Cosmological analysis of the DESI data to constrain the nature of dark energy and general relativity

Svyatoslav Trusov

Sorbonne Université
École Doctorale 560 STEP'UP
LPNHE

Exploitation des données du DESI Bright Galaxy Survey pour contraindre la nature de l'énergie noire et la relativité générale

Par Svyatoslav Trusov

Thèse de doctorat de Physique

Présentée et soutenue publiquement le 19/09/2024

Devant un jury composé de :

Christophe Balland	Directeur de thèse
Pauline Zarrouk	Directrice de thèse
Alice Pisani	Rapportrice
Jean-Paul Kneib	Rapporteur
Etienne Burtin	Examinateur
Delphine Hardin	Examinatrice
Shaun Cole	Invité

Contents

Résumé substantiel (Français)

> Si vous n'aimez pas la mer, si vous n'aimez pas la montagne, si vous n'aimez pas la ville...
>
> ---
>
> À bout de souffle, 1960

Dans les années 1990, une découverte fondamentale est faite grâce aux observations de la luminosité des Supernovæ de type Ia en fonction de leur distance : non seulement l'Univers est en expansion, mais cette expansion s'accélère les derniers 6 milliards d'années quand on estime l'âge de l'Univers à 13,8 milliards d'années. Cette expansion accélérée peut être décrite par une constante cosmologique, Λ, dans les équations de la Relativité Générale décrivant la dynamique de l'Univers. Toutefois, la nature de cette expansion accélérée est toujours un mystère et on appelle "énergie noire", le constituant inconnu responsable de cette accélération qui représente presque 70% du contenu énergétique actuel de l'Univers. Cette découverte a déclenché plusieurs générations de vastes programmes d'observation du ciel afin de sonder la nature de l'énergie noire, mais aussi de tester la validité de la Relativité Générale elle-même comme théorie de la gravitation aux échelles cosmologiques.

DESI (Dark Energy Spectroscopic Instrument, i.e. l'Instrument Spectroscopique de l'Énergie Noire) est un instrument spectroscopique attaché à un téléscope de 4m à l'observatoire Kitt Peak aux États-Unis. En collectant les spectres de plus de 40 millions de galaxies, on peut calculer leur distance radiale et ainsi créer une carte tri-dimensionnelle des grandes structures de notre Univers. Puisque la gravité est la force dominante à ces échelles, en mesurant la vitesse à laquelle la matière s'agglomère pour former des structures comme les galaxies, on peut contraindre les théories de la gravité. En même temps, en regardant assez loin dans l'Univers, on peut voir son passé distant, et cela nous permet de retracer l'évolution et l'histoire de l'expansion de l'Univers. Pour cela, on cherche à décrire les corrélations spatiales entre galaxies aux moyens d'outils statistiques comme la fonction de corrélation à deux points ou le spectre de puissance, qui sont reliés entre eux par une transformation de Fourier. Ces statistiques sont souvent appelées les statistiques d'agglomération, ou clustering statistics en anglais, d'où vient aussi le nom

d'analyse: analyse clustering. Grâce au développement de la théorie des perturbations cosmologiques, on sait prédire analytiquement ces quantités en fonction de paramètres cosmologiques que nous ajustons sur les données pour les déterminer. Dans le cadre de cette thèse, on s'intéresse en particulier autaux de croissance des structures, une quantité prédite par la théorie de la gravité et qui est directement liée à la nature de l'énergie noire.

DESI observe 4 types de traceurs: Bright Galaxy Survey ou BGS (le relevé de galaxies brillantes), Luminous Red Galaxies ou LRG (les galaxies lumineuses rouges), Emission Line Galaxies ou ELG (les galaxies à raie d'émission) et Quasi-Stellar Objects ou QSO (les objets quasi-stellaires). Cette thèse se concentre sur le relevé de galaxies brillantes (BGS), qui est l'échantillon de données le plus dense de DESI et qui est composé des galaxies les plus proches de nous ($z < 0.4$) et lesplus lumineuses (avec une coupure en magnitude apparente dans la bande r à $r < 19.5$).

Un problème avec une densité de galaxies aussi élevée à bas redshift est que l'information qu'on peut extraire d'un volume d'espace limité est aussi très limitée. Cela veut dire que les contraintes sur les paramètres cosmologiques ne sont pas dominées par la statistique de l'échantillon mais par une limite fondamentale liée à la quantité d'information accessible appelée 'variance cosmique'. Et pourtant, il y a une technique qui permet de contourner cet obstacle. On appelle cette technique l'analyse multi-traceur (ou "multi-tracer analysis" en anglais). Elle consiste à diviser le catalogue de données en deux (ou plusieurs) sous-catalogues différents avec des propriétés de clustering différentes et de prendre en compte les corrélations spatiales croisées entre les différents traceurs qui ajoutent une information supplémentaire, et permettent ainsi d'améliorer les contraintes sur certains paramètres cosmologiques.

Un autre problème qui se pose à cause de la densité élevée du BGS est celui de l'estimation des erreurs de mesure au moyen d'une matrice de covariance. Habituellement, pour estimer la matrice de covariance, on crée des milliers de simulations cosmologiques qui doivent imiter la relevé, en créant ainsi plusieurs réalisations de différents univers observables et en nous permettant d'estimer les erreurs sur la mesure de statistiques de 2-point. Pour le BGS, il faut produire des simulations avec une très bonne résolution en masse couvrant un volume cosmologique jusqu'à $z = 0.4$ suffisant pour inclure tout le relevé. Produire ce millier de simulations est très coûteux en temps de calcul et en mémoire. A ce jour, il n'en existe que 25, ce qui est insuffisant pour obtenir une matrice de covariance suffisamment précise. S'offrent à nous au moins deux options: (a) reduire les données considérées dans l'analyse (ce que la collaboration DESI a choisi de faire pour l'analyse des premières données collectées pendant 1 an); (b) trouver un moyen d'obtenir la matrice de covariance avec moins de simulation numériques. Dans cette thèse, nous proposons une nouvelle méthode, appelée FitCov, qui est une combinaison hybride entre la méthode d'estimation de covariance "jackknife" basée sur un échantillonnage des données, et la méthode classique basée sur des simulations.

Un troisième problème que nous avons rencontré au cours de cette thèse concerne plus généralement les analyses clustering. Dans l'approche standard, on compresse l'information sur les paramètres cosmologiques contenue dans la statistique à deux-points sous la forme de paramètres intermédiaires tels que le taux de croissance des structures déjà mentionné et les paramètres géométriques d'Alcock-Paczynski qui permettent de mesurer le taux d'expansion de l'univers par exemple. Cette compression nous fait perdre de l'information cosmologique lorsque l'analyse des grandes structures sondant l'Univers récent n'est pas combinée avec celle du fond diffus cosmologique qui sonde l'Univers primordial, mais elle accélère significativement la détermination des paramètres cosmologiques. Dans cette thèse, nous présentons un moyen d'accélerer l'inférence des paramètres cosmologiques directement à partir de la statistique à deux points, sans passer par les paramètres compressés. Pour cela, nous avons développé un réseau de neurones qui remplace la partie du modèle analytique qui prend beaucoup de temps à prédire l'évolution non-linéaire de la statistique à deux-points et nous avons montré que notre modèle hybride est aussi précis que le modèle analytique.

La structure de ce manuscrit de thèse est décrite ci-dessous. Dans l'introduction, je présente les bases de la cosmologie et comment extraire l'information cosmologique à partir des statistiques à deux-points des cartes tri-dimensionnelles de DESI.

Chapitre 1 est dédié à la présentation de l'instrument DESI et des données utilisées par la suite dans ce travail de thèse.

Dans le 2ème Chapitre, j'approfondis l'introduction à la théorie des perturbations, et je présente l'approche pour modéliser les statistiques à 2-points avec un réseau de neurones, que j'ai developpée.

Chapitre 3 décrit les méthodes de simulations numériques existantes pour simuler un relevé spectroscopique de galaxies. C'est aussi dans ce chapitre que je décris l'effort que j'ai mené afin de générer des simulations du BGS pour obtenir une matrice de covariance dans le cadre de l'analyse officielle avec un échantillon réduit.

Chapitre 4 détaille les moyens d'estimer la matrice de covariance pour une analyse clustering, et c'est aussi dans ce chapitre que je présente FitCov, la méthode hybride que j'ai développée pour l'estimation de covariance nécéssitant beaucoup moins de simulations.

Chapitre 5 présente le travail de thèse mené sur l'inférence cosmologique et les résultats obtenus avec les données 1 an duBGS en utilisant les techniques et les outils conçus pendant ma thèse. Je montre comment, au final, ils permettent d'augmenter la précision de près de 40% par rapport à l'analyse standard.

Enfin, dans la conclusion je décris quelques interprétations cosmologiques préliminaires que j'ai faites avec les résultats du Chapitre 5 et je termine avec des pistes d'amélioration et de nouvelles applications pour de futurs travaux sur le domaine.

Résumé court (Français)

DESI (Dark Energy Spectroscopic Instrument, i.e. l'Instrument Spectroscopique de l'Énergie Noire) est un instrument spectroscopique attaché à un téléscope de 4m à l'observatoire Kitt Peak aux États-Unis. En collectant les spectres de plus de 40 millions de galaxies, on peut calculer leur distance radiale et ainsi créer une carte tri-dimensionnelle des grandes structures de notre Univers. Puisque la gravité est la force dominante à ces échelles, en mesurant la vitesse à laquelle la matière s'agglomère pour former des structures comme les galaxies, on peut contraindre les théories de la gravité. Cette thèse se concentre sur le relevé de galaxies brillantes (BGS), qui est l'échantillon de données le plus dense de DESI et qui est composé des galaxies les plus proches de nous ($z < 0.4$) et les plus lumineuses (avec une coupure en magnitude apparente dans la bande r à $r < 19.5$). Un problème avec une densité de galaxies aussi élevée à bas redshift est que l'information qu'on peut extraire d'un volume d'espace limité est aussi très limitée. Cela veut dire que les contraintes sur les paramètres cosmologiques ne sont pas dominées par la statistique de l'échantillon mais par une limite fondamentale liée à la quantité d'information accessible appelée 'variance cosmique'. Et pourtant, il y a une technique qui permet de contourner cet obstacle. On appelle cette technique l'analyse multi-traceur (ou "multi-tracer analysis" en anglais). Elle consiste à diviser le catalogue de données en deux (ou plusieurs) sous-catalogues différents avec des propriétés de clustering différentes et de prendre en compte les corrélations spatiales croisées entre les différents traceurs qui ajoutent une information supplémentaire, et permettent ainsi d'améliorer les contraintes sur certains paramètres cosmologiques. Un autre problème qui se pose à cause de la densité élevée du BGS est celui de l'estimation des erreurs de mesure au moyen d'une matrice de covariance. Habituellement, pour estimer la matrice de covariance, on crée des milliers de simulations cosmologiques qui doivent imiter la relevé, en créant ainsi plusieurs réalisations de différents univers observables et en nous permettant d'estimer les erreurs sur la mesure de statistiques de 2-point. Produire ce millier de simulations est très coûteux en temps de calcul et en mémoire. A ce jour, il n'en existe que 25, ce qui est insuffisant pour obtenir une matrice de covariance suffisamment précise. Dans cette thèse, nous proposons une nouvelle méthode, appelée FitCov, qui est une combinaison hybride entre la méthode d'estimation de covariance "jackknife" basée sur un échantillonnage des données, et la méthode classique basée sur des simulations. Un troisième problème que nous avons rencontré au cours de cette

thèse concerne plus généralement les analyses clustering. Dans l'approche standard, on compresse l'information sur les paramètres cosmologiques contenue dans la statistique à deux-points sous la forme de paramètres intermédiaires tels que le taux de croissance des structures déjà mentionné et les paramètres géométriques d'Alcock-Paczynski qui permettent de mesurer le taux d'expansion de l'univers par exemple. Cette compression nous fait perdre de l'information cosmologique lorsque l'analyse des grandes structures, mais elle accélère significativement la détermination des paramètres cosmologiques. Dans cette thèse, nous présentons un moyen d'accélérer l'inférence des paramètres cosmologiques directement à partir de la statistique à deux points, sans passer par les paramètres compressés. Pour cela, nous avons développé un réseau de neurones qui remplace la partie du modèle analytique qui prend beaucoup de temps à prédire l'évolution non-linéaire de la statistique à deux-points et nous avons montré que notre modèle hybride est aussi précis que le modèle analytique.

Résumé substantiel (English)

In the 1990s, a fundamental discovery was made thanks to observations of the luminosity of type Ia supernovae as a function of their distance: not only is the Universe expanding, but this expansion is accelerating over the last 6 billion years, when the age of the Universe is estimated at 13.8 billion years.

This accelerated expansion can be described by a cosmological constant, Λ, in the equations of General Relativity describing the dynamics of the Universe. However, the nature of this accelerated expansion is still a mystery, and the unknown constituent responsible for this acceleration, which accounts for almost 70% of the Universe's current energy content, is known as 'dark energy'. This discovery triggered several generations of large-scale sky surveys to probe the nature of dark energy, but also to test the validity of General Relativity itself as a theory of gravitation at cosmological scales.

DESI (Dark Energy Spectroscopic Instrument) is a spectroscopic instrument attached to a 4m telescope at Kitt Peak Observatory in the United States. By collecting the spectra of more than 40 million galaxies, we can calculate their radial distances and thus create a three-dimensional map of the large-scale structures of our Universe. Since gravity is the dominant force at these scales, by measuring the speed at which matter clumps together to form structures such as galaxies, we can constrain theories of gravity. At the same time, by looking far enough into the Universe, we can see its distant past, enabling us to trace the evolution and history of the expansion of the Universe. To do this, we try to describe the spatial correlations between galaxies using statistical tools such as the two-point correlation function or the power spectrum, which are linked together by a Fourier transform. These statistics are often referred to as clustering statistics, hence the name of the analysis: clustering analysis. Thanks to the development of cosmological perturbation theory, we know how to predict these quantities analytically as a function of cosmological parameters that we fit to the data in order to determine them. In this thesis, we are particularly interested in the growth rate of structures, a quantity predicted by the theory of gravity and which is directly linked to the nature of dark energy.

DESI observes 4 types of tracers: Bright Galaxy Survey (BGS), Luminous Red Galaxies (LRG), Emission Line Galaxies (ELG) and Quasi-Stellar Objects (QSO). This thesis focuses on the Bright Galaxy Survey (BGS), which is DESI's densest data sample and is composed of the galaxies closest to us ($z < 0.4$) and the most luminous (with an

apparent magnitude cut-off in the r band at $r < 19.5$).

One problem with such a high density of galaxies at low redshift is that the information that can be extracted from a limited volume of space is also very limited. This means that the constraints on the cosmological parameters are not dominated by the statistics of the sample but by a fundamental limit linked to the amount of accessible information called the 'cosmic variance'. However, there is a technique that allows us to bypass this limitation. This technique is known as multi-tracer analysis. It consists of dividing the data catalogue into two (or more) different sub-catalogues with different clustering properties and taking into account the spatial correlations between the different tracers, which add extra information and thus make it possible to improve the constraints on certain cosmological parameters.

Another problem that arises because of the high density of the BGS is that of estimating measurement errors by means of a covariance matrix. Usually, to estimate the covariance matrix, we create thousands of cosmological simulations that must mimic the survey, thus creating several realisations of different observable universes and allowing us to estimate the errors in the measurement of 2-point statistics. For the BGS, we need to produce simulations with very good mass resolution covering a cosmological volume up to $z = 0.4$ sufficient to include the entire survey. Producing thousands of simulations is very costly in terms of computing time and memory. To date, only 25 of such simulations exist, which is not enough to obtain a sufficiently accurate covariance matrix. At least two options are open to us: (a) reduce the data considered in the analysis (which the DESI collaboration chose to do for the analysis of the first data collected over 1 year); (b) find a way of obtaining the covariance matrix with fewer numerical simulations. In this thesis, we propose a new method, called FitCov, which is a hybrid combination of the 'jackknife' covariance estimation method based on data resampling, and the classical method based on simulations.

A third problem we encountered during the course of this thesis relates more generally to clustering analyses. In the standard approach, the information on the cosmological parameters contained in the two-point statistic is compressed in the form of intermediate parameters such as the growth rate of the structures already mentioned and the Alcock-Paczynski geometric parameters that allow us to measure the expansion rate of the universe, for example. This compression means that we lose cosmological information when the analysis of the large structures probing the recent Universe is not combined with that of the cosmic microwave background probing the primordial Universe, but it significantly speeds up the determination of cosmological parameters. In this thesis, we present a way of accelerating the inference of cosmological parameters directly from the two-point statistic, without going through the compressed parameters. To this end, we have developed a neural network that replaces the part of the analytical model that takes a long time to predict the non-linear evolution of the two-point statistic, and we have shown that our hybrid model

is as accurate as the analytical model.

The structure of this thesis manuscript is described below. In the introduction, I present the basics of cosmology and how to extract cosmological information from the two-point statistics of three-dimensional DESI maps.

Chapter 1 is devoted to the presentation of the DESI instrument and the data used subsequently in this thesis work.

In Chapter 2, I give a more in-depth introduction to perturbation theory, and present the approach for modelling the 2-point statistics using a neural network, which I have developed.

Chapter 3 describes existing numerical simulation methods for simulating a spectroscopic survey of galaxies. It is also in this chapter that I describe the effort I made to generate BGS simulations to obtain a covariance matrix for the more traditional analysis with a reduced sample.

Chapter 4 details the means of estimating the covariance matrix for a clustering analysis, and it is also in this chapter that I present FitCov, the hybrid method I developed for covariance estimation requiring far fewer simulations.

Chapter 5 presents the thesis work on cosmological inference and the results obtained with 1-year BGS data using the techniques and tools developed during my thesis. I show how, in the end, they enable us to increase the accuracy by almost 40% with respect to the official DESI standard analysis.

Finally, in the conclusion I describe some preliminary cosmological interpretations that I have made with the results of Chapter 5 and I end with some ideas for improvement and new applications for future work in the field. .

Résumé court (Anglais)

DESI (Dark Energy Spectroscopic Instrument) is a spectroscopic instrument installed on a 4m telescope at Kitt Peak Observatory in the United States. By collecting the spectra of more than 40 million galaxies, we can calculate their radial distances and thus create a three-dimensional map of the large-scale structures of the Universe. Since gravity is the dominant force at these scales, by measuring the speed at which matter clumps together to form structures such as galaxies, we can constrain theories of gravity. This thesis focuses on the Bright Galaxy Survey (BGS), which is DESI's densest data sample and is composed of the galaxies closest to us ($z < 0.4$) and the most luminous (with an apparent magnitude cut-off in the r band at $r < 19.5$). One problem with such a high density of galaxies at low redshift is that the information that can be extracted from a limited volume of space is also very limited. This means that the constraints on the cosmological parameters are not dominated by the statistics of the sample but by a fundamental limit linked to the amount of accessible information called the 'cosmic variance'. However, there is a technique that allows us to bypass it. This technique is known as multi-tracer analysis. It consists of dividing the data catalogue into two (or more) different sub-catalogues with different clustering properties and taking into account the spatial correlations between the different tracers, which add extra information and thus make it possible to improve the constraints on certain cosmological parameters. Another problem that arises because of the high density of the BGS is related to the estimation of measurement errors by means of a covariance matrix. Usually, to estimate the covariance matrix, we create thousands of cosmological simulations that must mimic the survey, thus creating several realisations of different observable universes and allowing us to estimate the errors in the measurement of 2-point statistics. Producing this thousand simulations is very costly in terms of computing time and memory. To date, only 25 such simulations exist, which is insufficient to obtain a sufficiently accurate covariance matrix. In this thesis, we propose a new method, called FitCov, which is a hybrid combination of the jackknife covariance estimation method based on data resampling and the classical method based on simulations. A third problem that we attempt to tackle in this thesis is related the most optimal way of extracting information from a two-point galaxy clustering analysis. In the standard approach, the information on the cosmological parameters contained in the two-point statistics is compressed in the form of intermediate parameters such as the growth

rate of the structures that allows us to test gravity and the Alcock-Paczynski geometric parameters that allow us to measure the expansion rate of the universe, for example. This compression means that we lose cosmological information when analysing large scale structures without combining with cosmic microwave background, but it significantly speeds up the determination of cosmological parameters. In this thesis, we present a way of accelerating the inference of cosmological parameters directly from the two-point statistics, without going through the compressed parameters. To this end, we have developed a neural network model that replaces the part of the analytical model that takes a long time to predict the non-linear evolution of the two-point statistics, and we have shown that our hybrid model is as accurate as the analytical model. Eventually, we put both the hybrid covariance method and the neural network model that we have developed together to analyse the full DESI BGS Bright sample from 1 year of observation (BGS DR1). We perform both single- and multi-tracer analyses and we show that we can improve the cosmological constraints of the BGS DR1 sample by 40% with respect to the official DESI analysis.

List of Figures

List of Tables

Introduction

Полночных солнц к себе нас
манят светы... В колодцах
труб пытливый тонет взгляд.
Алмазный бег вселенные
стремят:
Системы звёзд, туманности,
планеты,

M. Voloshin, Translation

The study of the Universe as a whole from a physical perspective is a surprisingly new discipline. Evading mentioning myths of elephants and turtles, we will reserve ourselves to give an introduction to the basics of modern cosmology, focusing only on the most crucial components that are directly related to this PhD work.

General Relativity

We shall start this introduction by noticing a simple fact: according to modern beliefs, the Universe on large scales is dominated by gravity. Electromagnetic interaction is weak, as the Universe is on average electrically neutral and weak and strong forces are dominant at much smaller scales.

Therefore, we will shortly describe the most tested theory of gravity to this day: General Relativity. The basis for it is the common description of acceleration and gravitation. A good example of such a principle called the equivalency principle is a simple lift, in which, when moving upwards, we are getting ourselves heavier, and vice versa. From here we jump to the mathematical description of gravity by assuming that gravity is not a classical force but rather an effect caused by the deformation of space-time, such that an object standing still on Earth would feel the same effect as if it was moving up in a lift.

Now that we have decided on that, we can attempt to describe the curved space-time and its reaction to the presence of matter. For that, we will need a wide range of tools and concepts from differential geometry. Describing them here would make this section too lengthy and this is beyond the scope of this introduction chapter, so instead

we will direct you towards the wonderful book of [1]. As we are working with a physical system with infinite degrees of freedom, we will also employ the rich framework of theoretical mechanics [2] and classical field theory [3]. We will assume that the concepts of covariant derivative ∇_μ, Riemann tensor $R^\mu_{\nu\kappa\lambda}$, metric space, Ricci tensor $R_{\mu\nu}$ and scalar R, Christoffel symbols $\Gamma^\mu_{\nu\lambda}$ and minimal action principle are familiar to the reader, as we will use those intensively in our presentation of Einstein's General Relativity equations. Otherwise, you can also go directly to the Friedmann equations (equations 15 and 16).

Einstein-Hilbert action

Let us assume a physical system with an action S. This system consists of some very generalised purely physical content with action S_m, and the geometrical part with action S_g, both parts comprising an overall action $S = S_g + S_m$.

We will try to postulate the form of S_g. First of all, following Ostrograsky's theorem [4], we need to ensure that the resulting equations of motion, do not contain second derivatives. Secondly, we need to ensure that the assumed symmetries are preserved, and corresponding transformations do not change the form of the action. Thirdly, we need to choose which quantity we treat as a field. We will answer these questions starting from the end.

As a field, we choose the metric $g_{\mu\nu}$, describing the metric 4-space, where we have merged the usual 3d space and time into one 4d manifold. We can also characterise the metric by writing the interval as:

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{1}$$

Where dx^μ is a coordinate interval.

The traditional Minkowski metric space can be then characterised by $ds^2 = -dt^2 + \sum_{i=0}^{3} dx_i^2$ where dt is a time interval, and dx_i is a spatial interval. Therefore, the indices μ and ν can take values from 0 to 3. Remembering that for metric spaces any infinitesimal local region can be presented under some transformation as a space with Euclidian or pseudo-Euclidian local metric, we choose a local $SO(1,3)$[1] symmetry, imposing special relativity on any such region, however, it should be noted, that in general Einstein's equations can be (and sometimes actively are) applied also to metrics with other local symmetries. This choice of ours will not have any effect on the derivation of the action but will restrict us to the specific choices of ansatzes used when trying to get the answer for the metrics.

As for the symmetries, as we work with the non-flat non-homogeneous and non-isotropic space, as well as we might want to describe the behaviour of the system for all possible observers, not just the inertial ones, we will need the action to be invariant with

[1] $SO(1,3)$ is a group of orthogonal matrices leaving the Minkowski interval ds^2

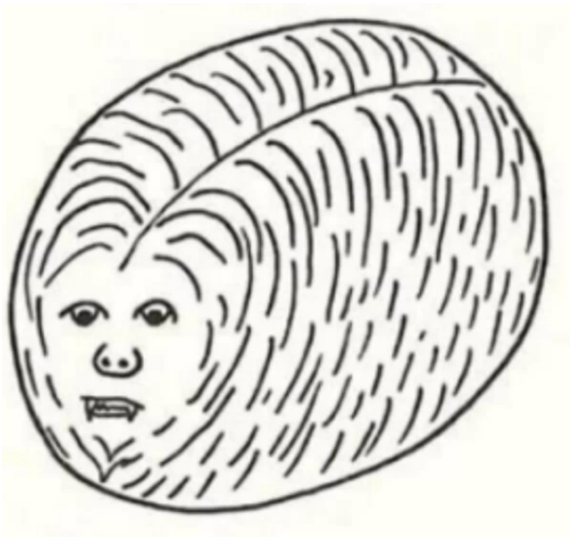

Figure 1: An illustration of the dog being smoothly transformed into a spherical shape. Taken from[5].

respects to diffeomorphisms, a set of all possible smooth transformations of the space (an example of such an action applied to a dog, whilst neglecting its internal structure and assuming all the holes to be closely shut, can be seen on Figure 1). We will then need to choose between different topological invariants as the basis for the action.

We need to also ensure that the volumes will be properly transformed under the diffeomorphisms, therefore we need to adjust the measure in each of the integrals $\int d^4x \rightarrow \int d^4x\sqrt{-g}$, where $g = det(g_{\mu\nu})$.

Following classic Hilbert's suggestion, we will choose Ricci scalar R, or to be more precise, its integral over the whole given space-time as our action. Defining the Lagrangian for matter as $\mathcal{L}_m$, and adding a constant for the gravitational action, we can write:

$$S = \frac{c^4}{8\pi G}\int d^4x\sqrt{-g}(R - 2\Lambda) + \int d^4x\sqrt{-g}\mathcal{L}_m \tag{2}$$

Where c is the speed of light and G is the gravitational constant, Λ is a constant, which we add now for generality, but which will become important as we proceed to solving the equations of motion. Note how gravity does not couple to the matter directly but rather through the $\sqrt{-g}$ term, responsible for the volume changes.

As it will be shown later, the given action does not yield any higher-order derivatives in the equations of motion.

It can be also shown, that any other possible combination of purely geometrical terms in the action will be equivalent [6].

Einstein equations

Following the minimal action principle, we need to require the variation of the action to be 0, to be able to find its minima $\delta S = 0$. First, we will vary the gravitational part of the action:

$$\frac{\delta S_g}{\delta g_{\mu\nu}} = \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}\left[\frac{\delta R}{\delta g^{\mu\nu}} + \frac{R-2\Lambda}{\sqrt{-g}}\frac{\delta\sqrt{-g}}{\delta g^{\mu\nu}}\right] \tag{3}$$

Varying the Ricci scalar might be the most complicated part of the derivation.

We start by reminding that $R = g^{\mu\nu}R_{\mu\nu}$.

We will then remind, that even though connection Γ is not a tensor on its own, the difference of two Γ's is one (the proof is trivial). This gives access to Palatini identity:

$$\delta R_{\mu\nu} = \nabla_\rho\left(\delta\Gamma^\rho_{\nu\mu}\right) - \nabla_\nu\left(\delta\Gamma^\rho_{\rho\mu}\right) \tag{4}$$

We will then note a useful property of covariant derivatives when multiplied by $\sqrt{-g}$:

$$\sqrt{-g}\nabla_\mu A^\mu = \partial_\mu\left(\sqrt{-g}A^\mu\right) \tag{5}$$

That will imply, that the terms containing Γ will simply disappear from the action being full divergences.

Last problem we need to overcome is the variation of the $\sqrt{-g}$. Using the Jacobi formula $\delta\sqrt{-g} = gg^{\mu\nu}\delta g_{\mu\nu}$

Then we get:

$$\delta\sqrt{-g} = -\frac{1}{2}\sqrt{-g}g_{\mu\nu}\delta g^{\mu\nu} \tag{6}$$

We have every component now to expand the Eq. 3 as follows:

$$\begin{aligned}\frac{\delta S_g}{\delta g_{\mu\nu}} &= \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}\left[R_{\mu\nu} - \frac{1}{2}(R-2\Lambda)g_{\mu\nu} + g^{\kappa\lambda}\left(\nabla_\rho\frac{\delta\Gamma^\rho_{\lambda\kappa}}{\delta g^{\mu\nu}} - \nabla_\kappa\frac{\delta\Gamma^\rho_{\rho\lambda}}{\delta g^{\mu\nu}}\right)\right] = \\ &= \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}\left[R_{\mu\nu} - \frac{1}{2}(R-2\Lambda)g_{\mu\nu} + \left(\nabla_\rho\frac{\delta\Gamma^{\rho\lambda}_{\lambda}}{\delta g^{\mu\nu}} - \nabla^\lambda\frac{\delta\Gamma^\rho_{\rho\lambda}}{\delta g^{\mu\nu}}\right)\right] = \\ &= \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}\left[R_{\mu\nu} - \frac{1}{2}(R-2\Lambda)g_{\mu\nu}\right]\end{aligned} \tag{7}$$

Remembering the classical definition of the momentum-energy tensor as:

$$T_{\mu\nu} = \frac{-2}{\sqrt{-g}}\frac{\delta(\sqrt{-g}\mathcal{L}_m)}{\delta g^{\mu\nu}} \tag{8}$$

We immediately notice that it will be the result of varying the matter density.

$$\frac{\delta S_m}{\delta g^{\mu\nu}} = \int d^4x\frac{\delta\sqrt{-g}\mathcal{L}_m}{\delta g^{\mu\nu}} = -\frac{1}{2}\int d^4x\delta\sqrt{-g}T_{\mu\nu} \tag{9}$$

Now we can gather our equations of motion:

$$0 = \frac{\delta S}{\delta g^{\mu\nu}} = \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}\left[R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} - \frac{8\pi G}{c^4}T_{\mu\nu}\right] \tag{10}$$

In the more common form these equations of motions are known as Einstein's equations, and are the main governing equations of the General Relativity:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu} \tag{11}$$

What we see here is that the geometric configuration of the space-time is intricately connected to the configuration of energy-matter inside. It can be shown, that as long as no non-trivial degrees of freedom are present, the Einstein equations are guaranteed to appear no matter the configuration of the geometrical action [6].

Friedman equations

The Einstein equations give us a framework to study the gravitational interactions, with respect to the matter-energy content. When applying it to the study of the Universe on large scales, we make several "common sense" assumptions.

First of all: the Universe is homogeneous. Of course, galaxies and stars exist, but if we look at very large scales, they become nothing but cosmic dust, similar to how tiny particles form the liquid.

Secondly, we assume that the Universe is isotropic, meaning that no matter the direction we are looking at, the laws of physics are the same.

Therefore, we need a metric, which will solve Einstein's equations, which depends on time only and not on spatial coordinates and which satisfies global $SO(3)$ (rotational) symmetry. It thus makes sense to work in the spherical coordinates for the spatial part and it gives the classical Friedman-Lemaitre-Robertson-Walker (FLRW) ansatz[7] for ds^2 which is defined by:

$$ds^2 = -dt^2 + a^2(t)\left(\sum \frac{1}{1-\frac{r^2}{K^2}}dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\phi^2\right) \tag{12}$$

where K is the curvature of the corresponding spatial hypersurface, $K = 0$ corresponds to a flat Universe, while $K > 0$ and $K < 0$ stand for spherical (closed Universe) and hyperbolic (open Universe) spaces. $a(t)$ is the scale factor, and we put it to be $a(t_{\text{now}}) = 1$, which will serve nicely as a boundary condition when we will solve the corresponding equations.

The coordinates in which the FLRW-metric takes form are called the comoving coordinates and correspond to:

$$ds^2 = -a^2(\tau)\left[d\tau^2 + \left(\sum \frac{1}{1-\frac{r^2}{K^2}}dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\phi^2\right)\right] \tag{13}$$

Now we need to parameterise the energy content. We will assume the content of the Universe to be a perfect fluid, therefore giving us the following expression for the energy-momentum tensor $T_{\mu\nu}$:

$$T_{\mu\nu} = (\rho(t) + p(t))u_\mu u_\nu + p g_{\mu\nu} \tag{14}$$

where u_μ is defined as the macroscopic speed of the medium, $\rho(t)$ is the energy density, and p is the isotropic pressure of such a perfect fluid.

Being generous to the reader, we allow them to perform the process of solving the Einstein's equations with the above ansatzes themselves, and we will not spoil this ~~joyful~~ experience.

In the end, the system is reduced to two equations, coming from the tt and rr components of Einstein's equations, where we see that the scale factor of the Universe is defined by the energy density in the Universe:

$$\left(\frac{\dot{a}(t)}{a(t)}\right)^2 = \frac{8\pi G}{3}\rho(t) + \frac{\Lambda}{3} - \frac{1}{K^2 a^2(t)} \tag{15}$$

$$\frac{\ddot{a}(t)}{a(t)} = -\frac{4\pi G}{3}(\rho(t) + 3p) + \frac{\Lambda}{3} \tag{16}$$

We should also mention, that often ρ_{species} is presented in the dimensionless form of Ω_{species}, which is defined as:

$$\Omega_{\text{species}} = \frac{\rho_{\text{species}}}{\rho_{\text{critical}}} \tag{17}$$

where the critical overdensity corresponds to the overall energy density of the flat Universe for a given and is thus defined by:

$$\rho_{\text{critical}} = \frac{3\dot{a}^2(t)}{8\pi G a^2(t)} \tag{18}$$

Equations 15 and 16 are called the Friedman equations, they govern the evolution of the isotropic homogeneous Universe[8]. We can accompany these equations with an equation of continuity for a perfect fluid, obtained by covariantly derivating the energy-momentum tensor such that it gives:

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + p) = 0 \tag{19}$$

If we assume a constant equation of state of the form $w = \frac{p}{\rho}$, we obtain the following relation between the scale factor $a(t)$ and the energy density $\rho(t)$:

$$\rho(t) \propto a(t)^{-3(1+w)} \tag{20}$$

We can further modify this equation by adding several fluids, representing different particle types in the Universe. For now, let us assume that the Universe is full of non-relativistic ($w = 0$) matter. Under the assumption of flatness $K = 0$, and the absence of Λ,

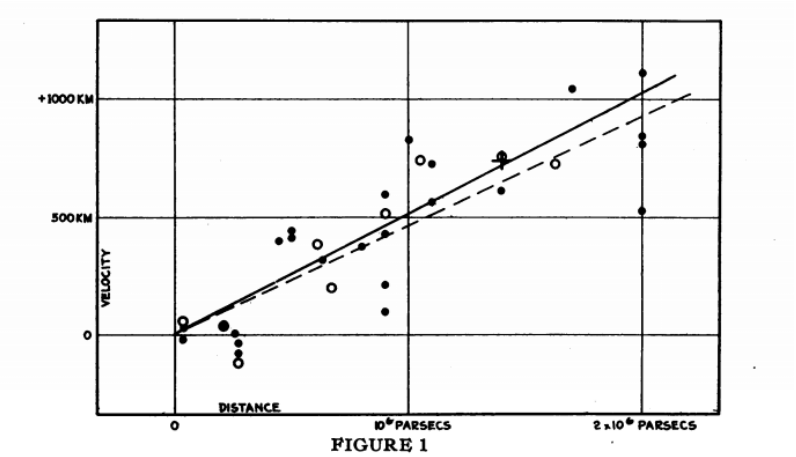

FIGURE 1
Velocity-Distance Relation among Extra-Galactic Nebulae.

Figure 2: Velocity of the Cepheids as a function of the distance from the observer. The proportionality is the sign of the expansion of the Universe. Taken from [9].

the solution of the Friedman equations scales such that:

$$a \propto t^{\frac{2}{3}} \tag{21}$$

That means that, as long as the Universe is dominated by non-relativistic matter, it is going to expand. We should note that this model is called CDM, an acronym for Cold Dark Matter.

The cosmic ladder and LCDM

Distance measurements

When we look at the Universe, what we usually see are the stars, galaxies, and other bright objects. If we assume that the Universe expands, naively speaking, we would expect them to move away from the observer on each point at least on average. Edwin Hubble in his outstanding discovery in 1929 measured the dependence of the velocity of Cepheides on the distance from Earth. The results are seen in Figure 2.

The result was... surprising to say the least. The velocities and distances were correlated! What does our model described in the previous section say about this?

Let us have a look at the interval ds^2 in our FLRW-metric (Eq. 12). We remind the reader that for light-like curves, or in other words, for trajectories of relativistic particles including photons, $ds^2 = 0$. Assuming a ray of light coming from an object to an observer in the centre of coordinates we get:

$$dt = a(t)\frac{1}{\sqrt{1-k^2r^2}}dr \tag{22}$$

We will also try to look at a kind of a Doppler's law. Let us assume the frequency of the incoming wave ω_{received} and introduce the concept of redshift z with the help of transmitted frequency $w_{\text{transmitted}}$:

$$\frac{\omega_{\text{transmitted}}}{\omega_{\text{received}}} = \frac{\delta t_{\text{received}}}{\delta t_{\text{transmitted}}} = 1 + z \tag{23}$$

Now, assuming $dt \approx \delta t$, we get:

$$1 + z = \frac{a(t_{\text{transmitted}})}{a(t_{\text{received}})} \tag{24}$$

We can try expanding the $a(t_{\text{transmitted}})$ into Taylor series around t_{received}, and after some simplifications and multiplying by c to connect to distances instead of times, we get at first order,

$$cz = \frac{\dot{a}(t_{\text{received}})}{a(t_{\text{received}})} c(t_{\text{transmitted}} - t_{\text{received}}) = HD \tag{25}$$

where D is the distance traversed by the light, and H is the Hubble parameter that is related to the scale factor $a(t)$ as:

$$H(t) = \frac{\dot{a}(t)}{a(t)} \tag{26}$$

What we see is that the redshift, which can be associated with the velocity of the galaxy escaping from Earth, is directly correlated with the distance to it with the Hubble expansion rate as a proportional factor!

In that manner, Hubble discovered the ongoing expansion of the Universe, which was completely predicted by General Relativity with the FLRW metric.

Cosmic history

Suddenly getting more and more data from the past of the Universe allowed humanity to discover many fascinating things about its history! Here we will only shortly present the commonly accepted scenario in a very compact version, as a detailed examination would require at least a separate book. An illustrative recap is shown in Figure 3.

The consensus model on the origins of the Universe is the hot Big Bang model, which postulates that the Universe was very small and extremely hot at the beginning [7]. This model has quite a few implications.

One of the first thing that is believed to have happened is inflation. It is an expansion of the Universe, which happened very rapidly, to the point where gaussian quantum fluctuations in the matter and energy distribution got "frozen" and formed the density fluctuations we are seeing in today's Universe, after which it continued cooling. For more information we refer the reader to [10].

Some seconds after that, the Universe cooled enough for the first atomic nuclei to be formed, mostly those of hydrogen and helium-4. This process is called the Big Bang

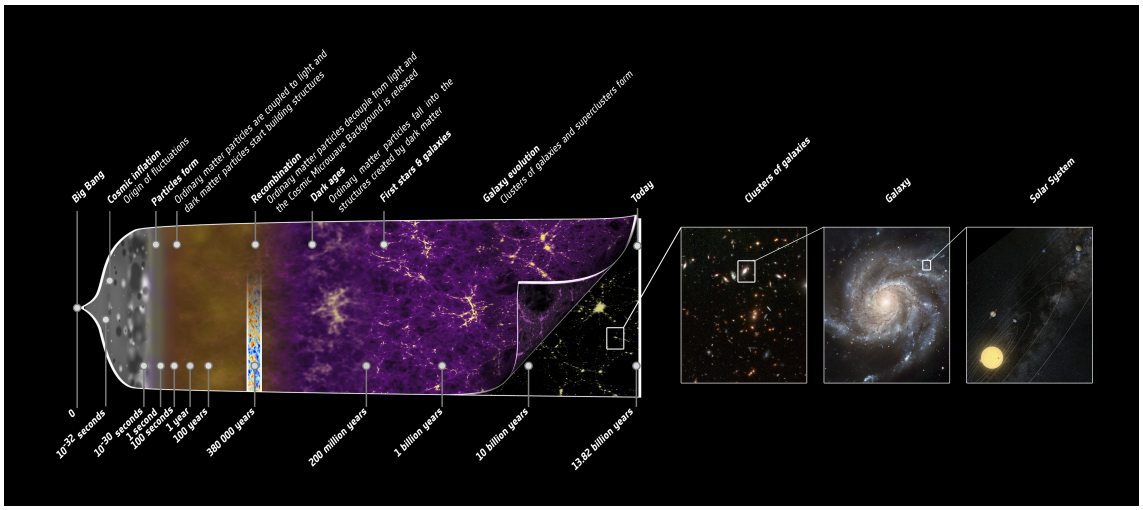

Figure 3: Illustration of the evolution of the Universe from the Big Bang until today. Taken from ESA.

nucleosynthesis [7]. However, it is still too early for the atoms to form, as the photons are too strongly coupled to the matter.

The Universe will need some thousands of years of additional cooling before reaching the matter-radiation decoupling. Once the photons decouple from the baryonic matter, the Universe finally becomes transparent, as the photons are able to travel through it freely. The first hydrogen atoms can be formed, a period which we call recombination, and from that epoch the Cosmic Microwave Radiation comes from, an isotropic black-body radiation, consisting of freely traveling photons. The moment when the baryons decouple as well and start travel freely is called the drag epoch. Additionally, the pressure waves within the electron-baryon plasma propagated until they got frozen sound horizon at drag epoch. They get embedded in the distribution of condensing matter until today. We call these waves the Baryon Acoustic Oscillations (BAO). As we will later see, with the continuation of the Universe growth, the fluctuations given by inflation and from BAO will not disappear but will continue to grow as well, allowing us to track the growth of Universe with time. And, finally, hundreds of millions of years after this the first galaxies start to form in the matter-dominated era.

Dark Energy

Until relatively recently, the consensus was held that the CDM model (with the addition of some baryonic effects) was the most accurate model of the Universe. However, in 1998 and 1999 that view was crushed by the evidence coming from type Ia supernova [12, 13]. Some of the high-redshift supernovae turned out to be further away than expected from a CDMmodel.

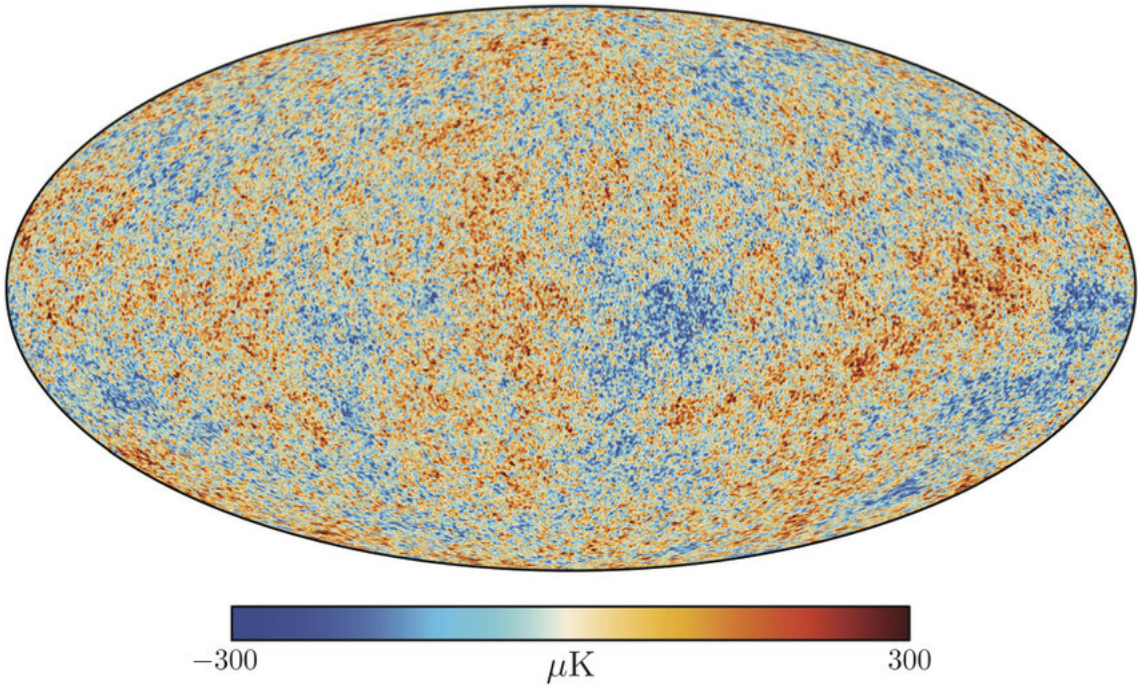

Figure 4: Temperature variations in the CMB sky from Planck. Taken from [11].

