spectrum of a light source occurs when it moves away from the observer, and this trend is general as the Universe is expanding. But as the source lays within a particular gravitational potential, the shift is also affected by its peculiar velocity with respect to the observer. This gives rise to the RSD. Being related to the matter distribution and gravitational potentials at large scales but also with the expansion of the Universe, its study is enormously rich as it can provide crucial insights into the growth of cosmic structure and the nature of dark energy and gravity. To change from real space r to redshift space s along the line of sight axis taking into account the expansion rate of the universe at a given redshift $H(z)$ and the peculiar velocity of the object along the sight axis $v_{\parallel}$ one has to apply (Kaiser, 1987):

$$s = r + \frac{v_{\parallel}(1+z)}{H(z)} \tag{8}$$

Even though QSOs are perfect targets for RSD studies, as they can be observed at great distances and have recognisable spectral features to compare with, the RSD is a fine observational target bound to different kinds of systematic uncertainties. For instance, there is wide variety of emission line widths in the spectra of quasars. According to the Active Galactic Nuclei model (AGN) (La Mura et al., 2017), the fast rotation of hot gas close to the central black hole can produce broad emission lines. Also, the radiation driven winds affect the gas emission lines misplacing them. The observing conditions and other astrophysical effects also contribute to this systematic error. Accurately modeling and fitting the uncertainties associated with the redshift smearing is thus of paramount importance, as it enables the mitigation of biases and the obtainment of reliable measurements of velocity-related quantities within the RSD study.

To quantify the redshift smearing uncertainty from observations, is necessary to analyse great samples of data and compare the different redshift estimations of same objects. This was done in Yu et al. (2023) for the QSOs present in the DESI One Percent First data realise. By studying the redshift difference from repeated observations of QSOs, the statistical redshift uncertainty was found to follow a Lorentzian like distribution. Here this behaviour of uncertainty in the redshift smearing of the simulated QSOs will be modeled by adding random noise to their z-axis peculiar velocities following a Lorentzian-like distribution and then changing to redshift space following 8. This will also be done with a Gaussian-like distribution for comparison.

Figure 6 illustrates the comparison between a Gaussian distribution and a Lorentzian distribution, along with their respective analytical expressions. The Lorentzian distribution, also referred to as the Cauchy distribution, exhibits a symmetric peak with heavy tails. Its analytical expression is given by:

$$L(p, \gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x-p)^2+\gamma^2}\right)$$

where p represents the peak location and γ represents the half-width at half-maximum (HWHM).

On the other hand, the Gaussian distribution, also known as the normal distribution, is characterized by its mean and standard deviation. Its analytical expression follows the form:

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

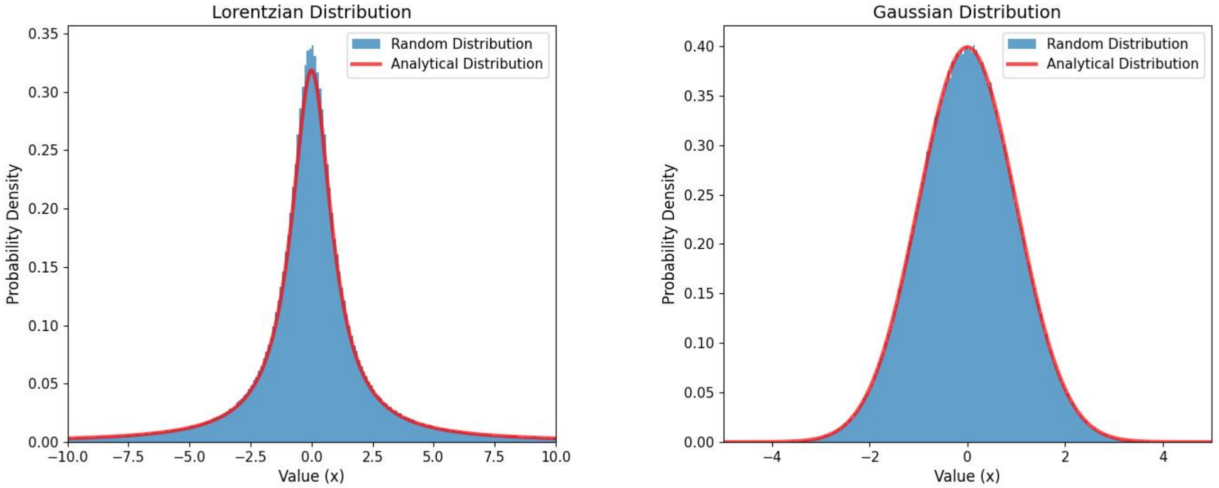

Figure 6: The vanilla Lorentzian and Gaussian probability distributions with their analytical curve.

where μ represents the mean and σ represents the standard deviation (Yu et al., 2023).

The Gaussian distribution in Figure 6 appears symmetric, centered around its mean, and gradually decreases towards its tails, demonstrating a characteristic bell-shaped curve. Also with a symmetric peak, the Lorentzian distribution exhibits in contrasts tails that extend infinitely, leading to more pronounced outliers. This is what makes it suitable to better fit QSOs redshift smearing uncertainties, although too large catastrophic redshift differences must be taken into account by setting a maximum cut to its tails, as in real surveys this catastrophic redshifts usually come from a reduction error or contamination and not from actual observational reasons (Smith et al., 2020).

Following Yu et al. (2023), for a fitting with physical base p and μ must be set to 0 for centered distributions while γ and σ are set to be $\nu_{smear} = 78\,\mathrm{km\,s^{-1}}$ by best observational fit for $z = 1.4$. Absolute peculiar velocity additions from the random noise distributions above the $2000\,\mathrm{km\,s^{-1}}$ are set to be outliers.

3.3 Galaxy Observables

The clustering measurements in cosmological simulations and surveys play a crucial role in its analysis, allowing for a quantitative understanding of the large-scale structure of the universe and in testing cosmological models. The 2PCF has been introduced in 2.2.2 and now other concepts as the Kaiser Factor and tools such as the Redshift Space Correlation Function, the Multipoles or the Projected Two Point Correlation Function are brought to the scene.

The 2PCF quantified the excess probability of finding pairs of objects at different separations. However, when changing to redshift space with Equation 8, peculiar velocities introduce anisotropic distortions. The Kaiser factor, denoted as K, is a correction term applied to the 2PCF in redshift space to account for the proper velocities at large separations. It takes into consideration the linear theory approximation for redshift space distortions caused by the coherent infall of galaxies towards overdense regions (Kaiser, 1987). The Kaiser factor can be written as in Gonzalez-Perez et al., 2020, i.e.:

$$K = 1 + \frac{2}{3}\frac{\Omega_{\mathrm{m}}^{\gamma'}}{b} + \frac{1}{5}\left(\frac{\Omega_{\mathrm{m}}^{\gamma'}}{b}\right)^2, \tag{9}$$

Where b represents the bias parameter, Ω_m is the matter content of the Universe at the given redshift and γ' is the growth index for which General Relativity predicts $\gamma' = 0.55$. The redshift space 2PCF shows a compression along the line of sight and a consequent enlargement of the clustering due to the peculiar velocities. This is known as the “Finger of God” effect for which the Kaiser formalism sets a first order relation as $\xi(s) = K(r)\xi(r)$, well establish only for large enough scales ($r > 1\,\mathrm{h}^{-1}\,\mathrm{Mpc}$) as it is based on linear perturbation theory.

The Redshift Space Correlation Function $\xi(s, \mu)$ is an extension of the 2PCF to account for the effects of redshift space distortions. It depends on two variables: s, which represents the line-of-sight separation between two objects, and μ, which is the cosine of the angle between the line-of-sight direction and the vector connecting the two objects. The redshift space correlation function can be estimated as in Avila et al. (2020):

$$\xi(s, \mu) = \frac{D_1 D_2(r, \mu)}{n\Delta V} - 1, \tag{10}$$

Where μ is the cosine of the angle with respect to the line of sight, DD is the number of objects pairs within the separation bin r + dr and orientation between $\mu + \Delta\mu$, n is the mean number density of objects, and ΔV is the volume of the spherical shell encircled by the r + dr separation bin. The Multipole Moments of the Redshift Space Correlation Function, denoted as $\xi_\ell(s)$, provide a decomposition of the correlation function into different angular momentum components. These moments can be obtained through a Legendre polynomial expansion of $\xi(s, \mu)$:

$$\xi_\ell(s) = (2\ell + 1) \int_0^1 \xi(s, \mu) L_\ell(\mu) d\mu, \tag{11}$$

where L_ℓ is the Legendre polynomial of degree ℓ. In linear theory, only the monopole ξ_0, quadrupole ξ_2 and hexadecapole ξ_4 are non-zero Avila et al. (2020). This all are sensible enough to unveil the effects of adding a redshift smear distortion.

The Projected Two Point Correlation Function $w_p(r_p)$ is a projection of the correlation function onto the plane of the sky, providing information about the clustering of objects as a function of their transverse separation r_p. It is obtained by integrating the redshift space correlation function over the line-of-sight separation r_π. Mathematically, the projected two point correlation function is given by (Yu et al., 2023):

$$w_p(r_p) = \int_{-r_{\pi_{\max}}}^{r_{\pi_{\max}}} \xi(r_p, r_\pi) dr_\pi, \tag{12}$$

where $\xi(r_p, r_\pi)$ is the correlation function at a given transverse and longitudinal distance and $\pm r_{\pi_{\max}}$ are the limits in the depth longitudinal coordinate r_π.

These methods and statistics are commonly used in the analysis of clustering in cosmological simulations and surveys, and they will be the main toll to investigate the effects of the redshift smear distortions modelled in this work. All of them can be directly calculated thanks to the build in functions from `abacusutils` or those from `pycorr`.

4 Results

4.1 Redshift Smearing Uncertainty Modelling

Figure 7 shows the actual Lorentzian and Gaussian distortions performed in the z-axis peculiar velocities of the mocked QSOs. Here the original velocities are plotted together with the distorted ones and the analytical shapes of the correspondingly added noise. A random velocity is chosen from the allowed noise profile and added to one of the QSO peculiar velocity in the z-axis before continuing with the next one until all have been perturbed. As the distribution main bodies lay within the original velocity profile of QSOs, only some of them end outside the