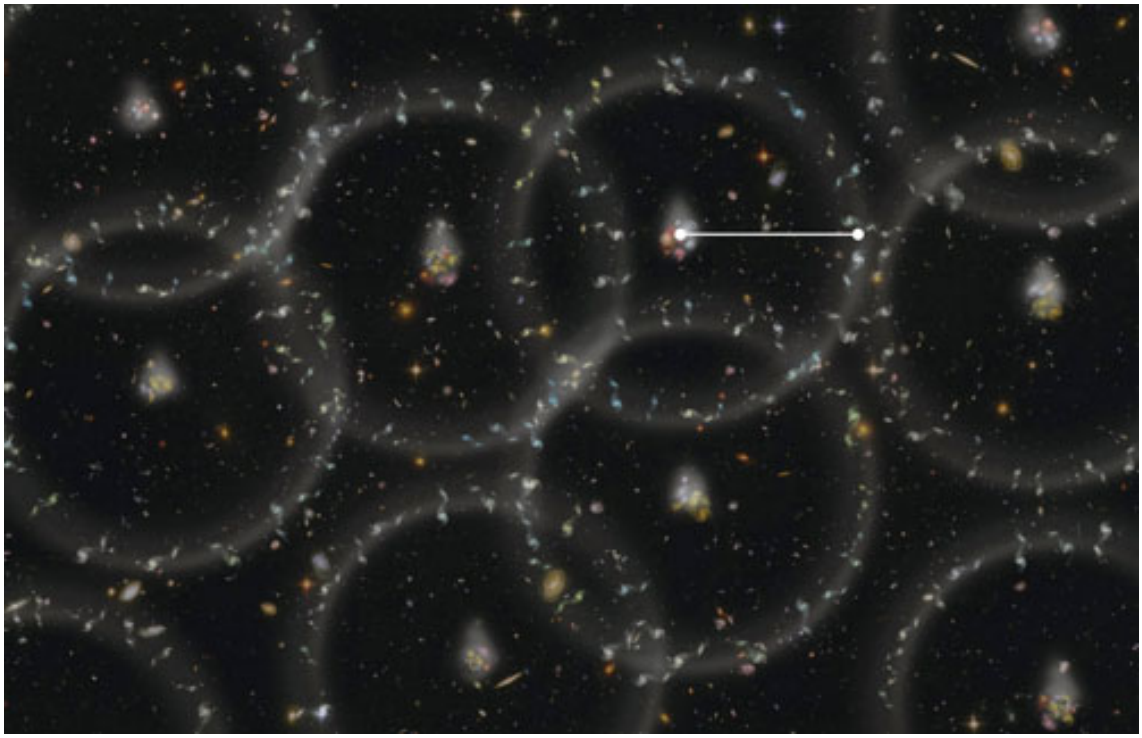

Figure 5: Illustration of the Baryon Acoustic Oscillations spreading in space, artistic view. Created by BOSS collaboration, adapted from here.

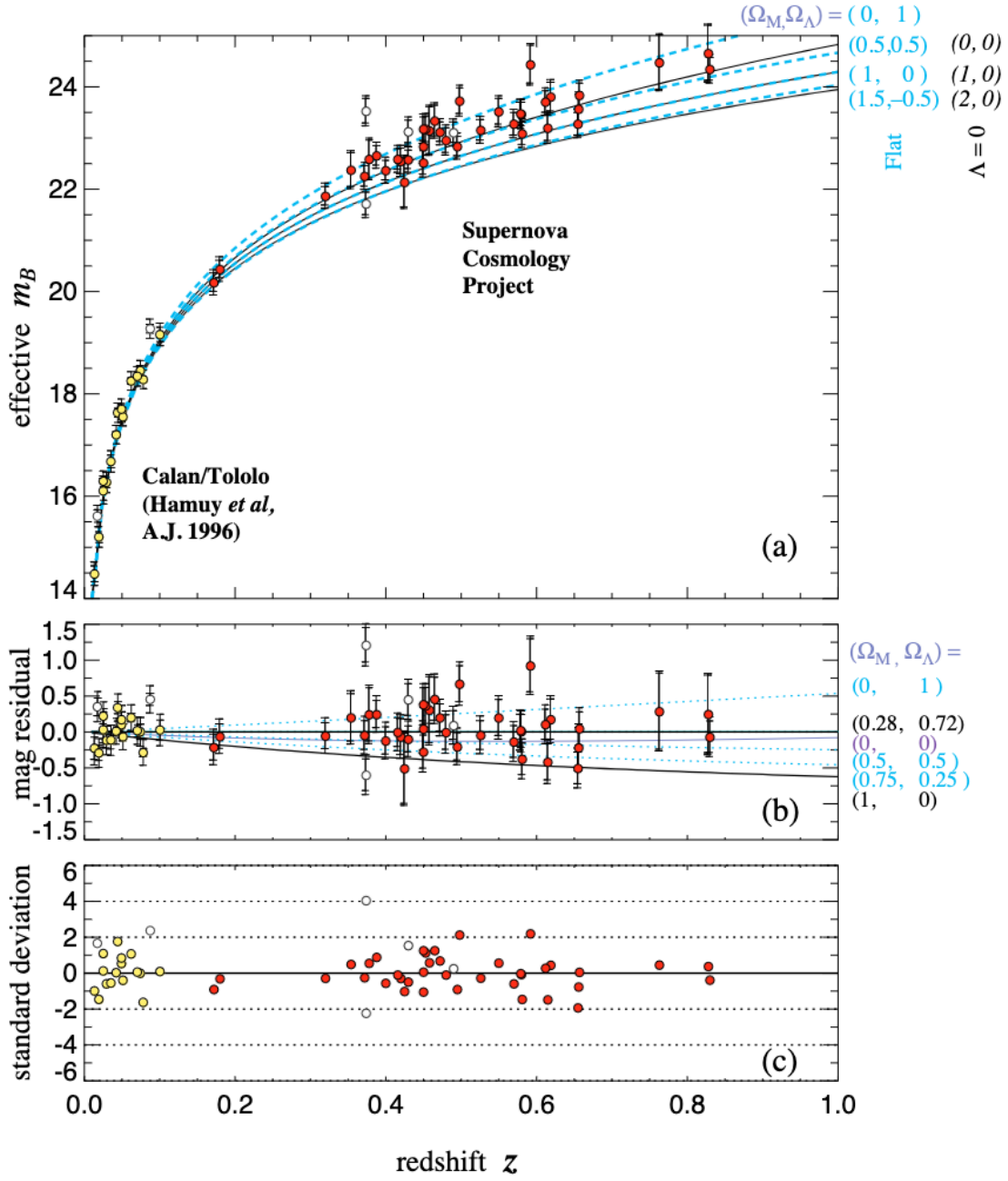

Figure 6: *(a)* Hubble diagram for 60 type Ia supernovae. *(b)* Magnitude residuals from the best fit cosmology for $\Omega_m = 0.28$, $\Omega_\Lambda = 0.72$. The dashed curves are for a range of the cosmological models. *(c)* uncertainty normalised results for the best-fit flat cosmology. Taken from [13].

They found a deviation of $> 5\sigma$ with CDM, and the best fit value of cosmological parameters turned out to be $\Omega_m = 0.28$ and $\Omega_\Lambda = 0.72$. The latter is the term responsible for the "energy density" of the Λ constant, which originally appeared in the Einstein's equations as the integration constant, which is back in the equations in order to explain the acceleration of the cosmic expansion. This mysterious energy is called the Dark Energy. But, what does it mean physically? Are we sure that we are not missing something in our understanding of the nature of gravity?

That is one of the main questions of modern cosmology. It is towards answering these questions that my work throughout the past years has been dedicated to. Numerous collaborations around the world, all tooled up with various probes. Table 1 summarises some of the experiments working towards this goal.

Project	Dates	Area/deg^2	Methods	Reference
BOSS	2008-2014	10000	BAO/RSD	[14]
KiDS	2011-2019	1500	WL/CL	[15]
eBOSS	2014-2018	7500	BAO/RSD	[16]
SuMIRE	2014-2024	1500	WL/CL/BAO/RSD	[17]
HETDEX	2017-2023	450	BAO/RSD	[18]
DESI	2020-2025	14000	BAO/RSD	[19]
LSST	2022-2032	20000	WL/CL/SN/BAO	[20]
Euclid	2022-2028	15000	WL/CL	[21]
WFIRST	2025-2030	2200	WL/CL/SN/BAO/RSD	[22]
ACT	2008-2022	19000	CMB/CL	[23]
Planck	2009-2013	41253	CMB/CL	[24]
ZTF	2018-2023	20000	SN	[25]
DES	2013-2019	5000	WL/CL/BAO/SN	[26]

Table 1: Different dark energy facilities with information from [27] and some additional updates with more recent experiments. WL stands for weak lensing, SN for Type Ia Supernovae and CL for clusters

Supernovae Type Ia (SN Ia)

Working on the same principle as the original Hubble observations, the SN Ia experiments build a cosmic ladder. However, instead of using Cepheids, supernovae of Type Ia are used. These are the thermonuclear explosions of white dwarfs. Calibrated on the low-redshift samples with high precision, by adding the information from the higher redshifts one can track the history of expansion of the Universe[28]. It is with this probe that the dark

energy was originally discovered, and it still plays a big role in unravelling the mysteries of Universe expansion. Examples of such experiments are the ongoing ZTF[25] and the upcoming Vera Rubin of LSST[20].

Cosmic Microwave Background (CMB)

Originating during recombination, this background radiation has been travelling in the Universe since then, thus serving as our main source of information about the primordial Universe, thus defining the starting conditions for the latter formation of large-scale structure. Though it is hypothesized to be isotropic and homogeneous, due to the circular movement of the observer, lensing from heavy objects, like galaxy clusters, as well as scaterring effects (Sunyaev-Zeldovich effect [29]) the anisotropies appear, therefore giving not only the information of the recombination era, but also about the later epochs as well. Such experiment include, but are not limited to Planck[24] and ACT[23].

Weak lensing (WL)

Light passing to the Earth from distant objects is deflected by other massive objects lying on its way. Often enough, this distortion of the image of source galaxies is too small to be detected from one image, however, with more emitters situated in a close enough region and passing near the same massive bodies, we can statistically notice the change. This effect is called the weak lensing (strong lensing is reserved for the distortions big enough to be detectable from a single source). Using that information we can infer information about the underlying matter distribution, and thus constrain the parameters of our cosmological theories such as ΛCDM [30]. Notable surveys using this probe are the previous generation survey DES[26] and the new generation surveys Euclid[21] and Vera Rubin of LSST[20]. An illustration of such an effect is shown in figure 7.

Galaxy clusters (CL)

The galaxy clusters are the most massive gravitationally-bound structures of the Universe. One can estimate their mass in the X-ray and with weak lensing surveys. The galaxy clusters are sensitive not only to the expansion of the Universe but also to the growth of structures, and serve as a unique prove for the hierarchical large-scale structure formation. By inferring the distribution of the underlying dark matter distribution over different redshifts, obtained thus from the positions and masses of the galaxy clusters, we can track the evolution of the Universe [31]. Notable surveys using that technique are Planck[24] and the new generation surveys Euclid[21] and Vera Rubin of LSST[20].

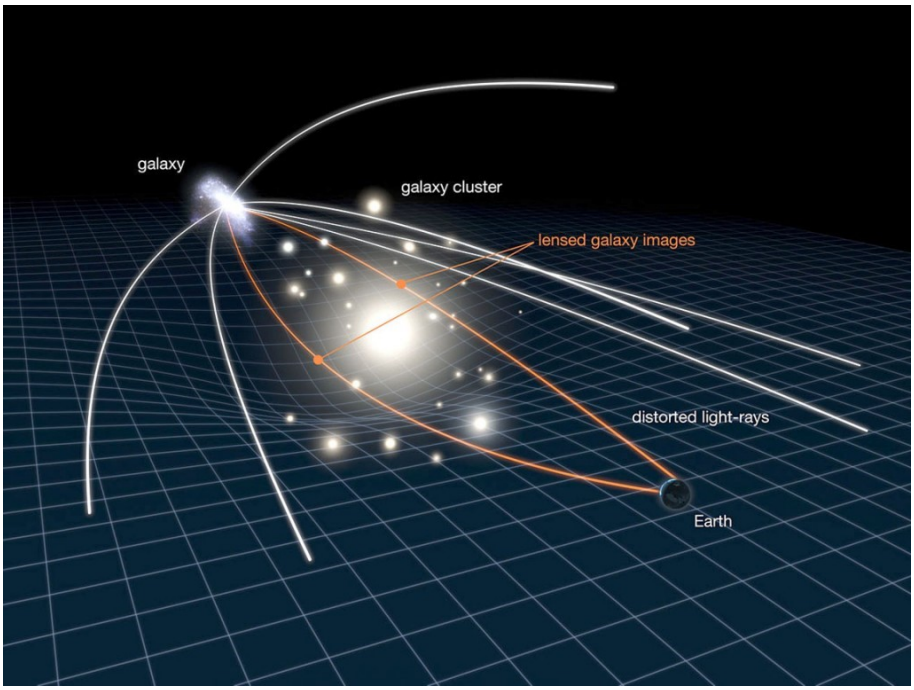

Figure 7: An illustration explaining the effect of lensing of images from distant galaxies. Taken from here, Copyright: NASA, ESA L. Calçada.

Ly-α forests

Light coming from the distant quasars on its way encounters hydrogen clouds. It should be noted, that by the time it reaches the clouds, the original α-peak from the hydrogen quasar emission becomes shifted, meaning that the absorption process creates a new one. Thus, on Earth what we see in the spectra is a multitude of absorption peaks which happened at different redshifts, thus allowing for another probe of the matter distribution, and as such, a possibility to infer the cosmological parameters[32]. An illustration of the process is shown in Figure 8. Notable surveys using this probe are the previous generation spectroscopic surveys BOSS[14] and eBOSS[16] and the new generation spectroscopic survey **DESI**[19].

Galaxy clustering

Galaxies are a natural way to track the matter density distribution. By estimating how the galaxies cluster and tracking the evolution of their clustering over time, we can infer the parameters of the cosmological models and the evolution of the Universe. This is the main probe for the **DESI**[19] experiment, and it is the probe we will be focusing on in this thesis.

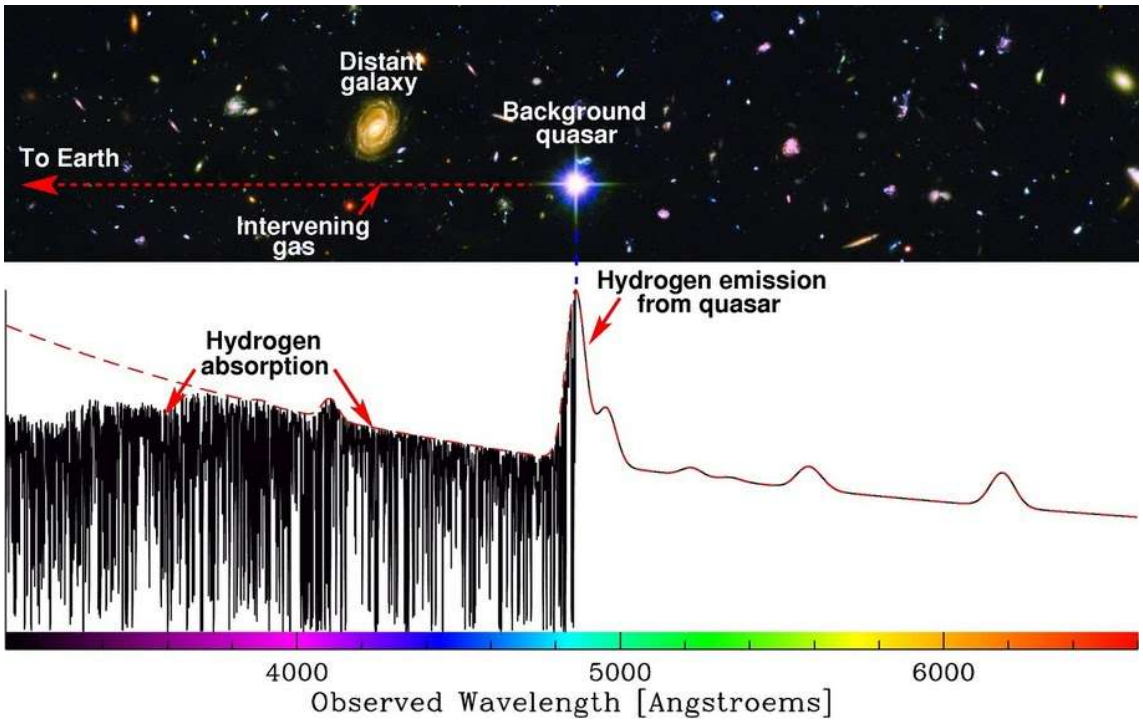

Figure 8: Illustrations of the process leading to the formation of the Lyman-α forest. As the light travels from the quasar to Earth, it gets the absorption peaks from the intervening gas. Courtesy of J. Webb and M. Murphy

Not so homogeneous Universe

As we have just mentioned, this thesis focuses on using galaxy clustering to estimate the cosmological parameters. However, one of the first assumptions we made when obtaining the Friedmann equations was the homogeneity of the Universe, with which galaxies do not really match. We are going to relax that assumption. Let us assume, that the most dominant part of the matter is distributed homogeneously, such that the inhomogeneities can be treated as perturbations. We therefore define these perturbations, also known as overdensities, by:

$$\delta(\tau, \boldsymbol{x}) = \frac{\rho(\tau, \boldsymbol{x}) - \overline{\rho}}{\overline{\rho}} \tag{27}$$

where $\overline{\rho}$ is the average matter density of the Universe.

Given that matter is non-relativistic (and pressureless), the equations of motion look like the traditional Newtonian ones. To achieve that, we modify the metric in the following manner, in the comoving coordinates:

$$ds^2 = -a^2(\tau)\left[(1+2\Psi)d\tau^2 + (1+2\Phi)\left(\sum_i dx_i^2\right)\right] \tag{28}$$

This is called the Newtonian gauge, and it includes all the perturbations represented as scalars. We refer the reader to[7] for a more complete description which includes vector

and tensor perturbations, why they are all independent from each other, why we can treat them separately, and why are they important, especially so for the other probes than galaxy clustering.

The classical way to derive the equations of motion for a perfect fluid in the given space usually starts with imposing the conservation of the particles law, resulting in the Vlasov equation. Combined with the gravitational equations derived earlier, it gives us the necessary set of equations. We will skip this derivation (which can be found here [7], and just write the Poisson equation directly:

$$\nabla^2\Phi(\tau,\boldsymbol{x}) = 4\pi G a^2\overline{\rho}\delta(\tau,\boldsymbol{x}) = \frac{3}{2}\mathcal{H}(\tau)^2\Omega_m\delta(\tau,\boldsymbol{x}) \tag{29}$$

where $\mathcal{H}$ is Hubble constant in comoving coordinates, defined as:

$$\mathcal{H} = aH \tag{30}$$

The fluid is also characterised by its velocities, which we define as $\boldsymbol{u}(\boldsymbol{x})$ (still in comoving coordinates). This requires us to also write the Euler equation:

$$\boldsymbol{u}'(\boldsymbol{x}) + \mathcal{H}\boldsymbol{u}(\tau,\boldsymbol{x}) + \boldsymbol{u}(\tau,\boldsymbol{x})\cdot\nabla\boldsymbol{u}(\tau,\boldsymbol{x}) = -\nabla\Phi(\tau,\boldsymbol{x}) \tag{31}$$

where $f' = \frac{\partial f}{\partial\tau}$, and we assume that the stress inside the fluid is negligible.

Finally, we can write the equation of continuity:

$$\delta'(\tau,\boldsymbol{x}) + \nabla\cdot[(1+\delta(\tau,\boldsymbol{x}))\boldsymbol{u}(\tau,\boldsymbol{x})] = 0 \tag{32}$$

Now we have all the components to proceed to more accurate computations. These are the basic equations of Eulerian perturbation theory. For a review, see [33].

Liner order perturbations

We focus on small, with respect to the background, perturbations of matter density, meaning that$\delta(\boldsymbol{x})$ and $\boldsymbol{u}(\boldsymbol{x})$ are considered to be small. Having that in mind, let us take the simplest possible approach for the moment and write the equations in a linear order.

$$\delta'(\tau,\boldsymbol{x}) + \nabla\cdot\boldsymbol{u}(\tau,\boldsymbol{x}) = 0 \tag{33}$$

$$\boldsymbol{u}'(\tau,\boldsymbol{x}) + \mathcal{H}(\tau)\boldsymbol{u}(\tau,\boldsymbol{x}) = -\nabla\Phi(\tau,\boldsymbol{x}) \tag{34}$$

Let us focus a bit on the equation 34. We apply $\nabla\cdot$ and $\nabla\times$ to it, to see what happens. Somewhere on the way, we notice that in the first case the right-hand side becomes Eq. 29, which we immediately substitute, ending up with the following:

$$(\nabla\cdot\boldsymbol{u}(\tau,\boldsymbol{x}))' + \mathcal{H}(\tau)\nabla\cdot\boldsymbol{u}(\tau,\boldsymbol{x}) + \frac{3}{2}\Omega_m\mathcal{H}(\tau)^2\delta(\tau,\boldsymbol{x}) = 0 \tag{35}$$

$$(\nabla \times \boldsymbol{u}(\tau, \boldsymbol{x}))' + \mathcal{H}(\tau)\nabla \times \boldsymbol{u}(\tau, \boldsymbol{x}) = 0 \tag{36}$$

We see, that due to the expansion of the Universe, the $\nabla \times \boldsymbol{u}$ decays quickly with the expansion of the Universe. This quantity is called the vorticity, and in the lienar regime we can neglect it such that the velocity field becomes a gradient field.

We can use equation 33 to substitute $\nabla \cdot \boldsymbol{u}(\tau, \boldsymbol{x})$ with the time derivative of $\delta(\tau, \boldsymbol{x})$, leading to:

$$\delta''(\tau, \boldsymbol{x}) + \mathcal{H}(\tau)\delta(\tau, \boldsymbol{x})' = \frac{3}{2}\Omega_m \mathcal{H}\delta(\tau, \boldsymbol{x}) \tag{37}$$

Note how there is no dependence on the spatial coordinates. That implies that the evolution of the perturbations is homogeneous, at least in the linear order.

We can introduce the linear growth function $D(\tau)$, which carries all the time-dependence of $\delta(\tau, \boldsymbol{x}) = D(\tau)\delta(\boldsymbol{x})$. Presenting the two solutions of the equation as $D^+(\tau)$, representing the faster growing mode, and $D^-(\tau)$, representing the slowest growing mode, we assume the solution to be of form:

$$\delta(\tau, \boldsymbol{x}) = D^+(\tau)A(\boldsymbol{x}) + D^-(\tau)B(\boldsymbol{x}) \tag{38}$$

As for the $\nabla \cdot \boldsymbol{u}(\boldsymbol{x})$, we obtain:

$$\nabla \cdot \boldsymbol{u}(\boldsymbol{x}) = -\mathcal{H}[fA(\boldsymbol{x}) + gB(\boldsymbol{x})] \tag{39}$$

where $f, g = \frac{1}{\mathcal{H}}\frac{d\log D^{+,-}}{d\tau}$. f is usually called the **linear growth rate of structure**. Now let us have a look at the most relevant cases.

For a matter-only Universe $\Omega_m = 1$ we obtain

$$D^+(\tau) = a(\tau) \tag{40}$$

$$D^-(\tau) = a^{-\frac{3}{2}}(\tau) \tag{41}$$

So, the density fluctuations grow as the scale factor.

For a Universe with dark energy with $\Lambda > 0$, there is an integral representation, which is not solvable analytically but can be approximated as [7]:

$$D^+ = \frac{5}{2}\frac{a\Omega_m}{\Omega_m^{\frac{4}{7}} - \Omega_\Lambda + (1 + \frac{1}{2}\Omega_m)(1 + \frac{1}{70}\Omega_\Lambda)} \tag{42}$$

$$D^- = \frac{\mathcal{H}}{a} \tag{43}$$

It is necessary to notice, that it is possible to go beyond the linear order in the Eulerian framework, and one can read more on that for example here [33, 34]. In chapter 2 we will however explore higher-than-linear terms in the Lagrangian Perturbation theory

framework, which is equivalent to the Eulerian Perturbation theory but whose quantity of interest is no longer the density and velocity fields but instead the displacement field [35–38].

Observables

As we aim at exploring the properties of a statistical distribution, we need to find a good quantity to observe. As we assumed the Universe to be mostly homogeneous, the mean of perturbations δ will be thus 0. We also assume that the density fluctuations are Gaussian, being the remnants of ancient quantum fluctuations frozen during inflation[10]. The most classical manner to deal with it is using the autocorrelation function. For the density field $\delta(\boldsymbol{x})$, it can be defined as:

$$\xi(\boldsymbol{r}) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle \tag{44}$$

Its Fourier counterpart is given by Fourier transform such that the density field $\delta(\boldsymbol{k})$ in Fourier space is defined by:

$$\delta(\boldsymbol{k}) = \int d^3x e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\delta(\boldsymbol{x}) \tag{45}$$

The power spectrum $P(\boldsymbol{k})$ is therefore defined as:

$$P(\boldsymbol{k}) = \langle \delta(\boldsymbol{k})\delta(\boldsymbol{k})\rangle \tag{46}$$

Let us derive a simple model of the linear power spectrum, assuming that the fluctuations originated during the inflationary epoch. We can define the transfer function $T(\boldsymbol{x})$ as the measure of the evolution of the potential from that epoch until today, and define, following [39], as:

$$\Phi(\tau, \boldsymbol{k}) = \Phi(\tau_0, \boldsymbol{k})T(\tau, \boldsymbol{k}) \tag{47}$$

In this context, the starting point is usually taken at the drag epoch, when baryon's optical depth reaches unity . An example of the computation of such a transfer function can be found for example in [39]. Some more advanced solutions like `class`[40] and `camb`[41], include the baryonic, BBN and neutrino effects even better.

In order to estimate the initial power spectrum at the drag epoch, we assume purely adiabatic scalar perturbations, which can be parameterised [39] as:

$$P(k) = \frac{2\pi^2}{k^3}A_s\left(\frac{k}{k_0}\right)^{n_s-1+\frac{1}{2}\frac{dn_s}{d\ln k}\ln\left(\frac{k}{k_0}\right)} \tag{48}$$

where A_s is the initial power spectrum normalisation, n_s is the spectral index, and A_s, n_s, k_0 and $\frac{dn_s}{d\ln k}$ are taken to be constant.

In order to get the power spectrum at the given redshift we just need to multiply the transfer function and account for the effect of the growth of structure:

$$P(\tau, k) = P(k)T^2(\tau, k) \tag{49}$$

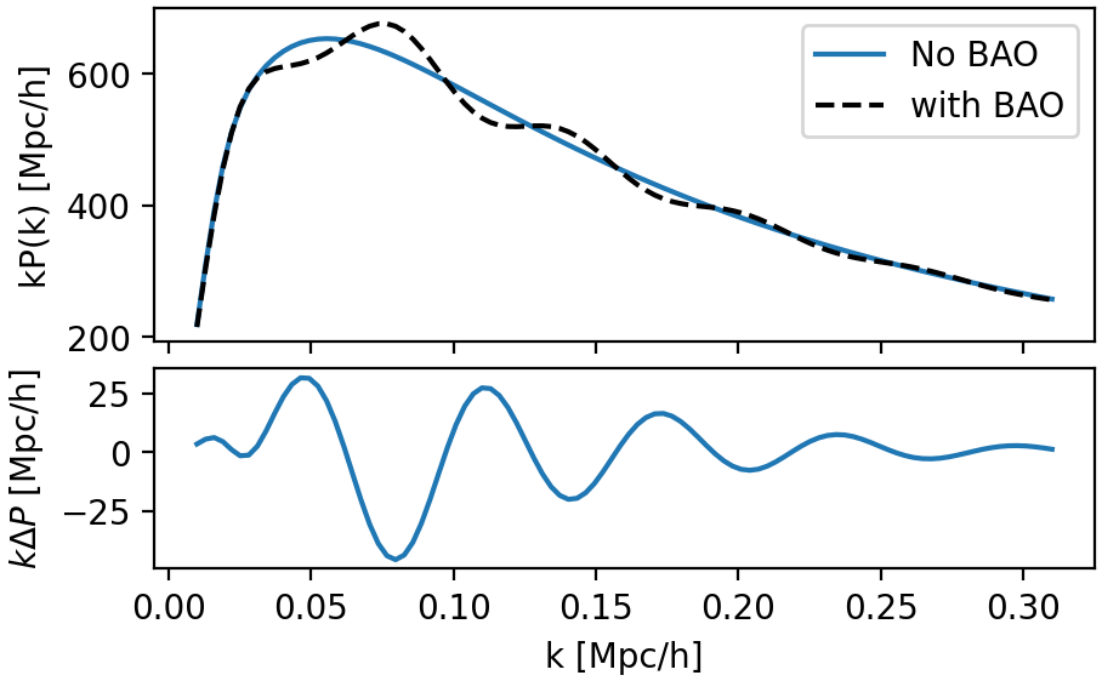

Figure 9: Real-space power spectrum with and without the BAO effect. The lower panel shows the difference between the two. Generated using the Eisenstein-Hu model[42].

A similar computation can be performed for the linear correlation function $\xi(s)$, or one can obtain it by performing a Fourier transform. Note however, that the power spectrum obtained is that of the overall matter distribution, not galaxies.

An important effect to be included in the linear power spectrum is the Baryon Acoustic Oscillations as introduced earlier. Those oscillations cause wiggles in the power spectrum that are shown in Figure 9. An example of computation of the transfer function with this effect included can be seen in [42].

In the radial mass profile, this effect appears as a peak which moves with time, as can be seen on Figure 10. As the peak appeared in the early times, when photons and baryonic matter were still coupled, the peak for both species appeared identical but as the decoupling between the two happened, the two shapes diverged. However, at the same time the baryonic matter starts collapsing inducing the peak on the dark matter distribution, as we can notice from the behaviour of the dark matter correlation function for the later times. As the peak from baryon acoustic oscillations is ”frozen” into the matter distribution, and only chnages due to an expansion of the Universe, providing us with a standard ruler that can be measured as a function of redshift.

Galaxy clustering

Now that we have figured out the clustering statistics for the overall matter distribution, we need to connect it to the observed density field of galaxies. It should be noted, that most

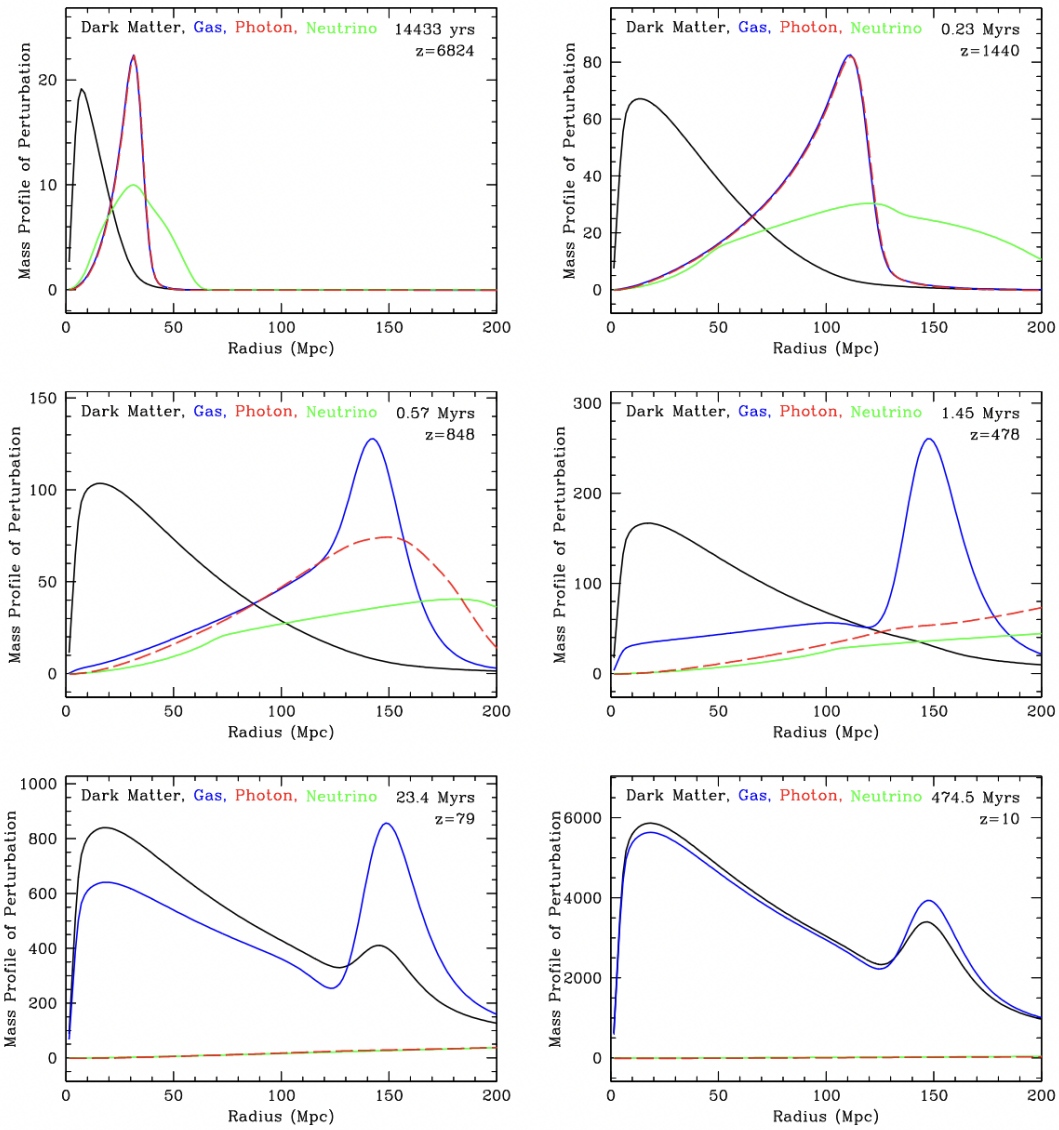

Figure 10: Evolution of the radial mass profile of initially point-like matter overdensities of various types of matter. Taken from [43].

of the matter in the Universe is actually non-observable dark matter [7], and only 20% is observable. Meaning that we need to connect the two, parametrising various effects specific to galaxies.

Galaxy bias

Despite the fact that the dark matter and galaxy distributions are different, we assume them to be gravitationally coupled, and thus, dependent. Thus, we will assume the bias expansion of the general form, assuming that it will depend on the underlying matter density field, to be:

$$\delta_g(\tau, \boldsymbol{x}) = \sum_i b_i(\tau) O_i[\delta](\tau, \boldsymbol{x}) \tag{50}$$

where $O[\delta](\tau, \boldsymbol{x})$ is an operator of the underlying matter distribution $\delta(\tau, \boldsymbol{x})$, relevant at a given order. If we continue describing the early linear stages of structure formation, the linear order bias relation becomes:

$$\delta_g(\tau, \boldsymbol{x}) = b_1(\tau)\delta(\tau, \boldsymbol{x}) \tag{51}$$

We will cover more extensive bias expansion models in the chapter 2 of this thesis.

Therefore the galaxy 2-point statistics (power spectrum or correlation function) is related to the matter 2-point statistics in the following manner:

$$P_g(\boldsymbol{k}) = b_1^2 P(\boldsymbol{k}) \tag{52}$$

$$\xi_g(\boldsymbol{r}) = b_1^2 \xi(\boldsymbol{r}) \tag{53}$$

where the bias b_1 is also called the linear bias, or linear Eulerian bias, as we follow the Eulerian approach to describe the matter distribution dynamics.

Redshift Space Distortions

The galaxies do not really form a smooth density field, but rather are seen by us as a myriad of point-like objects. The issue is that they have their own peculiar velocities, with respect to the underlying matter density. Moreover, the distance to galaxies is measured using the redshift, which accounts for the line of sight component of the galaxy peculiar velocity. Therefore, our estimation of distance diverges from the real one. The space in which galaxies have these "shifted" coordinates is called the redshift space, and the galaxy clustering in this redshift space appears distorted or anisotropic due to the galaxy peculiar velocities. The effect itself carries the name of "Redshift Space Distortions" (RSD).

Inferring the peculiar velocities from galaxies, though possible, is a complex task [44]. We will therefore go from the other side. We will try to get a correction in linear order, which simulates the RSD effect in the power spectrum.

We define the line of sight of the observer as $\hat{z}$, and we further define the directional cosine as $\mu = \hat{k} \cdot \hat{z}$.

That modifies the observed redshift $z_{\mathbf{obs}}$ as:

$$z_{\mathbf{obs}} = z + \dot{x} \cdot \hat{z} = z + \frac{1}{a}\boldsymbol{u} \cdot \hat{z} \tag{54}$$

where $\boldsymbol{u}$ is the velocity of the density elementin the comoving frame.

Assuming the redshift-space distance to be s, and the real space one r, we can write the law for the preservation of the density of matter:

$$d^3s(1 + \delta_s(\boldsymbol{s})) = d^3r(1 + \delta(\boldsymbol{r})) \tag{55}$$

We will also have the following transformation law between the two:

$$\boldsymbol{s} = \boldsymbol{r} + \frac{1}{\mathcal{H}}\boldsymbol{u} \cdot \hat{z} \tag{56}$$

where $\mathcal{H}$ is a conformal Hubble factor defined as $\mathcal{H} = aH$.

We can go to Fourier space, simultaneously using our results from the previous sectionand get the following derivation, assuming distant observer and plane-parallel approximations:

$$\begin{aligned}\delta_s(\boldsymbol{k}) = \int d^3s\delta(\boldsymbol{s})e^{i\boldsymbol{k}\cdot\boldsymbol{s}} = \int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{s}}(1 + \delta(\boldsymbol{x})) - \int d^3se^{i\boldsymbol{k}\cdot\boldsymbol{s}} = \\ = \int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{s}}(1 + \delta(x)) - \int d^3x\left|\frac{d^3s}{d^3x}\right|e^{i\boldsymbol{k}\cdot\boldsymbol{s}} = \int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{s}}(1 + \delta(x)) - \\ - \int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{s}}\left(1 + \frac{1}{\mathcal{H}}\frac{\partial u_z(\boldsymbol{x})}{\partial z}\right) = \int d^3x\left(\delta(\boldsymbol{x}) - \frac{1}{\mathcal{H}}\frac{\partial u_z(\boldsymbol{x})}{\partial z}\right)e^{i\boldsymbol{k}\cdot\boldsymbol{x} + ik\mu\frac{u_z}{\mathcal{H}}}\end{aligned} \tag{57}$$

We then note, that in Fourier space we have, remembering that we have discarded $\nabla \times \boldsymbol{u}$ due to a quick decay:

$$\int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{x}}\nabla \cdot \boldsymbol{u} = i\boldsymbol{k} \cdot \boldsymbol{u} = -\boldsymbol{k} \cdot \boldsymbol{k}\theta(\boldsymbol{k}) \tag{58}$$

such that:

$$\boldsymbol{u} = i\boldsymbol{k}\theta(\boldsymbol{k}) \tag{59}$$

Once again we take the linear order, in which the second term in the exponent disappears, in order to obtain:

$$\delta_s(\boldsymbol{k}) = \delta_m(\boldsymbol{k}) - \int d^3xe^{i\boldsymbol{k}\cdot\boldsymbol{x}}\frac{1}{\mathcal{H}}\frac{\partial u_z(\boldsymbol{x})}{\partial z} = \delta_m(\boldsymbol{k}) - \frac{1}{\mathcal{H}}(\boldsymbol{k} \cdot \hat{z})^2\theta(\boldsymbol{k}) = \delta_m(\boldsymbol{k}) - k^2\frac{1}{\mathcal{H}}\mu^2\theta(\boldsymbol{k}) \tag{60}$$

Then we can get the final answer:

$$\delta_s(\boldsymbol{k}) = \delta_m(\boldsymbol{k}) - k^2\frac{1}{\mathcal{H}}\mu^2\theta(\boldsymbol{k}) = \delta_m(\boldsymbol{k}) + f\mu^2\delta_m(\boldsymbol{k}) \tag{61}$$

If we repeat this computation taking the linear bias into account, we will get:

$$\delta_{g,s}(\boldsymbol{k}) = b_1\delta_m(\boldsymbol{k}) + f\mu^2\delta_m(\boldsymbol{k}) \tag{62}$$

where f is called a growth rate.

This modifies the linear power spectrum such that we obtain the Kaiser formula [45]:

$$P_g^s(k,\mu) = (b_1 + f\mu^2)^2 P_m(k) \tag{63}$$

The same can be done for the correlation function. We see that RSD creates anisotropy in the observed power spectrum. What is of the outmost importance for us, is that the growth rate of structure f contains plenty of physical information regarding the gravitational collapse that drives the structure formation. We can, through a convenient approximation [46] relate it to the Ω_m, as well as to the equation of state of Dark Energy ω_Λ as:

$$f \propto \Omega_m^\gamma \propto \Omega_m^{\frac{3(1-\omega_{\rm DE})}{5-6w_{\rm DE}}} \tag{64}$$

We note that in the case of the cosmological constant Λ taken as Dark Energy, we have $w_{\rm DE} = -1$ and thus $\gamma = 0.55$. We also note, that in case of alternative gravitational theories, γ might differ.

Alcock-Paczynski effect

As already mentioned, we are measuring the distances to observed galaxies using their redshifts, and converting them to coordinates. Same goes for the estimation of parameters like growth rate: a fiducial cosmology is required. However, in order to do it we need to assume some cosmology. What if we get the cosmology wrong? (Which is what is going to happen, anyway). Can we quantify how far from the truth we got?

For that we can use the BAO peak. Assuming the sound horizon scale r_s at the redshift of decoupling z_d, angular diameter distance $D_A(z) = \frac{r(z)}{1+z} =$, where $r(z)$ is the comoving distance to the object, and Hubble parameter $H(z)$.

We can then use the r_s as our main scale to look at its change with time. Assuming it to be isotropic, let us have a look at the transverse component as seen from Earth. It will have a small angular size of θ:

$$\theta = \frac{r_s}{D_A(z)} \tag{65}$$

Assuming the scaling to be small, we can define the transverse Alcock-Pasczynski (AP) parameter:

$$\alpha_\perp = \frac{D_A(z) r_s^{\rm fid}(z_d)}{D_A^{\rm fid}(z) r_s(z_d)} \tag{66}$$

Looking at the parallel component, we can assume that the distance change over time passed from the drag epoch to be $H(z)r_s(r_d)$, which we can use for the parallel AP parameter:

$$\alpha_{\parallel} = \frac{H^{\rm fid}(z) r_s^{\rm fid}(z_d)}{H(z) r_s(z_d)} \tag{67}$$

This will also imply, that the wavelengths in the Fourier space, transforming as the inverse of the separation, will be scaled as

$$k'_{\parallel} = \frac{k_{\parallel}}{\alpha_{\parallel}}, \qquad k'_{\perp} = \frac{k_{\perp}}{\alpha_{\perp}} \tag{68}$$

$$\tag{69}$$

We will then convert the parallel and perpendicular components of the wavevector to our usual angular form in terms of k and μ:

$$k' = \frac{k}{\alpha_{\perp}} \left[1 + \mu^2 \left(\frac{\alpha_{\perp}^2}{\alpha_{\parallel}^2} - 1 \right) \right]^{\frac{1}{2}} \tag{70}$$

$$\mu' = \mu \cdot \frac{\alpha_{\perp}}{\alpha_{\parallel}} \left[1 + \mu^2 \left(\frac{\alpha_{\perp}^2}{\alpha_{\parallel}^2} - 1 \right) \right]^{-\frac{1}{2}} \tag{71}$$

That will allow us to get the power spectrum in the "wrong" cosmology, accounting also for the difference in volume between the two cosmologies, as:

$$P(k,\mu) = \left(\frac{r_s^{\rm fid}}{r_s} \right)^3 \frac{1}{\alpha_{\perp}^2 \alpha_{\parallel}} P_{\rm fid}(k'(k,\mu), \mu'(\mu)) \tag{72}$$

That allows for a powerful agnostic cosmological test of ΛCDM, as now our presumption of the fiducial cosmology is fixed and serves as the reference point on how far what we observe is from what we have assumed, and carries the name of Alcock-Paczynski test [47].

It was shown that via a process called reconstruction [43], we can make the BAO peak and its estimation more precise, by trying to "reverse" the RSD effects, but this technique is unfortunately out of the scope of this thesis.

Clustering statistics in practice

After introducing the basic theoretical concepts of galaxy clustering and describing some of the effects that affect the power spectrum or correlation function, in this section we will describe how to estimate them in practise. We should also note that despite correlation function and power spectrum being configuration and Fourier space counterparts of each other, the way to estimate them is very different.

Correlation function estimation

Let us start with a set of coordinates $\boldsymbol{x}^i$ of galaxies with their respective weights (to account for potential sources of systematic effects) w^i. Intuitively we would assume, that the higher is the number of galaxies in a given space-time region, the higher is the overdensity $\delta(\boldsymbol{x})$ in this region.

We separate our survey volume into a grid with K cells, such that each cell contains either 0 or 1 galaxies, of which we have n. We then can write the expectation of finding an object in any of the cells:

$$\langle \nu \rangle = \frac{n}{K} \tag{73}$$

We look at two cases: correlated and uncorrelated. We start with the uncorrelated scenario where, for a given uncorrelated sample, the probability that two of the chosen cells contains a point is, for $i \neq j$ by construction:

$$\langle \nu_i \nu_j \rangle = \frac{n(n-1)}{K(K-1)} \tag{74}$$

We then write the expectation value for what we call the paircounts $RR(\boldsymbol{r})$: how many pairs separated by a separation $\boldsymbol{r}$ can be found in this uncorrelated sample:

$$\langle RR(\boldsymbol{r}) \rangle = \sum_{i<j}^{K} \langle \nu_i \nu_j \rangle \Theta_{ij}^{\boldsymbol{r}} = \frac{n(n-1)}{K(K-1)} \sum_{i<j}^{K} \Theta_{ij}^{\boldsymbol{r}} \tag{75}$$

where $\Theta_i j^{\boldsymbol{r}}$ is unity if cells i and j are separated by $\boldsymbol{r} \pm \frac{dr}{2}$, and 0 otherwise. Therefore, we define a geometry function $G_p(\boldsymbol{r})$ such that:

$$G_p(\boldsymbol{r}) = \frac{2}{K(K-1)} \sum_{i<j}^{K} \Theta_{ij}^{\boldsymbol{r}} \tag{76}$$

The uncorrelated paircounts thus become:

$$\langle RR(\boldsymbol{r}) \rangle = \frac{n(n-1)}{2} G_p(\boldsymbol{r}) \tag{77}$$

Now let us look at the correlated scenario. We define the correlation function of galaxies (so far, only statistically) as:

$$\langle \nu_i \nu_j \rangle = \frac{n(n-1)}{K(K-1)} (1 + \xi(\boldsymbol{r})) \tag{78}$$

Following the same reasoning as for the uncorrelated case, and taking in mind that the normalisation changes, which can be shown to be $C_n = 1 + \int d^3 r G_p(\boldsymbol{r}) \xi(\boldsymbol{r})$, we can obtain for the correlated paircounts $DD(\boldsymbol{r})$:

$$\langle DD(\boldsymbol{r}) \rangle = \frac{n(n-1)}{2} G_p(\boldsymbol{r}) \frac{1 + \xi(\boldsymbol{r})}{C_n} \tag{79}$$

We immediately see that to get rid of all the geometric effects, we can renormalise the paircounts by the corresponding $n(n-1)$, and write therefore:

$$\frac{DD_{\rm norm}}{RR_{\rm norm}} = \frac{1+\xi(\boldsymbol{r})}{C_n} \tag{80}$$

Assuming the linear order in $\xi(\boldsymbol{r})$ we get $C_n = 1$ and:

$$\xi(\boldsymbol{r}) = \frac{DD_{\rm norm}}{RR_{\rm norm}} - 1 \tag{81}$$

This is the original Peebles estimator of the correlation function [48]. Remembering that galaxies are biased tracers of the matter field, we can connect the two-point correlation function of galaxies with the underlying matter density correlation function $\xi_m(\boldsymbol{r})$ such that:

$$\xi(\boldsymbol{r}) = b_1^2 \xi_m(\boldsymbol{r}) \tag{82}$$

where b_1 is called the linear galaxy bias as defined in equation 53. A more complicated but also accurate bias model will be described later in Chapter 2.