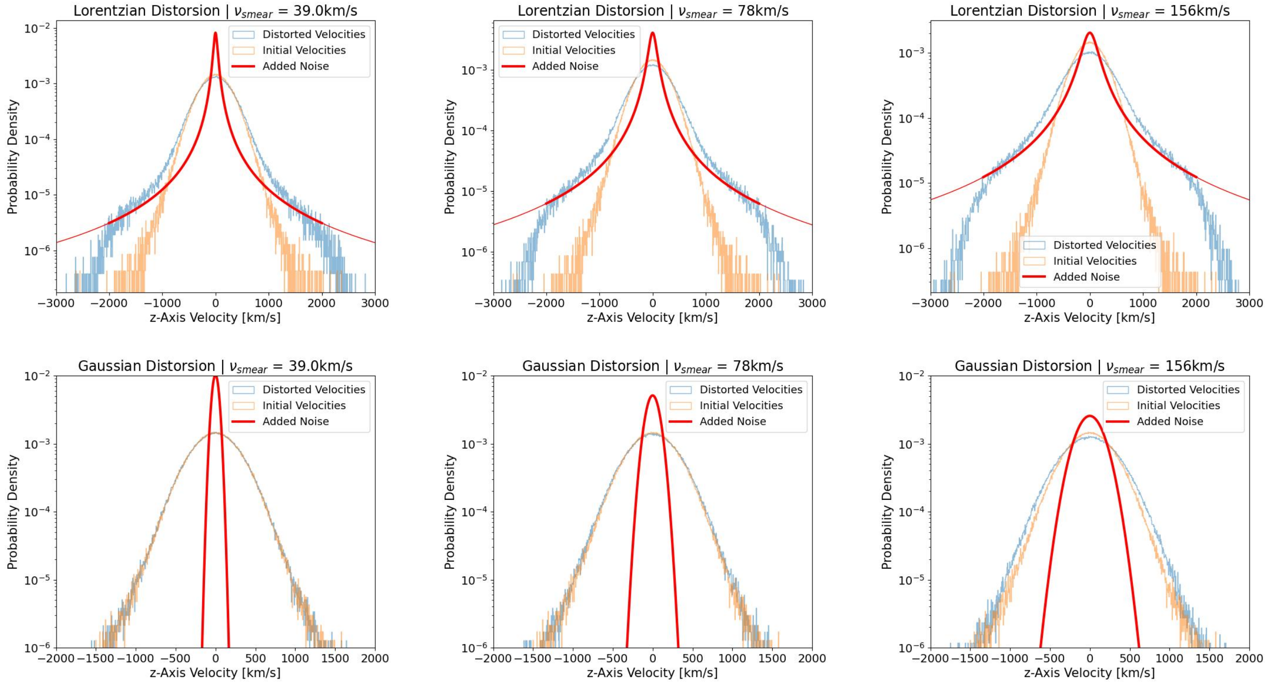

Figure 7: The Lorentzian and Gaussian distortions for the z-axis peculiar velocities of the mocked QSOs on top and bottom respectively. The original velocities are plotted together with the distorted ones and the analytical shapes of the correspondingly added noise. From left to right, ν_{smear} is increased with the best fit $\nu_{smear} = 78\,\mathrm{km\,s^{-1}}$ value at the center and with a factor of two difference to the sides

original Gaussian like distribution and the logarithmic scale becomes necessary to appreciate them.

The spatial contribution to the RSD from the proper velocities of QSOs in each case is shown in Figure 8 only for the best fit $\nu_{smear} = 78\,\mathrm{km\,s^{-1}}$ HWHM noise distributions. Again, the effects of the Lorentzian distortion are much more noticeable than in the Gaussian case, giving rise to an enhanced redshift space distortion. As a test, the maximum and minimum spatial coordinates reached by the QSOs in the sample are calculated to check that the periodic conditions of the box have been respected in the change of coordinates to redshift space with Equation 8.

4.2 Galaxy Clustering

In order to show the effect of the RSD and test the model, the 2PCF in real and redshift space for the QSOs with their original peculiar velocities is shown in Figure 9. The theoretical prediction for the correlation in redshift space as a function of the 2PCF in real space and the Kaiser factor described in Equation 9 is also shown and fulfilled for large scales. Under both correlations, the ratio between the two is noted with the intention of highlighting the Kaiser trend.

Next, the correlation functions in redshift space have been calculated by comparing the modified QSO catalogues according to the two different redshift uncertainty models, the Lorentzian and the Gaussian, with the correlation of the unperturbed distribution with a 35 distance binning between $0.2\,\mathrm{h^{-1}\,Mpc} \leq r \leq 60\,\mathrm{h^{-1}\,Mpc}$ and multiplying each clustering value by its correlated coordinate to enhance the scale differences. Each case has been repeated for the three values of the HWHM parameter selected to appreciate the effect of the noise modelling: $\nu_{smear} = 0.5 \cdot 78\,\mathrm{km\,s^{-1}}$, $\nu_{smear} = 78\,\mathrm{km\,s^{-1}}$ and $\nu_{smear} = 2 \cdot 78\,\mathrm{km\,s^{-1}}$.

Figure 10 contains the projected correlation function w_p calculated by the integration of

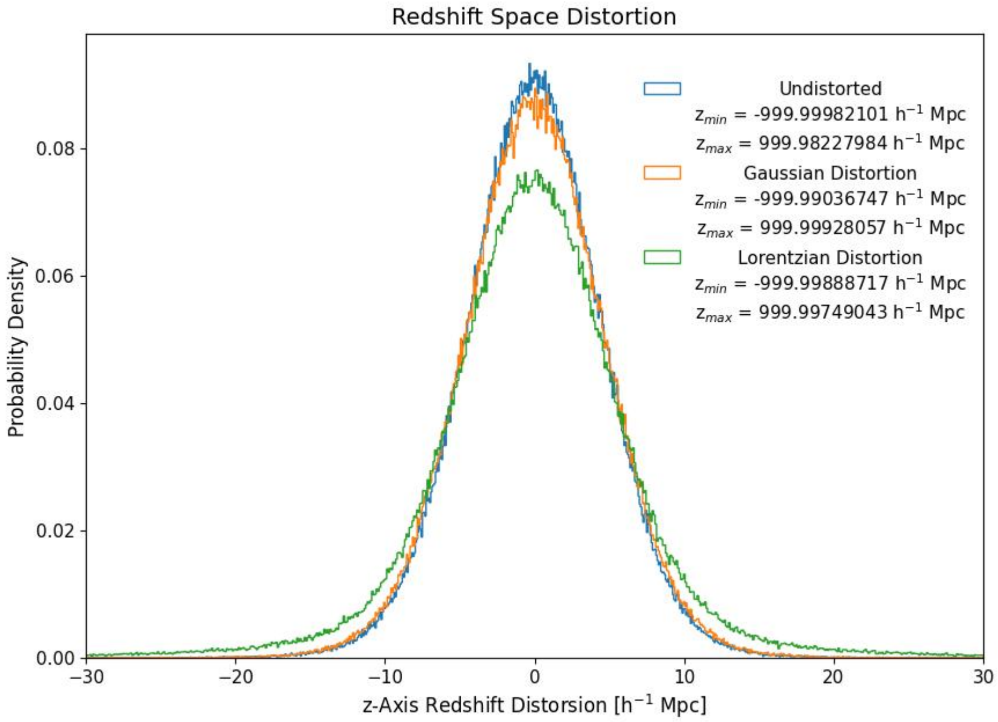

Figure 8: Spatial contribution to the RSD coming from the peculiar velocities of the disturbed and unperturbed mocked QSOs. The maximal z coordinates are consulted to check that the periodic conditions of the simulation box have been respected.

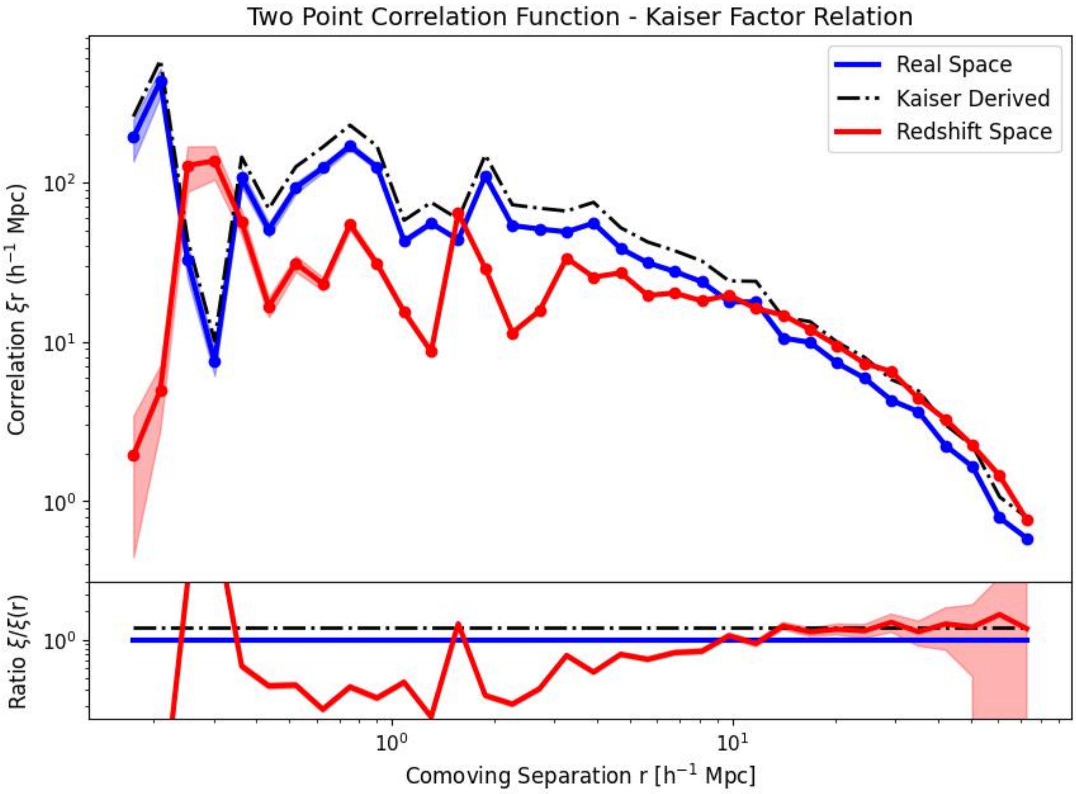

Figure 9: Real space and redshift space 2PCF multiplied by the corresponding distance among the 35 between $0.2\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 60\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ for which it is calculated. Below is depicted the ratio with respect to the real space 2PCF. The linear prediction followed from the Kaiser factor can be appreciated for large enough distances.

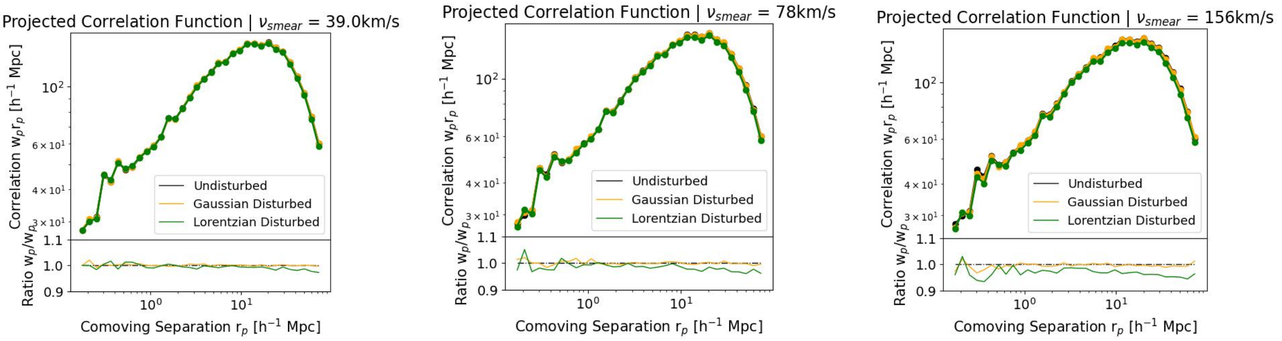

Figure 10: The w_p projected correlation for the Lorentzian and Gaussian distortions of the z-axis peculiar velocities of the mocked QSOs on top and its ratio with respect to the also plotted unperturbed QSO sample at the bottom. A 35 distance binning between $0.2\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 60\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ is used, multiplying each clustering value by its correlated coordinate. From left to right, ν_{smear} is increased with the best fit $\nu_{smear} = 78\,\mathrm{km\,s}^{-1}$ value at the center and with a factor of two difference to the sides.