It should be noted, that following [49] the Peebles estimator is not the most optimal one in terms of variance but instead the Landy-Szalay estimator performs better and is written as:

$$\xi(\boldsymbol{r}) = \frac{DD_{\rm norm}(\boldsymbol{r}) - DR_{\rm norm}(\boldsymbol{r}) + RR_{\rm norm}(\boldsymbol{r})}{2RR_{\rm norm}(\boldsymbol{r})} \tag{83}$$

where $DR_{\rm norm}(\boldsymbol{r})$ are the normalised paircounts between the uncorrelated catalogue, called also the randoms, and the correlated catalogue, which represents the galaxies in the survey and therefore is called data.

We note that the 3D-correlation function (and power spectrum as well) are usually not computed in Cartesian coordinates, but rather in spherical ones. To be more precise, as we assume a natural symmetry around the line of sight of the observer, we describe the pair by the distance (also called separation) s between two objects and by their angular separation μ as seen by the observer. In case of the isotropic picture, the angular dependence would not be present, but in realistic scenarios it can be seen from equation 72 that galaxy peculiar velociies make the clustering anisotropic through the RSD effect. We can therefore expand our 2-point correlation function into Legendre multipoles as following:

$$\xi_\ell(s) = \frac{2\ell+1}{2} \int d\mu \xi(s,\mu) \mathcal{L}_\ell(\mu) \tag{84}$$

where $\mathcal{L}_\ell(\mu) = \frac{1}{2^\ell \ell!} \frac{d^\ell}{d\mu^\ell}(\mu^2-1)^\ell$ is the Legendre polynomial. It can be shown that due to quadratic dependence in equation 72, the only relevant multipoles will be those with $\ell = 0, 2, 4$. In general case, one can complexify the model, showing that the information will also leak to odd and higher orderℓ, but so far the effect has been shown to be negligible [50, 51]. ξ_0 is called the monopole, ξ_2 - quadrupole, ξ_4 - hexadecupole.

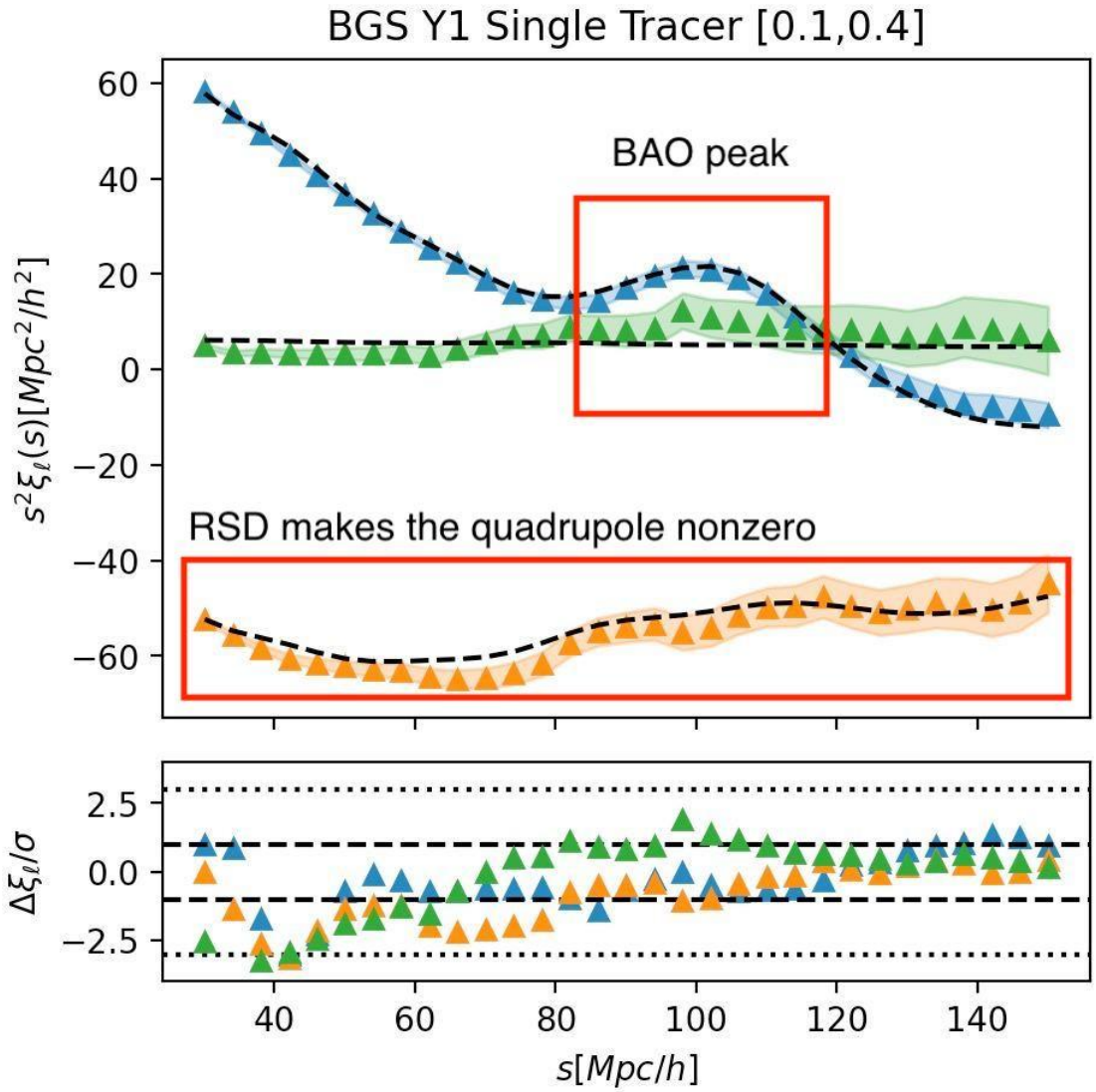

Figure 11: An example of the correlation function multipoles $\xi_\ell(s)$ multiplied by a square of separation s^2 and plotted against separation s, with also best-fit model prediction. Shaded regions represent the errorbars. This specific measurement is actually a mean of 25 measurements from 25 Abacus mocks, which we will describe in Chapter 3. The lower panel shows the difference of measured multipoles from the best-fit values.

We present the mean of 25 measurements of such multipoles taken from the DESI reference cosmological simulations (described in more detail in Chapter 3) in Figure 11. Note the presence of the BAO peak at $\sim$ 100 Mpc/h separation, as well as a very prominent quadrupole, resulting from the RSD. It is those two features that are highly important for the inference of the cosmological data.

Power spectrum estimation

The estimation of the power spectrum, Fourier space counterpart of the correlation function, is computed very differently from the correlation function.

At some point we will need to go to the Fourier space, which is usually done computationally via the use of the Fast Fourier Transform(FFT). The method works on grids. Therefore, we start by "painting" the galaxies onto a cubic grid of chosen size with N_g points, or defining a correspondence between the point of the grid and the averaged value of the field around that point. We will be approximately following [52].

Let us assume a regular grid of points $\boldsymbol{x}_i$, which are linearly spaced by Δx in all directions. The simplest way to paint the density is to sum up the points around each of the points in a cubic volume around the point and divide by the volume. This carries the name as the Nearest-Grid-Point (NGP) assignment scheme. However, that is not the most accurate way to do it. In a general case, we can define the contribution of each point for i-th cell D_i, such that, assuming N_p number of objects:

$$D_i = \sum_{j=1}^{N_p} W(x_i - x_j) \tag{85}$$

We can see now, that NGP scheme would correspond to the following definition of $W(s)$:

$$W^{(\mathrm{NGP})}(s) = \begin{cases} 1 & \text{if } |s| < \frac{1}{2}\Delta x \\ 0 & \text{otherwise} \end{cases} \tag{86}$$

From here, we can introduce other interpolation windows commonly used: Cloud-in-Cell (CIC) and Triangular-Shape-Cloud (TSC), such that their windows are defined as:

$$W^{(\mathrm{CIC})}(s) = \begin{cases} 1 - |s| & \text{if } |s| < \frac{1}{2}\Delta x \\ 0 & \text{otherwise} \end{cases} \tag{87}$$

$$W^{(\mathrm{TSC})}(s) = \begin{cases} \frac{3}{4} - s^2 & \text{if } \frac{1}{2}\Delta x < |s| < \frac{3}{2}\Delta x \\ \frac{1}{2}(\frac{3}{2} - |s|^2) & \text{if } |s| < \frac{1}{2}\Delta x \\ 0 & \text{otherwise} \end{cases} \tag{88}$$

Now we need to connect the following to the $\delta(\boldsymbol{x})$. Let us assume that the object density is presented by $\rho(\boldsymbol{x})$, such that:

$$\rho(\boldsymbol{x}) = \sum_{i=1}^{N_p} m\delta_D(\boldsymbol{x} - \boldsymbol{x}_i) \tag{89}$$

where m is the mass of the object, which we assume to be the same for simplicity for all of the objects present. We will then connect it to the overdensity as following:

$$\delta(\boldsymbol{x}) = \frac{1}{\overline{n}} \sum_{i=1}^{N_p} \delta_D(\boldsymbol{x} - \boldsymbol{x}_i) - 1 \tag{90}$$

where $\overline{n} = \frac{N_p}{V}$ is the global object density.

It can be shown [53] that we can relate the two as:

$$\tilde{\delta}(\boldsymbol{x_i}) = \frac{1}{\overline{n}} D_i - 1 \tag{91}$$

But, we need also to ensure that when transitioning to Fourier space, our estimator still gives the correct value. Going back to our idealised version of the discrete estimator 90, and performing a Fourier transform, we obtain:

$$\delta(\boldsymbol{k}_j) = \frac{1}{\overline{n}} \sum_{i=1}^{N_p} e^{-i\boldsymbol{k}_j \cdot \boldsymbol{x}_i} - \frac{\delta^c(\boldsymbol{k}_j)}{\Delta k^3} \tag{92}$$

where Δk is the fundamental frequency $\Delta k = \frac{2\pi}{L}$, L is the size of the box, and $\delta^c(x)$ is a Kronecker delta function, which is 1 with $x = 0$, otherwise 1.

Going to the power spectrum we obtain:

$$\langle \delta(\boldsymbol{k}_i)\delta(\boldsymbol{k}_j) \rangle = \frac{\delta^c(\boldsymbol{k}_i)\delta^c(\boldsymbol{k}_j)}{\Delta k^3} \left[P(\boldsymbol{k}_i) + \frac{1}{\overline{n}} \right] \tag{93}$$

The term $\frac{1}{\overline{n}}$ is called a shot noise term, and is the consequence of the discreteness of the tracer of the matter density.

However, it is not as simple as with 91, because the smoothing window function we have introduced affects the Fourier transform, so it needs to be corrected, resulting in:

$$\delta(\boldsymbol{k}_i) + \sum_{n \neq 0} w_n(\boldsymbol{k}_i)\delta(\boldsymbol{k}_i - \boldsymbol{n}k_s) \tag{94}$$

where $\boldsymbol{n}$ is an integer vector, $k_s = \frac{2\pi}{\Delta x}$ and w_n is defined such that:

$$w_n(\boldsymbol{k}) = \frac{W(\boldsymbol{k} - \boldsymbol{n}k_s)}{W(\boldsymbol{k})} \tag{95}$$

where $W(\boldsymbol{k})$ is a Fourier-space counterpart of the smoothing window function $W(\boldsymbol{x})$.

In order to fix that, a technique called interlacing is applied, whose details can be found in [54].

The approach we described might work for a cubic box, but assuming a realistic survey, we end up with two problems. First of all, the survey shape is not a cube, but has an irregular shape, and might even have a non constant $\overline{n}$.

We will intuitively do something similar to what we did for the correlation function, and compare a catalogue with correlations to the uncorrelated one, meaning randoms with density $\rho_r(\boldsymbol{x})$, such that there are $1/\alpha$ object more in random catalog that in the data, assuming the two cover the same volume.

We define the FKP estimator, defined in [55] as:

$$F(\boldsymbol{x}) = \frac{w(\boldsymbol{x})[\rho(\boldsymbol{x}) - \rho_r(\boldsymbol{x})]}{\left[\int d^3x\overline{n}^2(\boldsymbol{x})w^2(\boldsymbol{x})\right]} \tag{96}$$

where $w(\boldsymbol{x})$ is a weight function to be described later. Following [55], we get an estimator for the power spectrum to be:

$$P(\boldsymbol{k}) = \frac{1}{\int d^3x\overline{n}(\boldsymbol{x})w^2(\boldsymbol{x})}\int \frac{d\Omega}{4\pi}\left[d^3x_1d^3x_2F(\boldsymbol{x}_1)F(\boldsymbol{x}_2)e^{i\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}\right] - P_{\text{shot}} \tag{97}$$

where $P_{\text{shot}} = \frac{(1+\alpha)\int d^3x\overline{n}(\boldsymbol{x})w^2(\boldsymbol{x})}{\int d^3x\overline{n}(\boldsymbol{x})w^2(\boldsymbol{x})}$ is a shot noise term.

This estimator is unbiased with a small variance, however, there is a computational difficulty. As with the correlation function, power spectrum is also presented in the form of Legendre multipoles. With the problem, that the decomposition is presented in the configuration space, which is written using the FKP estimator as:

$$P_\ell(k) = \frac{2\ell+1}{\int d^3x\overline{n}(\boldsymbol{x})w^2(\boldsymbol{x})}\int \frac{d\Omega}{4\pi}\left[d^3x_1d^3x_2F(\boldsymbol{x}_1)F(\boldsymbol{x}_2)e^{i\boldsymbol{k}\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}\mathcal{L}_\ell(\hat{k}\cdot\hat{x_h})\right] \tag{98}$$

where $\boldsymbol{x_h} = \frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}$ and Ω is the solid angle. However, one can approximate $x_h \approx x_1$, and then define an FKP field with the Legendre multipole directly such that:

$$F_\ell(k) = \int \frac{d^3x}{(2\pi)^3}e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}F(\boldsymbol{x})\mathcal{L}_\ell(\hat{k}\cdot\hat{x}) \tag{99}$$

The last step is to decompose the Legendre polynomials into spherical harmonics $Y_{\ell m}$ such that:

$$\mathcal{L}_\ell(\hat{k}\cdot\hat{x}) = \frac{4\pi}{2\ell+1}\sum_{m=-\ell}^{\ell}Y_{\ell,m}(\hat{k})Y*_{\ell,m}(\hat{x}) \tag{100}$$

It allows us to decouple the term with configuration space component into a separate multiplier. Meaning that we will need to paint the density once on the mesh, and then Fourier transform it once for the monopole, and $2\ell+1$ times for each ℓ-th multipole, one time for each of the corresponding harmonic. Then we will just need to multiply the result by the Fourier space harmonic. The final integration over the angles to get to the power spectrum multipole estimator can be performed already in Fourier space:

$$P_\ell(k) = (2\ell+1)\int \frac{d\Omega}{4\pi}F_\ell(k)F_0(-k) \tag{101}$$

where $F_0 = F_{l=0}$. The obtained estimator carries the name of Yamamoto estimator[56].

Another important topic to be mentioned when talking about power spectrum estimation is the window function. Before, we have assumed the negligible geometric effects, roughly corrected by the introduction of the randoms to compare to. However, as the power spectrum is computed with a box encompassing the survey, in the Fourier space the so-called window effects (do not mix up with the smoothing window) arise once again.

There are two ways to correct them. One would be to estimate the window function and apply it to the output from the modelling, or deconvolve the power spectrum [57]. However, this is out of the scope of this thesis as we will mainly focus on configuration space analysis, so we will not discuss it much further.

FKP weights

One of the important effects needed to be taken into account when analysing spectroscopic surveys is the non-homogeneity of the average number density. In general, the $\overline{n}$ in the actual observed data would not be constant over the volume, but would rather depend on various factors, most importantly, redshift, which will make different subvolumes give different contributions to the variance of the measured statistics. This is the case both for the correlation function and the power spectrum. As it was shown in [55], this can be accounted for by using the weights $w(\boldsymbol{x}) = w_{\rm FKP}(\boldsymbol{x})$ such that:

$$w_{\rm FKP} = \frac{1}{1 + \overline{n}(z) P_0} \tag{102}$$

where $\overline{n}(z)$ is the number density evolving with redshift z, and P_0 is the value of the fiducial power spectrum, which is usually taken to be constant. We will discuss more in Chapter 5 how to generalise this expression to the situation of several species of objects present in the survey.

General inference scheme

In the previous sections we have presented a way to model the large-scale structures of the Universe analytically, as well as a way to predict the behaviour of matter density perturbations and their connection to the formation of galaxies. At the same time, we have defined the two-point clustering statistics, namely the two-point correlation function and the power spectrum from the realistic 3D maps of galaxies.

How do we properly compare the modelling and the measurements of the galaxy two-point clustering statistics?

Let us assume that the data, for example the correlation function multipoles can be described as a multivariate random vector ξ_i of size n, where we concatenated all the multipoles taken as non-negligible (usually $\ell = 0, 2, 4$) into one data vector.

Then, we define the probability of our data sample being described by a theoretical model $\xi_i^{\text{th}}(\boldsymbol{\theta})$ with parameters θ_k (which can be a set of cosmological and nuisance parameters, such as galaxy biases and additional theoretical terms, needed to model the correlation function) as $L(\boldsymbol{\theta}|\boldsymbol{\xi})$, which we will call the likelihood. It should be noted, that likelihood represents a conditional probability, thus Bayes theorem can be applied, which in the more general case of events A and B states [58]:

$$P(A|B)P(B) = P(B|A)P(A) \tag{103}$$

Where $P(X)$ is the probability of X, and $P(A|B)$ is the probability of A happening provided B is true. We can rewrite this equation, under the assumption that none of the probabilities is zero, as $P(A|B) = P(B|A)P(A)/P(B)$. In this analogy, $P(A|B)$ is called the posterior probability, $P(B|A)$ is the likelihood in this case, $P(A)$ and $P(B)$ are the unconditional probabilities of observing A and B respectively, and are known as prior probability and marginal probability.

Assuming that $\boldsymbol{\xi}$ is distributed as a multivariate Gaussian distribution with a covariance matrix Σ it can be shown [59] that:

$$\log L(\boldsymbol{\theta}|\boldsymbol{\xi}) = -\frac{1}{2}\sum_{i,j=1}^{n}(x_i - x_i^{\text{th}}(\boldsymbol{\theta}))\left(\Sigma^{-1}\right)^{ij}(x_j - x_j^{\text{th}}(\boldsymbol{\theta})) + \text{const} \tag{104}$$

We introduce also the quantity of χ^2 such that:

$$\chi^2 = \sum_{i,j=1}^{n}(x_i - x_i^{\text{th}}(\boldsymbol{\theta}))\left(\Sigma^{-1}\right)^{ij}(x_j - x_j^{\text{th}}(\boldsymbol{\theta})) = -2\log L(\boldsymbol{\theta}|\boldsymbol{\xi}) + \text{const} \tag{105}$$

which is a test statistic, describing how well does the given distirbution with the given set of parameters describe the obtained random values [60].

We can see now, that in order to maximize the likelihood, we need to find the values of $\boldsymbol{\theta}$ that minimize χ^2. The uncertainties are obtained by finding the surface such that if we define $\chi^2_{\text{best fit}}$ as the minimal value of $\chi 2$, we search for a hypersurface such that $\chi^2 = \chi^2_{\text{best fit}} + 1$, which will represent the 1σ confidence interval for the value [59]. There are various ways to do that. One can go with the Frequentist approach, assuming the existence of the exact "true" values of parameters of the distribution, and by descending following the gradient in a smart way [61], get to the minimum. We will use `iminuit`[61] in order to perform that type of inference. Another commonly approach is to employ the Bayesian inference and in particular Markov chain Monte Carlo (MCMC), which consists in building such a Markov chain that its equilibrium distribution (where equilibrium is when of values on each step are identical to the ones on the previous step) matches the target distribution, meaning that once the chain reaches equilibrium, it will emulate the target distribution allowing us to obtain an estimate of the distribution by sampling enough events. One of the classic algorithms to perform such inference is Metropolis-Hastings

algorithm[62, 63], for example. Throughout this thesis we will use however a more complex approach provided by the `emcee` package [64].

Cosmic variance

Following the classical definition, first we quantify the amount of information obtained from a specific measurement with probability of the model parameters p_i describing the data y_i as $L(y_i|p_i)$:

$$F_{ij} = -\left\langle \frac{\partial^2 \log L}{\partial p_i \partial p_j} \right\rangle \tag{106}$$

This quantity is often called the Fisher information.

It was shown that for single tracer analysis (where we analyse one single 2-point correlation function or power spectrum) modelled by the Kaiser formula for RSD (72), the Fisher matrix can be expressed as [65]:

$$F_{ij} = \int \frac{d^3k d^3x}{(2\pi)^3} \frac{d\log\mathcal{P}}{dp^i} \frac{d\log\mathcal{P}}{dp^j} \frac{1}{2} \left(\frac{\mathcal{P}}{1+\mathcal{P}} \right)^2 \tag{107}$$

where $\mathcal{P}$ is the effective power defined by $\mathcal{P} = \overline{n}(\boldsymbol{x}) \left[b + f(z)\mu_k^2 \right]^2 P(z,k)$, b is the linear bias and f is the linear growth rate. We can look at the Fisher information density $F(\boldsymbol{k}, \boldsymbol{x})$, defined as:

$$F(\boldsymbol{k}, \boldsymbol{x}) = \frac{1}{2} \left(\frac{\mathcal{P}}{1+\mathcal{P}} \right)^2 \tag{108}$$

We can notice that $F(\boldsymbol{k}, \boldsymbol{x}) < \frac{1}{2}$, which means that the Fisher information per volume of the phase-space is therefore limited.

The Rao-Cramer theorem states that [66, 67]:

$$\sigma^2(p_i) \geq \frac{1}{F_{ii}^2} \tag{109}$$

where $\sigma^2(p_i)$ is the uncertainty on the parameter p_i. The theorem means that this uncertainty is therefore limited by the information content of the survey. So there is an upper limit on how much information can be contained in a limited volume, such that even with increasing statistics inside the volume, at some point we will reach the limit of allowed information from single tracer analysis based on the 2-point statistics. It thus represents a fundamental limit in the precision we can achieve on a given parameter for single tracer standard analysis. This limit is also called cosmic variance in the literature.

Multitracer analysis

As was discussed earlier in the, one of the very fundamentals constraints on the standard full-shape analysis is cosmic variance. However, it was discovered recently, that it can

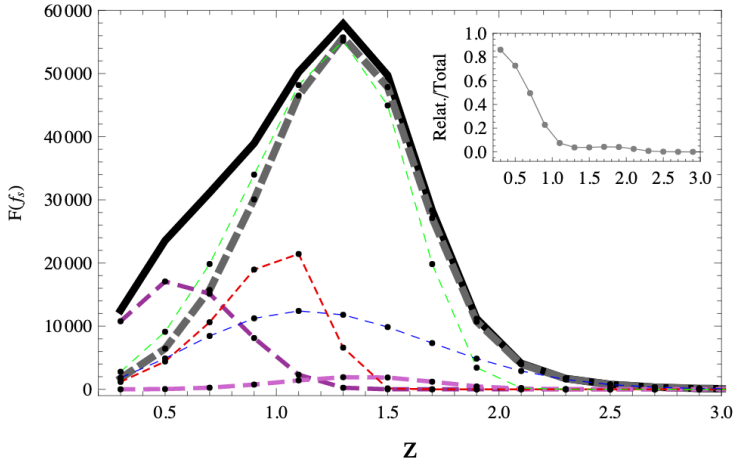

Figure 12: Information density plot for three tracers, their cross-correlations and a combined tracer taken from [69]

be circumvented by a multitracer analysis. It consists of using several tracers of varying clustering properties. We will start with showing where does the information come from.

Assume having two tracers, and correspondingly two autospectra from those. How do we define the most optimal estimator for those?

One can go back and generalise the Fisher matrix for the multi-tracer estimator, following [68]:

$$F_{\mu\nu;ij} = \sum_{\alpha\beta\gamma\sigma} \int d^3x d^3x' d^3x'' d^3x''' C^{-1}_{\alpha\beta}(x,x') \frac{\partial C_{\beta\gamma}(x',x'')}{\partial P_{\mu,i}} \times \\ \times C^{-1}_{\gamma\sigma}(x'',x''') \frac{\partial C_{\sigma\alpha}(x''',x)}{\partial P_{\nu,j}} \tag{110}$$

We can then proceed in the same way as we did earlier to obtain the Fisher information density in terms of the power spectrum:

$$\mathcal{F}_{\mu\nu}(x,k) = \frac{1}{4} \frac{\delta_{\mu\nu}\mathcal{P}_\mu \mathcal{P}(1+\mathcal{P}) + \mathcal{P}_\mu \mathcal{P}_\nu (1-\mathcal{P})}{(1+\mathcal{P})^2} \tag{111}$$

From here one can already notice how can cosmic variance be beaten by this approach. It can be further illustrated by the plot on Fig. 12.

Moving on in the same fashion as we derived the FKP estimator for the power spectrum we can now also figure out the multitracer weights, which maximize the obtained information:

$$w_{\sigma\alpha}(x,k) = \left[\delta_{\sigma\alpha} - \frac{\mathcal{P}_\sigma(x,k)}{1+\mathcal{P}(x,k}\right] \bar{n}_\alpha b_\alpha(x,k) \tag{112}$$

With $b_\alpha(x,k)$ being the effective bias. Such that the weighted density contrast becomes:

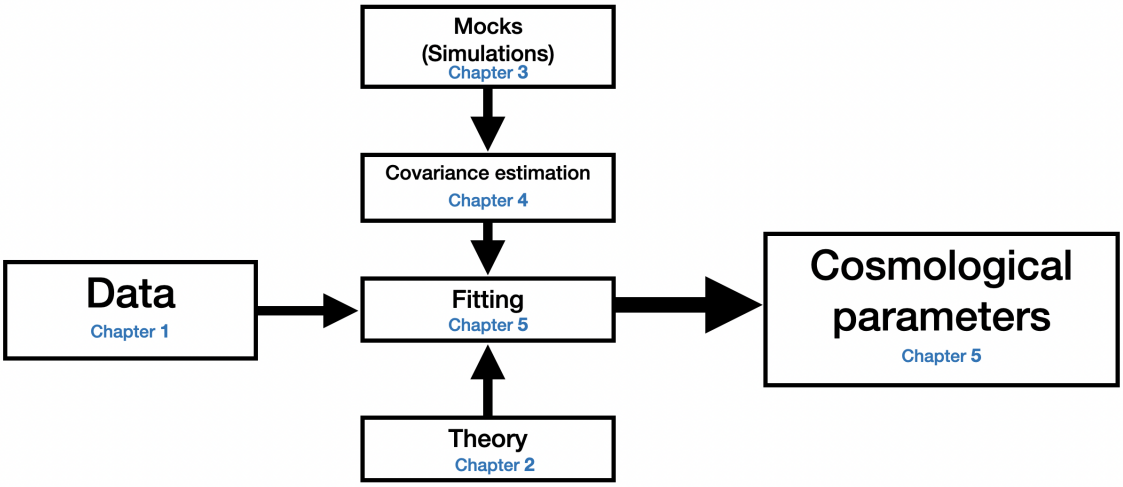

Figure 13: A diagram representing different parts of the analysis pipeline for the inference of the cosmological parameters from data with corresponding chapters of this thesis indicated.

$$f_\sigma(x,k) = \sum_\alpha w_{\sigma\alpha}(x,k)\delta_\alpha(x) \tag{113}$$

One might notice that essentially tracers interact with each other, as they actually exist in the same volume of space-time and are not completely independent, and that it is represented in the weighting scheme. In practice, we used the $\bar{n}$ of one tracer stretched over the counts of the other tracer to generate the corresponding tracers, with the addition of a normalization factor in order to input the data into the usual Yamamoto estimator.

Objectives of the thesis

Now we can build a scheme of the work needed to measure the cosmological parameters from a general survey, which is presented in a diagrammatic form in figure 13.

In Chapter 1 we will discuss shortly the instrument needed to obtain the positions of the galaxies to create a 3D map of the Universe: the Dark Energy Spectroscopic Instrument (DESI), and we will present our target dataset of bright galaxies: Bright Galaxy Survey (BGS).

In Chapter 2 we will go deeper into the theoretical modelling of the two-point statistics, introduce further order terms and significantly speed up the practical computation using machine learning techniques.

In Chapter 3 we will present different techniques for the numerical simulations of the Universe, which will allow us to study systematics and will serve us as realizations of the Universe for error estimation, and we will explain why the massive mock production for

BGS was unfeasible.

In Chapter 4 we will dive further into the errorbar estimation, where we will go over different approaches to obtain the covariance matrix, including the Fitcov approach developed specifically for the complicated case of BGS and we will explain the problem of cosmic variance and why it is so present for BGS. Finally, we will present different covariance matrices created by us for the analysis of the data,

Finally, in Chapter 5 we will present in more detail the 2-point full-shape analysis, we will discuss the problems of compression and present the multitracer approach, which allows to bypass the cosmic variance. Then we will present the results of massive tests of various tools (covariances, models, simulations) created by us, and we will finally reach the last point: we will present the results of the multi tracer analysis of the BGS DR1 data with FitCov covariance and NN-powered theoretical modelling, and compare it to the more traditional approach.

References

[1] B. A. Dubrovin et al. *Modern Geometry — Methods and Applications*. Springer New York, 1992. ISBN: 9781461243984. DOI: 10.1007/978-1-4612-4398-4. URL: http://dx.doi.org/10.1007/978-1-4612-4398-4.

[2] L.D. LANDAU and E.M. LIFSHITZ. *Mechanics*. Ed. by L.D. LANDAU and E.M. LIFSHITZ. Butterworth-Heinemann.

[3] Valery Rubakov. *Classical Theory of Gauge Fields*. Princeton University Press, Dec. 2009. ISBN: 9781400825097. DOI: 10.1515/9781400825097. URL: http://dx.doi.org/10.1515/9781400825097.

[4] M. Ostrogradsky. "Mémoires sur les équations différentielles, relatives au problème des isopérimètres". In: *Mem. Acad. St. Petersbourg* 6.4 (1850), pp. 385–517.

[5] H. Neill. "Concepts of modern mathematics, by Ian Stewart. Pp ix, 315. 80p. 1975. SBN 014 021849 1 (Penguin)". In: *The Mathematical Gazette* 61.415 (Mar. 1977), pp. 65–65. ISSN: 2056-6328. DOI: 10.2307/3617450. URL: http://dx.doi.org/10.2307/3617450.

[6] David Lovelock. "The Four-Dimensionality of Space and the Einstein Tensor". In: *Journal of Mathematical Physics* 13.6 (June 1972), pp. 874–876. ISSN: 1089-7658. DOI: 10.1063/1.1666069. URL: http://dx.doi.org/10.1063/1.1666069.

[7] Scott Dodelson. *Modern Cosmology*. Elsevier, 2003, pp. 180–215. ISBN: 9780122191411. DOI: 10.1016/b978-012219141-1/50026-7. URL: http://dx.doi.org/10.1016/b978-012219141-1/50026-7.

[8] A. Friedman. "Uber die Krummung des Raumes". In: *Zeitschrift fur Physik* 10.1 (Dec. 1922), pp. 377–386. ISSN: 1434-601X. DOI: 10.1007/bf01332580. URL: http://dx.doi.org/10.1007/BF01332580.

[9] Edwin Hubble. "A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae". In: *Proceedings of the National Academy of Science* 15.3 (Mar. 1929), pp. 168–173. DOI: 10.1073/pnas.15.3.168.

[10] J. Alberto Vazquez Gonzalez et al. "Inflationary cosmology: from theory to observations". In: *Revista Mexicana de Física E* 17.1 Jan-Jun (Jan. 2020), pp. 73–91. ISSN: 1870-3542. DOI: 10.31349/revmexfise.17.73. URL: http://dx.doi.org/10.31349/RevMexFisE.17.73.

[11] Daniel Baumann. "Primordial Cosmology". In: Sept. 2018, p. 009. DOI: 10.22323/1.305.0009.

[12] Adam G. Riess et al. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". In: *The Astronomical Journal* 116.3 (Sept. 1998), p. 1009. DOI: 10.1086/300499. URL: https://dx.doi.org/10.1086/300499.

[13] S. Perlmutter et al. "Cosmology from Type Ia supernovae". In: *Bull. Am. Astron. Soc.* 29 (1997), p. 1351. arXiv: astro-ph/9812473.

[14] Kyle S. Dawson et al. "The Baryon Oscillation Spectroscopic Survey of SDSS-III". In: 145.1, 10 (Jan. 2013), p. 10. DOI: 10.1088/0004-6256/145/1/10. arXiv: 1208.0022 [astro-ph.CO].

[15] J. T. A. de Jong et al. "The Kilo-Degree Survey". In: *The Messenger* 154 (Dec. 2013), pp. 44–46.

[16] Kyle S. Dawson et al. "The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data". In: 151.2, 44 (Feb. 2016), p. 44. DOI: 10.3847/0004-6256/151/2/44. arXiv: 1508.04473 [astro-ph.CO].

[17] Naoyuki Tamura et al. "Prime Focus Spectrograph (PFS) for the Subaru telescope: overview, recent progress, and future perspectives". In: ed. by Christopher J. Evans et al. SPIE, Aug. 2016. DOI: 10.1117/12.2232103. URL: http://dx.doi.org/10.1117/12.2232103.

[18] Karl Gebhardt et al. "The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Survey Design, Reductions, and Detections". In: 923.2, 217 (Dec. 2021), p. 217. DOI: 10.3847/1538-4357/ac2e03. arXiv: 2110.04298 [astro-ph.IM].

[19] DESI Collaboration et al. "The DESI Experiment Part I: Science,Targeting, and Survey Design". In: *arXiv e-prints*, arXiv:1611.00036 (Oct. 2016), arXiv:1611.00036. arXiv: 1611.00036 [astro-ph.IM].

[20] Ž. Ivezić et al. "LSST: From Science Drivers to Reference Design and Anticipated Data Products". In: 873, 111 (Mar. 2019), p. 111. DOI: 10.3847/1538-4357/ab042c. arXiv: 0805.2366.

[21] Euclid Collaboration et al. *Euclid. I. Overview of the Euclid mission*. 2024. arXiv: 2405.13491 [astro-ph.CO].

[22] D. Spergel et al. 2013. arXiv: 1305.5422 [astro-ph.IM].

[23] Thibaut Louis et al. "The Atacama Cosmology Telescope: two-season ACTPol spectra and parameters". In: *JCAP* 2017.06 (June 2017), pp. 031–031. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2017/06/031. URL: http://dx.doi.org/10.1088/1475-7516/2017/06/031.

[24] J. M. Lamarre et al. "The Planck High Frequency Instrument, a third generation CMB experiment, and a full sky submillimeter survey". In: 47.11-12 (Dec. 2003), pp. 1017–1024. DOI: 10.1016/j.newar.2003.09.006. arXiv: astro-ph/0308075 [astro-ph].

[25] Thomas Prince and Zwicky Transient Facility (ZTF) Project Team. "The Zwicky Transient Facility Galactic Plane Survey". In: *AAS Meeting Abstracts #231*. Vol. 231. AAS Meeting Abstracts. Jan. 2018, 348.18, p. 348.18.

[26] The Dark Energy Survey Collaboration. *The Dark Energy Survey*. 2005. arXiv: astro-ph/0510346 [astro-ph].

[27] David H. Weinberg et al. "Observational probes of cosmic acceleration". In: *Physics Reports* 530.2 (Sept. 2013), pp. 87–255. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2013.05.001. URL: http://dx.doi.org/10.1016/j.physrep.2013.05.001.

[28] Ariel Goobar and Bruno Leibundgut. "Supernova Cosmology: Legacy and Future". In: *Annual Review of Nuclear and Particle Science* 61.Volume 61, 2011 (2011), pp. 251–279. ISSN: 1545-4134. DOI: https://doi.org/10.1146/annurev-nucl-102010-130434. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-nucl-102010-130434.

[29] R. A. Sunyaev and Ya. B. Zel'dovich. "Microwave Background Radiation as a Probe of the Contemporary Structure and History of the Universe". In: *ARAA* 18.1 (Sept. 1980), pp. 537–560. ISSN: 1545-4282. DOI: 10.1146/annurev.aa.18.090180.002541. URL: http://dx.doi.org/10.1146/annurev.aa.18.090180.002541.

[30] Rachel Mandelbaum. "Weak Lensing for Precision Cosmology". In: *ARAA* 56.Volume 56, 2018 (2018), pp. 393–433. ISSN: 1545-4282. DOI: https://doi.org/10.1146/annurev-astro-081817-051928. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-astro-081817-051928.

[31] Steven W. Allen et al. "Cosmological Parameters from Observations of Galaxy Clusters". In: *Annual Review of Astronomy and Astrophysics* 49.Volume 49, 2011 (2011), pp. 409–470. ISSN: 1545-4282. DOI: https://doi.org/10.1146/annurev-astro-081710-102514. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-astro-081710-102514.

[32] Masami Ouchi et al. "Observations of the Lyman-α Universe". In: 58 (Aug. 2020), pp. 617–659. DOI: 10.1146/annurev-astro-032620-021859. arXiv: 2012.07960 [astro-ph.GA].

[33] F. Bernardeau et al. "Large-scale structure of the Universe and cosmological perturbation theory". In: *Physics Reports* 367.1 (2002), pp. 1–248. ISSN: 0370-1573. DOI: https://doi.org/10.1016/S0370-1573(02)00135-7. URL: https://www.sciencedirect.com/science/article/pii/S0370157302001357.

[34] F. R. Bouchet. *Introductory Overview of Eulerian and Lagrangian Perturbation Theories*. 1996. DOI: 10.48550/ARXIV.ASTRO-PH/9603013. URL: https://arxiv.org/abs/astro-ph/9603013.

[35] T. Buchert. "A class of solutions in Newtonian cosmology and the pancake theory". In: 223.1-2 (Oct. 1989), pp. 9–24.

[36] F. Moutarde et al. "Precollapse Scale Invariance in Gravitational Instability". In: 382 (Dec. 1991), p. 377. DOI: 10.1086/170728.

[37] E. Hivon et al. "Redshift distortions of clustering: a Lagrangian approach." In: 298 (June 1995), p. 643. DOI: 10.48550/arXiv.astro-ph/9407049. arXiv: astro-ph/9407049 [astro-ph].

[38] A. N. Taylor and A. J. S. Hamilton. "Non-linear cosmological power spectra in real and redshift space". In: *MNRAS* 282.3 (Oct. 1996), pp. 767–778. ISSN: 0035-8711. DOI: 10.1093/mnras/282.3.767. URL: https://doi.org/10.1093/mnras/282.3.767.

[39] Daniel J. Eisenstein and Wayne Hu. "Power Spectra for Cold Dark Matter and Its Variants". In: *The Astrophysical Journal* 511.1 (Jan. 1999), pp. 5–15. ISSN: 1538-4357. DOI: 10.1086/306640. URL: http://dx.doi.org/10.1086/306640.

[40] Diego Blas et al. "The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes". In: *JCAP* 2011.07 (July 2011), p. 034. DOI: 10.1088/1475-7516/2011/07/034. URL: https://dx.doi.org/10.1088/1475-7516/2011/07/034.

[41] Antony Lewis and Sarah Bridle. "Cosmological parameters from CMB and other data: A Monte Carlo approach". In: 66 (2002), p. 103511. DOI: 10.1103/PhysRevD.66.103511. arXiv: astro-ph/0205436 [astro-ph].

[42] Daniel J. Eisenstein and Wayne Hu. "Baryonic Features in the Matter Transfer Function". In: 496.2 (Mar. 1998), pp. 605–614. DOI: 10.1086/305424. arXiv: astro-ph/9709112 [astro-ph].

[43] Daniel J. Eisenstein et al. "Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak". In: 664.2 (Aug. 2007), pp. 675–679. DOI: 10.1086/518712. arXiv: astro-ph/0604362 [astro-ph].

[44] P. Andersen et al. "Cosmology with peculiar velocities: observational effects". In: *MNRAS* 463.4 (Sept. 2016), pp. 4083–4092. ISSN: 0035-8711. DOI: 10.1093/mnras/stw2252. eprint: https://academic.oup.com/mnras/article-pdf/463/4/4083/18514487/stw2252.pdf. URL: https://doi.org/10.1093/mnras/stw2252.

[45] Nick Kaiser. "Clustering in real space and in redshift space". In: *MNRAS* 227.1 (July 1987), pp. 1–21. ISSN: 0035-8711. DOI: 10.1093/mnras/227.1.1. eprint: https://academic.oup.com/mnras/article-pdf/227/1/1/18522208/mnras227-0001.pdf. URL: https://doi.org/10.1093/mnras/227.1.1.

[46] Eric V. Linder and Robert N. Cahn. "Parameterized beyond-Einstein growth". In: *Astroparticle Physics* 28.4 (2007), pp. 481–488. ISSN: 0927-6505. DOI: https://doi.org/10.1016/j.astropartphys.2007.09.003. URL: https://www.sciencedirect.com/science/article/pii/S0927650507001326.

[47] C. Alcock and B. Paczynski. "An evolution free test for non-zero cosmological constant". In: 281 (Oct. 1979), p. 358. DOI: 10.1038/281358a0.

[48] P. J. E. Peebles. *The large-scale structure of the universe*. 1980.

[49] Stephen D. Landy and Alexander S. Szalay. "Bias and Variance of Angular Correlation Functions". In: 412 (July 1993), p. 64. DOI: 10.1086/172900.

[50] Enea Di Dio and Uroš Seljak. "The relativistic dipole and gravitational redshift on LSS". In: *JCAP* 2019.04 (Apr. 2019), p. 050. DOI: 10.1088/1475-7516/2019/04/050. URL: https://dx.doi.org/10.1088/1475-7516/2019/04/050.

[51] Florian Beutler and Enea Di Dio. "Modeling relativistic contributions to the halo power spectrum dipole". In: *JCAP* 2020.07 (July 2020), pp. 048–048. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2020/07/048. URL: http://dx.doi.org/10.1088/1475-7516/2020/07/048.

[52] Weiguang Cui et al. "An Ideal Mass Assignment Scheme for Measuring the Power Spectrum with Fast Fourier Transforms". In: *The Astrophysical Journal* 687.2 (Nov. 2008), p. 738. DOI: 10.1086/592079. URL: https://dx.doi.org/10.1086/592079.

[53] E. Sefusatti et al. "Accurate estimators of correlation functions in Fourier space". In: *MNRAS* 460.4 (May 2016), pp. 3624–3636. ISSN: 1365-2966. DOI: 10.1093/mnras/stw1229. URL: http://dx.doi.org/10.1093/mnras/stw1229.

[54] R.W Hockney and J.W Eastwood. *Computer Simulation Using Particles*. CRC Press, Mar. 2021. ISBN: 9780367806934. DOI: 10.1201/9780367806934. URL: http://dx.doi.org/10.1201/9780367806934.

[55] Hume A. Feldman et al. "Power-Spectrum Analysis of Three-dimensional Redshift Surveys". In: 426 (May 1994), p. 23. DOI: 10.1086/174036. arXiv: astro-ph/9304022 [astro-ph].

[56] Kazuhiro Yamamoto et al. "A Measurement of the Quadrupole Power Spectrum in the Clustering of the 2dF QSO Survey". In: *Publications of the Astronomical Society of Japan* 58.1 (Feb. 2006), pp. 93–102. ISSN: 0004-6264. DOI: 10.1093/pasj/58.1.93. eprint: https://academic.oup.com/pasj/article-pdf/58/1/93/23993271/pasj58-0093.pdf. URL: https://doi.org/10.1093/pasj/58.1.93.

[57] Takahiro Sato et al. "Window effect in the power spectrum analysis of a galaxy redshift survey". In: *Int. J. Astron. Astrophys.* 3 (2013), pp. 243–256. DOI: 10.4236/ijaa.2013.33029. arXiv: 1308.3551 [astro-ph.CO].

[58] T. Bayes. "An essay towards solving a problem in the doctrine of chances". In: *Phil. Trans. of the Royal Soc. of London* 53 (1763), pp. 370–418.

[59] Robert D. Cousins. *Lectures on Statistics in Theory: Prelude to Statistics in Practice*. 2018. DOI: 10.48550/ARXIV.1807.05996. URL: https://arxiv.org/abs/1807.05996.

[60] William G. Cochran. "The χ^2 Test of Goodness of Fit". In: *AMS* 23.3 (1952), pp. 315–345. DOI: 10.1214/aoms/1177729380. URL: https://doi.org/10.1214/aoms/1177729380.

[61] Hans Dembinski and Piti Ongmongkolkul et al. "scikit-hep/iminuit". In: (Dec. 2020). DOI: 10.5281/zenodo.3949207. URL: https://doi.org/10.5281/zenodo.3949207.

[62] Nicholas Metropolis et al. *Equation of state calculations by fast computing machines*. Mar. 1953. DOI: 10.2172/4390578. URL: http://dx.doi.org/10.2172/4390578.

[63] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109. ISSN: 00063444. URL: http://www.jstor.org/stable/2334940 (visited on 05/27/2024).

[64] Daniel Foreman-Mackey et al. "¡tt¿emcee¡/tt¿: The MCMC Hammer". In: *PASP* 125.925 (Mar. 2013), pp. 306–312. ISSN: 1538-3873. DOI: 10.1086/670067. URL: http://dx.doi.org/10.1086/670067.

[65] Max Tegmark. "Measuring Cosmological Parameters with Galaxy Surveys". In: *Phys. Rev. Lett.* 79 (20 Nov. 1997), pp. 3806–3809. DOI: 10.1103/PhysRevLett.79.3806. URL: https://link.aps.org/doi/10.1103/PhysRevLett.79.3806.

[66] C. Radhakrishna Rao. "Information and the Accuracy Attainable in the Estimation of Statistical Parameters". In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Ed. by Samuel Kotz and Norman L. Johnson. New York, NY: Springer New York, 1992, pp. 235–247. ISBN: 978-1-4612-0919-5. DOI: 10.1007/978-1-4612-0919-5_16. URL: https://doi.org/10.1007/978-1-4612-0919-5_16.