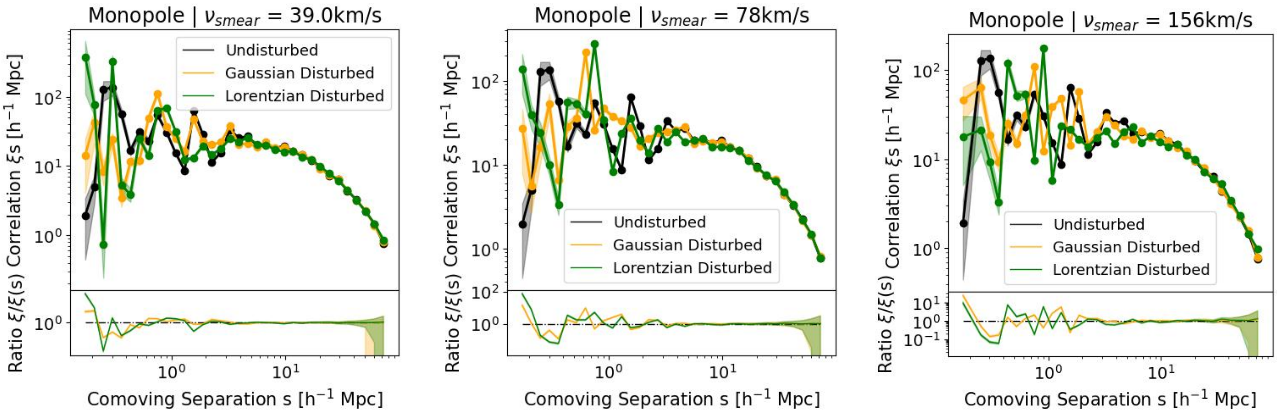

Figure 11: The ξ_0 monopole or redshift space 2PCF with the same display settings as in Figure 10.

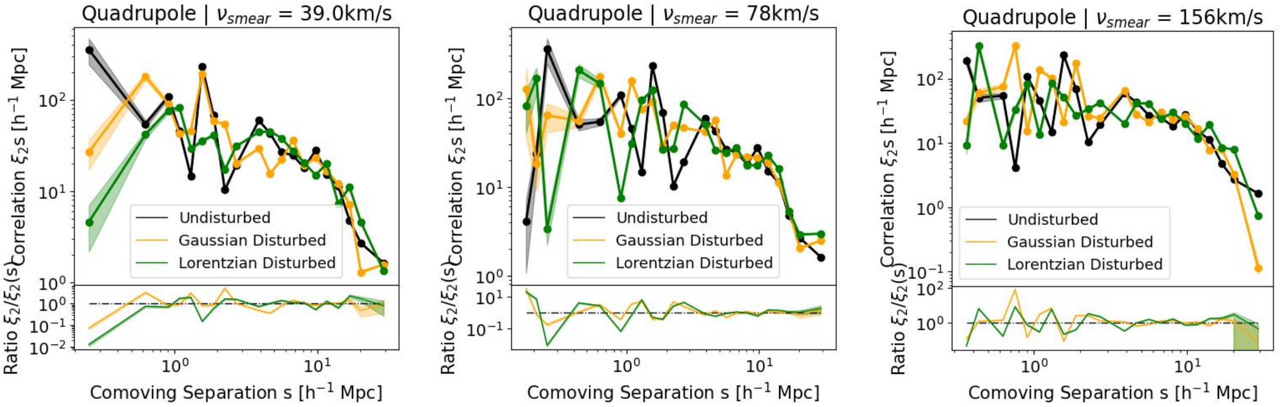

Figure 12: The ξ_2 quadrupole with the same display settings as in Figure 10.

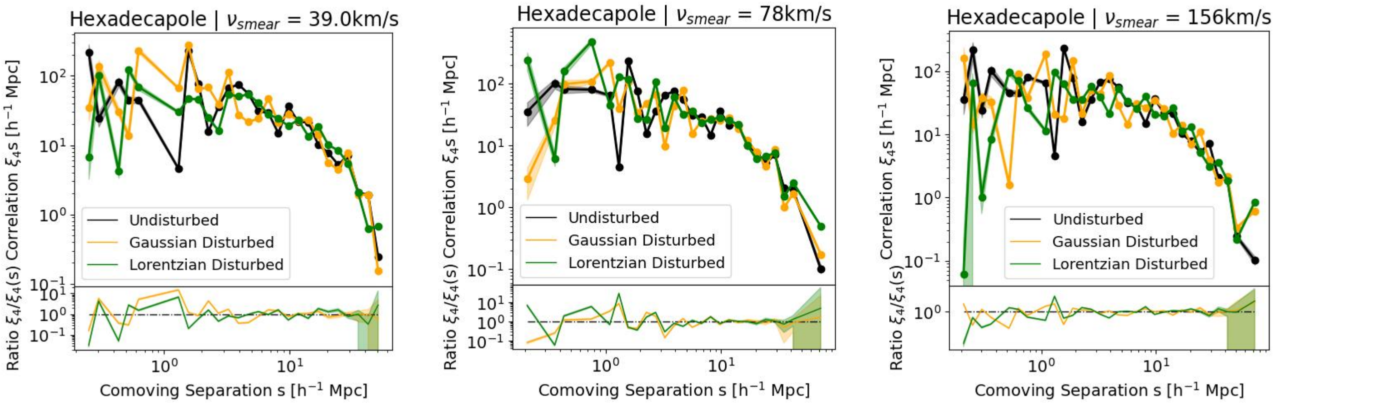

Figure 13: The ξ_4 hexadecapole with the same display settings as in Figure 10.