[67] Harald Cramér. *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton: Princeton University Press, 1946. ISBN: 9781400883868. DOI: doi:10.1515/9781400883868. URL: https://doi.org/10.1515/9781400883868.

[68] L. Raul Abramo et al. "Fourier analysis of multitracer cosmological surveys". In: *Mon. Not. R. Astron. Soc.* 455.4 (Feb. 2016), pp. 3871–3889. ISSN: 0035-8711. DOI: 10.1093/mnras/stv2588. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stv2588.

[69] L. R. Abramo and K. E. Leonard. "Why multitracer surveys beat cosmic variance". In: *Mon. Not. R. Astron. Soc.* 432.1 (June 2013), pp. 318–326. ISSN: 0035-8711. DOI: 10.1093/mnras/stt465. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stt465.

Chapter 1

DESI

О, пыль миров! О, рой
священных пчёл!
Я исследил, измерил, взвесил,
счёл, —
Дал имена, составил карты,
сметы...

V. Voloshin, Translation

Introduction

There are two categories of optical galaxy surveys: photometric and spectroscopic ones. The former observe the objects in the sky with a given magnitude limit, that depends on the depth of the survey, with different filters. However, the redshifts obtained for such galaxies are less precise than spectroscopic ones[1]. The spectroscopic surveys measure the spectra of observed objects. That does take longer, and as a result, the spectroscopic surveys feature less objects. However, for each galaxy observed we have a much more precise redshift, which allows us to build accurate 3D maps of the observable Universe.

The Dark Energy Spectroscopic Instrument (DESI, [2, 3]) is a fourth-generation spectroscopic survey, which I have been working on during this thesis work. It is a robotic fibre-fed highly multiplexed spectroscopic instrument. It is installed on the 4-meter Mayall telescope at Kitt Peak National Observatory in Arizona. Each of 5000 fibres is able to observe a separate spectrum, thus allowing the instrument to gather 5000 spectra at a time, covering for one exposure around 3deg^2 part of the sky. Currently, DESI aims to cover 14000deg^2 with 40 million objects observed within 5 years of observation. It started its survey operations in May 2021 and 2 more years of active survey data acquisition is left by the time of writing this thesis. This thesis work uses 1-year of DESI observation, called the Data Release 1 (DR1) hereafter.

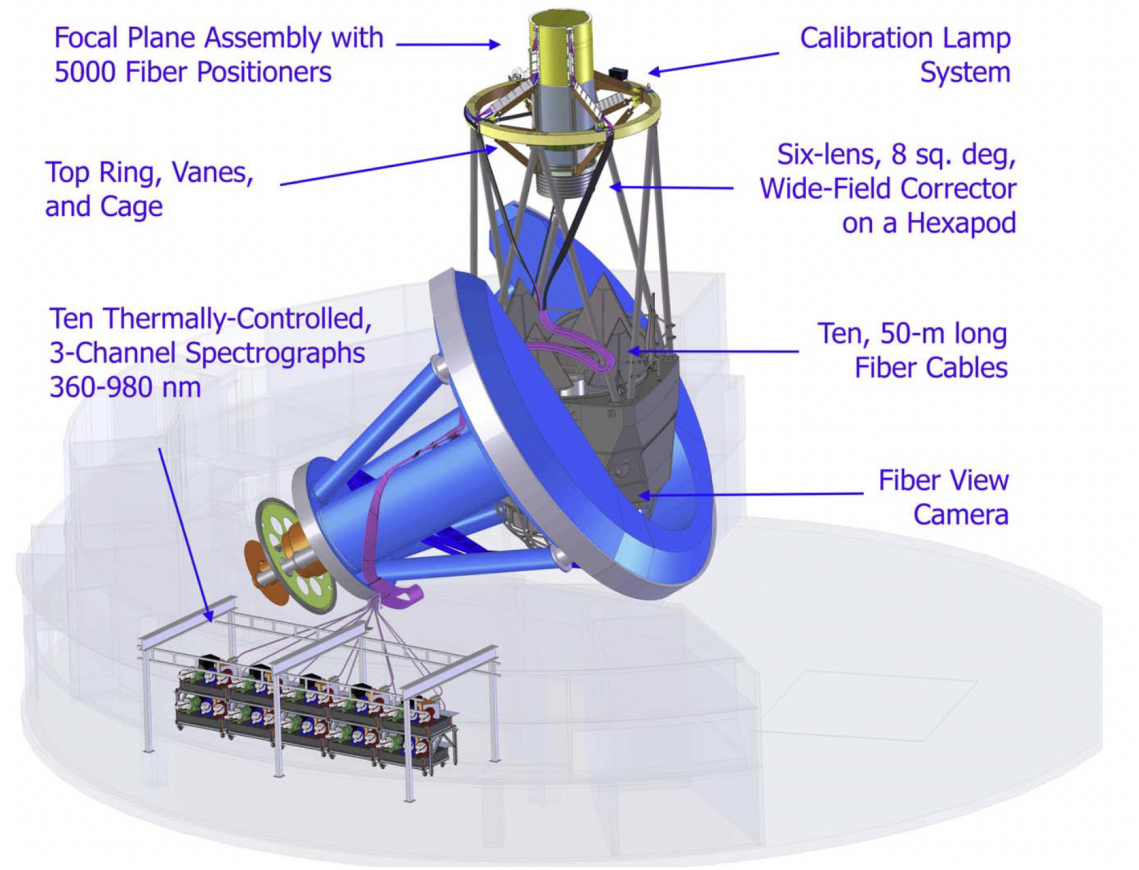

Figure 1.1: Model of the Mayall Telescope with the DESI instrumentation. Taken from [4].

1.1 Instrument overview and observations

1.1.1 Technical details

The DESI instrument is very complex, with a lot of technical details worth many papers on their own. An overview of the DESI instrument can be found in [4]. In this section, we will just briefly present the instrument. An illustration of the different components of the DESI instrument is presented in Figure 1.1.

The most important innovation of DESI is the 0.8 m diameter focal plane with 5000 robotically-positioned fibres. DESI's predecessor survey, SDSS eBOSS[5] which used SDSS[6], had its 1000 fibres positioned manually, which took much longer than DESI. Eash positioner has its own microcontroller and set of motors, thus coordinating the movement to evade collisions becomes an important task. More about them can be found here:[7].

The focal plane is composed of 10 "petals" (equal sized sectors of the round focal plane), each with a Guide Focus Alignment camera, which maintains optical alignment between the primary mirror and the optical corrector.

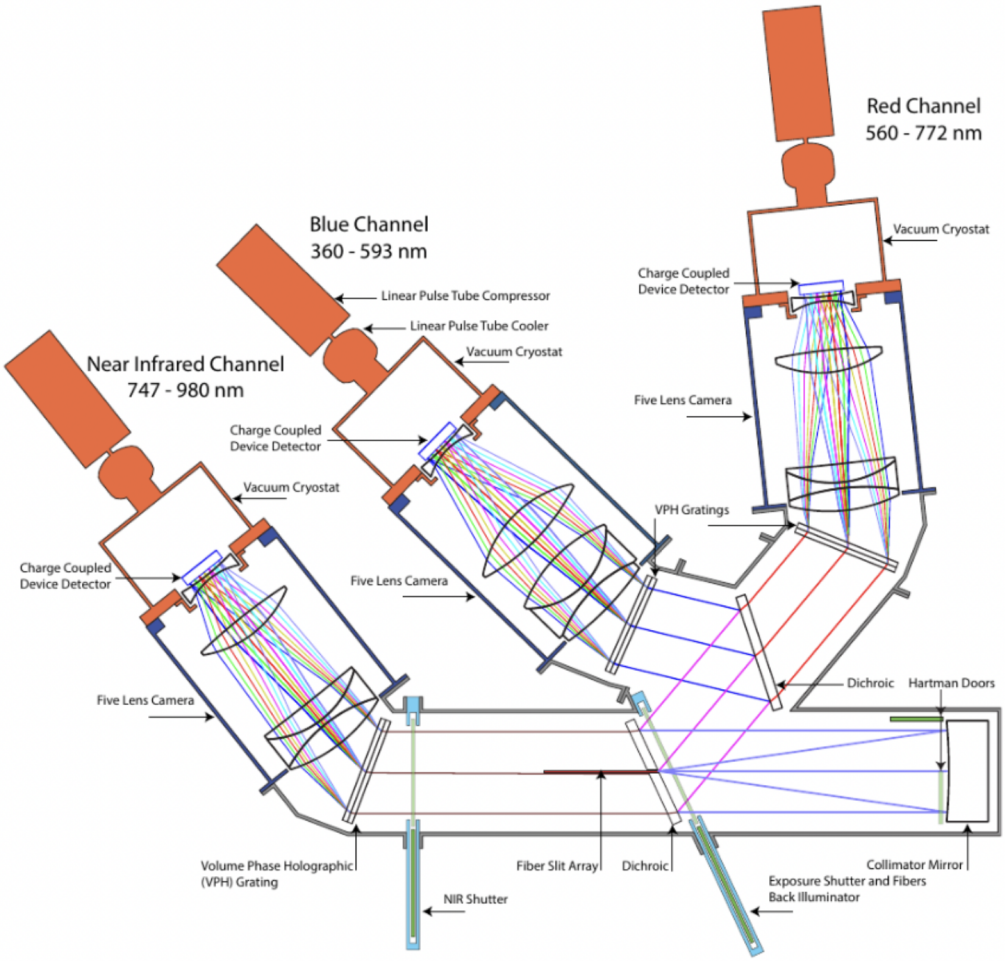

Figure 1.2: Schematic of one of the spectrographs of the DESI instrument. Taken from [2].

Channel	Spectral Range (Å)	Spectral resolution
Blue (B)	3600-5930	2000-3200
Red (R)	5600-7720	3200-4100
Near Infrared (Z)	7470-9800	4100-5100

Table 1.1: Spectral range and resolutions for each channel of the spectrographs. Taken from [3].

Each petal is connected to a single three-arm spectrograph, which covers a wavelength range from 3600 to 9800 Å. The scheme of the spectrograph is shown in Figure 1.2. The light is therefore divided into three wavelength channels: Blue(B), Red(R) and Near Infrared(Z), depending on the spectral range. The spectral resolution (which is defined as $SR = \frac{\lambda}{\Delta\lambda}$, where $\Delta\lambda$ is the smallest difference of wavelength that can be resolved at wavelength λ) of each of arms is shown in Table 1.1. This configuration optimises throughput, increases spectral coverage, with the resolution enough to resolve, for example, the [OII] doublet of ELGs and other fine spectral features.

Finally, the light is collected by CCD (charge-coupled device) sensors installed inside the vacuum cryostats, with the blue CCDs (4096×4096 STA4150) kept at $\sim 163K$ and the others (4114×4128 produced by LBL for DESI with wafers from Dalsa) kept at $\sim 140K$[3].

1.1.2 Observations

Target selection

Target selection is the process of choosing the objects in the sky based on their photometric properties from which we want to measure the spectra. In case of DESI, the data from the Data Release 9 of the Legacy Imaging Surveys program was used[8], which covered $\sim 19700\text{deg}^2$ of the sky in the three optical bands g, r and z. Additional optical bands were collected by different independent surveys[9–11].

Using the photometric data, the Target Selection algorithm is specific to each tracer but with a common strategy: 1) apply a morphological cut to select extended objects that correspond to galaxies or point-source objects, 2) apply a colour selection to select the required number density of a given type of galaxies in a given redshift range. The Target Selection pipeline for DESI is described in [12].

The main survey

Once the targets are selected, DESI performs the spectroscopic survey of the selected targets. There are two programs: the dark time program, which observes galaxies in the

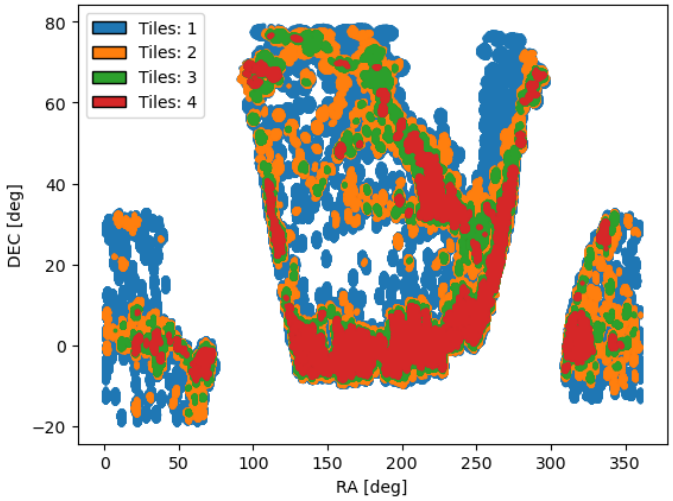

Figure 1.3: Bright galaxies observed by DESI throughout the first year of operations. The color indicates the number of tiles a galaxy could have been observed in. The footprint is comprised of round patches (tiles), which are often intersecting.

redshift range of $0.4 < z < 4$ typically and the bright time program, which observes bright galaxies and Milky Way stars. The second program is used when the observing conditions (typically the moon brightness) are not good enough for the dark program. There is also a backup program, which is considered when the observational conditions are too poor to observe the bright program. The combination of bright and dark programs is the main survey. The observational strategy for DESI is described in great detail in [13].

During the operations, the telescope is never completely shut down. The vacuum and the temperature have to be maintained to keep away the condensation from the CCDs, for example. The night usually starts with the verification of the correct functionality of all the components. Each night, the calibration is performed to subtract the background noise of CCDs, which is done by exposing with the dome being shut down, and ensuring there are no light sources inside. The calibration lamps inside the dome are then lit with a smooth spectrum, ensuring that each pixel of the spectrograph is illuminated with the same flux. After that, the procedure is repeated but now with a light with clearly defined emission lines. This is also how the spectral resolution is measured. Then, the observation starts.

A part of the sky observed by DESI with a single exposure is called a tile. The whole DESI footprint is in fact divided into such tiles, meaning that some targets can be inside several tiles, as they can overlap (Figure 1.3). The icohasedral tiling was chosen following [14], such that, taking into account the shape of the focal plane, there is no area inside the footprint which will be left uncovered through several passes, as every pass shifts and rotates the tile with respect to the previous one. For more detail on the tiling strategy we refer to [2].

At the end of each exposure, which usually lasts around 20 minutes, the raw CCD

spectra are obtained. It should be noted that the exposure time is dynamically allocated and varies from tile to tile and current sky conditions. Each of the exposures is manually verified by a support observer (a role which the author fulfilled for several dozens of nights), to ensure nothing went wrong, the noise on the CCDs stayed negligible and the guidance of the telescope was not compromised.

Redshift determination

Once the exposures are saved, the pipeline reduces, classifies and measures redshifts for all the targets. The details on the spectroscopic pipeline can be found in [15]. The spectra are classified by a software called `Redrock`[1], based on a template fitting method. The minimisation relies on a linear combination of spectral templates over the set of training templates. Depending on the class of a target, templates are constructed from previously observed objects of the same class. The parameters inferred through such a maximum likelihood fit are the redshift, uncertainty on it, the spectral class and the nuisance parameters. In figure 1.4 one can find an example of an Emission Line Galaxy spectra with some of the emission lines indicated, as well as the images from the DESI Legacy Imaging Surveys with the corresponding galaxies indicated.

1.2 DESI tracers

As mentioned earlier, the main survey of DESI consists of two parts: the dark and bright time surveys. The dark time survey can be further separated into three types of objects observed: Luminous Red Galaxies (LRGs), Emission Line Galaxies (ELGs), and Quasars (QSOs). The bright time is primarily dedicated to the follow up of bright galaxies with the Bright Galaxy Survey (BGS) and stars through the Milky-Way Survey. Each subsample has a different selection procedure, and constitutes a separate catalogue of objects for analysis. The DESI footprint is further separated into two galactic caps, disconnected regions on the sky usually defined by a $DEC \sim 30$deg line, and called the northern and the southern galactic caps (NGS and SGC respectively).

Bright Galaxy Survey

The Bright Galaxy Survey (BGS) [17] is a flux-limited sample of galaxies, which comes in two flavours: BGS Bright, with an apparent magnitude limit of $r < 19.5$, and BGS Faint with $19.5 < r < 20.175$.To discriminate between stars and galaxies the Gaia G-band magnitude is used [18], otherwise, if the object is not present in the Gaia catalogue, it is considered as a galaxy[17]. Due to their bright magnitudes, these are the galaxies which

[1] `https://github.com/desihub/redrock`

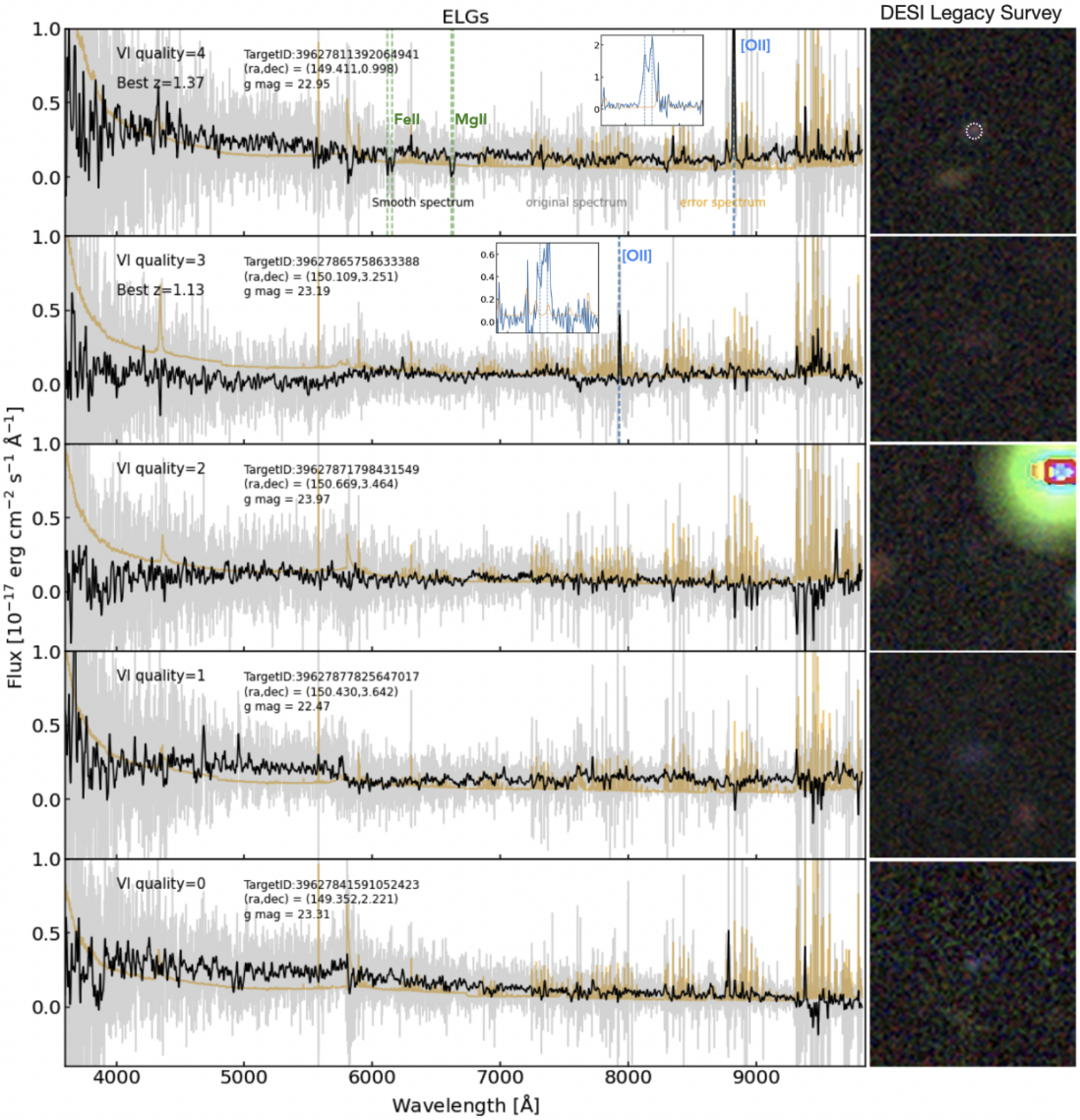

Figure 1.4: Examples of the ELG spectra (left panel) ordered by their quality, where quality goes from 0, represnting useless spectra with no signal detectable to quality 4 being the highest, representing two or more well resolved spectral features, and the corresponding target images from the photometric surveys (right panel). We can see that as the quality decreases, less and less features are observed in the spectrum. Adapted from [16].

are situated the closest to us. We will discuss this sample in more detail in a separate section as it is the dataset of interest for this thesis.

Luminous Red Galaxies

Luminous Red Galaxies (LRGs) are old elliptical galaxies in which the star formation process has already stopped. They are recognised by a $4000\mathring{A}$ break in their spectra in the rest frame [19]. We expect to collect 8 million LRGs in $0.4 < z < 1.0$. The details for the target selection of LRGs can be found here:[20].

Emission Line Galaxies

Emission Line Galaxies are typically the younger spiral galaxies with star-formation still going on. Thus, the noticeable emission lines in the galaxy spectra [21, 22], which gives the name to this tracer. We expect to collect 16 million ELGs in $0.6 < z < 1.6$. The details of the target selection for this tracer can be found in [23]. The ELG dataset is subdivided into two parts, `ELG_VLOP` and `ELG_LOP`, which tend to occupy different redshift ranges, $0.6 < z < 1.1$ and $1.1 < z < 1.6$ respectively, and have been assigned different priorities during the fibre assignment process (see [24] for more details).

Quasars

Quasi-stellar objects, often denoted as quasars (QSOs), are a type of active galactic nuclei (AGN), which can produce long jets of gas. The luminosity of a quasar powered by the accretion onto the Super Massive Black Hole at its centre therefore exceeds the luminosity of the host galaxy, and makes them as one of the brightest visible objects in the Universe, allowing to detect them even at high redshifts. We expect to collect 3 million QSOs in $0.9 < z < 4$: quasars between $0.9 < z < 2.1$ are used as direct tracers of the matter field in a similar fashion as the other tracers, while quasars at $z > 2.1$ are used for their Lyman-α forests. The latter correspond to the absorption of neutral hydrogen along the LOS by the intergalactic medium. The details on the quasar target selection can be found in [25].

1.3 Galaxy catalogues for clustering analysis

1.3.1 Systematics

Imaging systematics

The target density over the sky is not uniform, due to varying quality and depth of the photometry. Those inhomogeneities are due to many factors: presence of the Milky-Way, varying level of galactic extinction, contamination from stars biases the clustering

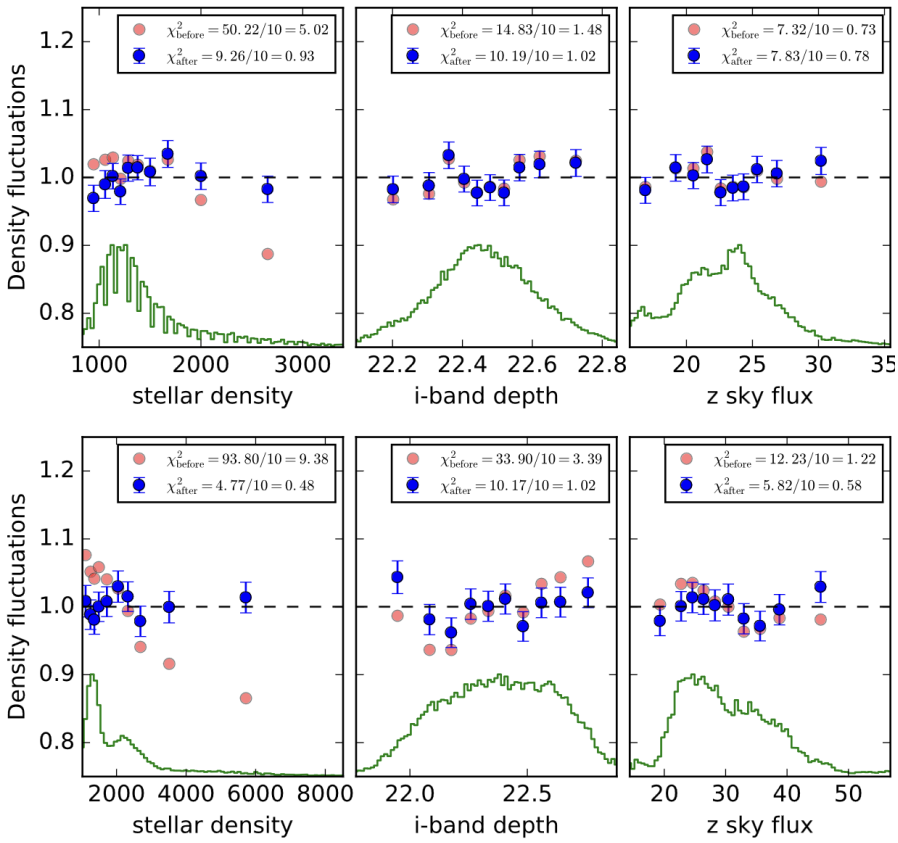

Figure 1.5: Fluctuations of the galaxy number density with respect to stellar density, i-band depth and z-band flux for eBOSS LRG data, for NGC (Upper panel) and SGC (Lower panel). Red points indicate the fluctuations before the correction, while the blue indicate those after. The green histograms show the distribution of the quantities. Taken from [26].

of obtained galaxies, imperfect star-galaxy separation that can bias the clustering signal of galaxies... In DESI we mitigate these imaging systematics by computing weights that minimise the dependency of the target selection density with those photometric conditions. For BGS and LRG targets, we use weights, obtained by creating a linear regression model accounting for non-cosmological density fluctuations, as described in [26]. They are not as affected by the imaging systematics as other fainter tracers [27]. An example of the effect of such weights on the density with respect to various photometric quantities can be seen in Figure 1.5, where the number density fluctuations are shown for the eBOSS LRG data before and after the corrections as red and blue points respectively. We can notice a noticeable improvement before and after the corrections, especially for the regions with high stellar density. Uncorrected, it otherwise impacts clustering on larger scales [26].

A package called `regressis`, based on random forests regression and observational feature templates has been developed for QSOs to mitigate those systematics and compute the corresponding photometric weights. It can also be used for ELGs, with more details

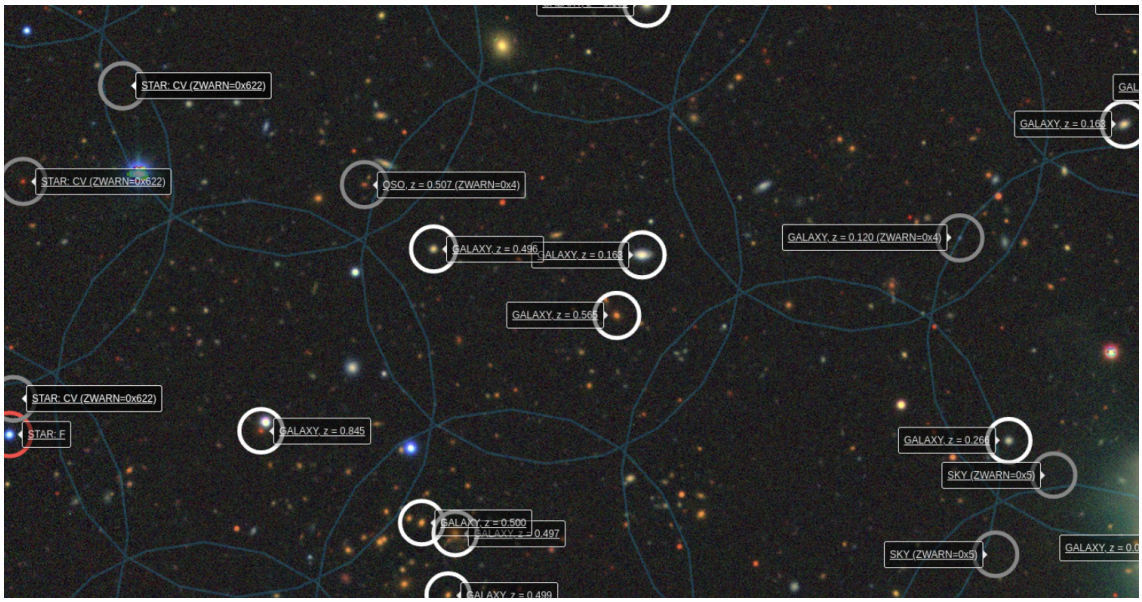

Figure 1.6: A Legacy Survey image annotated with specific DESI targets (white circles) and with the patrol radius of each fibre superimposed in light blue for a specific tile. One fibre can target only one galaxy in a given tile, thus in this example three passes are needed to cover the three targets reachable by one fibre only in the centre. Taken from [24].

in [28]. Another package called `Sysnet²`[29] based on neural networks also does the same and is used as default imaging weights for ELG. More details about the imaging systematics and their mitigation for DESI DR1 data can be found in [27].

Spectroscopic systematics

Similarly to imaging systematics, some contamination can arise from redshift uncertainty and/or spectroscopic failure. The corresponding weights are also generated, which are based on the redshift determination success rate, which is measured with respect to the redshift and flux (as we would expect fainter galaxies to have worse redshifts). The methodology is described in more detail in[23, 30–32]. Moreover, the impact of those spectroscopic systematics (redshift failure, redshift uncertainty and the dependence on the observation conditions of the redshift success rate) on the clustering statistics has been studied in detail in [31, 32].

Fibre assignment systematics

One of the most important systematics of DESI concerns the fiber assignment procedure. For a given tile, each fiber is able to observe only one target within its patrol radius, meaning that for the dense regions in the sky, some targets might remain unobserved if not enough passes have been made (an illustration of this is shown in Figure 1.6). Thus,

²`https://github.com/mehdirezaie/SYSNet`

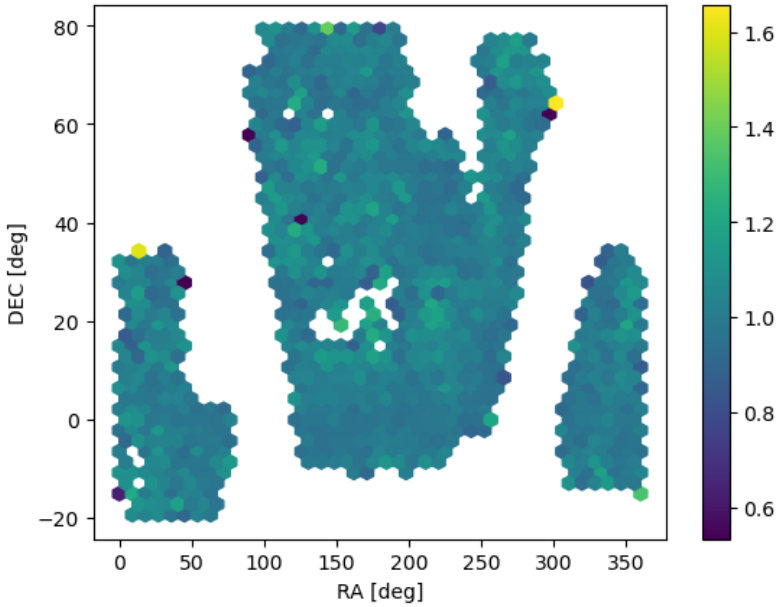

Figure 1.7: Spatial distribution of the systematic weight for DESI BGS DR1, with colour representing the mean weight value in the bin.

one needs to either pass multiple times over the dense region to obtain the spectra from all targets, or estimate the targeting incompleteness. This can be done by simulating the assignment of the fibers to the galaxies multiple times, generating the "alternative" observations, allowing for the estimations of the probability of the pair of objects to be observed (more details available in [33]). When computing clustering statistics, one way to account for this effect is to ignore the pairs with small angle separation when computing clustering statistics, which is the way DESI DR1 will deal with fibre collisions [34].

Systematics weights

The weights previously introduced (i.e. completeness $w_{\rm comp}$, imaging $w_{\rm im}$ and spectroscopic $w_{\rm zfail}$ weights) are combined by multiplication, to produce the final weights w:

$$w = w_{\rm im} w_{\rm zfail} w_{\rm comp} \tag{1.1}$$

These weights allow to achieve a less biased estimate of the 2-point clustering statistics [27]. The spatial distribution of systematic weights for BGS DR1 is presented in Figure 1.7.

Veto masks

Some parts of the footprint are removed all together from the clustering catalogues, due to the presence of bright objects like stars, or due to hardware issues during observations. Thus, $\sim 5\%$ of the footprint is removed using a veto mask. The veto mask is defined based on the tracer, for example for BGS it is based on the Legacy Surveys[8] and cuts the Milky Way stars. More details on that procedure can be found in [27].

1.3.2 Random catalogues

An important part of the clustering products for the analysis of any DESI galaxy data is the random catalogue, which, as shown in the introduction,accounts for the geometry of the survey. The random catalogue is created by generating a uniformly distributed footprint in RA and DEC first, which is then cut to the area observed by the DESI survey. Once that is done, the generated points are randomly assigned redshifts and other associated quantities from the observed galaxies, thus automatically matching the required number density. This approach is also known as shuffling and was used for SDSS as well [35, 36]. It has been shown that such shuffling method introduces a radial integral constraint [37]. More details about this effect and how to deal with it for DESI is described in [38]. Finally, the weights are normalised such that for each of the galactic caps the ratio of random data counts to observed data counts is the same. Additionally, the same veto masks are applied.

1.4 BGS specificities

This thesis focuses primarily on the Bright Galaxy Survey. Taking into account all the sub-selections combined (BGS Bright + BGS Faint + BGS Active Galactic Nuclei (AGN)), the BGS is more than ten times higher density of objects than that of the LOWZ SDSS-III BOSS sample ([39, 40]. It should be mentioned, that even though there are slightly denser surveys with comparable targets, like GAMA[41–43], the DESI BGS features a much larger area (50 larger than GAMA). The final target selection procedure for the BGS is described in [17]. As mentioned above, the star-galaxy separation uses GAIA magnitude which results in a stellar contamination being negligible (<1%). A fibre-magnitude cut is applied to remove spurious objects which are imaging artefacts or fragments of galaxies. For cosmological analysis, we restrict the analysis to the BGS Bright which will contain 13.5 million galaxies brighter than $r \lesssim 19.5$ over 5 years of observation. The target selection for the BGS bright was first described in [18] and the clustering properties of the BGS targets for cosmological analysis were first studied in [44]. The BGS catalogue also features many galaxy properties (apparent and absolute magnitudes, observed and rest-frame colours, ...) allowing for the multi-tracer analysis which is discussed in Chapter 5).

As we are interested in the galaxy clustering, we will be focusing on BGS Bright, which features a very high redshift success rate (Figure 1.9), and no additional physical cuts beyond the apparent magnitude one and the systematical ones, thus making it highly complete.

In this thesis, we will be interested in separating galaxies into a red and a blue sample. Hence we will need to convert their apparent colours to rest-frame colours. This can be done by using the k-correction between bands Q and R, K_{QR}, which is often defined

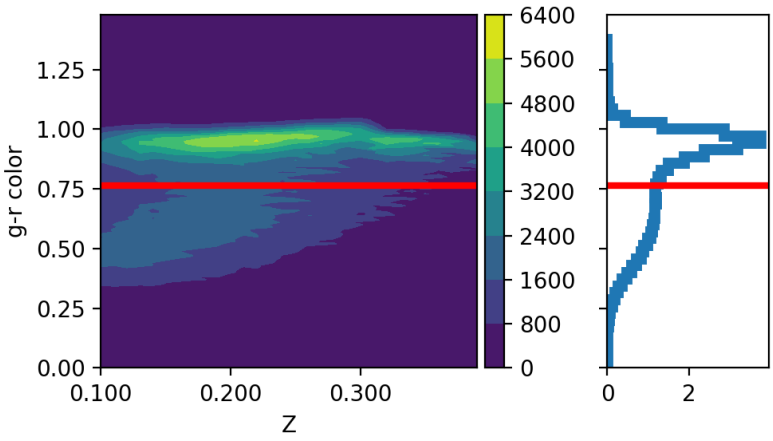

Figure 1.8: Distribution of BGS Bright DR1 galaxies in the redshift and rest-frame color plane, with an overall rest-frame color distribution on the left. The red line indicates the cut between red and blue galaxies, and the color indicates the number of galaxies in the bin.

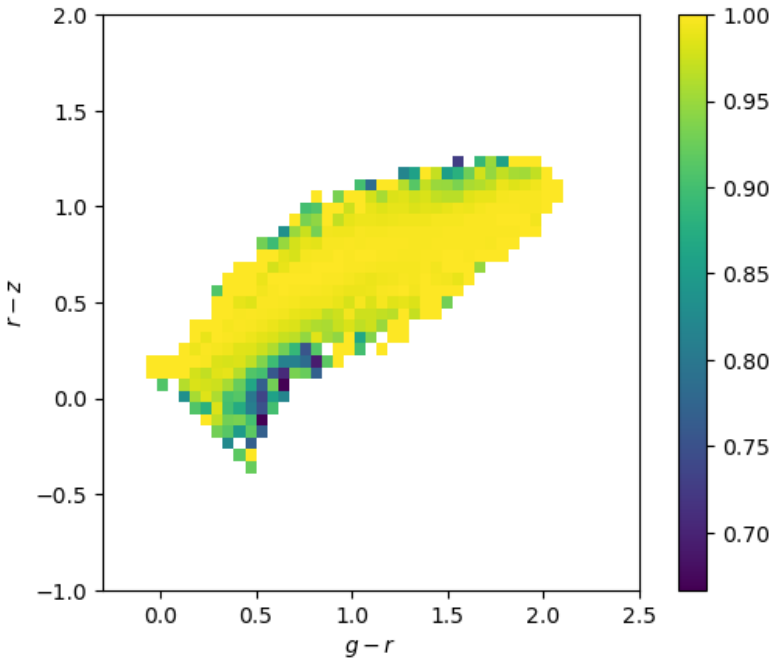

Figure 1.9: Binned redshift success rate (represented by color) as a function of observed (g-r) and (r-z) colors for DESI BGS Bright.

following [45] as:

$$m_R = M_Q + DM + K_{QR} \tag{1.2}$$

where m_R is the apparent magnitude in band R, M_Q is the absolute magnitude of a source in band Q, DM is a distance modulus, defined by the luminosity distance D_L as:

$$DM = 5\log_{10}\frac{D_L}{10\text{pc}} \tag{1.3}$$

Thus, the k-correction K_{QR} connects the observed apparent magnitude m_R in a bandpass R with an absolute magnitude (magnitude, which the source would have if the observer was 10 pc away from it), with a filter Q, all of that in the rest-frame of the source (thus, not redshifted). [45] provides:

$$K_{QR} = -2.5\log_{10}\left[\frac{1}{1+z}\frac{\int d\lambda_o\lambda_o f_\lambda(\lambda_o)R(\lambda_o)\int d\lambda_e\lambda_e g^Q_\lambda(\lambda_e)Q(\lambda_e)}{\int d\lambda_o\lambda_o g^R_\lambda(\lambda_o)R(\lambda_o)\int d\lambda_e\lambda_e f_\lambda([1+z]\lambda_e)Q(\lambda_e)}\right] \tag{1.4}$$

where $f_\lambda(\lambda)$ is the source flux density, $g^R_\lambda(\lambda)$ and $g^Q_\lambda(\lambda)$ are the standard-source flux densities, the bandpass function are $R(\lambda)$ and $Q(\lambda)$, λ is a wavelength and z is the redshift of the source. The flux densities can be related to apparent magnitudes as:

$$m_R = -2.5\log_{10}\frac{\int d\lambda_o\lambda_o f_\lambda(\lambda_o)R(\lambda_o)}{\int d\lambda_o\lambda_o g^R_\lambda(\lambda_o)R(\lambda_o)} \tag{1.5}$$

For AB magnitudes, such as the ones used in DESI, the standard source is taken to be a hypothetical constant source such that for all wavelengths $g^{AB} = 3631\times10^{-26}Wm^{-2}s$.

In DESI there are three photometric bands provided: r, g and z [17]. For each band the k-correction thus will be different, therefore separate k-corrections for g and r bands are needed to compute the rest-frame color. They can be inferred from the data following, for example, [46].

Another correction to be included is called an e-correction, and accounts for the evolution of the intrinsic luminosity of the galaxy. The form taken is usually the following, assuming the density does not evolve:

$$E(z) = -Q_0(z - z_{\text{ref}}) \tag{1.6}$$

where for r-band magnitudes Q_0 was empirically found to be $Q_0 = 0.97$[46]. It is substracted from the absolute magnitude.

This gives us all the tools needed to obtain rest-frame colors from observed fluxes. Figure 1.8 shows the plane $g-r$ colour vs z for the BGS DR1 data, with the indication of the cut between the red and blue galaxies. The example spectra of red a blue galaxies from BGS Bright are also shown in Figure 1.10.

The presence of colors is what will allow us later to attempt to bypass the cosmic variance in the otherwise overly dense catalogue.

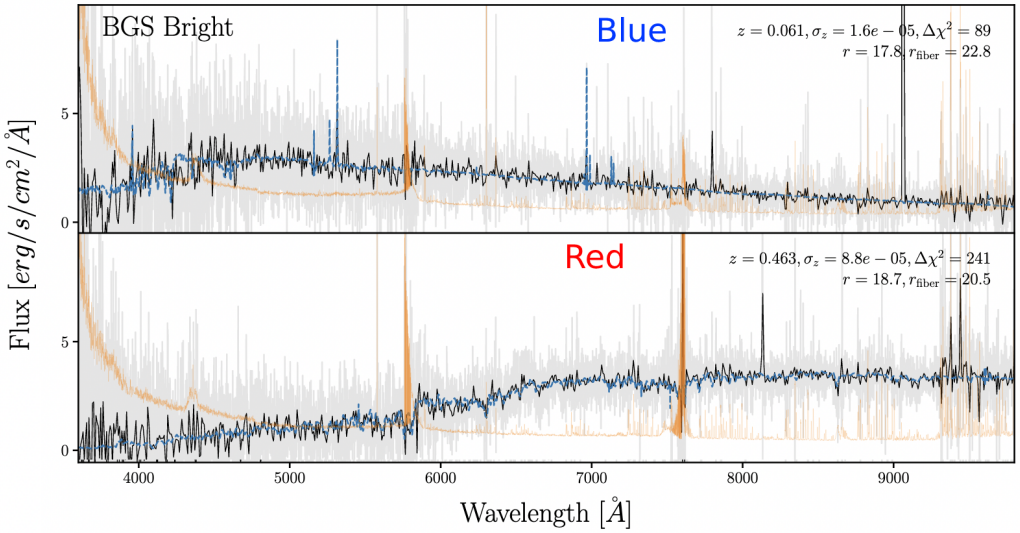

Figure 1.10: Optical spectrum of a blue (top panel) and a red (bottom panel) galaxy from BGS Bright DR1 in grey. The orange line represents the uncertainty, the black line is a spectra rebinned to a coarser wavelength, and the blue line represents the best-fit value from the `Redrock` template used to measure the redshift.

References

[1] Jeffrey A. Newman and Daniel Gruen. "Photometric Redshifts for Next-Generation Surveys". In: *Annual Review of Astronomy and Astrophysics* 60.Volume 60, 2022 (2022), pp. 363–414. ISSN: 1545-4282. DOI: https://doi.org/10.1146/annurev-astro-032122-014611. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-astro-032122-014611.

[2] DESI Collaboration et al. "The DESI Experiment Part I: Science,Targeting, and Survey Design". In: *arXiv e-prints*, arXiv:1611.00036 (Oct. 2016), arXiv:1611.00036. arXiv: 1611.00036 [astro-ph.IM].

[3] DESI Collaboration et al. *The DESI Experiment Part II: Instrument Design*. 2016. arXiv: 1611.00037 [astro-ph.IM].

[4] B. Abareshi et al. "Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument". In: *The Astronomical Journal* 164.5 (Oct. 2022), p. 207. ISSN: 1538-3881. DOI: 10.3847/1538-3881/ac882b. URL: http://dx.doi.org/10.3847/1538-3881/ac882b.

[5] Kyle S. Dawson et al. "The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data". In: 151.2, 44 (Feb. 2016), p. 44. DOI: 10.3847/0004-6256/151/2/44. arXiv: 1508.04473 [astro-ph.CO].

[6] James E. Gunn et al. "The 2.5 m Telescope of the Sloan Digital Sky Survey". In: *The Astronomical Journal* 131.4 (Apr. 2006), pp. 2332–2359. ISSN: 1538-3881. DOI: 10.1086/500975. URL: http://dx.doi.org/10.1086/500975.

[7] Joseph Harry Silber et al. "The Robotic Multiobject Focal Plane System of the Dark Energy Spectroscopic Instrument (DESI)". In: 165.1, 9 (Jan. 2023), p. 9. DOI: 10.3847/1538-3881/ac9ab1. arXiv: 2205.09014 [astro-ph.IM].

[8] D. Schlegel et al. "DESI Legacy Imaging Surveys Data Release 9". In: *AAS Meeting Abstracts*. Vol. 237. AAS Meeting Abstracts. Jan. 2021, 235.03, p. 235.03.

[9] R. M. Cutri and et al. *VizieR Online Data Catalog: WISE All-Sky Data Release (Cutri+ 2012)*. Apr. 2012.

[10] Gaia Collaboration et al. "The Gaia mission". In: *AA* 595 (2016), A1. DOI: 10.1051/0004-6361/201629272. URL: https://doi.org/10.1051/0004-6361/201629272.

[11] Edward L. Wright et al. "The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance". In: 140.6 (Dec. 2010), pp. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM].

[12] Adam D. Myers et al. "The Target-selection Pipeline for the Dark Energy Spectroscopic Instrument". In: *The Astronomical Journal* 165.2 (Jan. 2023), p. 50. ISSN: 1538-3881. DOI: 10.3847/1538-3881/aca5f9. URL: http://dx.doi.org/10.3847/1538-3881/aca5f9.

[13] E. F. Schlafly et al. *Survey Operations for the Dark Energy Spectroscopic Instrument*. 2024. arXiv: 2306.06309 [astro-ph.CO].

[14] R.H Hardin et al. *Tables of Spherical Codes with Icosahedral Symmetry*. 2021.

[15] J. Guy et al. "The Spectroscopic Data Processing Pipeline for the Dark Energy Spectroscopic Instrument". In: *The Astronomical Journal* 165.4 (Mar. 2023), p. 144. ISSN: 1538-3881. DOI: 10.3847/1538-3881/acb212. URL: http://dx.doi.org/10.3847/1538-3881/acb212.

[16] Ting-Wen Lan et al. "The DESI Survey Validation: Results from Visual Inspection of Bright Galaxies, Luminous Red Galaxies, and Emission-line Galaxies". In: *The Astrophysical Journal* 943.1 (Jan. 2023), p. 68. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aca5fa. URL: http://dx.doi.org/10.3847/1538-4357/aca5fa.

[17] ChangHoon Hahn et al. 2022. DOI: 10.48550/ARXIV.2208.08512. URL: https://arxiv.org/abs/2208.08512.

[18] Omar Ruiz-Macias et al. "Characterizing the target selection pipeline for the Dark Energy Spectroscopic Instrument Bright Galaxy Survey". In: *MNRAS* 502.3 (Feb. 2021), pp. 4328–4349. ISSN: 0035-8711. DOI: 10.1093/mnras/stab292. eprint: https://academic.oup.com/mnras/article-pdf/502/3/4328/39112755/stab292.pdf. URL: https://doi.org/10.1093/mnras/stab292.

[19] B. M. Poggianti and G. Barbaro. "Indicators of star formation: 4000 Å break and Balmer lines." In: 325 (Sept. 1997), pp. 1025–1030. DOI: 10.48550/arXiv.astro-ph/9703067. arXiv: astro-ph/9703067 [astro-ph].

[20] Rongpu Zhou et al. "Target Selection and Validation of DESI Luminous Red Galaxies". In: *The Astronomical Journal* 165.2 (Jan. 2023), p. 58. ISSN: 1538-3881. DOI: 10.3847/1538-3881/aca5fb. URL: http://dx.doi.org/10.3847/1538-3881/aca5fb.

[21] John Moustakas et al. "Optical star formation rate indicators". en. In: *Astrophys. J.* 642.2 (May 2006), pp. 775–796.

[22] G. Favole et al. *Characterizing the ELG luminosity functions in the nearby Universe*. 2023. arXiv: 2303.11031 [astro-ph.GA].

[23] A Raichoor et al. "Target selection and validation of DESI Emission Line galaxies". In: *Astron. J.* 165.3 (Mar. 2023), p. 126.

[24] Antoine Rocher et al. "Halo occupation distribution of Emission Line Galaxies: fitting method with Gaussian processes". In: *JCAP* 2023.05 (May 2023), p. 033. DOI: 10.1088/1475-7516/2023/05/033. URL: https://dx.doi.org/10.1088/1475-7516/2023/05/033.

[25] Edmond Chaussidon et al. "Target selection and validation of DESI quasars". In: *Astrophys. J.* 944.1 (Feb. 2023), p. 107.

[26] Julian E. Bautista et al. "The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations at Redshift of 0.72 with the DR14 Luminous Red Galaxy Sample". In: *The Astrophysical Journal* 863.1 (Aug. 2018), p. 110. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aacea5. URL: http://dx.doi.org/10.3847/1538-4357/aacea5.

[27] A. J. Ross et al. 2024. arXiv: 2405.16593 [astro-ph.CO]. URL: https://arxiv.org/abs/2405.16593.

[28] Edmond Chaussidon et al. "Angular clustering properties of the DESI QSO target selection using DR9 Legacy Imaging Surveys". In: *MNRAS* 509.3 (Nov. 2021), pp. 3904–3923. ISSN: 0035-8711. DOI: 10.1093/mnras/stab3252. eprint: https://academic.oup.com/mnras/article-pdf/509/3/3904/41446828/stab3252.pdf. URL: https://doi.org/10.1093/mnras/stab3252.

[29] Mehdi Rezaie et al. "Improving galaxy clustering measurements with deep learning: analysis of the DECaLS DR7 data". In: 495.2 (May 2020), pp. 1613–1640. DOI: 10.1093/mnras/staa1231. arXiv: 1907.11355 [astro-ph.CO].

[30] DESI Collaboration Et Al. 2023. DOI: 10.5281/ZENODO.7964162. URL: https://zenodo.org/record/7964162.

[31] Jiaxi Yu et al. "ELG Spectroscopic Systematics Analysis of the DESI Data Release 1". In: *arXiv e-prints*, arXiv:2405.16657 (May 2024), arXiv:2405.16657. DOI: 10.48550/arXiv.2405.16657. arXiv: 2405.16657 [astro-ph.CO].

[32] A. Krolewski et al. "Impact and mitigation of spectroscopic systematics on DESI DR1 clustering measurements". In: *arXiv e-prints*, arXiv:2405.17208 (May 2024), arXiv:2405.17208. DOI: 10.48550/arXiv.2405.17208. arXiv: 2405.17208 [astro-ph.CO].

[33] J. Lasker et al. 2024. arXiv: 2404.03006 [astro-ph.CO].

[34] M. Pinon et al. *Mitigation of DESI fiber assignment incompleteness effect on two-point clustering with small angular scale truncated estimators*. 2024. arXiv: 2406.04804 [astro-ph.CO]. URL: https://arxiv.org/abs/2406.04804.

[35] Beth Reid et al. "SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues". In: *MNRAS* 455.2 (Nov. 2015), pp. 1553–1573. ISSN: 0035-8711. DOI: 10.1093/mnras/stv2382. eprint: https://academic.oup.com/mnras/article-pdf/455/2/1553/18511627/stv2382.pdf. URL: https://doi.org/10.1093/mnras/stv2382.

[36] Ashley J Ross et al. "The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Large-scale structure catalogues for cosmological analysis". In: *MNRAS* 498.2 (Sept. 2020), pp. 2354–2371. ISSN: 0035-8711. DOI: 10.1093/mnras/staa2416. eprint: https://academic.oup.com/mnras/article-pdf/498/2/2354/33777098/staa2416.pdf. URL: https://doi.org/10.1093/mnras/staa2416.

[37] Arnaud de Mattia and Vanina Ruhlmann-Kleider. "Integral constraints in spectroscopic surveys". In: *JCAP* 2019.08 (Aug. 2019), pp. 036–036. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2019/08/036. URL: http://dx.doi.org/10.1088/1475-7516/2019/08/036.

[38] Edmond Chaussidon and DESI Collaboration. (in prep.) 2024.

[39] Daniel J. Eisenstein et al. "SDSS-III: MASSIVE SPECTROSCOPIC SURVEYS OF THE DISTANT UNIVERSE, THE MILKY WAY, AND EXTRA-SOLAR PLANETARY SYSTEMS". In: *The Astronomical Journal* 142.3 (Aug. 2011), p. 72. ISSN:

1538-3881. DOI: 10.1088/0004-6256/142/3/72. URL: http://dx.doi.org/10.1088/0004-6256/142/3/72.

[40] Kyle S. Dawson et al. "THE BARYON OSCILLATION SPECTROSCOPIC SURVEY OF SDSS-III". In: *The Astronomical Journal* 145.1 (Dec. 2012), p. 10. ISSN: 1538-3881. DOI: 10.1088/0004-6256/145/1/10. URL: http://dx.doi.org/10.1088/0004-6256/145/1/10.

[41] Simon P. Driver et al. "GAMA: towards a physical understanding of galaxy formation". In: *Astronomy and Geophysics* 50.5 (Oct. 2009), pp. 5.12–5.19. DOI: 10.1111/j.1468-4004.2009.50512.x. arXiv: 0910.5123 [astro-ph.CO].

[42] S. P. Driver et al. "Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release". In: 413.2 (May 2011), pp. 971–995. DOI: 10.1111/j.1365-2966.2010.18188.x. arXiv: 1009.0614 [astro-ph.CO].

[43] Simon P Driver et al. "Galaxy And Mass Assembly (GAMA): Data Release 4 and the z
lt; 0.1 total and z
lt; 0.08 morphological galaxy stellar mass functions". In: *MNRAS* 513.1 (Mar. 2022), pp. 439–467. ISSN: 0035-8711. DOI: 10.1093/mnras/stac472. eprint: https://academic.oup.com/mnras/article-pdf/513/1/439/43426380/stac472.pdf. URL: https://doi.org/10.1093/mnras/stac472.

[44] Pauline Zarrouk et al. "Preliminary clustering properties of the DESI BGS bright targets using DR9 Legacy Imaging Surveys". In: *MNRAS* 509.1 (Oct. 2021), pp. 1478–1493. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2814. eprint: https://academic.oup.com/mnras/article-pdf/509/1/1478/41161113/stab2814.pdf. URL: https://doi.org/10.1093/mnras/stab2814.

[45] David W. Hogg et al. *The K correction*. 2002. arXiv: astro-ph/0210394 [astro-ph]. URL: https://arxiv.org/abs/astro-ph/0210394.

[46] Tamsyn McNaught-Roberts et al. "Galaxy And Mass Assembly (GAMA): the dependence of the galaxy luminosity function on environment, redshift and colour". In: *MNRAS* 445.2 (Oct. 2014), pp. 2125–2145. ISSN: 0035-8711. DOI: 10.1093/mnras/stu1886. eprint: https://academic.oup.com/mnras/article-pdf/445/2/2125/18198298/stu1886.pdf. URL: https://doi.org/10.1093/mnras/stu1886.

Chapter 2

Theory of galaxy clustering

Мы шепчем всем ненужные
признанья,
От милых рук бежим к
обманным снам,
Не видим лиц и верим именам,
Томясь в путях напрасного
скитанья.

M. Voloshin, Translation

Introduction

We will start this section with an obvious statement: the data does not carry any value without comparing it to the theory, which serves as a mathematical representation of our beliefs about the Universe. A brief overview of the current theoretical framework used for studying the expansion of the Universe is needed. I will also share some new techniques I have used in order to use that framework to compare it to the data, with the potential of pushing our theoretical predictions even further in terms of accuracy using the recent developments in the area of machine learning.

Equipped with an intuitional understanding given by the Eulerian perturbation theory developed in the Introduction, I will start with a shift in formulation of the theory already presented to a Lagrangian perturbation theory framework (LPT), a bit more convenient in actual calculations but slightly less intuitive from a naive point of view, in terms of which I will formulate the redshift space distortions and present the different ways of expanding the non-linear power spectrum into a perturbative series. I will conclude by presenting the effort that I and other people have done in order to push further the speed and accuracy of the theoretical frameworks using the latest developments in data science techniques,

including the presentation of the neural network emulator for 2-point clustering statistics, presented in [1].

2.1 Lagrangian perturbation theory

Lagrangian perturbation theory is an approach developed in [2–12]. It is characterised by the fact that, instead of trying to follow the naive approach of following the positions of the fluid elements $\boldsymbol{x}(\tau)$, it instead tracks the displacement field $\boldsymbol{\Psi}(\boldsymbol{x}_0, \tau)$ with respect to the starting position $\boldsymbol{x}_0$, which can be defined as:

$$\boldsymbol{x}(\tau) = \boldsymbol{x}_0 + \boldsymbol{\Psi}(\boldsymbol{x}_0, \tau) \tag{2.1}$$

That fully defines the evolution of the fluid elements. The density field then can be formulated in configuration and Fourier space as follows:

$$1 + \delta(\boldsymbol{x}, \tau) = \int d^3\boldsymbol{x}_0 \, \delta_D(\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{\Psi}(\boldsymbol{x}_0, \tau)) \tag{2.2}$$

$$(2\pi)^3 \delta_D(\boldsymbol{k}) + \delta(\boldsymbol{k}, \tau) = \int d^3\boldsymbol{x}_0 \, e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}_0 + \boldsymbol{\Psi}(\boldsymbol{x}_0, \tau))} \tag{2.3}$$

where $\delta_D(x)$ is the Dirac delta-function.

The evolution equations for $\boldsymbol{\Psi}(\boldsymbol{x}_0, \tau)$ can be then written (following [11]) as:

$$\ddot{\boldsymbol{\Psi}}(\boldsymbol{x}_0, \tau) + \mathcal{H}\dot{\boldsymbol{\Psi}}(\boldsymbol{x}_0, \tau) = -\nabla\Phi(\boldsymbol{x}_0 + \boldsymbol{\Psi}(\boldsymbol{x}_0, \tau)) \tag{2.4}$$

Where $\Phi(\boldsymbol{x})$ is the gravitational potential from Introduction.

From now on, we will attempt the perturbative solution of these equations. However, it should be noted that the small scales are unfortunately unsolvable. So, following the traditional approach [11] we will first filter them out, and deal with them in a different manner (ex. Equation 2.10).

We therefore introduce a filter $W_R(\boldsymbol{x}, \boldsymbol{x}')$, and separate the displacements into $\boldsymbol{\Psi}_S$ and $\boldsymbol{\Psi}_L$ for short-wavelength and long-wavelength components correspondingly, such that:

$$\boldsymbol{\Psi}_L(\boldsymbol{x}_0, \tau) = \int d^3\boldsymbol{x}_0' W_R(\boldsymbol{x}_0, \boldsymbol{x}_0') \Psi(\boldsymbol{x}_0', \tau) \tag{2.5}$$

$$\boldsymbol{\Psi}_S(\boldsymbol{x}_0, \tau) = \boldsymbol{\Psi}(\boldsymbol{x}_0, \tau) - \boldsymbol{\Psi}_L(\boldsymbol{x}_0, \tau) \tag{2.6}$$

We can see, that defined in such a manner $\int d^3\boldsymbol{x}_0' W_R(\boldsymbol{x}_0', \boldsymbol{x}_0) \Psi_S(\boldsymbol{x}_0', \tau) = 0$.

By analogy we also define the long-wavelength density contrast δ_L and Ψ_L, which allows us [11] to rewrite the equation 29 for long-wavelength components:

$$\nabla^2 \Phi_L = \frac{3}{2} \mathcal{H}^2 \Omega_m \delta_L \tag{2.7}$$

We can then rewrite the equation 2.4 as:

$$\ddot{\boldsymbol{\Psi}}_L(\boldsymbol{x}_0,\tau)+\mathcal{H}\dot{\boldsymbol{\Psi}}_L(\boldsymbol{x}_0,\tau)=-\int d^3\boldsymbol{x}_0 W_R(\boldsymbol{x}_0',\boldsymbol{x}_0)\nabla\Phi(\boldsymbol{x}_0'+\boldsymbol{\Psi}(\boldsymbol{x}_0',\tau))$$
$$=-\int d^3\boldsymbol{x}_0 W_R(\boldsymbol{x}_0',\boldsymbol{x}_0)\nabla\Phi_S(\boldsymbol{x}_0'+\boldsymbol{\Psi}(\boldsymbol{x}_0',\tau))-\int d^3\boldsymbol{x}_0 W_R(\boldsymbol{x}_0',\boldsymbol{x}_0)\nabla\Phi_L(\boldsymbol{x}_0'+\boldsymbol{\Psi}(\boldsymbol{x}_0',\tau))$$
$$=-\nabla\Phi_L(\boldsymbol{x}_0+\boldsymbol{\Psi}(\boldsymbol{x}_0,\tau))+\alpha_S(\boldsymbol{x}_0+\Psi_L(\boldsymbol{x}_0,\tau)) \quad (2.8)$$

We have added the term $\boldsymbol{\alpha}_s(\boldsymbol{q},\boldsymbol{\Psi}_L(\boldsymbol{q},\boldsymbol{\tau}))$ to represent the sources of displacement from the small-scale modes, which may not be well captured by the perturbation theory, following [11].

The solution to such an equation can be written in a perturbative form, where an n^{th}order contribution in the long-wavelength regime can be represented as:

$$\boldsymbol{\Psi}_L^{(n)}(\boldsymbol{k})=\frac{iD^n}{n!}\int\frac{d^3k_1}{(2\pi)^3}...\frac{d^3k_n}{(2\pi)^3}\delta_D(\sum_j\boldsymbol{k}_j-\boldsymbol{k})\boldsymbol{L}_n(\boldsymbol{k}_1,...,\boldsymbol{k}_n)\delta_0(\boldsymbol{k}_1)...\delta_0(\boldsymbol{k}_n) \quad (2.9)$$

where D is the linear growth rate and the perturbative kernels $\boldsymbol{L}_n$ can be found in [6].

Additional contributions come from the $\boldsymbol{\alpha}_s$ term we will parameterize by a number of the counter-terms, which will govern the small-scale mode evolution. This way of modelling the smaller scales is called Effective Field Theory(EFT). Accounting for only the lowest order contributions, the displacement is then written perturbatively as:

$$\boldsymbol{\Psi}(\boldsymbol{x}_0,\tau)=\boldsymbol{\Psi}_L^{(1)}(\boldsymbol{x}_0,\tau)+\boldsymbol{\Psi}_L^{(2)}(\boldsymbol{x}_0,\tau)+\boldsymbol{\Psi}_L^{(3)}(\boldsymbol{x}_0,\tau)+...+\frac{1}{2}\alpha_1\nabla\delta_0+\mathcal{S}+... \quad (2.10)$$

where $\mathcal{S}$ is a term uncorrelated with δ_0, leading to the shot noise term for the power spectrum.

If we define now the quantity $K(\boldsymbol{x}_0,\boldsymbol{k})=\left\langle e^{i\boldsymbol{k}\cdot\boldsymbol{\Delta}}\right\rangle$, where $\boldsymbol{\Delta}(\boldsymbol{x_0},\tau)=\boldsymbol{\Psi}(\boldsymbol{x}_0,\tau)-\boldsymbol{\Psi}(\boldsymbol{0},\tau)$, then following[6] the power spectrum $P(k)$ and the correlation function $\xi(r)$ can be presented as:

$$P(k)=\int d^3x_0\, e^{i\boldsymbol{x}_0\cdot\boldsymbol{k}}(K(\boldsymbol{x}_0,\boldsymbol{k})-1) \quad (2.11)$$

$$1+\xi(r)=\int\frac{d^3x_0\,d^3k}{(2\pi)^3}\,e^{i\boldsymbol{k}\cdot(\boldsymbol{x}_0-\boldsymbol{r})}K(\boldsymbol{x}_0,\boldsymbol{k}) \quad (2.12)$$

$K(\boldsymbol{x}_0,\boldsymbol{k})$ can be expanded using the cumulant theorem [7] as:

$$\log K(\boldsymbol{x}_0,\boldsymbol{k})=-\frac{1}{2}k_ik_jA_{ij}(\boldsymbol{x}_0)+\frac{i}{6}k_ik_jk_lW_{ijl}(\boldsymbol{x}_0)+... \quad (2.13)$$

with:

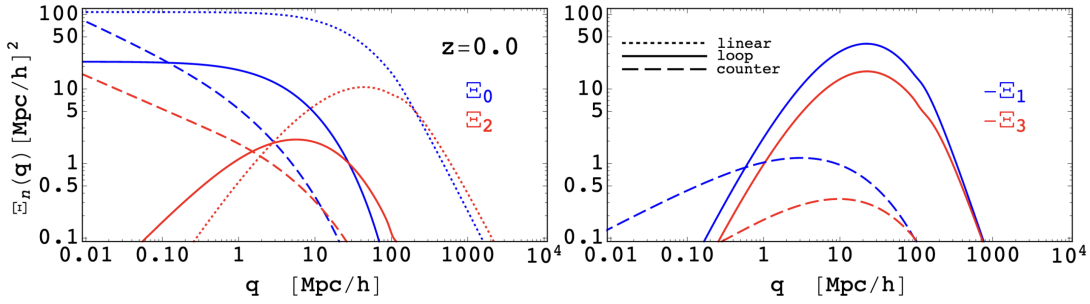

Figure 2.1: The terms entering into the cumulants of Ψ, divided into contributions from different orders and the counter-terms, assuming $\alpha_n = 1$ and $z = 0$. Taken from [11].

$$A_{ij}(\boldsymbol{x}) = 2\langle\Psi_i(\boldsymbol{0})\Psi_j(\boldsymbol{0})\rangle - 2\langle\Psi_i(\boldsymbol{x}_1)\Psi_j(\boldsymbol{x}_2)\rangle \equiv 2(\Sigma^2\delta_{ij} - \eta_{ij}) = \langle\Delta_i\Delta_j\rangle \tag{2.14}$$

$$W_{ijl}(\boldsymbol{x}) = \langle\Psi_{\{i}(\boldsymbol{x_1})\Psi_j(\boldsymbol{x_2})\Psi_{l\}}(\boldsymbol{x_2})\rangle - \langle\Psi_{\{i}(\boldsymbol{x_2})\Psi_j(\boldsymbol{x_1})\Psi_{l\}}(\boldsymbol{x_1})\rangle = \langle\Delta_i\Delta_j\Delta_l\rangle \tag{2.15}$$

where $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$, and $\Delta_i = \Psi_i(\boldsymbol{0}) - \Psi_i(\boldsymbol{x})$

Given that W_{ijk} contains the product of two displacement fields evaluated at the same point, it adds additional counter-terms, which we can obtain using symmetry arguments [13], so we can write the expansion of the product of two displacements as:

$$\Psi_i(\boldsymbol{x})\Psi_j(\boldsymbol{x}) = \Psi_i^{(1)}(\boldsymbol{x})\Psi_j^{(1)}(\boldsymbol{x}) + \Psi_i^{(1)}(\boldsymbol{x})\Psi_j^{(2)}(\boldsymbol{x}) + \Psi_i^{(2)}(\boldsymbol{x})\Psi_j^{(1)}(\boldsymbol{x}) + \ldots + \frac{1}{3}\alpha_0\delta_{ij} + \alpha_2\delta_{ij}\nabla_l\Psi_l^{(1)} + \alpha_3\left[\nabla_i\Psi_j^{(1)} + \nabla_j\Psi_i^{(1)}\right] + \ldots \tag{2.16}$$

Following [11] and using their notation and the integrals they have defined as Ξ_ℓ, we can decompose A_{ij} and W_{ijk} as:

$$A_{ij} = \frac{2}{3}\delta_{ij}(\Xi_0(0) + \alpha_0 - \Xi_0(q)) + \left(\hat{q}_i\hat{q}_j - \frac{1}{3}\delta_{ij}\right)\Xi_2(q) \tag{2.17}$$

$$W_{ijk}(q) = \frac{2}{5}\hat{q}_{\{i}\delta_{jk\}}\Xi_1(q) + \frac{3}{5}(5\hat{q}_i\hat{q}_j\hat{q}_k - \hat{q}_{\{i}\delta_{jk\}})\Xi_3(q) \tag{2.18}$$

Using this decomposition, we can estimate the contributions from different orders and magnitudes in the final expression, which can be seen in Figure 2.1.

Eventually, we can separate the terms A_{ij} into a linear part coming from the linear order contributions and a non-linear part, e.g.:

$$A_{ij}(\boldsymbol{x}) = A_{ij}^{\text{lin}}(\boldsymbol{x}) + A_{ij}^{\text{nonlin}}(\boldsymbol{x}) \tag{2.19}$$

The matter power spectrum $P(k)$ then can be shown to be, gathering everything we have so far:

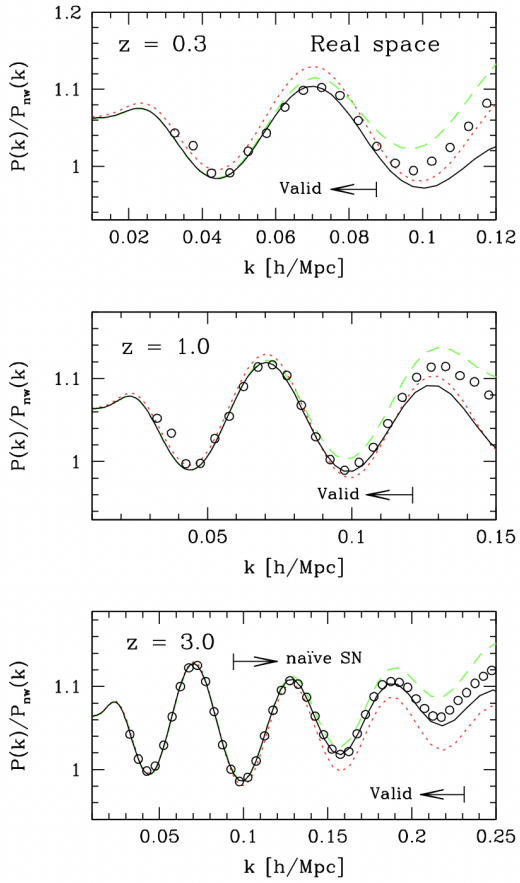

Figure 2.2: Comparison of the power spectra to the N-body simulations at various redshifts. Open circles: N-body simulations, black line: LPT power spectrum, red(dotted) line: linear theory, green theory(dashed) theory: EPT similar to that used in [14]. Taken from [6]

$$P(k) = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q} - (1/2)k_i k_j A_{ij}^{\rm lin}} \left[1 - \frac{1}{2}k_i k_j A_{ij}^{\rm nonlin} + \frac{I}{6}k_i k_j k_k W_{ijk} + \ldots\right] \tag{2.20}$$

This gives us all of the ingredients to obtain a 1-loop corrected matter power spectrum in real space. You can find more details on the computation in [7, 11]. We can see the performance of such a model of a power spectrum in Figure 2.2.

2.2 Galaxy clustering

2.2.1 Lagrangian biases

Cosmological surveys observe discrete tracers such as galaxies rather than the underlying matter distribution. Therefore, one has to take that into account when trying to match the clustering of galaxies to that of the matter field. One of the ways to describe such a

connection would be a bias model, which will describe the connection between the two. In Lagrangian framework first we need to include the so-called bias functional $F[\delta_0(\boldsymbol{q})]$ in the initial conditions:

$$(2\pi)^3\delta_D(\boldsymbol{k}) + \delta_g(\boldsymbol{k}) = \int d^3qF[\delta_0(\boldsymbol{q})]e^{-i\boldsymbol{k}\cdot(\boldsymbol{q}+\Psi(\boldsymbol{q}))} \tag{2.21}$$

which can be further expanded as the polynomial of overdensity $\delta_0(q)$ as:

$$F[\delta_0(\boldsymbol{q})] = 1 + b_1\delta_0 + \frac{1}{2}b_2(\delta_0^2(\boldsymbol{q}) - \langle\delta_0^2\rangle) + b_s(s_0^2(\boldsymbol{q}) - \langle s_0^2\rangle) + b_3O_3(\boldsymbol{q}) \tag{2.22}$$

where $s_0 = (\partial_i\partial_j/\partial^2 - \delta_{ij}/3)\delta_0$ is the initial shear tensor, b_1 can be connected to the linear Eulerian bias as $b_{1,e} = 1 + b_1$. We follow the notation of [15].

We can therefore write the real-space power spectrum, repeating steps of the previous section as:

$$P_{\rm biased}(k) = \int d^3q\ \left\langle F[\delta_0(\boldsymbol{q}_1)]F[\delta_0(\boldsymbol{q}_2)]e^{ik\cdot\Delta}\right\rangle \tag{2.23}$$

We should note, however, that the presence of the galaxy biases modifies the original expansion of eq 2.20. In order to describe it in a more concise manner, we will define some new cumulant quantities:

$$U_i^{mn} = \langle\delta^m(\boldsymbol{x}_0)\delta^n(\boldsymbol{x}_0\Delta_i\rangle \tag{2.24}$$

$$A_{ij}^{mn} = \langle\delta^m(\boldsymbol{x}_0)\delta^n(\boldsymbol{x}_0)\Delta_i\Delta_j\rangle \tag{2.25}$$

Together with quantities $\Upsilon, V_i^{mn}, \theta, \zeta$ etc., expressions for which can be found in, for example, [15], we can write an expression for the real-space biased power spectrum:

$$\begin{aligned}P_{\rm biased}(k) = \int d^3\boldsymbol{q}\, e^{i\boldsymbol{k}\boldsymbol{q}}e^{-\frac{1}{2}k_ik_jA_{ij}^{lin}}\Big\{1 - \frac{1}{2}k_ik_jA^{\rm loop}ij + \frac{i}{6}k_ik_jk_kW_{ijk}\\ + b_1\left(2ik_iU_i - k_ik_jA_{ij}^{10}\right) + b_1^2\left(\xi_{\rm lin} + ik_iU_i^{11} - k_ik_jU_i^{\rm lin}U_j^{\rm lin}\right) +\\ + \frac{1}{2}b_2^2\xi_{\rm lin}^2 + 2ib_1b_2\xi_{\rm lin}k_iU_i^{\rm lin} - b_2\left(k_ik_jU_i^{\rm lin}U_j^{\rm lin} + ik_iU_i^{20}\right) +\\ + b_s\left(-k_ik_j\Upsilon_{ij} + 2ik_iV_i^{10}\right) + 2ib_1b_sk_iV_i^{12} + b_2b_s\chi + b_s^2\zeta +\\ +2ib_3k_iU_{b_3,i} + 2b_1b_3\theta + \alpha_Pk^2 + ...\Big\} + R_h^3\end{aligned} \tag{2.26}$$

2.2.2 Redshift space distortions in LPT

Another important component in the power spectrum of galaxies are redshift space distortions, that we introduced in Introduction. We can model them by boosting the displacement field along the line of sight $\hat{z}$, using the definition of the displacement 2.1 as follows:

$$\boldsymbol{\Psi}_s = \boldsymbol{\Psi} + \frac{\hat{z}(\boldsymbol{v}\cdot\hat{z})}{\mathcal{H}} = \boldsymbol{\Psi} + \frac{\hat{z}\cdot\dot{\boldsymbol{\Psi}}}{H}\hat{z} \tag{2.27}$$

where $\boldsymbol{v} = a\dot{\boldsymbol{x}}$ is the galaxy peculiar velocity.

Following the time-independent approximation, where $\dot{\boldsymbol{\Psi}}^{(n)} \propto D^n$, one can notice that the time derivative of the perturbative kernels becomes:

$$\dot{\boldsymbol{\Psi}}^{(n)} = nHf\boldsymbol{\Psi}^{(n)} \tag{2.28}$$

that allows us to express the redshift-space kernels without time derivatives as:

$$\boldsymbol{\Psi}_s^{(n)} = \boldsymbol{\Psi}^{(n)} + nf(\hat{z}\cdot\boldsymbol{\Psi}^{(n)})\hat{z} \tag{2.29}$$

This transformation is therefore nothing more than a linear mapping of the displacement vector of each order. We can further express it using the n-th order redshift space distortion tensor $R_{ij}^{(n)}$, defined as:

$$R_{ij}^{(n)} = \delta_{ij} + nf\hat{z}_i\hat{z}_j \tag{2.30}$$

that allows us to rewrite Eq. 2.29 as:

$$\boldsymbol{\Psi}_s^{(n)} = R^{(n)}\boldsymbol{\Psi}^{(n)} \tag{2.31}$$

Therefore, the redshift-space density contrast in Fourier space $\delta_s(\boldsymbol{k})$ becomes:

$$(2\pi)^3\delta^D(\boldsymbol{k}) + \delta_s(\boldsymbol{k}) = \int d^3q\, F(\boldsymbol{q})e^{i\boldsymbol{k}\cdot(q+\boldsymbol{\Psi}(\boldsymbol{q})+\hat{z}\cdot\dot{\boldsymbol{\Psi}}\hat{z}/H)} \tag{2.32}$$

Eventually, we define the pairwise displacement field in redshift space as $\Delta_s = \Psi_s(_1) - \Psi_s(_2)$, and we can obtain the redshift-space galaxy power spectrum [15] as:

$$P_{g,s}(\boldsymbol{k}) = \int d^3\boldsymbol{q}\left\langle e^{i\boldsymbol{k}\cdot\Delta_s}F(\boldsymbol{q}_1)F(\boldsymbol{q}_2)\right\rangle_{\boldsymbol{q}=\boldsymbol{q}_1-\boldsymbol{q}_2}. \tag{2.33}$$

where $F(\boldsymbol{q}) = F[\delta_0(\boldsymbol{q})]$.

We have now almost all of the components necessary to build the galaxy power spectrum in redshift space.

2.3 Moment expansion

Having built the foundation in section 2.1, and adding descriptions of some galaxy-related effects for LPT in section 2.2, we can obtain the non-linear redshift-space galaxy power spectrum.

We will then define the generating functional $M(\boldsymbol{J},\boldsymbol{k})$ as:

$$M(\boldsymbol{J},\boldsymbol{k}) = \frac{k^3}{2\pi^2}\int d^3q\, e^{i\boldsymbol{k}\cdot\boldsymbol{q}}\langle F(\boldsymbol{q}_1)F(\boldsymbol{q}_2)e^{i\boldsymbol{k}\cdot\Delta+i\boldsymbol{J}\cdot\dot{\Delta}}\rangle_{\boldsymbol{q}=\boldsymbol{q}_1-\boldsymbol{q}_2} \tag{2.34}$$

One can immediately notice that if we set $\boldsymbol{J} = \boldsymbol{0}$, we get:

$$M(0, \boldsymbol{k}) = \frac{k^3}{2\pi^2} \int d^3q \, \langle F[\delta_0(\boldsymbol{q}_1)] F[\delta_0(\boldsymbol{q}_2)] e^{i\boldsymbol{k}\cdot\Delta} \rangle = \frac{k^3}{2\pi^2} P_{\text{biased}}(\boldsymbol{k}) \tag{2.35}$$

Even more so, we can notice that some other quantities describing the behaviour of the density field can be obtained from the generating functional. For example, one can obtain the velocity auto-spectrum by simply taking a derivative of $M(\boldsymbol{J}, \boldsymbol{k})$ and setting $\boldsymbol{J} = \boldsymbol{0}$:

$$\begin{aligned}
\hat{\boldsymbol{n}} \left. \frac{\partial M(\boldsymbol{J}, \boldsymbol{k})}{\partial \boldsymbol{J}} \right|_{\boldsymbol{J}=\boldsymbol{0}} &= \frac{ik^3}{2\pi^2} \int d^3q \, e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \langle \dot{\Delta} F(\boldsymbol{q}_1) F(\boldsymbol{q}_2) e^{i\boldsymbol{k}\cdot\Delta} \rangle_{\boldsymbol{q}=\boldsymbol{q}_1-\boldsymbol{q}_2} = \frac{k^3}{2\pi^2} \hat{\boldsymbol{n}} \boldsymbol{v}_{\boldsymbol{12}}(\boldsymbol{k}) = \\
&= \hat{n}_i \frac{k^3}{2\pi^2} \int d^3\boldsymbol{q} e^{i\boldsymbol{k}\boldsymbol{q}} e^{-\frac{1}{2}k_i k_j A^{lin}_{ij}} \left\{ ik_j \dot{A}_{ji} - \frac{1}{2} k_j k_k \dot{W}_{jki} + \right. \\
&\quad + 2b_1 \left(\dot{U}_i - k_k U^{\text{lin}}_k k_j \dot{A}^{\text{lin}}_{ji} + k_j \dot{A}^{10}_{ji} \right) + \\
&\quad + b_1^2 \left(2ik_j U^{\text{lin}}_j \dot{U}^{\text{lin}}_i + i\xi_{\text{lin}} k_j \dot{A}^{\text{lin}}_{ji} + \dot{U}^{11} \right) + \\
&\quad + b_2 \left(\dot{U}^{20} + 2ik_j U^{\text{lin}}_j \right) + 2b_1 b_2 \xi_{\text{lin}} \dot{U}^{\text{lin}}_i + 2b_s \left(\dot{V}^{10}_i + ik_j \dot{\Upsilon}_{ji} \right) + \\
&\quad \left. + 2b_1 b_s \dot{V}^{12}_i + 2b_3 \dot{U}_{b_3,i} + \alpha_v k_i + \right\} \ldots + R_h^4 \sigma_v
\end{aligned} \tag{2.36}$$

This quantity is often called the first moment, while the real-space power spectrum corresponds to the zeroth moment. This way of computing the non-linear power spectrum using the momentum functional and its expansions in terms of $\boldsymbol{J}$, which we will later use to add the redshift-space distortions into the mix, is called Moment Expansion. We will also show here the second moment, which is called the velocity dispersion, as it is also an important ingredient for the model we use in this thesis:

$$\begin{aligned}
\sigma_{12,ij}(\boldsymbol{k}) = \int d^3\boldsymbol{q} \, e^{i\boldsymbol{k}\boldsymbol{q}} e^{-\frac{1}{2}k_i k_j A^{lin}_{ij}} &\left\{ \ddot{A}_{ij} + ik_n \ddot{W}_{nij} - k_n k_m \dot{A}^{\text{lin}}_{ni} \dot{A}^{\text{lin}}_{mj} + \right. \\
&+ b_1 \left(2ik_n U^{\text{lin}}_n \ddot{A}^{\text{lin}}_{ij} + 2ik_n \left[\dot{A}^{\text{lin}}_{ni} \dot{U}^{\text{lin}}_j + \dot{A}^{\text{lin}}_{nj} \dot{U}^{\text{lin}}_i \right] + 2\ddot{A}^{10}_{ij} \right) + \\
&+ b1^2 \left(\xi_{\text{lin}} \ddot{A}^{\text{lin}}_{ij} + 2\dot{U}^{\text{lin}}_i \dot{U}^{\text{lin}}_j + \right) + 2b_s \ddot{\Upsilon}_{ij} + \\
&\left. + \alpha_\sigma \delta_{ij} + \beta_\sigma \xi^2_{0,L} \left(\hat{q}_i \hat{q}_j - \frac{1}{3} \delta_{ij} \right) + \ldots \right\} + R_h^3 s_v^2 \delta_{ij}
\end{aligned} \tag{2.37}$$

The expression can be decomposed into transverse and longitudinal components as:

$$\sigma_{ij} = \sigma_0(k) \delta_{ij} + \frac{3}{2} \sigma_2(k) \left(\hat{k}_i \hat{k}_j - \frac{1}{3} \delta_{ij} \right) \tag{2.38}$$

Finally, we can notice that, by setting $\boldsymbol{J} = \boldsymbol{k}$ the factor in the exponent in 2.34 becomes a redshifted displacement difference Δ_s [16], same one as in Eq 2.33.

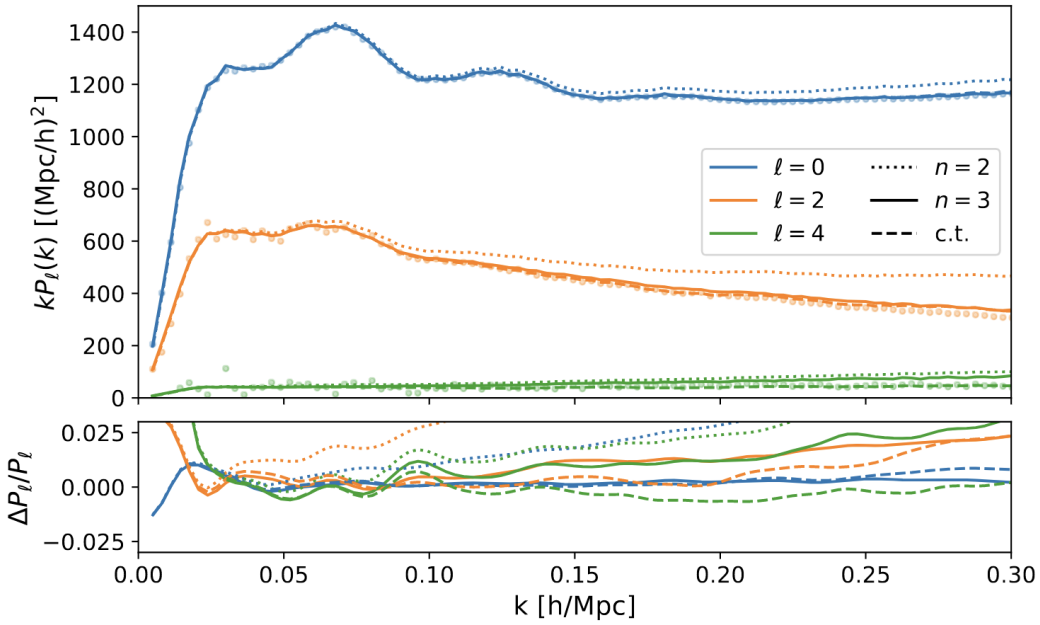

Figure 2.3: Convergence of the moment expansion at $z = 0.8$ for the monopole(blue), quadrupole(orange) and hexadecupole(green) of the power spectrum with n-th order of expansion being used. Dots are representing the results from N-body simulation. Taken from [15].

The last thing which is left to do, is to expand the generating moment functional, which becomes, for $\boldsymbol{J} = \boldsymbol{k}$, an expression of the redshift space galaxy power spectrum, in terms of $\hat{z}\boldsymbol{k}$. Grouping the counter-terms α_i, stochastic terms and the FOG effect σ_v^2, the total galaxy redshift-space power spectrum obtained using the Moment Expansion $P_{g,s}^{\mathrm{ME}}(\boldsymbol{k})$ is therefore given by:

$$\begin{aligned} P_{g,s}^{\mathrm{ME}}(\boldsymbol{k}) = \left(P(k) + i(k\mu)v_{12,\hat{n}}(\boldsymbol{k}) - \frac{(k\mu)^2}{2}\sigma^2_{12,\hat{n}\hat{n}}(\boldsymbol{k})\right) + \\ + \left(\alpha_0 + \alpha_2\mu^2 + \alpha_4\mu^4 + \ldots\right) k^2 P_{\mathrm{lin,Zel}}(k) + R_h^3(1 + \sigma_v^2(k\mu)^2 + \ldots) \end{aligned} \tag{2.39}$$

The performance of such a model in comparison with simulations can be seen on Figure 2.3

There exist other ways of computing the redshift-space galaxy power spectrum, such as streaming models [17], as well as more recent approaches such as the one presented in [18]. However, for the scope of this thesis, we have used the Moment Expansion approach.

In this thesis, we consider the public state-of-the-art Effective Field Theory (EFT) code named `velocileptors` [1] [15, 18] as our reference theory. This model is one

[1] `https://github.com/sfschen/velocileptors`

of the EFT models used in DESI for the Full-Shape analysis of the DR1 galaxy samples. More precisely, we focus on the moment expansion model as implemented in the `MomentExpansion` module of `velocileptors`.

2.4 An emulator for clustering

Despite the high advancements in analytic evaluation of the redshift-space galaxy 2-point statistics, unfortunately, many of the integrals in the formulas presented in the previous sections require numerical integration or lots of cumbersome Fourier Transforms, which often take significant computational time.

For the needs of cosmological inference (which we will describe in more detail in Chapter 5) hundreds of thousands of spectra have to be computed, with different cosmologies and nuisance parameters marginalised over (biases, counterterms, stochastic terms). To make that feasible from the computational point of view, different techniques have been developed in order to accelerate the theoretical predictions or the inference itself or both.

We should also point out that the perturbative approach to modelling that I described in the previous sections is still limited to ~~large scales~~ the quasi-linear regime. With the growth of the wavenumber k, the power spectra from any perturbative approach start to deviate from those obtained in simulations and are not able to describe the observed data. We will also discuss how techniques developed for the speedup of the analytic codes can be used to boost the accuracy of the original theory.

The most natural choice of the quantity to emulate will be the statistic itself [19]. However, this approach shows its limitations when it comes to reusing the same emulator. Usually, the existing codes struggle with redshift evolution and need to be retrained for each given redshift, thus limiting their reusability. Plus, usually, the increase in the number of parameters harms precision, so complicated models with lots of nuisance parameters might experience difficulties in reaching the expected level of accuracy, as the growth of the complexity of the model makes it harder for the neural network of a fixed size to retrain.

One can also emulate a 2-point statistics in the linear order, or some other simplified form, and add certain features by multiplying the output of the emulator with that approximate version [20]. That approach is sometimes used when specific features from, for example, beyond ΛCDM models [21], are necessary, however, it often exhibits similar problems.

We decided to take a slightly different approach, when developing our emulator [1]. In practice, generating a non-linear power spectrum consists of several steps. First, the cosmological parameters are used to create the linear power spectrum in real space. There are many pieces of software which achieve that, most notably class[22] and camb[23,

Table 2.1: Definitions and ranges of the parameters of the training set for the emulator.

Parameter	Interpretation	Prior range
$\omega_{\rm cdm}$	Physical cold dark matter density parameter	[0.05, 0.30]
$\omega_{\rm b}$	Physical baryon density parameter	[0.01, 0.04]
$\log\left[10^{10}A_s\right]$	Normalization of the matter power spectrum	[2, 4]
n_s	Spectral index of the primordial power spectrum	[0.8, 1.1]
h	Normalized Hubble constant at $z = 0$	[0.5,0.8]
w_0	Static part of the Dark Energy equation of state	[-2, -0.5]
w_a	Dynamic part of the Dark Energy equation of state	[-3, 0.3]
z	Redshift	[0.0, 1.4]

24]. That can also take some significant time. We have decided to use cosmoprimo package with class, to achieve that. After that, we have used the velocileptors tool[15, 18], to compute the bias invariant terms for moment expansion. Finally, the multiplication with the nuisance parameters is a simple arithmetic operation, which does not require any specific software to be used.

Looking at equations 2.26, 2.36, 2.37, we can see that the nuisance parameters enter the equations multiplied with perturbation theory cumulants defined in 2.18, 2.17. These terms do not depend on biases or counterterms. So, we chose to emulate them instead from the input cosmological parameters.

2.4.1 Architecture

Having identified the possible candidates for emulation, we can assemble them in a format suitable for emulation. In total that gives us 30 quantities, which we emulate in 50 k-bins. What helps is that the angular dependence is also multiplicative, which decreases the number of degrees of freedom potentially needed for emulation. That gives us a data-vector of 1500. We use 6 cosmological parameters as an input for the model, which are summarised in Table 2.1.

We have chosen to go for the neural network approach to replace the generation of the linear power spectrum and the bias-invariant terms. The two approaches are summarised in a diagram in Figure 2.4

Usually a fully connected neural network (which is the one we are using) can be used to approximate a function f such that $\mathbf{y} = f(\mathbf{x}|\boldsymbol{\theta})$, where $\mathbf{x}$ represents the features of the data set, $\mathbf{y}$ the desired outputs, and $\boldsymbol{\theta}$ the free parameters of the network which also can be referred to as trainable parameters. The optimal function f is defined by the set of parameter values $\boldsymbol{\theta}$ that minimises the loss function (the form of which is discussed below). The loss function provides a measure of the performance of the model when

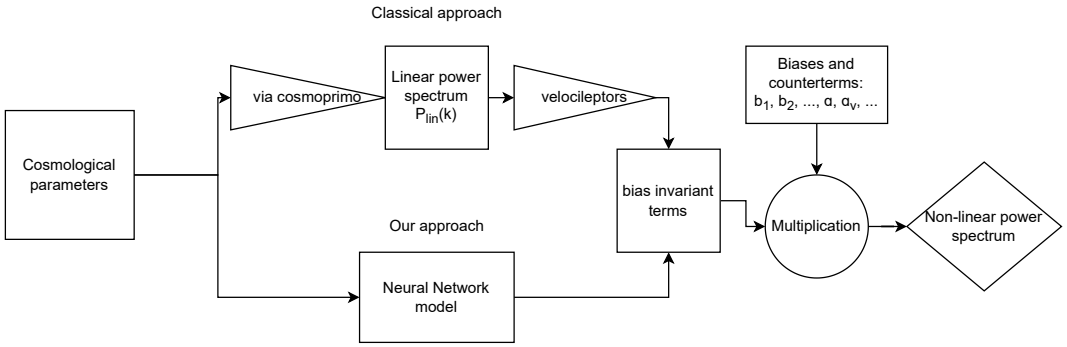

Figure 2.4: Schematic of the theory module in the analysis pipeline: the classical approach consists in first predicting the linear power spectrum and then computing the non-linear power spectrum with an EFT model such as velocileptors. We propose to replace the computation of the linear power spectrum and of the bias-invariant terms in the PT model by a neural-network emulator. These bias-invariant terms that depend only on the cosmological model are combined with the a set of bias and nuisance parameters, common to the velocileptors and NN pipeline, to predict the non-linear redshift-space galaxy power spectrum.

evaluated on the data set.