$\xi(r_p, r_\pi)$ along the chosen z-axis line of sight. Figure 11 shows the redshift space 2PCF, better know as the ξ_0 monopole, followed by the ξ_2 quadrupole and the ξ_4 hexadecapole. Its worth noting that the small and large scale regimes present more relevant Poisson errors as the numbers of available galaxy pairs for correlation diminish.

5 Summary and Conclusions

Throughout this work, a comprehensive characterisation of the cosmological simulation `AbacusSummit_base_c000_ph000` at redshift $z = 1.4$ has been performed, using the Halo Mass Function (HMF) and the halo bias to understand the distribution and clustering of dark matter halos based on its masses. In accordance with this analysis, a Halo Occupation Distribution (HOD) model has been applied to generate a simulated catalogue of Quasars (QSOs) as described in Yuan et al. (2023). By plotting the mass function of the QSOs and the mean HOD, the effectiveness of the HOD model in capturing the population and distribution of QSOs in the simulated universe has been successfully demonstrated. Next, the real-space and redshift correlation functions applicable to the study of the QSO catalogue samples were exposed, also ensuring their theoretical consistency by testing their relation through the Kaiser factor on large scales. Having reached this point, the main objective of the study has been consistently addressed. This was to model the uncertainty in the redshift smearing measurements of QSOs by applying distortions to their peculiar velocities and to further understand its affects on the observational statistics of these tracers. To achieve this, Gaussian and Lorentzian noise has been introduced into the peculiar velocity of the QSOs as described in Yu et al. (2023) before transforming their position along the line of sight to redshift space. To assess the impact of these velocity distortions, the projected correlation function w_p and the 2PCF in redshift space along with the quadrupole and the hexadecapole correlations have been used for various noise profiles and widths in the velocities.

Analysing the results presented in Section 4.1 it can be first concluded that the redshift uncertainty model has been correctly implemented within the QSOs mock catalogues from the vanilla HOD model. Furthermore, from Section 4.2 it can be assured that the clustering of the QSOs redshift space shows valuable observable features with respect to the application of the redshift smearing uncertainty model that make it suitable for observational testing and comparison.

Visualising Figure 9 one can appreciate the trend of the 2PCF in redshift space towards the correlation values obtained by the Kaiser factor for sufficiently large scales starting at distances around $10\,\mathrm{h}^{-1}\,\mathrm{Mpc}$. While Poisson uncertainty dominates at small scales, it cannot explain the fluctuations at scales of the order of $1\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ present for the 2PCF in redshift space and which move it away from the trend described by the Kaiser factor at these scales when one would expect them to be large enough to appreciate the effect (e.g. Gonzalez-Perez et al., 2011; Cabré and Gaztanaga, 2009). This may just be a sign of some peculiarity at these scales for the given tracer, HOD and simulation employed, as the consistency of the other measures in Section 4.1 dissipates the probability of an implementation error.

Looking for example at the distribution of peculiar velocities of the QSOs in Figure 7 it can be concluded that they are correctly centred and within the expected range from Yu et al. (2023). Once distorted, the velocities extend their range of possible values according to the kind of noise generated and its defining parameters. It should be evident the large difference between Gaussian and Lorentzian noise due to the amplitude of their tails. While the Gaussian noise profile redistributes the velocities following a shape similar to the original one, the Lorentzian redistribution creates a new profile with a larger number of QSOs with substantially high apparent proper velocities. As the ν_{smear} half-width at half-maximum (HWHM) parameter value changes, the range of possible noise varies with prominent effects on the Lorentzian distorted sample, while for the Gaussian noise the change has a low-profile impact, at least in this regime where the best fit $\nu_{smear} = 78\,\mathrm{km\,s^{-1}}$ HWHM value and its variations determine a noise which actually fits inside the actual QSO velocity distribution, not really mocking any evident distortion but mostly randomly redistributing the peculiar velocities. Thus the Lorentzian noise profile for the peculiar velocities results in a better mocking of a serious uncertainty in the QSOs redshift smear, as expected from Yu et al. (2023).

While the main velocity dispersion is simulated with a physical sense according to Yu et al. (2023), the other configurations have been taken into account for comparison and further characterisation of the effect by varying the width parameter of the noise distribution, doubling or halving it, directly affecting the RSD. This operation significantly modifies the contribution of the peculiar velocities to the RSD as shown in Figure 8. The distortion of the distances in the redshift space due to the peculiar velocities is in the order of $5\,\mathrm{h^{-1}\,Mpc}$, but when the corresponding Lorentzian noise is added it can increase its maximum effect up to the $20\,\mathrm{h^{-1}\,Mpc}$ concerning a larger number of extremal QSOs as shown by the extended width of the distributions with noise. The indicator of the minimum and maximum coordinates reached by the QSOs in the redshift space refers to the coordinates calculated after adding the baseline reported noise shown in Figure 7. As indicated, the periodic conditions of the simulation volume have been correctly taken into account.

Studying now the 2PCF in redshift space for the samples of QSOs with distorted and undisturbed velocities in Figure 11, some interesting insights can be appreciated. First, at small distances dominated by Poisson uncertainty, it is noticeable the presence of more pronounced fluctuations in the correlation of the perturbed QSOs, fluctuations that extend beyond $1\,\mathrm{h^{-1}\,Mpc}$ with diminished Poisson uncertainty. These fluctuations relax until the $10\,\mathrm{h^{-1}\,Mpc}$ large scales, where the three clusterings end up coinciding, although with grater uncertainties. A trend can be identified when studying the response of the 2PCF to an intensification of the redshift smearing uncertainty as the fluctuations develop to larger scales when doing so. Greater peculiar velocities add to grater redshift space distortions and thus to more persistent noise fluctuations with scale in the 2PCF in redshift space.