Fig. 2.5 presents the architecture of our neural network. The input parameters of the ΛCDM model are $\boldsymbol{C} = \left\{\omega_{\rm cdm}, \omega_{\rm b}, \log\left[10^{10}A_s\right], n_s, h\right\}$ and the redshift z. The training set comprises of 3000 samples drawn from a Latin hypercube featuring the six cosmological parameters, and the redshift in the range $0 < z < 1.4$. All the datasets are generated using `MomentExpansion` module of `velocileptors`. Table 2.1 summarises the very broad flat un-informative priors used for the cosmological parameters when defining the Latin hypercube. We group the training data into a 31×50 matrix, where 31 is the number of bias-independent terms described in Section 2.3 and 50 is our fiducial choice for the number of k bins.

A feed-forward fully-connected model based on the machine learning framework `pytorch`[2] is created for each such matrix. We use 2 hidden layers of 16 328 neurons and the Gaussian Error Linear Units (GELU) activation function [25], which can be represented as

$$\mathrm{GELU}(x) = 0.5x\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{2.40}$$

The outputs and the inputs x_i are normalised $x_i \in [-1, 1]$.

The training is done in batches of 128 for 1000 epochs, meaning that the training dataset is divided into groups of 128, where the elements of each group are then simultaneously passed through the neural network, and after that the weights are adjusted using

[2] `https://github.com/pytorch/pytorch`

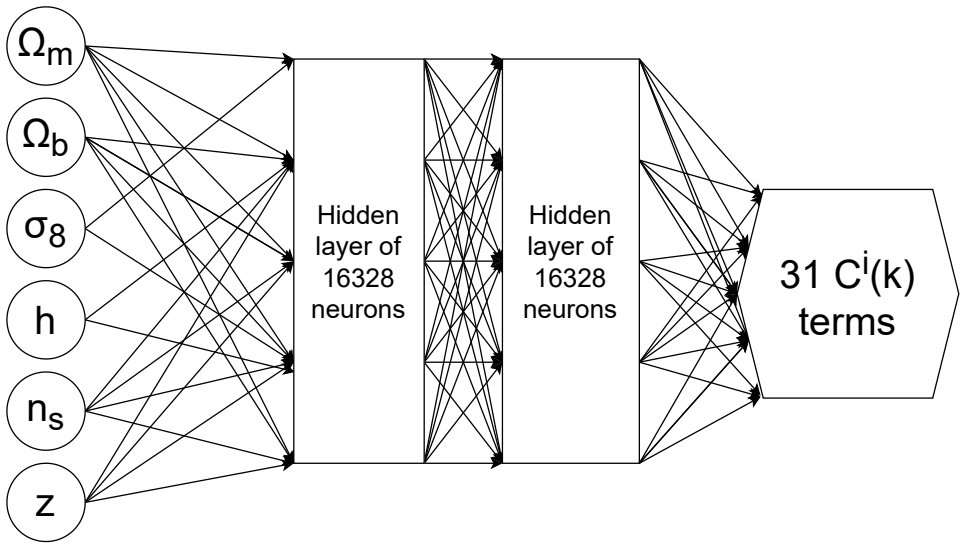

Figure 2.5: Architecture of the neural network emulator: The 6 ΛCDM cosmological parameters and the redshift z are used as input parameters of a fully-connected neural network model composed of 2 hidden layers. The output of the neural network emulator are the predicted 31 bias-invariant terms that enter the PT predictions, binned in 50 bins of $k = [0.0, 0.3]$

backpropagation. These groups are called batches, and we do this until all of the possible groups have been used. That constitutes an epoch. This procedure is therefore repeated 1000 times. The validation dataset consists of 1000 samples, constituting a hypercube with the same parameters as the training data. We minimise the L1 norm loss function (ref here) defined by:

$$\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N} |y^i_{\rm true} - y^i_{\rm predicted}|, \tag{2.41}$$

with optimisation performed using `pytorch` realisation of the Adam optimiser [26]. The learning rate is set to 4×10^{-7}. We stop the training after 100 epochs when the validation loss is not improving.

2.4.2 Testing the emulator performance

First, we assess the performance of the emulator in predicting the Legendre multipoles of the power spectrum defined as:

$$P_\ell(k) = \frac{(2\ell+1)}{2}\int_{-1}^{1} d\mu\, P(k,\mu)\mathcal{L}_\ell(\mu) \tag{2.42}$$

where $L_\ell(\mu)$ is the Legendre polynomial of order ℓ. In this work, we consider the monopole $\ell = 0$, the quadrupole $\ell = 2$ and the hexadecapole $\ell = 4$.

Table 2.2: Ranges of the parameters used for the multipole testing.

Parameter	Range
$\omega_{\rm cdm}$	[0.10, 0.14]
$\omega_{\rm b}$	[0.01, 0.03]
$\log\left(10^{10}A_s\right)$	[2.5, 3.5]
w_0	[-2, -0.5]
w_a	[-3, 0.3]
h	[0.64, 0.72]
n_s	[0.9, 1.0]
b_1	[-1, 3]
b_2	[-10, 10]
b_s	[-20, 20]
b_3	[-20, 20]

In order to assess the performance of the emulator at the level of the multipoles, we generate $N = 10\,000$ sets of the cosmological and nuisance parameters taken from the ranges given in Table 5.5. Then, we produce the multipoles using both the original `velocileptors` code and our emulator. Fig. 2.6 shows the ratio of the neural network LPT emulator multipole $P_{l,\rm NN}$ to the theoretical prediction from `velocileptors` $P_{l,\rm th}$ for the monopole (top), quadrupole (middle) and hexadecapole (bottom). The dashed curves show the 3σ scatter. Up to = 0.25, the overall multipoles computed from the emulator agree with the ones from the reference analytic version at below 0.5% at 3σ, which means below 0.2% at 1σ.

Fig. 2.7 shows the same information as Fig. 2.6 but for $w_o w_a$CDM model. We recover a similar accuracy even for this extended cosmological model with the emulator predicting the multipoles at a precision below 0.2% at 1σ up to = 0.25.

We test the improvement in speed to compute the power spectrum multipoles by generating 50 batches of multipoles with the variable number of multipoles in each $n_b = [0, 10, 25, 50, 100, 200]$, such that we can estimate the performance boost as a ratio between the elapsed time for their production with the original code to the time taken by the emulator. The corresponding proportion is then plotted against the number of multipoles in a single batch in Fig. 2.8. We attribute most of the speed growth with increasing the batch size as due to the parallelisation over Graphics Processing Units (GPU), a feature that the original `velocileptors` software does not support due to the very sequential nature of its code.

As was mentioned earlier, the main usage of our emulator was in cosmological inference. That is why the most important test we did was on how well the pipeline using it as a theory yields the expected results, and how much can we trust the output. The additional

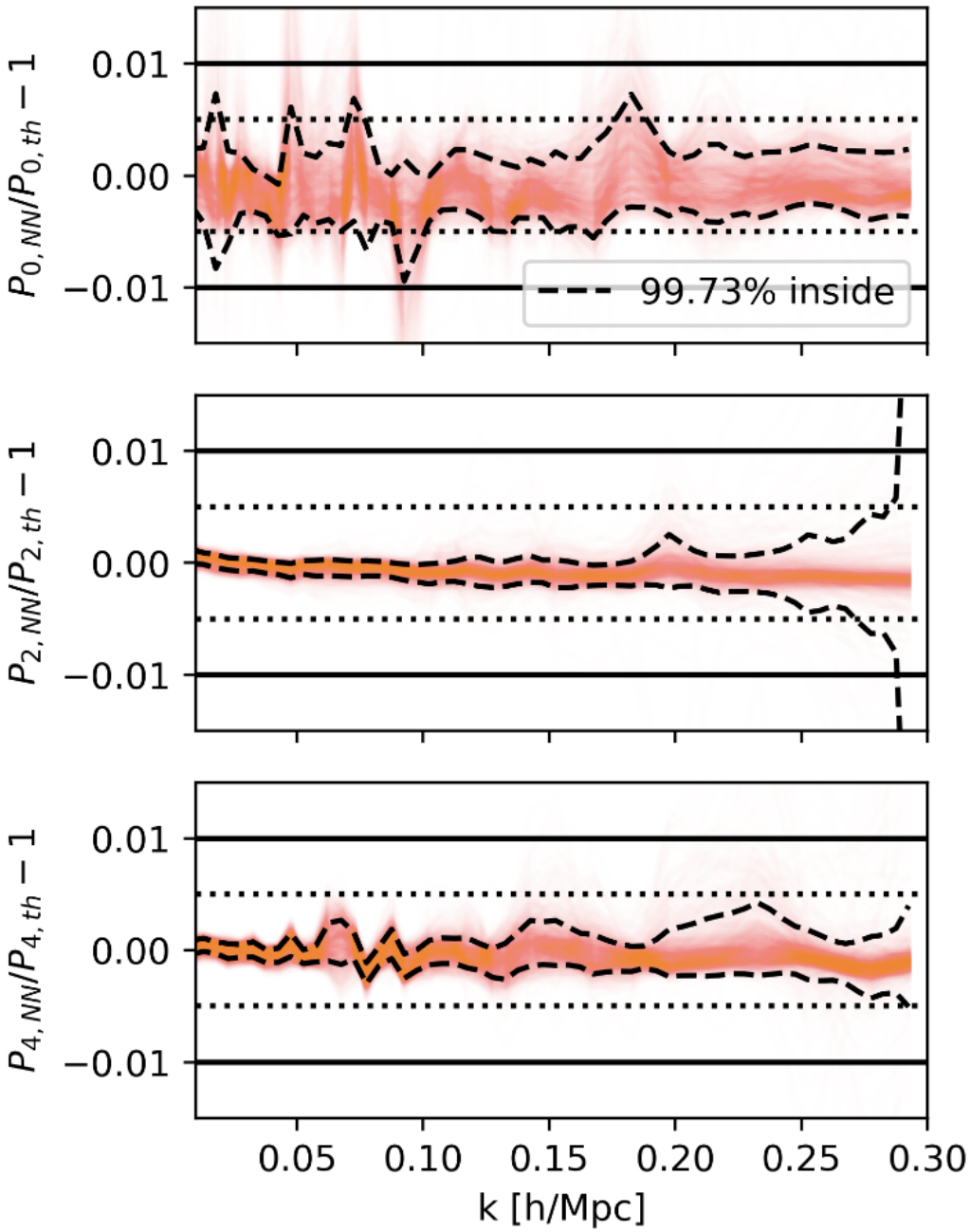

Figure 2.6: Comparison between the galaxy redshift space power spectrum multipoles of the emulator $P_{\ell,\mathrm{NN}}$ and of the theoretical version $P_{\ell,\mathrm{th}}$ for $\ell = 0$ (top), $\ell = 2$ (middle) and $\ell = 4$ (bottom). The dashed curves represent the 3σ scatter and the red curves the individual realisations.

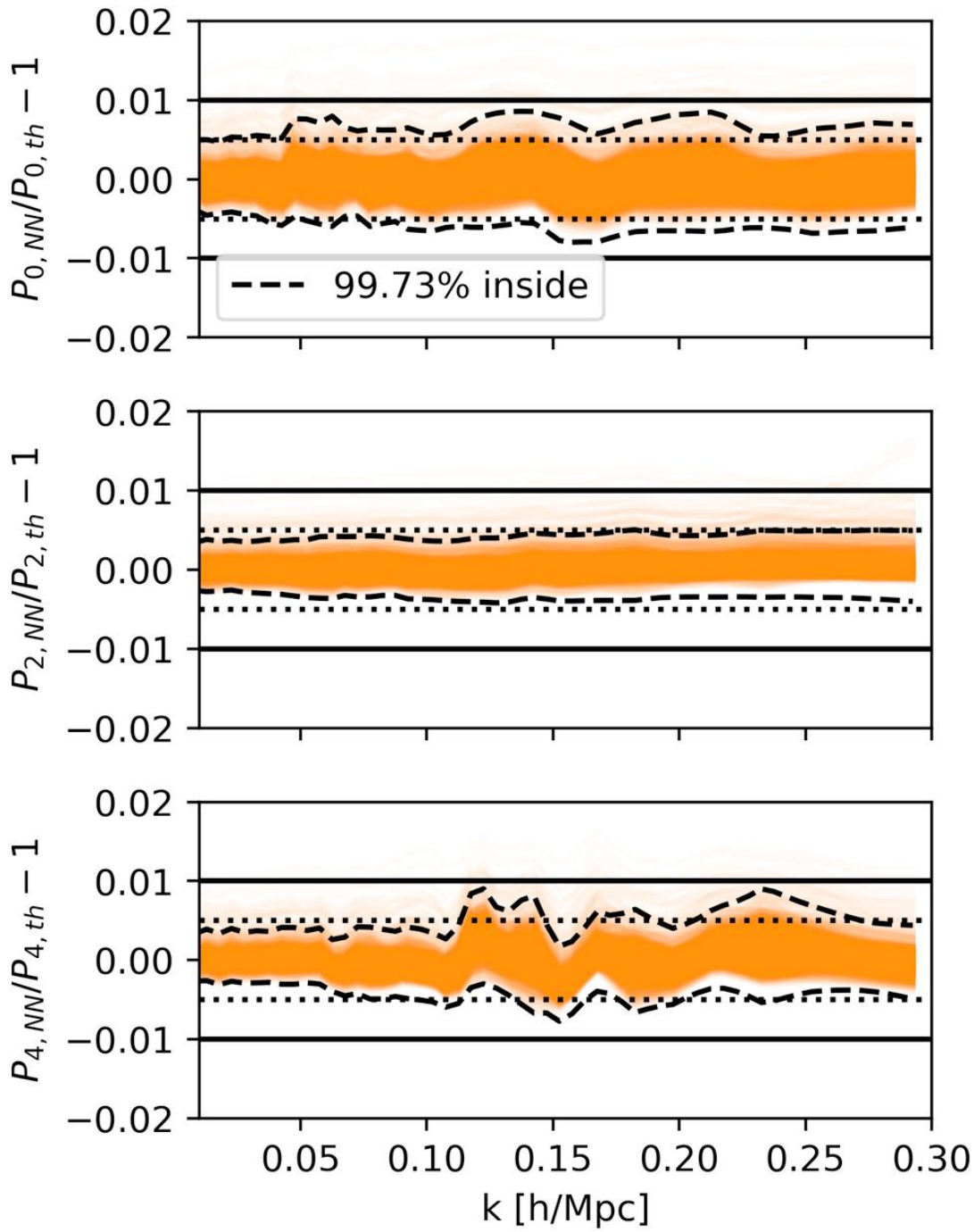

Figure 2.7: Same as for Fig. 2.6 but for $w_o w_a$CDM. We recover a performance slightly worse than that for the ΛCDM case, due to the 2 additional parameters, but still $\sim 0.5\%$ in precision at 3σ.

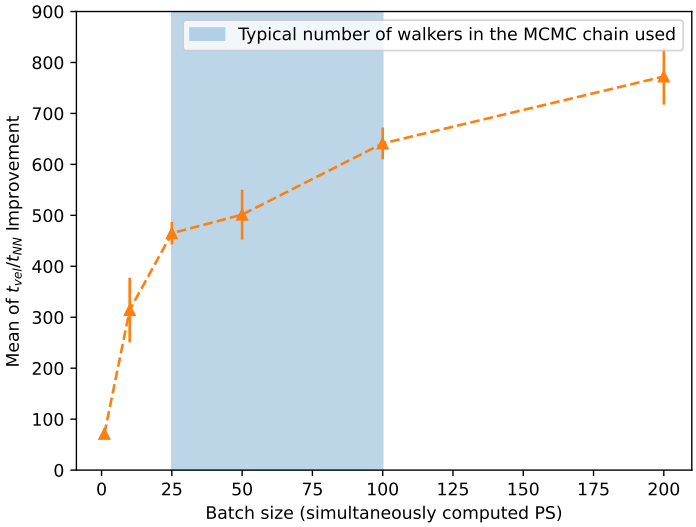

Figure 2.8: Speed performance of the neural network emulator with respect to the original code as a function of the number of simultaneously computed multipoles. The ratio of computation time for the time with original code to that of our emulator is plotted against the batch size: number of simultaneously computed non-linear power spectra.

tests on the cosmological inference will be described in Chapter 5.

Having all those tests done, we managed to obtain an instrument which will make some of the analyses previously too demanding computationally viable. It will be used extensively for different types of analysis performed in Chapter 5.

References

[1] Svyatoslav Trusov et al. 2024. arXiv: 2403.20093 [astro-ph.CO].

[2] T. Buchert. "A class of solutions in Newtonian cosmology and the pancake theory". In: 223.1-2 (Oct. 1989), pp. 9–24.

[3] F. Moutarde et al. "Precollapse Scale Invariance in Gravitational Instability". In: 382 (Dec. 1991), p. 377. DOI: 10.1086/170728.

[4] E. Hivon et al. "Redshift distortions of clustering: a Lagrangian approach." In: 298 (June 1995), p. 643. DOI: 10.48550/arXiv.astro-ph/9407049. arXiv: astro-ph/9407049 [astro-ph].

[5] A. N. Taylor and A. J. S. Hamilton. "Non-linear cosmological power spectra in real and redshift space". In: *MNRAS* 282.3 (Oct. 1996), pp. 767–778. ISSN: 0035-8711. DOI: 10.1093/mnras/282.3.767. URL: https://doi.org/10.1093/mnras/282.3.767.

[6] Takahiko Matsubara. "Resumming cosmological perturbations via the Lagrangian picture: One-loop results in real space and in redshift space". In: *Phys. Rev. D* 77 (6 Mar. 2008), p. 063530. DOI: 10.1103/PhysRevD.77.063530. URL: https://link.aps.org/doi/10.1103/PhysRevD.77.063530.

[7] Takahiko Matsubara. "Nonlinear perturbation theory with halo bias and redshift-space distortions via the Lagrangian picture". In: *Phys. Rev. D* 78 (8 Oct. 2008), p. 083519. DOI: 10.1103/PhysRevD.78.083519. URL: https://link.aps.org/doi/10.1103/PhysRevD.78.083519.

[8] Jordan Carlson et al. "Convolution Lagrangian perturbation theory for biased tracers". In: *MNRAS* 429.2 (Dec. 2012), pp. 1674–1685. ISSN: 0035-8711. DOI: 10.1093/mnras/sts457. eprint: https://academic.oup.com/mnras/article-pdf/429/2/1674/18465345/sts457.pdf. URL: https://doi.org/10.1093/mnras/sts457.

[9] Vladislav Zheligovsky and Uriel Frisch. "Time-analyticity of Lagrangian particle trajectories in ideal fluid flow". In: *Journal of Fluid Mechanics* 749 (2014), pp. 404–430. DOI: 10.1017/jfm.2014.221.

[10] Takahiko Matsubara. "Recursive solutions of Lagrangian perturbation theory". In: *Phys. Rev. D* 92 (2 July 2015), p. 023534. DOI: 10.1103/PhysRevD.92.023534. URL: https://link.aps.org/doi/10.1103/PhysRevD.92.023534.

[11] Zvonimir Vlah et al. "Lagrangian perturbation theory at one loop order: Successes, failures, and improvements". In: *Phys. Rev. D* 91 (2 Jan. 2015), p. 023508. DOI: 10.1103/PhysRevD.91.023508. URL: https://link.aps.org/doi/10.1103/PhysRevD.91.023508.

[12] Zvonimir Vlah et al. "A Lagrangian effective field theory". In: *JCAP* 2015.09 (Sept. 2015), p. 014. DOI: 10.1088/1475-7516/2015/09/014. URL: https://dx.doi.org/10.1088/1475-7516/2015/09/014.

[13] Rafael A. Porto et al. "The Lagrangian-space Effective Field Theory of large scale structures". In: *JCAP* 2014.05 (May 2014), p. 022. DOI: 10.1088/1475-7516/2014/05/022. URL: https://dx.doi.org/10.1088/1475-7516/2014/05/022.

[14] Bhuvnesh Jain and Edmund Bertschinger. "Second-Order Power Spectrum and Nonlinear Evolution at High Redshift". In: 431 (Aug. 1994), p. 495. DOI: 10.1086/174502. arXiv: astro-ph/9311070 [astro-ph].

[15] Shi-Fan Chen et al. "Consistent modeling of velocity statistics and redshift-space distortions in one-loop perturbation theory". In: *JCAP* 2020.07 (July 2020), p. 062. DOI: 10.1088/1475-7516/2020/07/062. URL: https://dx.doi.org/10.1088/1475-7516/2020/07/062.

[16] Zvonimir Vlah and Martin White. "Exploring redshift-space distortions in large-scale structure". In: *JCAP* 2019.03 (Mar. 2019), p. 007. DOI: 10.1088/1475-7516/2019/03/007. URL: https://dx.doi.org/10.1088/1475-7516/2019/03/007.

[17] Zvonimir Vlah et al. "The Gaussian streaming model and convolution Lagrangian effective field theory". In: *JCAP* 2016.12 (Dec. 2016), p. 007. DOI: 10.1088/1475-7516/2016/12/007. URL: https://dx.doi.org/10.1088/1475-7516/2016/12/007.

[18] Shi-Fan Chen et al. "Redshift-space distortions in Lagrangian perturbation theory". In: *JCAP* 2021.03 (Mar. 2021), p. 100. DOI: 10.1088/1475-7516/2021/03/100. URL: https://dx.doi.org/10.1088/1475-7516/2021/03/100.

[19] Carolina Cuesta-Lazaro et al. "Galaxy clustering from the bottom up: a streaming model emulator I". In: 523.3 (Aug. 2023), pp. 3219–3238. DOI: 10.1093/mnras/stad1207. arXiv: 2208.05218 [astro-ph.CO].

[20] Raul E Angulo et al. "The BACCO simulation project: exploiting the full power of large-scale structure for cosmology". In: *MNRAS* 507.4 (July 2021), pp. 5869–5881. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2018. eprint: https://academic.oup.com/mnras/article-pdf/507/4/5869/40419371/stab2018.pdf. URL: https://doi.org/10.1093/mnras/stab2018.

[21] Renate Mauland et al. 2023. arXiv: 2309.13295 [astro-ph.CO].

[22] Diego Blas et al. "The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes". In: *JCAP* 2011.07 (July 2011), p. 034. DOI: 10.1088/1475-7516/2011/07/034. URL: https://dx.doi.org/10.1088/1475-7516/2011/07/034.

[23] Antony Lewis et al. "Efficient computation of CMB anisotropies in closed FRW models". In: 538 (2000), pp. 473–476. DOI: 10.1086/309179. arXiv: astro-ph/9911177 [astro-ph].

[24] Cullan Howlett et al. "CMB power spectrum parameter degeneracies in the era of precision cosmology". In: 1204 (2012), p. 027. DOI: 10.1088/1475-7516/2012/04/027. arXiv: 1201.3654 [astro-ph.CO].

[25] Dan Hendrycks and Kevin Gimpel. "Gaussian Error Linear Units (GELUs)". In: *arXiv e-prints*, arXiv:1606.08415 (June 2016), arXiv:1606.08415. DOI: 10.48550/arXiv.1606.08415. arXiv: 1606.08415 [cs.LG].

[26] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014), arXiv:1412.6980. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs.LG].

Chapter 3

Simulations

И пусть кругом грохочут
глухо громы,
Пусть веет вихрь сомнений и
обид, —
Явь наших снов земля не
истребит!

M. Voloshin, Translation

Introduction

Despite major progress in the development of the cosmological theoretical framework, there are still many processes which can not be properly described analytically. Perturbation theory has its limits, and breaks down on small scales. Hence the conventional analytical ways to model the behaviour of the matter clustering start to fall drastically in accuracy, as presented in the previous chapter. However, we are not restricted to only analytical means.

That is where the cosmological simulations come into play. They serve many functions besides modelling the clustering of the matter field, as they can also capture the noise caused by the statistical uncertainty in the survey, thus making them crucial for the error estimation of the cosmological parameters inferred. Moreover, simulations are used as a benchmark to test the analysis pipelines.

In this chapter we will describe different types of cosmological simulations for galaxy clustering, starting with the most accurate N-body ones and then slightly simpler variants using particle meshes. After that, we will describe the ways of populating the obtained dark matter distributions with galaxies, together with several features required for the analysis of the DESI BGS such as magnitudes and optical colours. We will then present the simplified mock generation techniques, which generate the galaxies directly, bypassing

the necessity for modelling the galaxy-halo connection, before concluding with an overall comparison of the mock generation techniques. In particular, I will present the mocks I have worked with and developed for different purposes that mimic the clustering of the DESI BGS.

3.1 N-body simulations

An N-body simulation is a simulation which treats the matter density as a set of particles gravitationally interacting with each other, making it a numerical solution to an N-body problem, hence the name. This makes them, compared to other approaches, the most accurate when it comes to the modelling of the density field evolution. They are usually run with periodic boundary conditions to simulate the infinite size of the Universe, when used for cosmology. Eventually, the density is obtained by averaging over the distribution of the simulated particles. The evaluation of the forces acting upon the particles and solving their equations of motion are the two principal components of the simulation.

3.1.1 Equations of motion

The equations which govern the evolution of the particles are usually obtained by starting with the standard Newtonian equations of motion in their comoving coordinates. Coupled with the gravitational potential ϕ which characterises the gravitational field, we get the following set of expressions:

$$\ddot{\boldsymbol{x}} + 2\frac{\dot{a}}{a}\dot{\boldsymbol{x}} = -\frac{1}{a^2}\nabla\phi \tag{3.1}$$

And the Poisson equation for the gravitational potential:

$$\nabla^2\phi = 4\pi G\rho(t)a^2\delta = \frac{3}{2}H_0\Omega_0\frac{\delta}{a} \tag{3.2}$$

where $\boldsymbol{x}$ denotes the position of the particle. They describe the behaviour of the non-relativistic matter at scales smaller than the Hubble radius [1].

We can notice, thanks to the first derivative term $\dot{x}$, that the expansion of the Universe acts as a viscous force, which opposes gravitational infall, and results in a slower growth of density perturbations in an expanding Universe. The equations are usually solved in steps, the size of which decreases as we get close to $z = 0$.

One should recall, however, that due to the discrete nature of the particles, we obtain certain granularity artefacts. The way to deal with it is usually by assigning a certain volume to a particle, thus blurring the mass in the volume around it. The procedure is characterised by the smoothing kernel $W(r, h)$, where h is a characteristic length of the particle volume, also known as the softening length, and r is the distance from the particle The most commonly used form of such function used is the cubic spline of [2]:

$$W(r,h) = \frac{8}{\pi h^3} \begin{cases} 1 - 6\left(\frac{r}{h}\right)^2 + 6\left(\frac{r}{h}\right)^3, & 0 \leq \frac{r}{h} < 0.5 \\ 2\left(1 - \frac{r}{h}\right), & 0.5 \leq \frac{r}{h} < 1 \\ 0, & 1 \leq \frac{r}{h} \end{cases} \tag{3.3}$$

This modifies the gravitational force between two particles of masses m to be approximately:

$$\boldsymbol{F} = -G\frac{m^2}{(r^2 + h^2)^{\frac{3}{2}}}\boldsymbol{r} \tag{3.4}$$

This introduces a bias into the interaction on small scales but also smooths the interaction. At the same time, the softening length h can not be made arbitrarily small, as it will reintroduce the discreteness effects. Therefore, a certain compromise is usually made when choosing this quantity.

One can immediately notice, that for the simulations to have a combination of a good resolution, hence enough particles, and large volume one will have to deal with evaluating the forces, which will be very heavy computationally. There exist many ways of dealing with or bypassing this problem.

3.1.2 N-body approaches

Particle-particle approach

If we decide to treat every particle pair as we are supposed to, we are soon to notice that the pairwise summation is of complexity $O(N^2)$, where N is the number of particles. One of the ways to deal with that problem was implemented in the `Abacus` code [3], and originally developed in [4]. It relies on separating the given volume into a 3-D cartesian grid, and computing the force for a given particle depending on whether the other particle is closer than a chosen threshold, and is called the near-field computation. Otherwise, the multipole decomposition is used for the gravitational potential. It is computed for each cell up to a given order p. As the Cartesian grid features discrete translational symmetry, the summation of forces from cells further than the chosen threshold can be performed as a convolution using discrete Fourier transforms, which significantly speeds up the process, and also allows to painlessly introduce periodic boundary conditions [3, 4], which used to be a problem before [5]. Another notable code using the multipole decomposition for the far region is `SWIFT`[6].

Near-field force kernel is a simple $\sim \frac{1}{r^2}$ law that is easily computed, given the repeated geometric structure of the near-field domain, allowing for a great acceleration due to co-processing potential. Overall, the combination of those two tricks and the usage of modern GPUs allows for a viable solution for Particle-Particle N-body simulations, which were rarely used for cosmology before. The illustration on how the particle interactions

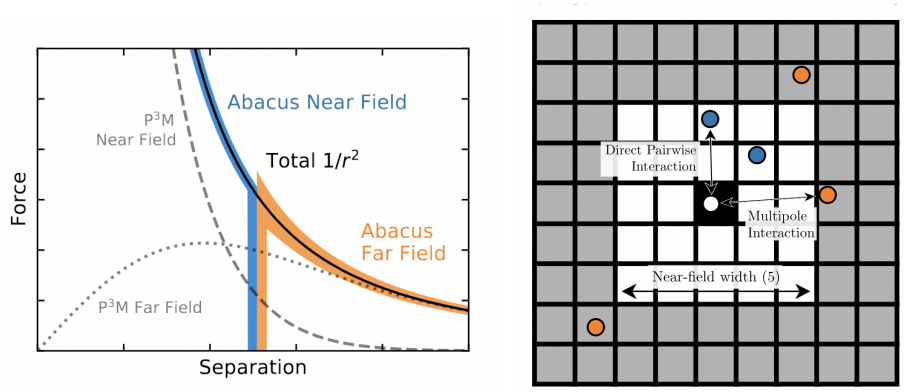

Figure 3.1: *Left panel*: a schematic illustration of the exact near-field/far-field separation of the force computation in `Abacus`. The grey lines represent other schemes, like particle mesh, where the far-field is compensated by the near field on the smaller scales. As for the `Abacus` the force is given by one of the near field and far field, represented by the shaded lines. *Right panel*:The domain decomposition. Forces for the particles in the black cell are computed in the near-field mode from the particles in the white cells, and with a far-field approximation for the particles in the grey cells.

are separated into the far-field and near-field regimes, and how does the force looks like with respect to the pure particle mesh approach is in Figure 3.1.

Particle Mesh approach

In order to escape the detailed computation of the individual pair-wise forces, one can build the mesh to estimate the gravitational potential in a given point. The particles are painted onto the mesh using a given window function $W(r)$. Then, the Poisson equation for gravitational potential is solved on it 3.2. The computation can be additionally fastened by the use of Fast Fourier Transforms. More on that can be found in [7]. An example of the application of such an approach can be the `GLAM` code, which will be used and presented in more details later in this chapter.

The use of FFTs also enables to implement the periodic boundary conditions straightforwardly. One of the drawbacks from such an approach is the dependence of the simulation on the mesh size. However, it does allow to create simulation with the largest number of particles with respect to other approaches.

The mesh can be additionally optimised by using the adaptive mesh refinement, where the mesh is refined in the domains of larger density, allowing for more accurate results. That is the case, for example, for the `RAMSES` code [8]. An example of such a refined mesh from a cosmological N-Body simulation is shown in Figure 3.2, where different degrees of refinement are marked by different colors.

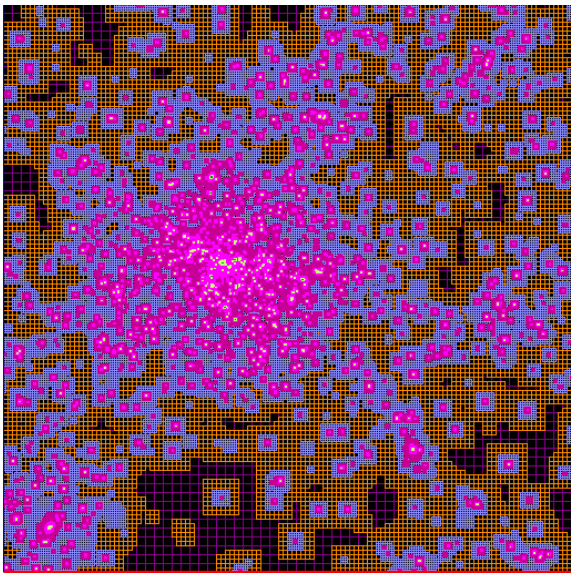

Figure 3.2: Example of an adaptive mesh grid, obtained during a cosmological simulation, where each color corresponds to a given level of refinement. Taken from `https://irfu.cea.fr/Phocea/Vie_des_labos/Ast/ast_sstechnique.php?id_ast=904`

P^3M approach

To solve the problem of computational efficiency, sometimes, similarly as for the approach used in Abacus simulations, the simulation volume is divided into cubic cells, and the far-field interactions are treated differently, often with a particle mesh approach instead.

Such an approach provides a hybrid treatment, ensuring accuracy on small scales by exactly solving the near-field interactions, yet still ensuring faster computing times than that of Particle-Particle analogues by relying on Particle Mesh for far-field interactions [9].

Tree approach

Another way to ease the computational costs of the Particle-Particle approach is by using the tree structures [10]. The simulation volume is recursively divided into cubic regions until each of the regions contains only one particle. The largest cells are used to define the groups of particles, and serve for the calculation of the interactions. The groups are based on cell acceptance criterion [10].

One can combine the tree approach with P^3M approach by employing it for the near-field computations in dense regions. That is for example the case for the `GreeM` code [11, 12], which is the code used to create the Uchuu simulation [13], which will be also used and presented in more detail later in this chapter. Another very popular code used for modelling of dark matter clustering is `Gadget`[14].

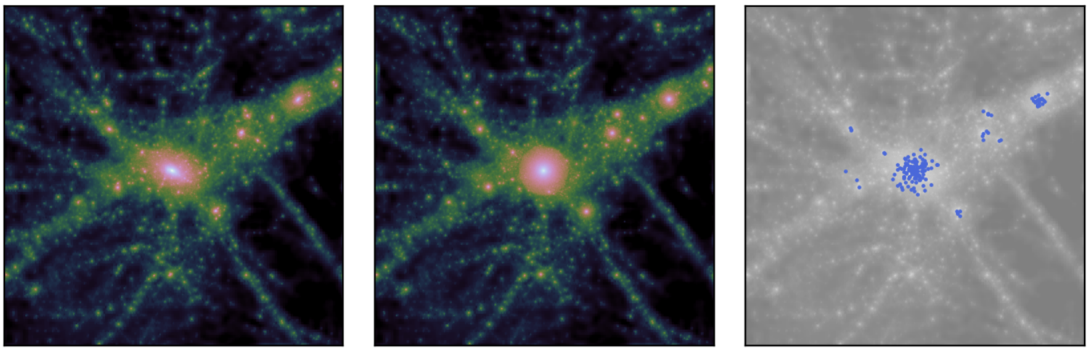

Figure 3.3: *Left panel*: The matter density field in a $25 \times 25 \times 5(h^{-1}Mpc)^3$ region of an N-body simulation, centered on a massive halo at $z = 0$. *Central panel*: The matter density field represented by identified spherical halos. *Right panel*: An example on how galaxies could populate the given halo distribution. Taken from [15].

3.2 Halo model

N-Body simulations typically yield as an output a collection of particles, which represent the dark matter density field. However, to further process the simulation into a realistic replica of the Universe, one needs to reduce the complexity of the simulated data. This can be achieved by approximating the cosmic structures by dark matter halos. The study of those resulting halos and their connection to the observable matter, which is called halo model, allows to connect them to the galaxy distribution, allowing for a powerful tool of studying the matter distribution.

The halo model is based on the assumption that the complex structure of the cosmic web can be described by separating it into massive clumps (halos) with given properties, such as the halo profile, mass distribution, and bias with respect to the underlying matter distribution. We will also use the halo model later in order to populate the dark matter simulations with galaxies. An example of the halos representing cosmic structure using the N-body simulation is shown in Figure 3.3.

3.2.1 Halo formation

If we consider a Universe with $\Lambda = 0$, the spherical overdensity can be described in a pure Newtonian framework.

$$\ddot{r} = -\frac{GM}{r^2} \tag{3.5}$$

where r is the radius of the sphere and M is the enclosed mass. We can solve this equation parametrically:

$$r = A(1 - cos\theta) \tag{3.6}$$

$$t = B(\theta - sin\theta) \tag{3.7}$$

The solution can also be presented in the form of $r(t)$ as a series:

$$r(t) = \frac{A}{2}\left(\frac{6t}{B}\right)^{2/3}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{2/3} + ...\right] \tag{3.8}$$

where A and B are integration constants.

The overdensity within the sphere will be:

$$\delta(t) = \frac{3}{20}\left(\frac{6t}{B}\right)^{2/3} \tag{3.9}$$

Initially the sphere grows as the Universe expands, however the overdensity enclosed within the sphere slows down that process. At the turnaround point when $\theta = \pi$, the sphere reaches its maximum size $r = 2A$, stops growing and starts to collapse. At the end of the collapse time of $t = 2\pi B$ (corresponding to $\theta = 2\pi$), the collapse can be assumed complete, and the critical overdensity becomes $\delta_c = 1.69$, following equation 3.9. This model is called the spherical collapse model [16, 17].

In the real Universe, of course, the collapse never leads to the overdensity contracting to a point ($r = 0$ for $\theta = 2\pi$). Rather, the sphere reaches a virial equilibrium called virialisation, where the kinetic energy and potential energy are related through $V = -2K$.

3.2.2 Halo finding

Thanks to the spherical collapse model, we know have an idea about how dark matter halos form overdense gravitationally bound systems. We can now attempt to identify them using different halo finder techniques (an extensive review can be found here: [18]). Here I will only mention and discuss some of them.

Spherical overdensity

Historically the first halo-finding algorithm was presented in [19]. It is based on finding the spherical regions in a simulation having a certain mean overdensity, with the assumption that the mass inside the region spehircally collapsed. First the local density is computed for each particle by finding the distance r_N to the Nth nearest neighbour. Then, the density ρ_i of i-th particle is defined as:

$$\rho_i = \frac{3(N+1)}{4\pi r_N^3} \tag{3.10}$$

The particles are sorted by density. The particle with the highest density is assumed to be at the centre for the first sphere, whose radius is increased until the mean overdensity falls lower than a certain limit. The centre of mass of particles is then taken as the new

centre of a potential halo, and the process is repeated until the centre shift reaches a certain allowed minimum. The particles in the sphere are then removed, and the process is repeated for the particles left in the simulation volume until all of the halos are found. The halos can intersect in the case of this approach, and therefore the halo-merging procedure has to be performed. Usually, if the centre of a smaller sphere is inside a larger sphere, the smaller sphere is then merged with the larger sphere, leaving the larger radius intact.

Friends-Of-Friends

The Friends-of-Friends algorithm [20] (often denoted as FoF) is based on connecting the neighbouring particles in a simulation. Two particles are linked if they are closer than a certain linking length b in units of the mean inter-particle separation. This creates a net of linked particles which are then assigned to be in the same halo. By construction, this ensures that the particles are uniquely assigned to a halo, and no halo-merging is needed. However, when two structures are too close, they can form a "FoF bridge", that might result in unusual shapes.

CompaSO

Developed as a group-finding algorithm for Abacus N-Body code, CompaSO[21] (COMPetitive Assigment to Spherical Overdensities) runs on-the-fly. It is using a combination of the previous two approaches. First, it measures the local density using a kernel of the form $W = 1 - \frac{r^2}{(0.4b)^2}$, where b is the linking length defined in the previous section about FoF. The so-called L0 halos are then created with FoF from particles with a density higher than a given threshold. Inside those L0 halos, the main L1 halos are formed. The particle with the highest density becomes the first halo nucleus, and following the Spherical Overdensity method, the sphere with radius R_{L_1} is formed. The particles inside 80% of R_{L_1} are removed from further consideration. The remaining particles of the L0 group are then used to potentially form other halos using the same principle. In case one particle can be assigned to more than one halo, the competitive assignment is performed, which only allows the reassignment of the particle to the new halo if its enclosed density with respect to that particle is at least twice that of the old one. The halo-merging algorithm is also present, though works slightly differently than in the original approach.

ROCKSTAR

ROCKSTAR[22] (Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement) is a halo-finding algorithm which uses both the position and velocity of the particles.

It starts by creating FoF groups in real space, using a slightly larger linking length than usual. The phase-space metric is then computed in 6D-phase space for particles in each

of the FoF groups, which is defined as:

$$d = \left(\frac{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^2}{\sigma_x^2} + \frac{|\boldsymbol{v}_1 - \boldsymbol{v}_2|^2}{\sigma_v^2} \right) \tag{3.11}$$

where $\overline{x}_i$ is the position of i-th particle, $\overline{v}_i$ is its velocity, and σ_x^2, σ_v^2 are the dispersions of the corresponding positions and velocities inside the given group. The particles are then linked with a phase-space linking length, which is chosen depending on the standard deviation of the particle positions and velocities, forcing 70% of them to form a subgroup. The process is repeated inside each subgroup until a minimum size of particles in the subgrouop (usually 10) is achieved and a hierarchical set of structures is achieved. The final level will contain the halo-seeds, and each particle will be assigned to their closest seed. Once all particles are distributed, the user can optionally remove all of the particle which are not considered gravitationally bounded. The code defines halo masses using the Spherical Overdensity criteria, with virial mass as default, and only the subset of the innermost particles is used to compute centres and velocities.

3.2.3 Halo Occupation Distribution

Once we have approximated the complex structure of the cosmic web with halos, and we have managed to identify them in the simulations, there is still one step left for our simulation to be physical. As we observe not the dark matter distribution, but the galaxies in the sky, we need to somehow connect the two. It will be more natural for the bigger halos to have more baryons, thus, meaning that we can assume a higher probability of having more galaxies in more massive dark matter halos. The probability $P(N|M)$ of having N galaxies in the halo of mass M is called the Halo Occupation Distribution, which we will further denote as HOD. It is usually divided into two parts: one for so-called central galaxies, which occupy the centre of the halo and share its peculiar velocity (there can be no more than one per halo), and another one is for the number of satellite galaxies.

A popular form of the HOD [23], using the division into central and satellite galaxies[24] which was used to populate the LRG galaxies [25], can be given by the following empirical formula:

$$\langle N_{cen}(M) \rangle = \frac{A_c}{2} \left[1 + \mathrm{erf} \left(\frac{\log M - \log M_{min}}{\sigma_{\log M}} \right) \right] \tag{3.12}$$

$$\langle N_{sat}(M) \rangle = \langle N_{cen}(M) \rangle \left(\frac{M - M_0}{M_1'} \right)^{\alpha} \tag{3.13}$$

where M_{min} can be interpreted as the mass of the haloes for which only a half will be populated with a central galaxy, $\sigma_{\log M}$ becoming the step-size, M_0 is the cutoff mass scale, whereas M_1' is considered as a normalisation term, whilst α is there to introduce the power law slope. A_c is the constant used to set the height of the step function. It should be

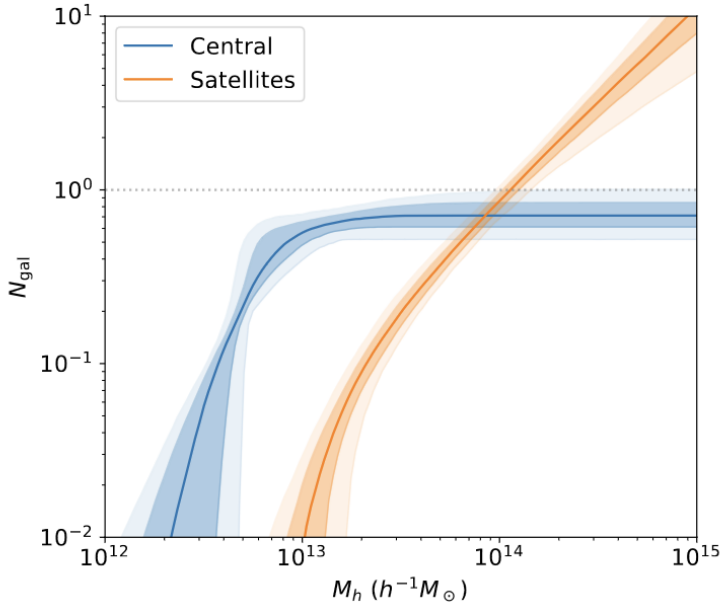

Figure 3.4: The best-fit HOD using the 6-parameter HOD model presented in equation 3.34 for LRG sample of DESI in the redshift range $0.4 < z < 0.6$, where the shaded regions correspond to 1σ and 2σ posteriors. Taken from [25].

noted, that $\sigma_{\log M}$ is also assumed to describe the Gaussian scatter of the in the logarithm of galaxy baryonic masses, which is not always the case [26]. An example of such HOD model with uncertainties for the DESI LRG Abacus mocks is shown in Figure 3.4. We can see that there can be no more than 1 central galaxy and that they dominate the numbers when it comes to low-mass halos.

These formulas can be extended by making their parameters dependent on redshift, luminosities, and other quantities, allowing for the creation of more realistic mocks with photometric quantities. That was done, for example, for the BGS MXXL mock [27] and later on for other BGS mocks. The details on that extension for BGS will be presented in [28].

It should also be noted, that other HOD prescriptions exist. For example, for the ELG sample a modified High Mass Quenched model[29] (mHMQ) was developed and used, which provides the following expressions for central and satellite galaxies:

$$\langle N_{cen}(M)\rangle = \frac{A_c}{\sqrt{2\pi}\sigma_M} e^{-\frac{(\log_{10} M - \log_{10} M_c)^2}{2\sigma^2_{\log M}}} \left[1 + \left(\frac{\gamma(\log_{10} M - \log_{10} M_c)}{\sqrt{2}\sigma_{\log M}}\right)\right] \tag{3.14}$$

$$\langle N_{sat}(M)\rangle = A_s \left(\frac{M - M_0}{M_1}\right)^{\alpha} \tag{3.15}$$

The HOD is obtained by fitting the clustering statistics from the data to the one in the mocks by varying the HOD parameters. The process can be significantly sped up by

using emulators, for example based on Gaussian processes for modelling the likelihood for the fitting, instead of generating the galaxies for each step and computing the 2-point statistics from the resulting simulation [30]. The process can also be sped up by the employment of the tabulated HOD method, developed in [31], which consists of dividing the halos by a given property (halo mass for example) in narrow bins, after which the pair counts are computed. It allows for the on-fly changing of the HOD, as the only thing needed to be changed are the weights resulting from the employed HOD. Convolving the halo occupation with this clustering can significantly speed up the computation of the correlation functions with different HODs.

It was shown, using N-Body simulations by [32] that the distribution of mass in a halo is in fact characterised only by the halo mass M and the concentration parameter c, while the central density ρ_s and the scale radius r_s can be derived. Therefore the mass density profile of halos $\rho(r)$ follows a Navarro-Frenk-White (NFW [32]) profile:

$$\rho(r) = \frac{\rho_s}{(r/r_s)(1 + r/r_s)^2} \tag{3.16}$$

$$V_c(r) = V_{vir}\sqrt{\frac{f(cr/r_{vir})}{r/r_{vir}f(c)}} \tag{3.17}$$

where $f(x) = \ln(1 + x) + \frac{x}{1+x}$ and $V_c(r)$ is the circular velocity.

An example of the performance of such a density profile can be seen in Figure 3.5 where $t = 0$ and $t = 40$ represent the time of the formation of the halo and 40 crossing times, which is the time of circular velocity reaching its maximum at the radius of $r_{max} = 2.163r_s$ [33].

Once the halo profile is obtained, we can position the satellite galaxies following the density profile and assign their velocities randomly, which is drawn from a distribution with a velocity dispersion $\sigma_v^2(M) = \frac{GM}{2r_{vir}}$, such that they follow the NFW profile. The central galaxies however, inherit the position and the velocity of the host halo.

That procedure allows us to place the galaxies on top of the halos obtained from our simulations, and create in such a manner a realistic simulation volume with galaxies that cluster in a similar statistical way as the one seen in galaxy surveys.

3.2.4 Abundance matching

There is another way to connect galaxies to their host halos. It starts with an assumption that one particular galaxy property is monotonically related to a halo property (for example, M) by matching their abundance[24, 26, 34–36]. One of the galaxy properties often used to perform such a technique is galaxy luminosity. A major strength of such an approach is that it can also use the infromation the properties of substructures inside the halos. There is a version of it, which performs such a matching on subhalos instead of halos, and is

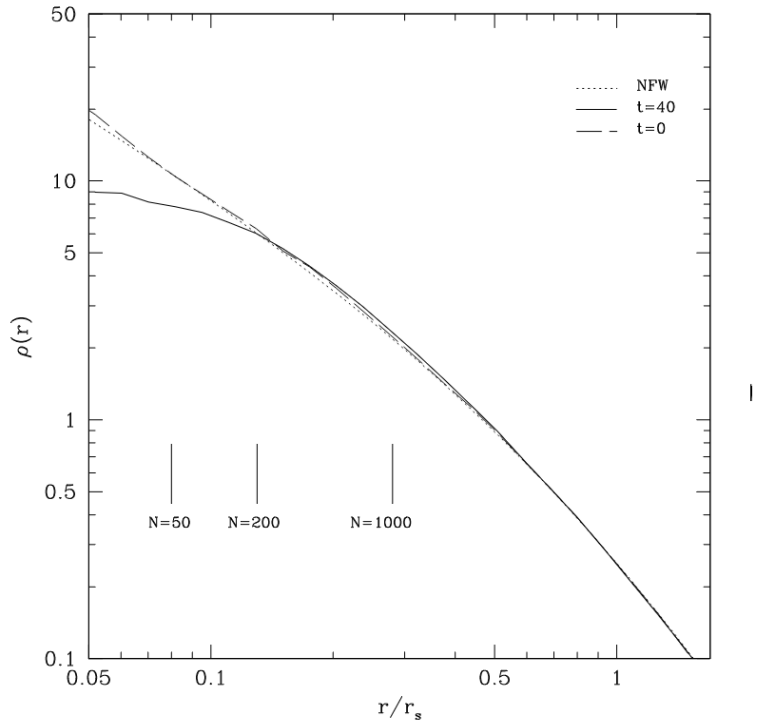

Figure 3.5: The density profile of the NFW halo. The solid curve shows the density profile after 40 crossing times, where the circular velocity reaches the maximum at the radius of $r_{max} = 2.163 r_s$, the dashed line shows the profile at the forming time of the halo, and the dotted line shows the NFW profile. Taken from [33], there the details on the simulations where this halo was produced can be found

called SHAM (Sub-Halo Abundance Matching)[37–39], which is what has been used for the DESI BGS Uchuu mocks[40].

3.3 Mocks without halos

For certain tasks, such as the conventional covariance matrix generation for clustering analysis, thousands of mocks are required. Given that modern surveys probe unprecedented high number densities of objects and large volumes, producing accurate N-Body simulations which match the observations becomes a very challenging task, even with modern computational resources and all the advancements in generating N-Body simulations. However, other techniques were developed to produce approximate mocks, which are often precise enough to build a precise covariance matrix for certain summary statistics (we will discuss that in more detail in the next chapter), or to perform certain tests where large statistics are needed.

3.3.1 Lognormal mocks

One of the simplest ways to create a galaxy survey mock is to assume a pure Gaussian field to model the density field. Mocks, produced using that approach are called the lognormal mocks.

The lognormal distributed density contrast $\delta(\vec{x})$ is related to a Gaussian field $G(\vec{x}) = \ln[1+\delta(\vec{x})] - \langle \ln[1+\delta(\vec{x})] \rangle$ as:

$$\delta(\vec{x}) = e^{-\langle G^2 \rangle + G(\vec{x})} - 1 \tag{3.18}$$

The two-point correlation function $\xi(r)$ is related to the correlation function of the Gaussian field $\xi_G(r)$ as:

$$\xi_G(r) = \log[1+\xi(r)] \tag{3.19}$$

So, a fiducial power spectrum $P(k)$ can be transformed into the correlation function $\xi(r)$, which is then converted to the correlation function of the Gaussian field using eq. (3.19). We Fourier transform it to the power spectrum $P_G(k)$ and eventually generate the Fourier space Gaussian field $G(k)$ as:

$$G(k) = \sqrt{\frac{P_G(k)V}{2}}(\theta_r + i\theta_i) \tag{3.20}$$

where θ_r, θ_i are Gaussian random variables with unit variance and zero mean, and V is the volume of the simulation [41]. After simulating the Fourier space Gaussian field $G(k)$ on the grid, we then use Fast Fourier Transform (FFT) to transform it and obtain the regular configuration space Gaussian field $G(x)$. This is then transformed into the over-density

field using eq. (3.18). The expectation value for the number of galaxies in a particular cell is computed given a fixed mean number density $\bar{n}$, and galaxies are then drawn using the Poisson distribution and placed randomly in the cell. Velocities are then assigned using the linearised continuity equation33, but in proper coordinates:

$$a(t)\frac{\partial\delta(\vec{x})}{\partial t} + \vec{\nabla}\cdot\vec{v}(\vec{x}) = 0 \tag{3.21}$$

where $a(t)$ is a scale factor. This equation is solved using Zeldovich approximation [42].

Eventually, the RSD effect is modelled at a chosen redshift using the velocity information by affecting the coordinates of the galaxy x^i as:

$$x^i_{\text{rsd}} = x^i + f(\boldsymbol{n}\cdot\boldsymbol{v})n^i \tag{3.22}$$

where x^i_{rsd} are the redshift-distorted coordinates, f is the linear growth rate of structure, $\vec{v}$ is the velocity of the galaxy, and n^i is the line of sight.

The main advantage of such a modelling technique is that it is extremely cheap computationally, compared to other approaches, which enables the creation of thousands of independent mocks with different galaxy number densities using relatively low computational costs. However, the small-scale accuracy is very poor, as basically no small-scale computations or physics are accounted for in the process of their creation.

3.3.2 Effective Zeldovich mocks (EZmock)

We can add a bit more complexity to the mock creation and try to include more physical, less linear effects while keeping the process relatively cheap computationally, using an approach developed in [43, 44].

Following LPT, we consider the first-order solution, which is given by the Zeldovich approximation as:

$$\boldsymbol{\Psi}(\boldsymbol{q}) = \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \frac{i\boldsymbol{k}}{k^2}\delta(\boldsymbol{k}) \tag{3.23}$$

The Zeldovich density field is then painted on a mesh of a selected size. Once the density field is simulated, the simulation volume is populated with galaxies. In order to do that, we define a general bias function B which connects the density of galaxies ρ_g to the dark matter density ρ_m:

$$\rho_t = B(\rho_m) \tag{3.24}$$

Then, we notice that to form gravitational bound systems, a minimum local density ρ_c is required to overcome the background expansion. That corresponds to the first term in our bias function $\theta(\rho_m - \rho_c)$, where $\theta(x)$ is a step function. One can then introduce the density saturation ρ_{sat}, and an exponential term with ρ_{exp} responsible for the exponential

cut-off of the halo bias relation, and therefore, for the galaxies as well. Equation 3.24 then becomes:

$$\rho_g = \theta(\rho_m - \rho_c)\rho_{sat}\left[1 - e^{-\frac{\rho_m}{\rho_{exp}}}\right] B_s \tag{3.25}$$

where B_s stands for the stochastic bias term, which serves as a random rescaling factor. This bias is defined as:

$$B_s = \begin{cases} 1 + G(\lambda), & G(\lambda) \geq 0; \\ \exp(G(\lambda)), & G(\lambda) < 0. \end{cases} \tag{3.26}$$

where $G(\lambda)$ stands for a random number drawn from a Gaussian distribution centred at 0 and with a standard deviation of λ. The exponential form is used to avoid negative bias values.

In order to further correct the galaxy number density ρ_g and map it to the number of galaxies per grid cell, we model the probability distribution function for them by a power-law relation:

$$P(n_g) = Ab^{n_g} \tag{3.27}$$

This will serve as an additional bias description.

Of course, one of the important components of every galaxy mock is the velocity distribution, which will govern the redshift-space distortions. In this approach, the peculiar velocity of the galaxy $\boldsymbol{u}_g$ is generated with the help of the velocity of the interpolated densty field $\boldsymbol{u}_m$ as:

$$\boldsymbol{u}_g = \boldsymbol{u}_m + G(\nu) \tag{3.28}$$

Where $G(\nu)$ is a random number taken from a Gaussian distribution with a variance of ν.

Thus, we come to the scheme with 6 parameters, where in practice λ and ρ_{sat} are fixed due to their degeneracy with others. This leaves 4 parameters left. These parameters are calibrated using the survey data or the N-body simulations by eye. Once the matching parameters are found, the last additional step is to ensure the correct variance of the Alcock-Paczynski by adjusting the parameter responsible for the amlitude of the BAO peak.

EZmock has been shown to provide a fast and efficient manner to obtain thousands of mocks for precise covariance matrices, which have higher accuracy than lognormal mocks. An example of such mocks performance for BOSS/eBOSS data is presented in figure 3.6.

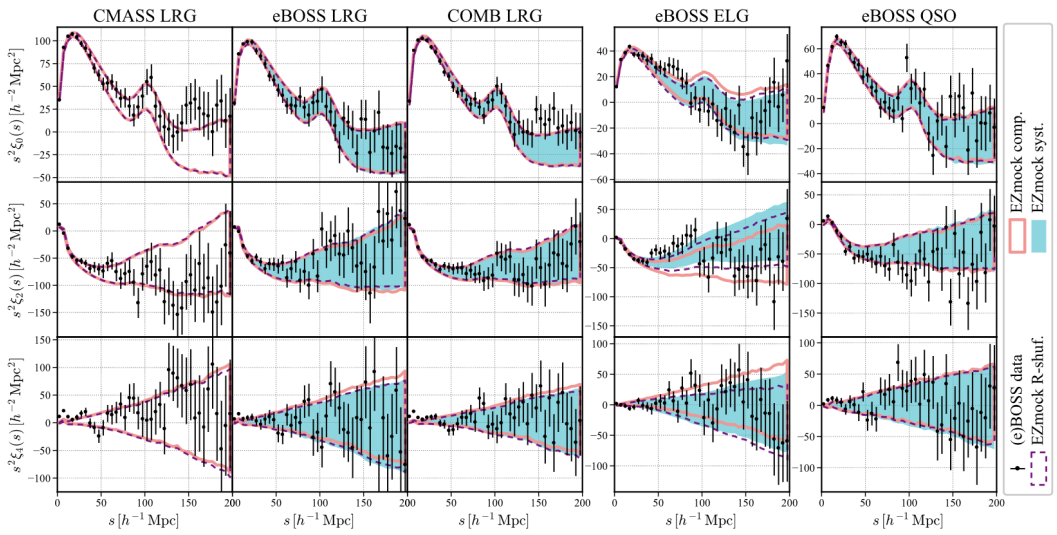

Figure 3.6: Correlation function multipoles of the BOSS/eBOSS data and the corresponding EZmock catalogues in the Southern Galactic Cap. The shaded regions and solid/dashed envelopes indicate 1σ regions evaluated from 1000 realisations, with different systematic effects applied. Taken from [44].

3.4 Mock species for DESI BGS

Several mock species were produced in order to mimic the DESI BGS, each one with different purposes in mind and using various techniques. In this section we will start with the common features of those mocks that they need to have, and then we will discuss each mock variety in detail. The summary of mocks basic parameters can be seen in Table 3.1.

Mock	Technique	Sample	№mocks	Main use
Uchuu	TreePM, SHAM	BGS Bright	1	Systematics, Halos, Small scales
Abacus	PP, HOD	BGS Bright	25	Systematics, Halos, Test fits
GLAM	PM, HOD	$Mr < -21.5$ only	1000	Covariance, Test fits
EZmock	Effective Zeldovich	$Mr < -21.5$ only	1000	Covariance

Table 3.1: Summary table of 4 different mock species used to model BGS for different purposes. PP stands for N-body particle-particle approach and PM stands for N-body particle mesh approach.

3.4.1 BGS specificities

The DESI BGS presents several specificities that make this sample particularly challenging to simulate:

- as mentioned in Chapter 1, it is an extremely dense sample with $n \sim 10^{-2}h^3/Mpc^{-3}$, which is at least 2 orders of magnitude denser than the LRGs and ELGs.
- The mass range spanned by the dark matter halos that host BGS galaxies is very broad from roughly 10^{10} to $10^{15}M_\odot/h$, which put also tight constraints on the mass resolution of the simulation.
- eventually, given that we aim at performing a multi-tracer analysis between blue and red galaxies of the BGS, we also need to assign some photometric properties to the dark matter halos that host BGS galaxies, such as magnitude and colour. We will describe how to perform this assignment in the following when describing each type of BGS mocks.

Systematics

Once we have the simulations whose footprint matches the data survey and which are populated with realistic galaxies, we are still missing one last step to properly mimic the data sample. We need to take into account the observational systematics that affect the data. In particular, we need to simulate the fibre-assignment in order to account for fibre incompleteness as discussed into more details in Chapter 1.

In order to do that, DESI uses a tool called `fast-fiberassign`, which is an emulator of the fiber collisions on the mocks. It uses Friend-of-Friends algorithm to detect the nearby lying galaxies. Then it uses the data of the placements from different fibres and the detected blobs of galaxies to simulate the effect. More on that can be found here:[45].

3.4.2 Abacus

AbacusSummit suite of simulations was created specifically for DESI, in order to meet our scientific requirements. It uses the PP approach, described in the beginning of this chapter. The suite is composed of 150 simulation boxes, covering 97 cosmological models. The base simulations feature 6912^3 particles with a mass of $\sim 2 \cdot 10^9 M_\odot/h$ in a 2 Gpc/h box. The fiducial cosmology for the 25 of them is Planck2018. More details about the AbacusSummit suite of simulations and the other cosmologies featured can be found here [46]. The plot presenting all of the Abacus cosmologies can be found in Figure 3.7. We note that for the BGS modelling we have used 25 boxes with box size of $2000Mpc/h$ and `Planck2018` cosmology[47]

Realistic BGS mocks are produced using AbacusSummit boxes by [28], where HODs are fitted to the clustering of MXXL mock[27]. In order to properly infer the information on photometric quantities, such as magnitudes, the HOD parameters are made dependent on the magnitude. The HOD fits are then performed separately for magnitudes ${}^{0.1}M_r = [-22, -18]$ with the step of $\Delta^{0.1}M_r = 0.5$. The HODs for such a bestfit, performed for

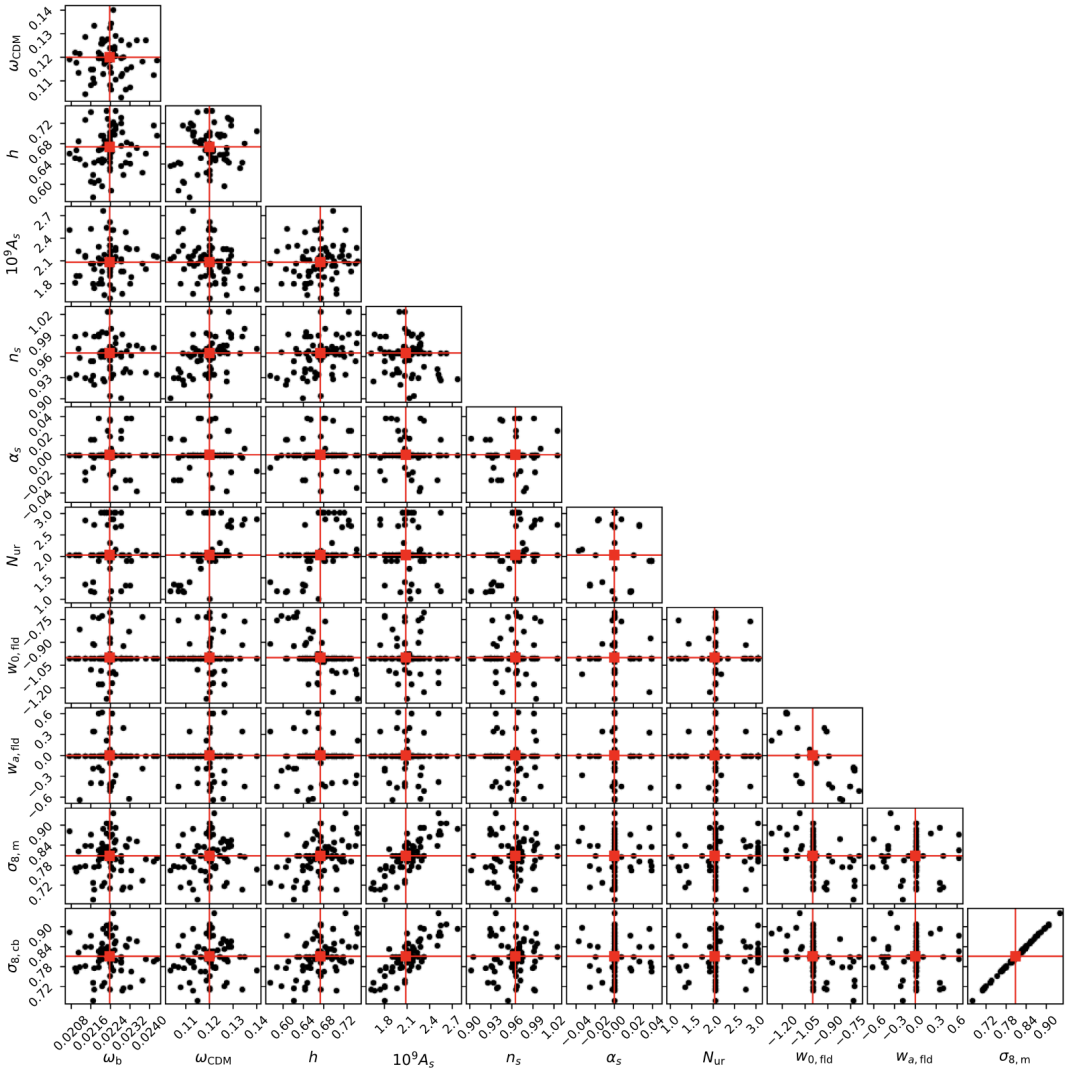

Figure 3.7: The distribution of cosmological parameters of different Abacus simulations. Taken from [46].

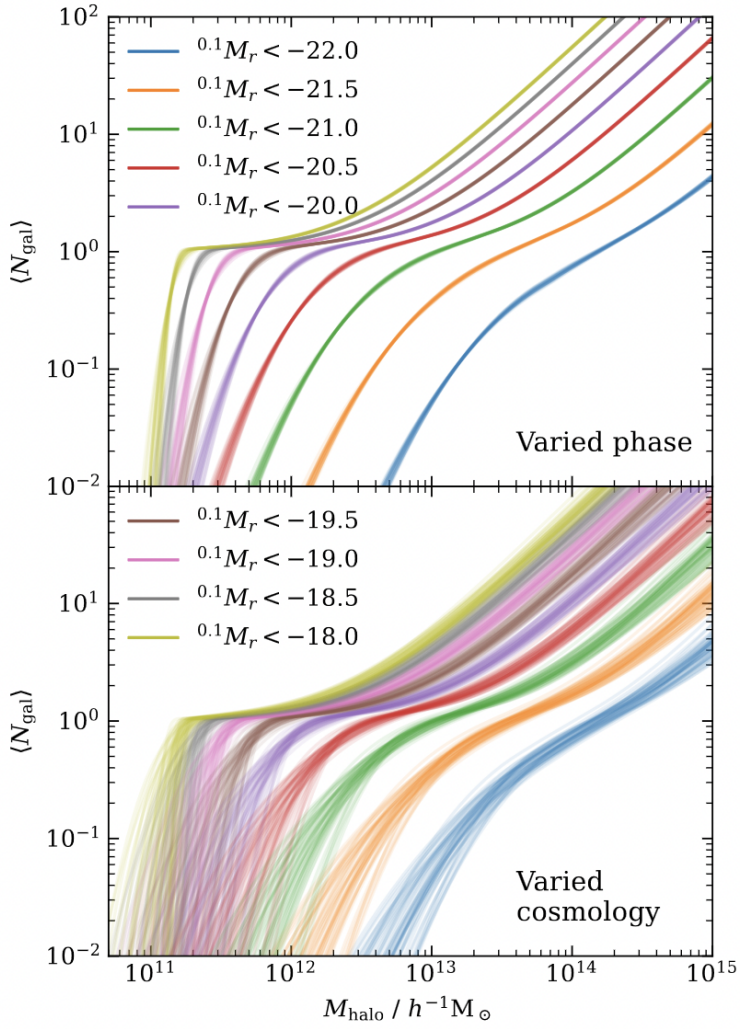

Figure 3.8: Different HODs with magnitude dependence measured with simulations with various phases and cosmologies. Taken from [28].

mocks with varying phase and cosmology are shown in Figure 3.8. It should be noted that the fitted parameters are actually 17 meta-parameters, introducing dependence for the HOD parameters on the magnitude M_r in the following fashion:

$$\log M_{\rm min} = 12 + A_{\rm min} + B_{\rm min} M_r' + C_{\rm min} M_r'^2 + D_{\rm min} M_r'^3 \tag{3.29}$$

$$\sigma_{\log M} = A_\sigma + \frac{B_\sigma - A_\sigma}{1 + \exp(C_\sigma (M_r' + D_\sigma))} \tag{3.30}$$

$$\log M_0 = 11 + A_0 + B_0 M_r' \tag{3.31}$$

$$\log M_1 = 12 + A_1 + B_1 M_r' + C_1 M_r'^2 + D_1 M_r'^3 \tag{3.32}$$

$$\alpha = A_\alpha + B_\alpha^{-M_r' + C_\alpha} \tag{3.33}$$

where $M_r' = {}^{0.1} M_r + 20$. The resulting evolution of the best-fit HODs is shown in Figure 3.9.

The resulting HOD describes the number of galaxies brighter than a given magnitude. The change however is introduced to the HOD of the central galaxies, in order to prevent the crossing of HODs from different magnitudes, such that:

$$\langle N_{cen}(M, L) \rangle = \frac{1}{2} \left[1 + S \left(\frac{\log M - \log M_{min}}{\sigma_{\log M}} \right) \right] \tag{3.34}$$

with,

$$S(x, \mu, \sigma) = \frac{4}{3\sqrt{12}\sigma} \text{spline} \left(\frac{x - \mu}{\sqrt{12}\sigma} \right) \tag{3.35}$$

Where spline is defined as:

$$\text{spline}(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & |x| \leq 0; \\ 2(1 - |x|)^3 & 0.5 \leq |x| \leq 1; \\ 0 & |x| > 1. \end{cases} \tag{3.36}$$

For each halo, a random number x is drawn from the spline kernel distribution $S(x, \mu, \sigma)$, defined in equation with $\mu = 0$ and $\sigma = 1$, which introduced the scatter in the luminosity of the central galaxy. The luminosity L required to produce such scatter is given by solving [27]:

$$\frac{x \sigma_{\log M}(L)}{\sqrt{2}} = \log M - \log M_{\rm min}(L) \tag{3.37}$$

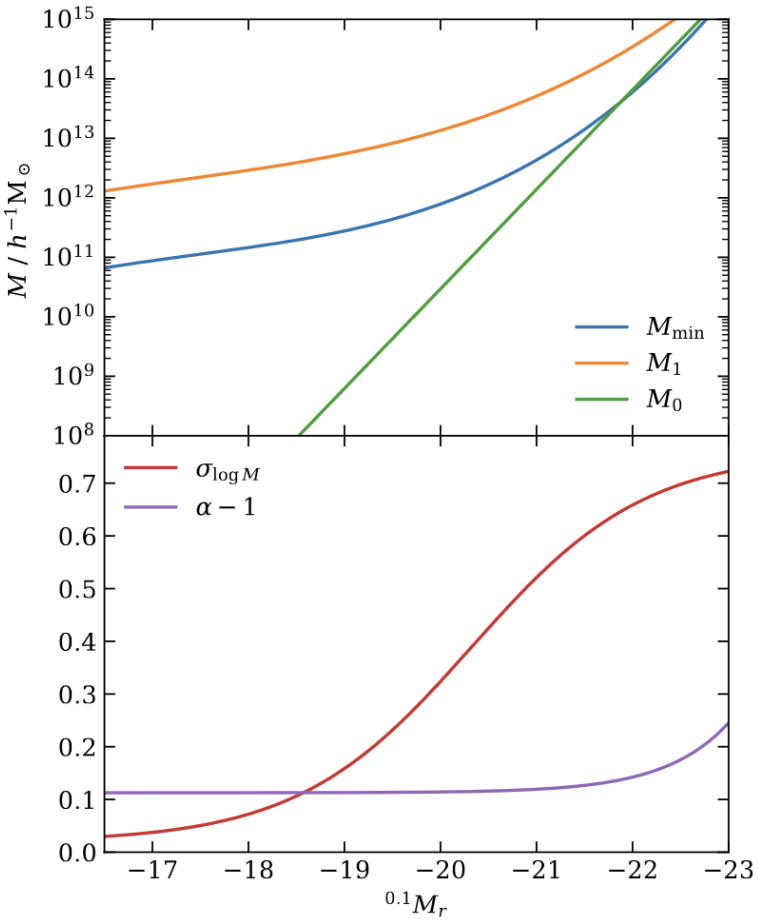

Figure 3.9: The evolution of the HOD parameters as a function of magnitude for the best-fit fitted meta-parameters. Taken from [28].

We remind that luminosity L can be connected to the absolute magnitude M_r in approximation as:

$$L = 10^{\frac{(M_S - M_r)}{2.5}} \tag{3.38}$$

where $M_s = 4.76$ is the nominal solar luminosity, following the convention used in the literature.

When generating the satellite galaxies, the minimum luminosity $L_{\min}$ is chosen, such that it corresponds to an apparent magnitude slightly fainter than $r = 20.175$. The number of satellites is then drawn from a corresponding HOD. A random number $0 < u < 1$ is then drawn, and the luminosity assigned to the galaxy is such that:

$$u = \frac{N_{\text{sat}}(> L)}{N_{\text{sat}}(> L_{\min})} \tag{3.39}$$

Then, the galaxies are positioned and velocities are assigned according to the NFW profile following the standard procedure described earlier. This allows us to obtain the catalogue of galaxies with assigned absolute magnitudes. The resulting box catalogues are then cut into spherical layers and combined to form a full-sky catalogue, creating a cutsky mock, if only one snapshot on a given redshift is used, or a lightcone, if for each spherical layer, a snapshot with the corresponding redshift is used. Using the redshift of the galaxy and its absolute magnitude, we can then assign the rest-frame colors[48], by mimicking the GAMA color-magnitude diagram [27] and apparent magnitude, using the k-correction $k(z)$, which also depends on colors:

$$r = M_r + 5 \log_{10} D_L(z) + 25 + k(z) \tag{3.40}$$

In case of Abacus BGS mocks, the GAMA[49] k-correction was used.

The Abacus BGS mock is a cutsky mock produced using $z = 0.2$ box catalogues. No replications were needed for the sample until $z = 0.4$, as the volume the box provides is enough to cover the entirety of the BGS DR1 footprint. The clustering of BGS DR1 Abacus mocks with the magnitude limit $Mr < -21.5$ in redshift range of $z = [0.1, 0.4]$ compared to a similarly chosen sample of data with the same cuts is shown in Figure 3.10.

3.4.3 Uchuu

Uchuu simulation is run using the `GreeM` simulation software with 12800^3 particles and a particle mass resolution of $3.27 \cdot 10^8 M_\odot/h$, making it the best resolved simulation from those presented here. Rockstar halo finder is used for halo-finding. The cosmology employed is `Planck2015`[50].

The production of the galaxy catalogue based on Uchuu starts with employing SHAM. In this case, the matching is done between the peak circular velocity V_{peak} (peak maximum circular velocity over the history of subhalo) with luminosity. The algorithm goes as follows:

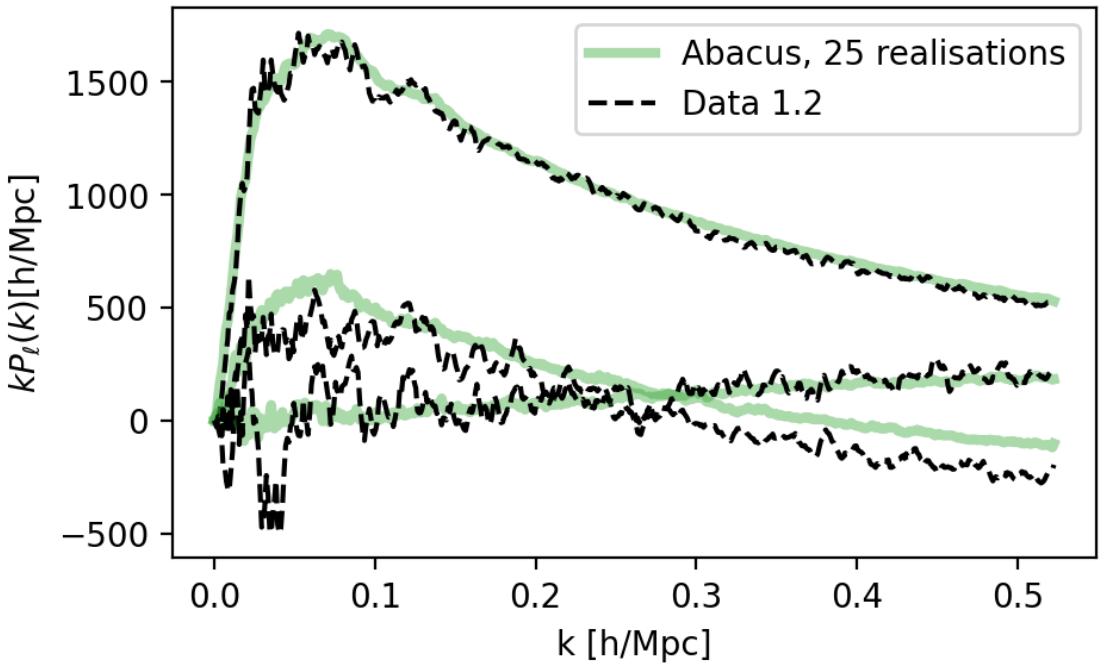

Figure 3.10: The mean of power spectrum multipoles $\ell = 0, 2, 4$ measured from fully-processed (footprint-cut and systematics modelled) Abacus BGS mock and their correspondence to those obtained from the DESI DR1 data, both with $Mr < -21.5$

1. The subhalos are sorted by $V_{\rm peak}$ in descending order to get the cumulative number density $n_h(V_{\rm peak})$.
2. The fiducial, "unscattered" value of magnitude $Mr_{\rm unscat}$ is assigned to each halo such that $n_h(V_{\rm peak}) = n_h(Mr_{\rm unscat})$
3. The magnitude is then scattered by adding a random number from a normal distribution $Mr = Mr_{\rm unscattered} + \mathcal{N}(0, \sigma)$, where σ is measure from the data.
4. The subhalos are once again sorted but now by scattered magnitude Mr
5. The galaxies are then placed in subhalos based on the ranking of the unscattered magnitudes.

Once the catalogue with assigned absolute luminosities is obtained, the same procedure as for Abacus follows in order to obtain colours and apparent magnitudes. We highlight that it is the same procedure as the one first used for the production of SDSS-Uchuu lightcones presented in [51] and for which I performed the standard RSD analysis at the beginning of my PhD, which validated my pipeline for single-tracer analysis.

The big difference from Abacus mocks arises from the fact, that Uchuu is a proper lightcone, for the generation of which boxes with redshifts $z = 0, 0.093, 0.19, 0.3, 0.43, 0.49$ have been used, thus better modelling the redshift evolution of the clustering. More on the

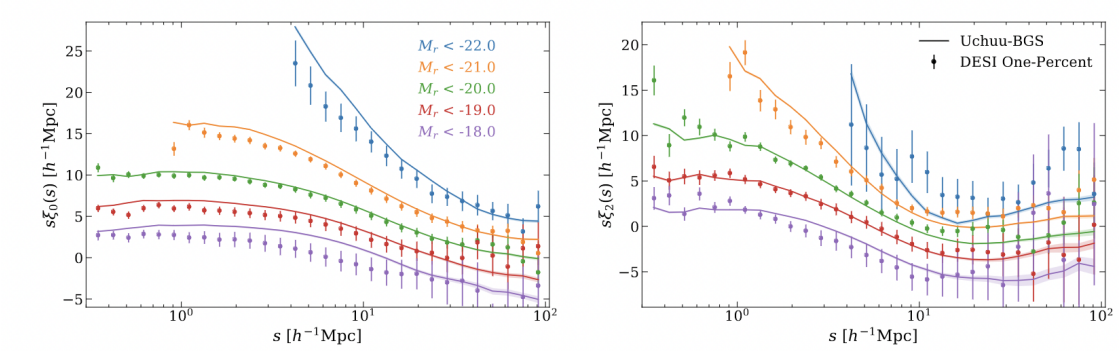

Figure 3.11: Correlation function multipoles for the Uchuu mocks mimicking 1-percent DESI in comparison to the data for various cuts in absolute magnitude M_r. *Left panel*: monopole $\ell = 0$. *Right panel*: quadrupole $\ell = 2$. Taken from [40].

production of the Uchuu mock can be found in [40]. Unfortunately, by the time of writing of this thesis, Uchuu has not been passed through the LSS pipeline, so despite remaining a formidable mock, it will not be used in the comparison presented later.

In Figure 3.11 one can see the comparison of the clustering of the correlation function multipoles from Uchuu with those from 1-percent DESI data for the BGS split into different absolute magnitude thresholds.

3.4.4 GLAM mock

GLAM is a particle-particle mesh code, which was created for the massive production of medium-quality N-body mocks, written in Fortran, and whose detailed description can be found in [52].

When I started my PhD, only the BGS Abacus mocks were created and they are not enough to obtain a precise mock-based covariance matrix as further explained in the next chapter. Therefore, my supervisor and I have started a collaboration with Francisco Prada (University of Granada) and Anatoly Klypin (University of New Mexico State) and I have been in charge of creating and testing the BGS GLAM mocks.

In order to model the DESI BGS, I have tested several simulation settings (box sizes and associated mass resolutions) for the boxes created by Anatoly Klypin, all using `Planck2015` cosmology[50]. In what follows, I will focus on two settings. The first one, internally called the E1 series, has a high resolution which allows us to tackle the modelling of the full BGS Bright with its high number density, even with colour and magnitude information. However, the limited box size of 500 Mpc/h, which was a necessary compromise, did bring unwanted effects on the covariance, which we will describe in more detail in the next chapter. The second series of the GLAM simulations that I created are denoted as F3 series. They featured a much poorer resolution, allowing however for a much bigger box size of $2Gpc/h$, which was more than enough to simulate

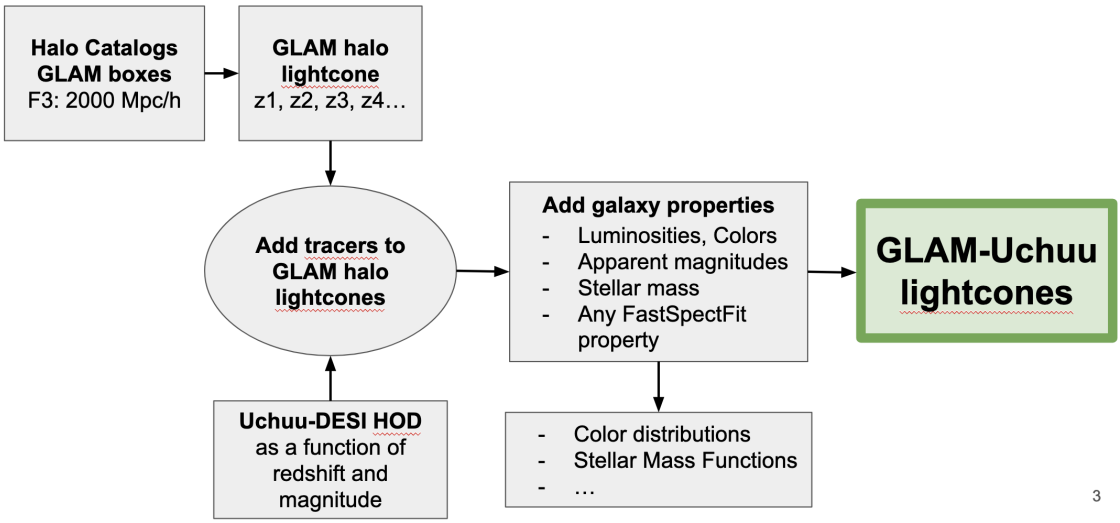

Figure 3.12: A diagram illustrating the production of GLAM BGS mocks with a pipeline that I have developed.

the entire BGS DR1 footprint. However, that came with the price of being able to model only a version of BGS with a severe absolute magnitude cut of $Mr < -21.5$. In order to produce those mocks, I have created a pipeline, which can be summarised in Figure 3.12.

First, the GLAM boxes are cut into spherical shells, and combined in a halo lightcone. We use the boxes with redshifts $z = 0, 0.03, 0.09, 0.2, 0.31, 0.49$. They are then populated with galaxies using HODs measured directly from Uchuu. This allows us to avoid using any parameterization as for the Abacus mocks, but directly measure them as a function of redshift and magnitude. Once that is done, additional properties are added such as colours and apparent magnitudes, using the same technique as for the Abacus mocks, described earlier. That leaves us with a very agnostic w.r.t. parameterizations mock, which preserves all of the quantities featured of the "parent" Uchuu mock, such as absolute and apparent magnitudes and colors, although we had to find a compromise in terms of box size vs resolution. At the time of writing this thesis, I was able to produce xx BGS GLAM mocks. More boxes are currently being produced by Anatoly Klypin such that we obtain 1000 in total in order to use them for covariance estimates (see next chapter).

The comparison of the power spectrum from the produced mock and the BGS DR1 data is presented in figure 3.13.

3.4.5 EZmock

The EZmocks for the DESI BGS are created by calibrating the EZmock parameters described in subsection 3.3.2 by comparing the clustering of the output mocks to that of the mean of 25 Abacus BGS boxes. The same compromise was made as with F3 series of the GLAM mocks, meaning that in order to have a sufficiently large box size, the

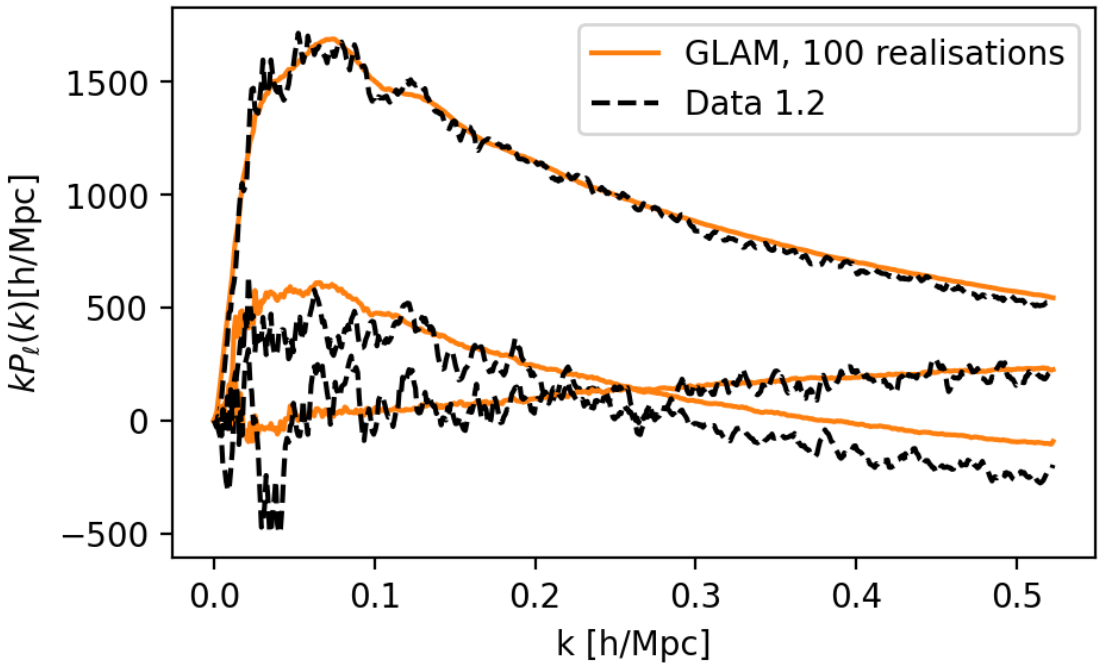

Figure 3.13: The mean of power spectrum multipoles $\ell = 0, 2, 4$ measured from fully-processed (footprint-cut and systematics modelled) GLAM BGS mock and their correspondence to those obtained from the DESI DR1 data, both with $Mr < -21.5$

resolution had to be lowered, with respect to what would be needed to model a full BGS Bright without absolute magnitude cut.

The EZmock for BGS are calibrated on the mean clustering of 25 Abacus mocks. I performed the first tests and settings and Cheng Zhao who produced those mocks for the other DESI galaxy samples, produced the final version. As expected, the mocks perform well on the large scales, while deviating on the smaller scales as seen in Figure 3.14, where the multipoles of the power spectrum for the 100 EZmocks and the data are shown.

Overall, 1000 EZmocks for BGS DR1 are produced in order to obtain a mock-based covariance matrix, and they have been used in the BAO analysis presented here [53].

3.4.6 Comparison of mocks for BGS DR1

In the end, of the 4 mock species presented, 3 reached the final stage: they were cut to the footprint of the survey andpassed through the LSS pipeline of DESI, which simulates the fibre-assignment and compute completeness weights. We have computed 2-point statistics both in Fourier and configuration spaces.

For the tests presented below,we consider the BGS DR1 sample with $Mr < -21.5$, as EZmock could not be produced with the full BGS Bright number density due to computational difficulties, and GLAM mocks could only reproduce the full number density of the BGS if the box size is reduced to $500 Mpc/h$. However, EZmock and GLAM mocks

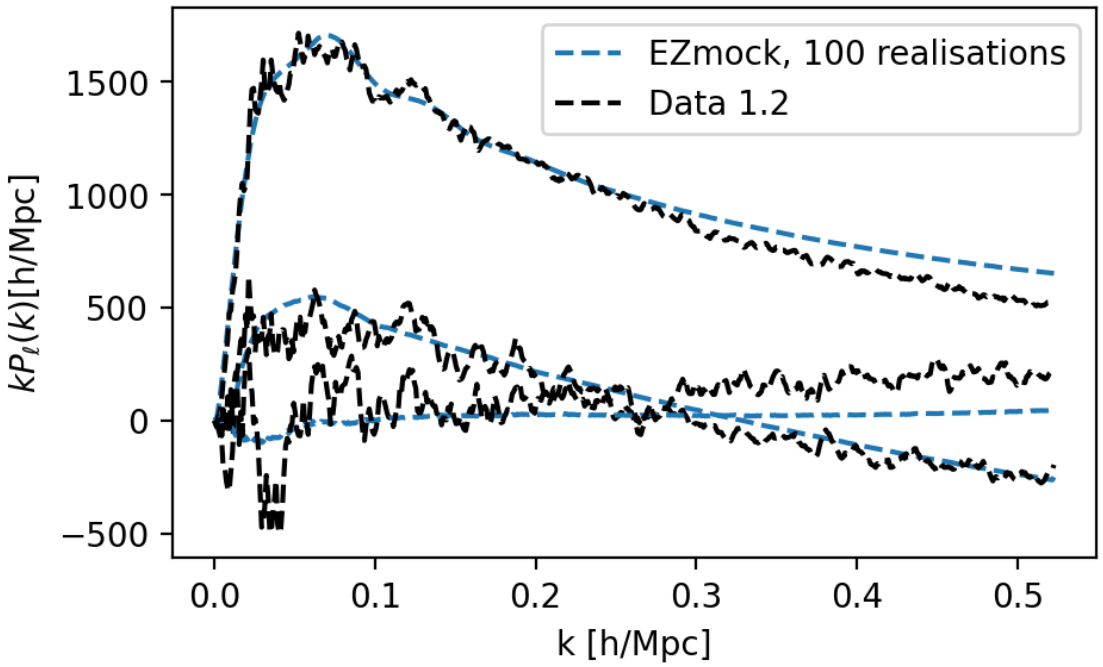

Figure 3.14: The mean of power spectrum multipoles $\ell = 0, 2, 4$ measured from fully-processed (footprint-cut and systematics modelled) EZmock BGS mock and their correspondence to those obtained from the DESI DR1 data with $Mr < -21.5$

are planned to be used for the covariance matrix estimation, and therefore the limited boxsize creates the replication effects, which we will discuss in more detail in the next chapter.