Meanwhile the quadrupole and the hexadecapole described in Figures 12 and 13 correlations seem to be much more chaotic and sensible to the redshift smearing uncertainty. Although fluctuations are spread all over the studied separation regime, their amplitude seems to diminish from the intermediate distance regime around the $1\,\mathrm{h^{-1}\,Mpc}$ the the extremes and more clearly focus around the undisturbed clustering here as the intensity of the distortion increases. This may be a trace of an homogeneity effect of the redshift smearing noise in the multipole correlations just around the distance scales more affected by the noise addition, as seen in Figure 8.

Finally, the projected correlation in Figure 10 is analysed. It is noticeable at first glance the high agreement in the shape of w_p between the perturbed and unperturbed QSO catalogues at all scales. This is largely due to the smoothing effect on the peculiar noise that has integrating $\xi(r_\mathrm{p}, \pi)$ along the line of sight to compute w_p. However, there is a slightly difference in the clustering power more or less constant at all scales noticeable in the plotted ratios mainly due to the correlation-weakening effect of perturbing the velocities of the QSOs along only one direction, the line of sight, leaving the rest unperturbed. This trend is accentuated with the intensity of the modelled redshift smearing uncertainty. Moreover, the difference between the Gaussian w_{rmp} and the Lorentzian is very low here and almost imperceptible at large scales.

These findings emphasise the importance of properly accounting for uncertainties in redshift measurements and provide information to guide future research as well as a simple method to simulate noise in the observational redshift space. A pending task would be to characterise in more detail the difference between the addition of Gaussian or Lorentzian noise, which could perhaps be achieved by studying other more complex Gaussian distributions as done in Smith et al. (2020) to better exploit the possibilities of this probability profile. Another inviting path would be to try to corroborate the effect of the error modelling in the RSD with observations from large galaxy mappings such as DESI to test its imitation power. In addition, the question remains open about the low correlation of the QSOs with respect to the prediction made by the Kaiser factor for distances between $1\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq s \leq 10\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ in this particular model and simulation as shown in Figure 9.

All in all, the study successfully characterised the given cosmological simulation using the HMF and halo bias, applied a HOD model to generate a catalogue of simulated QSOs, and investigated the impact of a self-made mocked redshift smearing uncertainty model on the clustering and RSD of QSOs by analysing their statistical correlations in redshift space.

6 Bibliography

abacusorg. abacusutils. `https://github.com/abacusorg/abacusutils`, 2023. Documentation available at `https://abacusutils.readthedocs.io/en/latest/`.

N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro, N. Bartolo, S. Basak, et al. Planck 2018 results-vi. cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020.

Y. Ali-Haimoud and C. M. Hirata. Hyrec: A fast and highly accurate primordial hydrogen and helium recombination code. *Physical Review D*, 83(4):043513, 2011.

S. Avila, V. Gonzalez-Perez, F. G. Mohammad, A. de Mattia, C. Zhao, A. Raichoor, A. Tamone, S. Alam, J. Bautista, D. Bianchi, E. Burtin, M. J. Chapman, C. H. Chuang, J. Comparat, K. Dawson, T. Divers, H. du Mas des Bourboux, H. Gil-Marin, E. M. Mueller, S. Habib, K. Heitmann, V. Ruhlmann-Kleider, N. Padilla, W. J. Percival, A. J. Ross, H. J. Seo, D. P. Schneider, and G. Zhao. The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: exploring the halo occupation distribution model for emission line galaxies. , 499(4): 5486–5507, Dec. 2020. doi: 10.1093/mnras/staa2951.

S. Bonoli, F. Shankar, S. D. White, V. Springel, and J. S. B. Wyithe. On merger bias and the clustering of quasars. *Monthly Notices of the Royal Astronomical Society*, 404(1):399–408, 2010.

S. Bose, D. J. Eisenstein, B. Hadzhiyska, L. H. Garrison, and S. Yuan. Constructing high-fidelity halo merger trees in abacussummit. *Monthly Notices of the Royal Astronomical*

Society, 512(1):837–854, 03 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac555. URL https://doi.org/10.1093/mnras/stac555.

A. Cabré and E. Gaztanaga. Clustering of luminous red galaxies–ii. small-scale redshift-space distortions. *Monthly Notices of the Royal Astronomical Society*, 396(2):1119–1131, 2009.

Cosmodesi. cosmoprimo. https://github.com/cosmodesi/cosmoprimo, 2023. Documentation available at https://cosmoprimo.readthedocs.io/en/latest/index.html.

M. Davis, G. Efstathiou, C. S. Frenk, and S. D. White. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 292, May 15, 1985, p. 371-394. Research supported by the Science and Engineering Research Council of England and NASA.*, 292:371–394, 1985.

K. S. Dawson, J.-P. Kneib, W. J. Percival, S. Alam, F. D. Albareti, S. F. Anderson, E. Armengaud, É. Aubourg, S. Bailey, J. E. Bautista, et al. The sdss-iv extended baryon oscillation spectroscopic survey: overview and early data. *The Astronomical Journal*, 151(2):44, 2016.

L. Garrison. Notes on the random-random term in autocorrelations. https://corrfunc.readthedocs.io/en/master/modules/rr_autocorrelations.html#rr-autocorrelations, 2023.

L. H. Garrison, D. J. Eisenstein, D. Ferrer, N. A. Maksimova, and P. A. Pinto. The abacus cosmological N-body code. *Monthly Notices of the Royal Astronomical Society*, 508(1):575–596, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2482. URL https://doi.org/10.1093/mnras/stab2482.

Y. Gong et al. Growth factor parametrization and modified gravity. *Physical Review D*, 78 (12):123010, 2008.

V. Gonzalez-Perez, C. M. Baugh, C. G. Lacey, and J. W. Kim. Massive, red galaxies in a hierarchical universe - II. Clustering of Extremely Red Objects. , 417(1):517–531, Oct. 2011. doi: 10.1111/j.1365-2966.2011.19294.x.

V. Gonzalez-Perez, W. Cui, S. Contreras, C. M. Baugh, J. Comparat, A. J. Griffin, J. Helly, A. Knebe, C. Lacey, and P. Norberg. Do model emission line galaxies live in filaments at z $\sim$ 1? , 498(2):1852–1870, Oct. 2020. doi: 10.1093/mnras/staa2504.

B. Hadzhiyska, D. Eisenstein, S. Bose, L. H. Garrison, and N. Maksimova. CompaSO: A new halo finder for competitive assignment to spherical overdensities. *Monthly Notices of the Royal Astronomical Society*, 10 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2980. URL https://doi.org/10.1093/mnras/stab2980. stab2980.

N. Kaiser. Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society*, 227(1):1–21, 1987.

G. La Mura, G. Busetto, S. Ciroi, P. Rafanelli, M. Berton, E. Congiu, V. Cracco, and M. Frezzato. Relativistic plasmas in agn jets: From synchrotron radiation to γ-ray emission. *The European Physical Journal D*, 71:1–10, 2017.

J. Lesgourgues. The cosmic linear anisotropy solving system (class) i: overview. *arXiv preprint arXiv:1104.2932*, 2011.

M. Levi, C. Bebek, T. Beers, R. Blum, R. Cahn, D. Eisenstein, B. Flaugher, K. Honscheid, R. Kron, O. Lahav, et al. The desi experiment, a whitepaper for snowmass 2013. *arXiv preprint arXiv:1308.0847*, 2013.

B. W. Lyke, A. N. Higley, J. McLane, D. P. Schurhammer, A. D. Myers, A. J. Ross, K. Dawson, S. Chabanier, P. Martini, H. D. M. Des Bourboux, et al. The sloan digital sky survey quasar catalog: Sixteenth data release. *The Astrophysical Journal Supplement Series*, 250(1):8, 2020.

N. A. Maksimova, L. H. Garrison, D. J. Eisenstein, B. Hadzhiyska, S. Bose, and T. P. Satterthwaite. AbacusSummit: a massive set of high-accuracy, high-resolution N-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):4017–4037, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2484. URL `https://doi.org/10.1093/mnras/stab2484`.

F. G. Mohammad and W. J. Percival. Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function. *Monthly Notices of the Royal Astronomical Society*, 514(1):1289–1301, 2022. As implemented by Cosmodesi in `https://github.com/cosmodesi/pycorr`. Documentation available at `https://py2pcf.readthedocs.io/en/latest/user/building.html`.

G. Sato-Polito, A. D. Montero-Dorta, L. R. Abramo, F. Prada, and A. Klypin. The dependence of halo bias on age, concentration, and spin. *Monthly Notices of the Royal Astronomical Society*, 487(2):1570–1579, 2019.

M. Sinha and L. Garrison. Corrfunc: Blazing fast correlation functions with avx512f simd intrinsics. In A. Majumdar and R. Arora, editors, *Software Challenges to Exascale Computing*, pages 3–20, Singapore, 2019. Springer Singapore. ISBN 978-981-13-7729-7. URL `https://doi.org/10.1007/978-981-13-7729-7_1`.

M. Sinha and L. H. Garrison. CORRFUNC - a suite of blazing fast correlation functions on the CPU. , 491(2):3022–3041, Jan 2020. doi: 10.1093/mnras/stz3157.

A. Smith, E. Burtin, J. Hou, R. Neveux, A. J. Ross, S. Alam, J. Brinkmann, K. S. Dawson, S. Habib, K. Heitmann, et al. The completed sdss-iv extended baryon oscillation spectroscopic survey: N-body mock challenge for the quasar sample. *Monthly Notices of the Royal Astronomical Society*, 499(1):269–291, 2020.

R. H. Wechsler and J. L. Tinker. The connection between galaxies and their dark matter halos. *Annual Review of Astronomy and Astrophysics*, 56:435–487, 2018.

J. Yu, C. Zhao, V. Gonzalez-Perez, C.-H. Chuang, A. Brodzeller, A. de Mattia, J.-P. Kneib, A. Krolewski, A. Rocher, A. Ross, et al. The desi one-percent survey: Exploring a generalized sham for multiple tracers with the unit simulation. *arXiv preprint arXiv:2306.06313*, 2023.

S. Yuan, L. H. Garrison, B. Hadzhiyska, S. Bose, and D. J. Eisenstein. AbacusHOD: a highly efficient extended multitracer HOD framework and its application to BOSS and eBOSS data. *Monthly Notices of the Royal Astronomical Society*, 510(3):3301–3320, 11 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab3355. URL `https://doi.org/10.1093/mnras/stab3355`.

S. Yuan, H. Zhang, A. J. Ross, J. Donald-McCann, B. Hadzhiyska, R. H. Wechsler, Z. Zheng, S. Alam, V. G. Perez, J. N. Aguilar, et al. The desi one-percent survey: Exploring the halo occupation distribution of luminous red galaxies and quasi-stellar objects with abacussummit. *arXiv preprint arXiv:2306.06314*, 2023.

Z. Zheng, A. L. Coil, and I. Zehavi. Galaxy evolution from halo occupation distribution modeling of deep2 and sdss galaxy clustering. *The Astrophysical Journal*, 667(2):760, 2007.

Appendix

An essential part of this work was to implement the code for the redshift smearing uncertainty model, the statistical correlation analysis, the graphics and much more details. All this can be accessed in the following Jupyter Notebook.