In figure 3.15 we present the clustering of 3 mock species compared to the data in configuration space. The data is quite noisy. We see that GLAM and Abacus mocks are providing fully consistent between themselves clustering on the small scales, and start to deviate slightly on the larger scales. EZMock however deviates quite noticeably in the separation range of $[25, 75]$ from the Abacus mocks, on which it was calibrated. The conclusion is consistent in the Fourier space picture presented in Figure 3.16.

On figure 3.16 we present the clustering of 3 mock species compared to the data in Fourier space. We plot additionally here a hexadecapole, where it becomes obvious that, unfortunately, that part of the power spectrum EZmock is not able to reproduce properly. The monopole is successfully reproduced by all the mock species up to $k = 0.3h/Mpc$, however after this limit, EZmock starts to deviate from the others.

As mentioned earlier, Abacus mocks are the only ones from the species presented, which can mimic full BGS Bright. The comparison of BGS Bright clustering in configuration space with the one from the mean of 25 Abacus mocks is presented in figure 3.15.

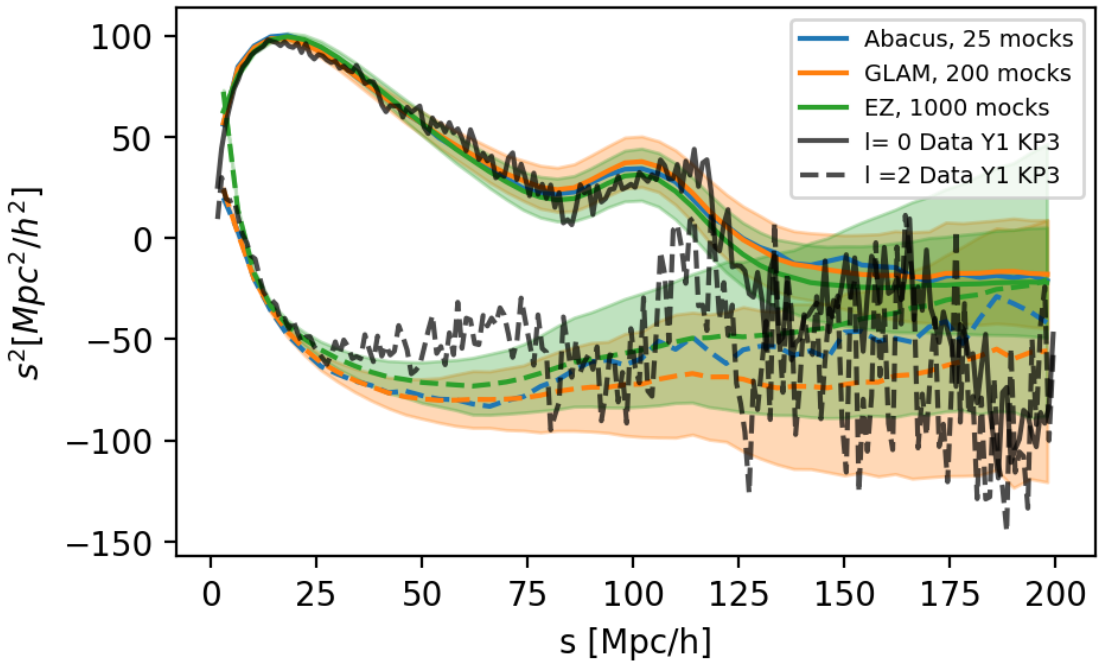

Figure 3.15: The mean of correlation function multipoles $\ell = 0, 2, 4$ measured from fully-processed (footprint-cut and systematics modelled) GLAM, EZmock and Abacus BGS mock and their correspondence to those obtained from the DESI DR1 data, all with $Mr < -21.5$. The respectively colored shaded regions correspond to uncertainties, orange for GLAM and green for EZmock.

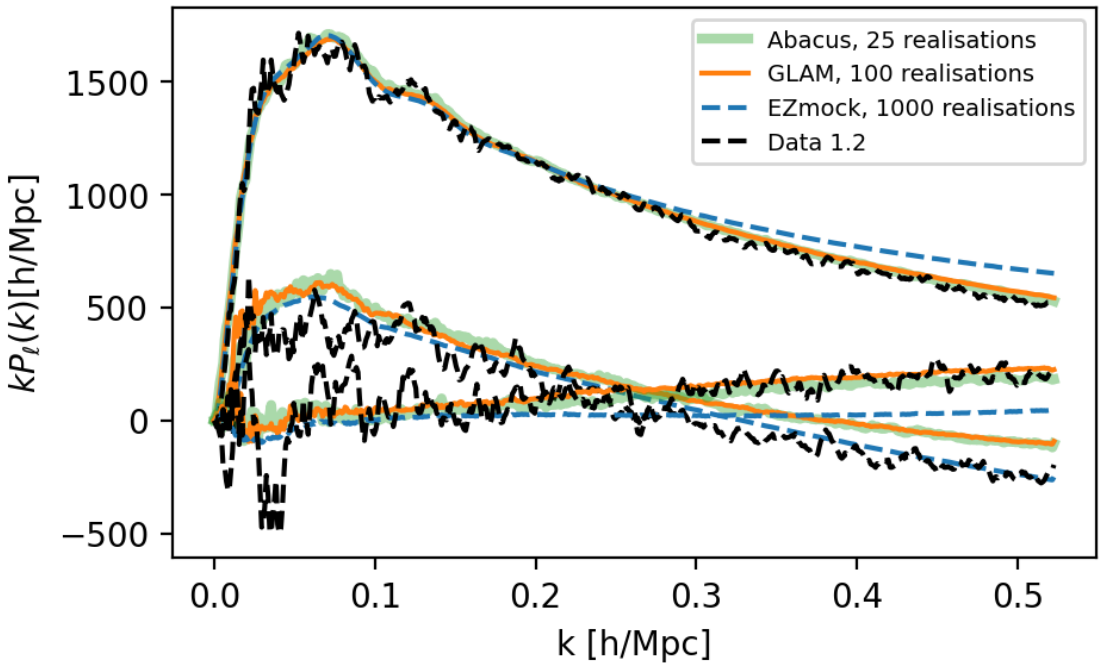

Figure 3.16: The mean of power spectrum multipoles $\ell = 0, 2, 4$ measured from fully-processed (footprint-cut and systematics modelled) GLAM, EZmock and Abacus BGS mock and their correspondence to those obtained from the DESI DR1 data, all with $Mr < -21.5$

References

[1] P. J. E. Peebles. *The large-scale structure of the universe*. 1980.

[2] J. J. Monaghan and J. C. Lattanzio. "A refined particle method for astrophysical problems". In: 149.1 (Aug. 1985), pp. 135–143.

[3] Lehman H Garrison et al. "The abacus cosmological N-body code". In: *MNRAS* 508.1 (Sept. 2021), pp. 575–596. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2482. eprint: https://academic.oup.com/mnras/article-pdf/508/1/575/40458823/stab2482.pdf. URL: https://doi.org/10.1093/mnras/stab2482.

[4] Marc Victor Leonard Metchnik. "A fast N-body scheme for computational cosmology". PhD thesis. University of Arizona, Jan. 2009.

[5] J S Bagla and T Padmanabhan. "CosmologicalN-body simulations". In: *Pramana* 49.2 (Aug. 1997), pp. 161–192. ISSN: 0973-7111. DOI: 10.1007/bf02845853. URL: http://dx.doi.org/10.1007/BF02845853.

[6] Matthieu Schaller et al. "SWIFT: A modern highly-parallel gravity and smoothed particle hydrodynamics solver for astrophysical and cosmological applications". In: (Mar. 2024). DOI: 10.1093/mnras/stae922. arXiv: 2305.13380 [astro-ph.IM].

[7] R.W Hockney and J.W Eastwood. *Computer Simulation Using Particles*. CRC Press, Mar. 2021. ISBN: 9780367806934. DOI: 10.1201/9780367806934. URL: http://dx.doi.org/10.1201/9780367806934.

[8] Teyssier, R. "Cosmological hydrodynamics with adaptive mesh refinement - A new high resolution code called RAMSES". In: *AA* 385.1 (2002), pp. 337–364. DOI: 10.1051/0004-6361:20011817. URL: https://doi.org/10.1051/0004-6361:20011817.

[9] G. Efstathiou et al. "Numerical techniques for large cosmological N-body simulations". In: 57 (Feb. 1985), pp. 241–260. DOI: 10.1086/191003.

[10] Guohong Xu. "A New Parallel N-Body Gravity Solver: TPM". In: 98 (May 1995), p. 355. DOI: 10.1086/192166. arXiv: astro-ph/9409021 [astro-ph].

[11] Tomoaki Ishiyama et al. "GreeM: Massively Parallel TreePM Code for Large Cosmological N-body Simulations". In: *Publications of the Astronomical Society of Japan* 61.6 (Dec. 2009), pp. 1319–1330. ISSN: 0004-6264. DOI: 10.1093/pasj/61.6.1319. URL: https://doi.org/10.1093/pasj/61.6.1319.

[12] Tomoaki Ishiyama et al. "4.45 Pflops Astrophysical N-Body Simulation on K computer – The Gravitational Trillion-Body Problem". In: *arXiv e-prints*, arXiv:1211.4406 (Nov. 2012), arXiv:1211.4406. DOI: 10.48550/arXiv.1211.4406. arXiv: 1211.4406 [astro-ph.CO].

[13] Tomoaki Ishiyama et al. "The Uchuu simulations: Data Release 1 and dark matter halo concentrations". In: 506.3 (Sept. 2021), pp. 4210–4231. DOI: 10.1093/mnras/stab1755. arXiv: 2007.14720 [astro-ph.CO].

[14] Volker Springel et al. "Simulating cosmic structure formation with the gadget-4 code". In: *MNRAS* 506.2 (July 2021), pp. 2871–2949. ISSN: 0035-8711. DOI: 10.1093/mnras/stab1855. eprint: https://academic.oup.com/mnras/article-pdf/506/2/2871/39271725/stab1855.pdf. URL: https://doi.org/10.1093/mnras/stab1855.

[15] Marika Asgari et al. "The halo model for cosmology: a pedagogical review". In: *The Open Journal of Astrophysics* 6 (Nov. 7, 2023). DOI: 10.21105/astro.2303.08752.

[16] Kenji Tomita. "Formation of Gravitationally Bound Primordial Gas Clouds". In: *Progress of Theoretical Physics* 42.1 (July 1969), pp. 9–23. ISSN: 0033-068X. DOI: 10.1143/PTP.42.9. eprint: https://academic.oup.com/ptp/article-pdf/42/1/9/5357758/42-1-9.pdf. URL: https://doi.org/10.1143/PTP.42.9.

[17] James E. Gunn and III Gott J. Richard. "On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution". In: 176 (Aug. 1972), p. 1. DOI: 10.1086/151605.

[18] Alexander Knebe et al. "Haloes gone MAD: The Halo-Finder Comparison Project". In: 415.3 (Aug. 2011), pp. 2293–2318. DOI: 10.1111/j.1365-2966.2011.18858.x. arXiv: 1104.0949 [astro-ph.CO].

[19] William H. Press and Paul Schechter. "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation". In: 187 (Feb. 1974), pp. 425–438. DOI: 10.1086/152650.

[20] M. Davis et al. "The evolution of large-scale structure in a universe dominated by cold dark matter". In: 292 (May 1985), pp. 371–394. DOI: 10.1086/163168.

[21] Boryana Hadzhiyska et al. "compaso: A new halo finder for competitive assignment to spherical overdensities". In: *MNRAS* 509.1 (Oct. 2021), pp. 501–521. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2980. eprint: https://academic.oup.com/mnras/article-pdf/509/1/501/41110692/stab2980.pdf. URL: https://doi.org/10.1093/mnras/stab2980.

[22] Peter S. Behroozi et al. "The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores". In: 762.2, 109 (Jan. 2013), p. 109. DOI: 10.1088/0004-637X/762/2/109. arXiv: 1110.4372 [astro-ph.CO].

[23] Zheng Zheng et al. "Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies". In: *The Astrophysical Journal* 633.2 (Nov. 2005), p. 791. DOI: 10.1086/466510. URL: https://dx.doi.org/10.1086/466510.

[24] Andrey V. Kravtsov et al. "The Dark Side of the Halo Occupation Distribution". In: 609.1 (July 2004), pp. 35–49. DOI: 10.1086/420959. arXiv: astro-ph/0308519 [astro-ph].

[25] Sihan Yuan et al. "The DESI one-percent survey: exploring the halo occupation distribution of luminous red galaxies and quasi-stellar objects with AbacusSummit". In: *MNRAS* 530.1 (Apr. 2024), pp. 947–965.

[26] Argyro Tasitsiomi et al. "Modeling Galaxy-Mass Correlations in Dissipationless Simulations". In: *The Astrophysical Journal* 614.2 (Oct. 2004), pp. 533–546. ISSN: 1538-4357. DOI: 10.1086/423784. URL: http://dx.doi.org/10.1086/423784.

[27] Alex Smith et al. "A lightcone catalogue from the Millennium-XXL simulation". In: *MNRAS* 470.4 (June 2017), pp. 4646–4661. ISSN: 0035-8711. DOI: 10.1093/mnras/stx1432. eprint: https://academic.oup.com/mnras/article-pdf/470/4/4646/19287368/stx1432.pdf. URL: https://doi.org/10.1093/mnras/stx1432.

[28] A. Smith et al. 2023. arXiv: 2312.08792 [astro-ph.CO].

[29] In: *JCAP* 2023.10 (Oct. 2023), p. 016. DOI: 10.1088/1475-7516/2023/10/016. URL: https://dx.doi.org/10.1088/1475-7516/2023/10/016.

[30] Antoine Rocher et al. "Halo occupation distribution of Emission Line Galaxies: fitting method with Gaussian processes". In: *JCAP* 2023.05 (May 2023), p. 033. DOI: 10.1088/1475-7516/2023/05/033. URL: https://dx.doi.org/10.1088/1475-7516/2023/05/033.

[31] Zheng Zheng and Hong Guo. "Accurate and efficient halo-based galaxy clustering modelling with simulations". In: *MNRAS* 458.4 (Mar. 2016), pp. 4015–4024. ISSN: 0035-8711. DOI: 10.1093/mnras/stw523. eprint: https://academic.oup.com/mnras/article-pdf/458/4/4015/13454467/stw523.pdf. URL: https://doi.org/10.1093/mnras/stw523.

[32] Julio F. Navarro et al. "The Structure of Cold Dark Matter Halos". In: 462 (May 1996), p. 563. DOI: 10.1086/177173. arXiv: astro-ph/9508025 [astro-ph].

[33] Anatoly Klypin et al. "Halo abundance matching: accuracy and conditions for numerical convergence". In: *MNRAS* 447.4 (Jan. 2015), pp. 3693–3707. DOI: 10.1093/mnras/stu2685. URL: https://doi.org/10.1093/mnras/stu2685.

[34] Qi Guo et al. "How do galaxies populate dark matter haloes?" In: 404.3 (May 2010), pp. 1111–1120. DOI: 10.1111/j.1365-2966.2010.16341.x. arXiv: 0909.4305 [astro-ph.CO].

[35] Rachel M. Reddick et al. "The Connection between Galaxies and Dark Matter Structures in the Local Universe". In: 771.1, 30 (July 2013), p. 30. DOI: 10.1088/0004-637X/771/1/30. arXiv: 1207.2160 [astro-ph.CO].

[36] Andrey V. Kravtsov. "The Size-Virial Radius Relation of Galaxies". In: 764.2, L31 (Feb. 2013), p. L31. DOI: 10.1088/2041-8205/764/2/L31. arXiv: 1212.2980 [astro-ph.CO].

[37] Charlie Conroy et al. "Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time". In: 647.1 (Aug. 2006), pp. 201–214. DOI: 10.1086/503602. arXiv: astro-ph/0512234 [astro-ph].

[38] F. Shankar et al. "New Relationships between Galaxy Properties and Host Halo Mass, and the Role of Feedbacks in Galaxy Formation". In: *The Astrophysical Journal* 643.1 (May 2006), pp. 14–25. ISSN: 1538-4357. DOI: 10.1086/502794. URL: http://dx.doi.org/10.1086/502794.

[39] A. Vale and J. P. Ostriker. "Linking halo mass to galaxy luminosity". In: *MNRAS* 353.1 (Sept. 2004), pp. 189–200. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2004.08059.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2004.08059.x.

[40] F. Prada et al. 2023. arXiv: 2306.06315 [astro-ph.CO].

[41] Jeong Donghui. "Halo occupation distribution of Emission Line Galaxies: fitting method with Gaussian processes". In: (Aug. 2010).

[42] Ya. B. Zel'dovich. "Gravitational instability: An approximate theory for large density perturbations." In: 5 (Mar. 1970), pp. 84–89.

[43] Chia-Hsun Chuang et al. "EZmocks: extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics". In: *MNRAS* 446.3 (Nov. 2014), pp. 2621–2628. ISSN: 0035-8711. DOI: 10.1093/mnras/stu2301. eprint: https://academic.oup.com/mnras/article-pdf/446/3/2621/10448149/stu2301.pdf. URL: https://doi.org/10.1093/mnras/stu2301.

[44] Cheng Zhao et al. "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: 1000 multi-tracer mock catalogues with redshift evolution and systematics for galaxies and quasars of the final data release". In: *MNRAS* 503.1 (Feb. 2021), pp. 1149–1173. DOI: 10.1093/mnras/stab510. URL: https://doi.org/10.1093/mnras/stab510.

[45] DESI Collaboration Et Al. 2023. DOI: 10.5281/ZENODO.7964161. URL: https://zenodo.org/record/7964161.

[46] Nina A Maksimova et al. "AbacusSummit: a massive set of high-accuracy, high-resolution N-body simulations". In: *MNRAS* 508.3 (Sept. 2021), pp. 4017–4037. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2484. eprint: https://academic.oup.com/mnras/article-pdf/508/3/4017/40811763/stab2484.pdf. URL: https://doi.org/10.1093/mnras/stab2484.

[47] N. Aghanim et al. "iPlanck/i2018 results". In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. URL: https://doi.org/10.1051/0004-6361/201833910.

[48] Ramin A. Skibba and Ravi K. Sheth. "A halo model of galaxy colours and clustering in the Sloan Digital Sky Survey". In: *MNRAS* 392.3 (Jan. 2009), pp. 1080–1091. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2008.14007.x. eprint: https://academic.oup.com/mnras/article-pdf/392/3/1080/3647280/mnras0392-1080.pdf. URL: https://doi.org/10.1111/j.1365-2966.2008.14007.x.

[49] Tamsyn McNaught-Roberts et al. "Galaxy And Mass Assembly (GAMA): the dependence of the galaxy luminosity function on environment, redshift and colour". In: *MNRAS* 445.2 (Oct. 2014), pp. 2125–2145. ISSN: 0035-8711. DOI: 10.1093/mnras/stu1886. eprint: https://academic.oup.com/mnras/article-pdf/445/2/2125/18198298/stu1886.pdf. URL: https://doi.org/10.1093/mnras/stu1886.

[50] Planck Collaboration et al. "Planck 2015 results - XIII. Cosmological parameters". In: *AA* 594 (2016), A13. DOI: 10.1051/0004-6361/201525830. URL: https://doi.org/10.1051/0004-6361/201525830.

[51] C. A. Dong-Páez et al. "The Uchuu–SDSS galaxy light-cones: a clustering, redshift space distortion and baryonic acoustic oscillation study". In: *Mon. Not. Roy. Astron. Soc.* 528.4 (2024), pp. 7236–7255. DOI: 10.1093/mnras/stae062. arXiv: 2208.00540 [astro-ph.CO].

[52] Anatoly Klypin and Francisco Prada. "Dark matter statistics for large galaxy catalogues: power spectra and covariance matrices". In: 478.4 (Aug. 2018), pp. 4602–4621. DOI: 10.1093/mnras/sty1340. arXiv: 1701.05690 [astro-ph.CO].

[53] DESI Collaboration et al. 2024. arXiv: 2404.03002 [astro-ph.CO].

Chapter 4

Covariance

Не замкнут круг. Заклятья
недопеты. . .
Когда для всех сапфирами
лучей
Сияет день, журчит в полях
ручей, —
Для нас во мгле слепые бродят
светы,!

M. Voloshin, 75

Introduction

The last component that is required for having a complete inference pipeline of the cosmological parameters, and which we have not covered yet is the error estimation. Throughout this thesis, as well as in the overwhelming majority of cosmological inference literature, we intrinsically assume the data vectors, meaning the values of multipoles for different scales of the correlation functions and power spectrums to be distributed according to multivariate Gaussian distributions. This assumption allows us to describe the nature of our statistical errors in the form of the covariance matrix, which also enable cross-correlation terms between different scales and multipoles. In this chapter, we will focus on the standard single-tracer analysis, assuming that our data vector is composed of the multipoles of a single correlation function or power spectrum. In the next chapter we will also discuss the multitracer analysis and the covariance for such a kind of analysis.

4.1 Analytic covariance

In this section, we present the derivation of the analytic covariance for the power spectrum in the Gaussian approximation following the approach in [1]. For the derivation of the analytic covariance in configuration space please refer to [2] and to [3] for its application to DESI DR1. The reason for that is mainly that I was mostly working with analytical covariances for the power spectrum, and has no direct experience producing those for correlation function, thus creating such an exception for this section.

Before we start, we will introduce some new notation, which will facilitate the derivation.

Assuming that our survey is covering only a part of the sky, and we use a set of weights $w^i(\boldsymbol{x})$ to correct for systematics, we introduce the window matrix W_{ij} as:

$$W_{ij}(\boldsymbol{x}) = \overline{n}^i(\boldsymbol{x}) w^j(\boldsymbol{x}) \tag{4.1}$$

The normalization factors for these windows will then become:

$$I_{ij} = \int d^3\, \overline{n}^i(\boldsymbol{x}) w^j(\boldsymbol{x}) \tag{4.2}$$

We then estimate the density perturbation at the survey level δ_{N_g} as:

$$\delta_{N_g} = \frac{\int_{\boldsymbol{x}} \delta(\boldsymbol{x}) W_{10}(\boldsymbol{x})}{I_{10}} \tag{4.3}$$

We can then rewrite the FKP estimator, previously defined in eq. 96, following [1] as:

$$\hat{\delta}^{\mathrm{FKP}}(\boldsymbol{x}) = F(\boldsymbol{x}) = \frac{1}{\sqrt{I_{22}}} \frac{\delta_W(\boldsymbol{x})}{(1+\delta_{N_g})^{\frac{1}{2}}} \tag{4.4}$$

where,

$$\delta_W(\boldsymbol{x}) = W_{11}(\boldsymbol{x}) \delta(\boldsymbol{x}) \tag{4.5}$$

4.1.1 Real space

In the real space we can write the expression for the power spectrum as:

$$P(\boldsymbol{k}) = \int d^3x e^{i\boldsymbol{k}\boldsymbol{x}} \left\langle \delta(\boldsymbol{x_1}) \delta(\boldsymbol{x_2}) \right\rangle_{x=x_2-x_1} \tag{4.6}$$

We can also transfer the overdensities to the Fourier space obtaining:

$$\langle \delta(\boldsymbol{k}_1) \delta(\boldsymbol{k}_2) \rangle = (2\pi)^3 \delta_D(\boldsymbol{k}_1 + \boldsymbol{k}_2) P(k_1) \tag{4.7}$$

Going further, assuming the isotropy of the real-space and integrating over angles, while substituting the FKP overdensity, we end up with:

$$\hat{P}^{FKP}(k) = \frac{1}{I_{22}} \int \frac{1}{V_k} d^2\hat{k} \frac{|\delta_(\boldsymbol{k})|^2}{1+\delta_{N_g}} \tag{4.8}$$

We would want to obtain the covariance matrix $C(k_1, k_2) = \langle P(k_1)P(k_2)\rangle - \langle P(k_1)\rangle\langle P(k_2)\rangle$, so the main components we will need will be the expectation value for the power spectrum and the correlator of the two of them [1]:

$$\langle \hat{P}(k)\rangle = \frac{1}{I_{22}} \int d^2\hat{k} \int d^3k' |W_{11}(\boldsymbol{k}')|^2 P(\boldsymbol{k}-\boldsymbol{k}') = \frac{1}{I_{22}} \int d^2\hat{k} P(\boldsymbol{k}) \int d^3k' |W_{11}(\boldsymbol{k}')|^2 = \hat{P}(k) \tag{4.9}$$

$$\begin{aligned}\langle \hat{P}(k_1)\hat{P}(k_2)\rangle &= \frac{1}{I_{22}^2} \int d^2\hat{k}_1 \int d^2\hat{k}_2 \langle \delta_W(\boldsymbol{k}_1)\delta_W(-\boldsymbol{k}_1)\delta_W(\boldsymbol{k}_2)\delta_W(-\boldsymbol{k}_2)\rangle \\ &= \frac{1}{I_{22}^2} \int d^2\hat{k}_1 \int d^2\hat{k}_2 \int d^3p_1 \int d^3p_1' \int d^3p_2 \int d^3p_2' W_{11}(k_1-p_1)W_{11}(-k_1-p_1')\times \\ &\quad \times W_{11}(k_2-p_2)W_{11}(-k_2-p_2')\langle \delta(\boldsymbol{p}_1)\delta(\boldsymbol{p}_1')\delta(\boldsymbol{p}_2)\delta(\boldsymbol{p}_2')\rangle \end{aligned} \tag{4.10}$$

From this we can now assemble our analytic covariance matrix for 1D power spectrum in real space.

4.1.2 Redshift space

However, we will need to infer the parameters from the data located in the redshift space. The redshift space analytic covariance is significantly more nontrivial that in the real space. First, we can not enjoy the isotropy anymore, as RSD does introduce LOS-angle dependent clustering. Therefore, the 3D power-spectrum will not be fully described by the 1-dimensional part anymore.

Denoting the redshift-space distorted overdensity as $F_W(\boldsymbol{x}) = \int d^3x e^{-i\boldsymbol{kx}} W_{11}(\boldsymbol{x})\delta(\boldsymbol{x})$, we can write the expression for the power spectrum multipoles as:

$$P_\ell(k)\frac{(2\ell+1)}{I_{22}} \int d^2\hat{k} F_{W,\ell}(\boldsymbol{k}) F_{W,0}(-\boldsymbol{k}) \tag{4.11}$$

where

$$F_{W,\ell}(\boldsymbol{k}) \int d^3x e^{-i\boldsymbol{kx}} W_{11}(\boldsymbol{x})\delta(\boldsymbol{x})\mathcal{L}_\ell(\hat{k}\cdot\hat{x}) \tag{4.12}$$

Now that we have divided the power spectrum into angular components, we can write the the expectation value of the power spectrum multipole $\langle P_\ell(k)\rangle$:

$$\langle P_\ell(k)\rangle = \frac{(2\ell+1)}{I_{22}} \int d^2\hat{k} \int d^3x \int d^3x' e^{-i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')} \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}')\rangle W_{11})\boldsymbol{x}) W_{11}(\boldsymbol{x}')\mathcal{L}_\ell(\hat{x}\cdot\hat{k}) \tag{4.13}$$

Remembering that $\langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}')\rangle = \xi(\boldsymbol{s}, \boldsymbol{x}_+)$, where $s = \boldsymbol{x}' - \boldsymbol{x}$ and $x_+ = (\boldsymbol{x}' + \boldsymbol{x})/2$, we can rewrite the expectation value as:

$$\langle P_\ell(k)\rangle = \frac{(2\ell+1)}{I_{22}}\int d^2\hat{k}\int d^3x\int d^3s e^{-ik\boldsymbol{s}}\xi(\boldsymbol{s},\boldsymbol{x}_+)W_{11}(\boldsymbol{x}-\boldsymbol{s})W_{11}(\boldsymbol{x})\mathcal{L}_\ell(\hat{x}\cdot\hat{k}) \tag{4.14}$$

We will now introduce the redshift-space local power spectrum:

$$P_{local}(\boldsymbol{k},\boldsymbol{x}) = \int d^3s\xi(\boldsymbol{s},\boldsymbol{x})e^{-i\boldsymbol{k}\cdot\boldsymbol{s}} \tag{4.15}$$

The integral can be then rewritten as:

$$\int d^3s\xi(\boldsymbol{s},\boldsymbol{x})W_{11}(\boldsymbol{x}-\boldsymbol{s}) = \int d^3s\int d^3k'\int d^3q e^{-i\boldsymbol{k}\cdot\boldsymbol{s}}P_{local}(\boldsymbol{k}',\boldsymbol{x})W_{11}(\boldsymbol{q})e^{i[\boldsymbol{s}\cdot\boldsymbol{k}'+(\boldsymbol{x}-\boldsymbol{s})\cdot\boldsymbol{q}]} =$$
$$= \int d^3\boldsymbol{q}P_{local}(\boldsymbol{k}+\boldsymbol{q},\boldsymbol{x})W_{11}(\boldsymbol{q})e^{i\boldsymbol{x}\cdot\boldsymbol{q}} = [\text{assuming } k\gg q\ ] = W_{11}(\boldsymbol{x})P_{local}(\boldsymbol{k},\boldsymbol{x}) \tag{4.16}$$

Now, substituting 4.16 into 4.14 gives us the desired answer for $\langle P_{W,\ell}(\boldsymbol{k})\rangle$, where one needs to also assume the limit of $kd\to\infty$, where d is the distance to the galaxy, which is equivalent to the plane-parallel approximation, and we will expand the local power spectrum implicitly in multipoles $P_{local}(\boldsymbol{k},\boldsymbol{d}) = \sum_{\ell'}P_{\ell'}(k)\mathcal{L}_{\ell'}(\hat{d}\cdot\hat{k}) + O(kd)^{-2}$, yielding in the end:

$$\langle P_{W,\ell}(\boldsymbol{k})\rangle = \frac{(2\ell+1)}{I_{22}}\int d^2\hat{k}\int d^3x\mathcal{L}_\ell(\hat{x}\cdot\hat{k})P_{local}(\boldsymbol{k},\boldsymbol{x})W_{22}(\boldsymbol{x}) = P_\ell(\boldsymbol{k}) \tag{4.17}$$

Now we need to find the correlator of two power-spectra $\langle P_{W,\ell_1}(\boldsymbol{k}_1)P_{W,\ell_2}(\boldsymbol{k}_2)\rangle$ in order to obtain the covariance matrix. Using Wick's theorem we obtain:

$$\langle P_{W,\ell_1}(\boldsymbol{k}_1)P_{W,\ell_2}(\boldsymbol{k}_2)\rangle =$$
$$= \frac{(2\ell_1+1)(2\ell_2+1)}{I_{22}^2}\int d^2\hat{k}_1\int d^2\hat{k}_2\langle F_{W,\ell_1}(\boldsymbol{k}_1)F_{W,0}(-\boldsymbol{k}_1)F_{W,\ell_2}(\boldsymbol{k}_2)F_{W,0}(-\boldsymbol{k}_2)\rangle =$$
$$= \frac{(2\ell_1+1)(2\ell_2+1)}{I_{22}^2}\int d^2\hat{k}_1\int d^2\hat{k}_2\left[\langle F_{W,\ell_1}(\boldsymbol{k}_1)F_{W,0}(-\boldsymbol{k}_1)\rangle\langle F_{W,\ell_2}(\boldsymbol{k}_2)F_{W,0}(-\boldsymbol{k}_2)\rangle + \right.$$
$$\left. + \langle F_{W,\ell_1}(\boldsymbol{k}_1)F_{W,\ell_2}(-\boldsymbol{k}_2)\rangle\langle F_{W,0}(\boldsymbol{k}_1)F_{W,0}(-\boldsymbol{k}_2)\rangle + \langle F_{W,\ell_1}(\boldsymbol{k}_1)F_{W,0}(-\boldsymbol{k}_2)\rangle\langle F_{W,\ell_2}(\boldsymbol{k}_2)F_{W,0}(-\boldsymbol{k}_1)\rangle\right] \tag{4.18}$$

We can now expand the terms in the expression using 4.12:

$$\langle P_{W,\ell_1}(\boldsymbol{k}_1)P_{W,\ell_2}(\boldsymbol{k}_2)\rangle =$$
$$= \frac{(2\ell_1+1)(2\ell_2+1)}{I_{22}^2}\int d^2\hat{k}_1\int d^2\hat{k}_2\int d^3x_1\int d^3x_1'\int d^3x_2\int d^3x_2'\times$$
$$\times e^{-i\boldsymbol{k}_1\cdot(\boldsymbol{x}_1-\boldsymbol{x}_1')-i\boldsymbol{k}_2\cdot(\boldsymbol{x}_2-\boldsymbol{x}_2')}\langle\delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2')\rangle\langle\delta(\boldsymbol{x}_1')\delta(\boldsymbol{x}_2)\rangle W_{11}(\boldsymbol{x}_1)W_{11}(\boldsymbol{x}_1')W_{11}(\boldsymbol{x}_2)W_{11}(\boldsymbol{x}_2')\times$$
$$\times\mathcal{L}_{\ell_1}(\hat{x}_1\cdot\hat{k}_1)\left[2\mathcal{L}_{\ell_2}(\hat{x}_2\cdot\hat{k}_2)+\mathcal{L}_{\ell_2}(-\hat{x}_2'\cdot\hat{k}_2)\right] \tag{4.19}$$

Intoducing relative coordinates $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, approximating LOS as before and substituting 4.16, we obtain:

$$\begin{aligned}\langle P_{W,\ell_1}(\boldsymbol{k}_1)P_{W,\ell_2}(\boldsymbol{k}_2)\rangle = \\ = \frac{(2\ell_1+1)(2\ell_2+1)}{I_{22}^2}\int d^2\hat{k}_1\int d^2\hat{k}_2\int d^3x_1\int d^3x_2\times \\ \times e^{-i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}P_{local}(\boldsymbol{k}_1,\boldsymbol{x}_2)P_{local}(\boldsymbol{k}_2,\boldsymbol{x}_1)W_{22}(\boldsymbol{x}_1)W_{22}(\boldsymbol{x}_2)\times \\ \times\mathcal{L}_{\ell_1}(\hat{x}_1\cdot\hat{k}_1)\left[2\mathcal{L}_{\ell_2}(\hat{x}_2\cdot\hat{k}_2)+\mathcal{L}_{\ell_2}(-\hat{x}_2'\cdot\hat{k}_2)\right] \quad (4.20)\end{aligned}$$

Using the multipoles expansion and, as earlier, cutting to leafing order in $(kd)^{-1}$, and substracting $\langle P_{W,\ell_1}(\boldsymbol{k}_1)\rangle\langle P_{W,\ell_2}(\boldsymbol{k}_2)\rangle$ we obtain the Gaussian covariance:

$$\begin{aligned}C_{\ell_1,\ell_2}(k_1,k_2) = \frac{(2\ell_1+1)(2\ell_2+1)}{I_{22}^2}\sum_{\ell_1',\ell_2'}P_{\ell_1'}(k_1)P_{\ell_2'}(k_2)\int d^2\hat{k}_1\int d^2\hat{k}_2\int d^3x_1\int d^3x_2\times \\ \times e^{-i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot(\boldsymbol{x}_1-\boldsymbol{x}_2)}P_{local}(\boldsymbol{k}_1,\boldsymbol{x}_2)P_{local}(\boldsymbol{k}_2,\boldsymbol{x}_1)W_{22}(\boldsymbol{x}_1)W_{22}(\boldsymbol{x}_2)\times \\ \times\mathcal{L}_{\ell_1'}(\hat{x}_2\cdot\hat{k}_1)\mathcal{L}_{\ell_2'}(\hat{x}_1\cdot\hat{k}_2)\mathcal{L}_{\ell_1}(\hat{x}_1\cdot\hat{k}_1)\left[\mathcal{L}_{\ell_2}(\hat{x}_2\cdot\hat{k}_2)+\mathcal{L}_{\ell_2}(-\hat{x}_2'\cdot\hat{k}_2)\right] = \\ = \sum_{\ell_1',\ell_2'}P_{\ell_1'}(k_1)P_{\ell_2'}(k_2)\mathcal{W}_{\ell_1\ell_2\ell_1'\ell_2'} \quad (4.21)\end{aligned}$$

However, there is one component which we have not yet mentioned, but which becomes very important for realistic surveys. It is accounting for the finite nature of the observed objects. This contribution to the covariance is often called the shot noise. There are different approaches to tackle that contribution, for analytic computation of it we recommend taking a look at [1]. In practice analytic approach to the computation of the covariance matrix allows, as seen from the 4.21 a covariance matrix with not so much computational effort, requiring for the input the computed kernels $\mathcal{W}_{\ell_1\ell_2\ell_1'\ell_2'}$ and the target power spectra.

However, the non-gaussian contributions to the covariance, as well as contributions coming from higher-dimensional statistics start to play a bigger role for bigger values of k, meanwhile being quite tricky to figure out, as can be seen on Fig 4.1, comparing the signal-to-noise ratio for monopole and quadrupole of the power spectrum with and without the non-gaussian terms. Adding to this a certain challenge in properly accounting for instrumentation effects in such covariances makes them of relatively limited use.

Starting from the next section we will switch back to correlation function as our target statistics.

4.2 Mock covariance

There is only one Universe that we can observe, which unfortunately limits the observable samples of our data to 1. Nevertheless, we can try to simulate other universes, and we

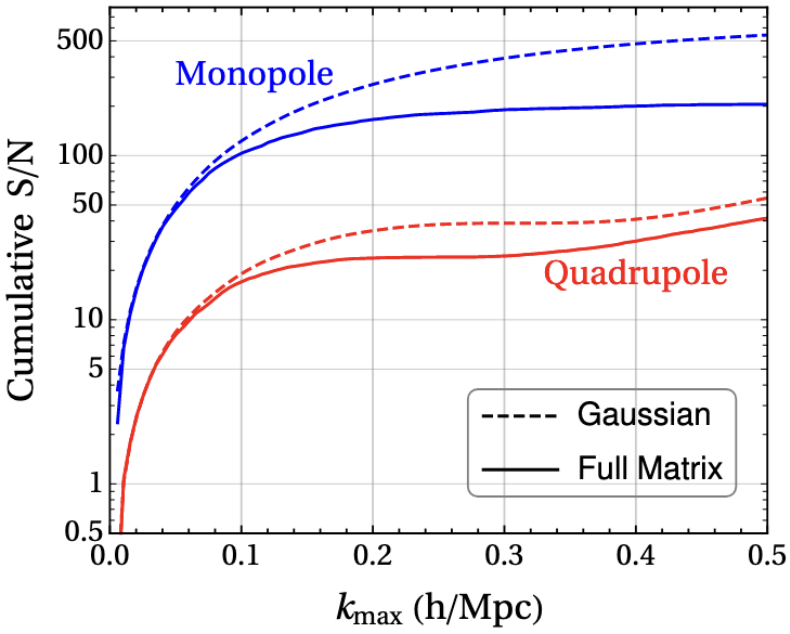

Figure 4.1: The signal to noise ration for the power spectrum multipoles in the redshift space for the covariance matrix produced in Gaussian approximation and beyond (Full matrix). We can see that the contribution of the non-Gaussian terms becomes more and more significant as we go to the higher k's. Taken from [1].

presented the cosmological simulations in the previous chapter. Thanks to those simulations, we can have as many independent realisations of the survey as our computational resources enable. We can write the estimator for the covariance matrix from N_m mocks as:

$$C_{ij} = \frac{1}{N_{\rm m} - 1} \sum_{k=1}^{N_{\rm m}} \left[X_i^{[k]} - \langle X_i \rangle \right] \left[X_j^{[k]} - \langle X_j \rangle \right], \tag{4.22}$$

where $X_i^{[k]}$ is the $i^{\rm th}$ bin of either the correlation function or power spectrum of the $k^{\rm th}$ mock, and $\langle X_i \rangle$ is the mean over the $N_{\rm m}$ mocks of the $i^{\rm th}$ bin of the correlation function or power spectrum. The datavector assumed, in case of clustering analysis, to be a concatenation of several multipoles of power spectrum or correlationf function. It should be noted however, that first of all, the number of mocks N_m has to be larger than the length of the data-vector, leading to at least hundreds of mocks needed for a non-singular covariance. Furthermore, as we will see in the following chapter, covariance matrix is used in its inversed form in the likelihood estimation. That is where an additional caveat arises. The inverse of a statistical unbiased estimator is not unbiased on its own. Therefore, [4] showed that by rescaling the inverse covariance matrix C_*^{-1}, we can obtain an unbiased estimator of the inverse covariance matrix $C^{-1}_{\ell_1\ell_2}$ as:

$$C^{-1}_{\ell_1\ell_2} = \frac{N_m - p - 2}{N_m - 2} {C_*^{-1}}_{\ell_1\ell_2} \tag{4.23}$$

where p is the number of independent degrees of freedom in the analysis, usually the number of bins in the data-vector minus the number of fitted parameters (both the cosmological and nuisance parameters).

Therefore, it is straightforward to see that the accuracy of such a covariance matrix depends on the number of simulations produced and on the accuracy of the simulations themselves. Usually, in the surveys [5] thousands of mocks are used for covariance production, which is a big challenge for the DESI BGS, given its extremely high density of galaxies in the survey volume.

4.3 Jackknife covariance

Having mock covariance can be very expensive, while analytic covariance can not capture all of the necessary details such as non-linearities and observational systematics. We can try to resample the data itself in a smart way to capture the covariance from the deviations inside the survey volume itself, and maybe to apply some scaling to fix potential bias. This data resampling approach is called the jackknife [6].

Jackknife is a data resampling approach that involves creating multiple sub-samples of the same dataset by systematically excluding regions of the data. When applied to the

cosmological surveys, the footprint is divided into regions of similar area and it is these that are systematically excluded to make the multiple sub-samples.

This approach has the advantage of making no assumptions regarding non-linear evolution and non-standard physics, and at the same time is extremely cheap from the computational perspective, as it does not require expensive production of thousands of mocks. Assuming we have cut our dataset into $N_{\rm jk}$ pieces, the covariance matrix is:

$$C_{ij} = \frac{N_{\rm jk}-1}{N_{\rm jk}} \sum_{k=1}^{N_{\rm jk}} \left[\xi_i^{[k]} - \langle\xi_i\rangle\right] \left[\xi_j^{[k]} - \langle\xi_j\rangle\right], \tag{4.24}$$

where $\xi_i^{[k]}$ is the $i^{\rm th}$ bin of the correlation function of the $k^{\rm th}$ jackknife region, and $\langle\xi_i\rangle$ is its mean over all the $N_{\rm jk}$ jackknife regions. The coefficient on the right-hand side is larger than the corresponding factor in Eq. 4.22 as it compensates for the reduction in the covariance due to the overlap between the subsamples.

In practice, we consider the galaxy 2-point correlation function and the DD, DR and RR pair counts mentioned in the Landy-Szalay estimator defined in Eq. 83.

4.3.1 Standard approach

We will assume the number of sub-samples is $N_{\rm jk}$ and work in terms of pair counts rather than correlation functions. For simplicity, we will denote as AA_k the auto-counts that are contributed by pairs of galaxies that both reside in the $k^{\rm th}$ area of the survey (the areas that are systematically excluded to form the jackknife sub-samples) and CC_k the cross-counts between galaxies in this $k^{\rm th}$ area and those in the jackknife sub-sample that is made by excluding this area. The counts in the jackknife sub-sample TT_k are related to the overall number of counts in the full survey $TT_{\rm tot}$ and the above quantities by

$$TT_k = TT_{\rm tot} - AA_k - CC_k, \tag{4.25}$$

where in defining each of these pair counts we count each unique pair only once. The total number of auto- and cross-pairs can be related to their means over the jackknife samples by

$$AA^{\rm tot} = N_{\rm jk}\overline{AA} \tag{4.26}$$

and, as we account for double counting with the cross-pairs only while looking at the full sample, we need to divide the obtained estimate by 2 to be consistent with the auto-pairs:

$$CC^{\rm tot} = \frac{N_{\rm jk}}{2}\overline{CC}, \tag{4.27}$$

where $\overline{AA} = \frac{1}{N_{\rm jk}}\sum_{k=1}^{N_{\rm jk}} AA_k$ and $\overline{CC} = \frac{2}{N_{\rm jk}}\sum_{k=1}^{N_{\rm jk}} CC_k$.

Following [7], we choose to define an estimator of the normalised auto-pairs $\theta_{\mathrm{a},k}$ in a specific realisation, such that $\overline{\theta}_{\mathrm{a}} = \overline{AA}$ by

$$\theta_{\mathrm{a},k} = \frac{1}{N_{\mathrm{jk}} - 1}\left(N_{\mathrm{jk}}\overline{AA} - AA_k\right) \tag{4.28}$$

and the estimator of the normalised cross-pairs $\theta_{c,k}$ such that $\overline{\theta}_{\mathrm{c}} = \overline{CC}$ by

$$\theta_{c,k} = \frac{2}{N_{\mathrm{jk}} - 2}\left(\frac{N_{\mathrm{jk}}}{2}\overline{CC} - CC_k\right), \tag{4.29}$$

where it was taken into account that the cross-pairs contribute to the total estimate twice, while the auto-pairs only once.

We can then further compute for each jackknife realization the deviation from the mean value of the auto paircounts

$$\theta_{a,k} - \overline{\theta_a} = \frac{1}{N_{\mathrm{jk}} - 1}\left(\overline{AA} - AA_k\right) \tag{4.30}$$

and cross paircounts

$$\theta_{c,k} - \overline{\theta_c} = \frac{2}{N_{\mathrm{jk}} - 2}\left(\overline{CC} - CC_k\right). \tag{4.31}$$

We can now express how the covariance of each type of pair count can be represented in terms of the estimators above, if we assume the following definition for the covariance, where DD_t are just some pair counts of type t:

$$\mathrm{cov}(DD_1, DD_2) = \sqrt{\frac{\overline{DD}_1\overline{DD}_2}{DD_1^{\mathrm{tot}}DD_2^{\mathrm{tot}}}}\frac{1}{N_{\mathrm{jk}} - 1}\sum_{k=1}^{N_{\mathrm{jk}}}(DD_{1k} - \overline{DD}_1)(DD_{2k} - \overline{DD}_2) \tag{4.32}$$

By replacing (DD_1, DD_2) by (AA, AA) or (CC, CC) or (CC, AA) in Eq. 4.32 and using Eqs. 4.30 and 4.31, one obtains:

$$\mathrm{cov}(AA, AA) = \frac{N_{\mathrm{jk}} - 1}{N_{\mathrm{jk}}}\sum_{k=1}^{N_{\mathrm{jk}}}\left(\theta_{\mathrm{a},k} - \bar{\theta}_{\mathrm{a}}\right)^2 \tag{4.33}$$

$$\mathrm{cov}(CC, CC) = \frac{(N_{\mathrm{jk}} - 2)^2}{2N_{\mathrm{jk}}(N_{\mathrm{jk}} - 1)}\sum_{k=1}^{N_{\mathrm{jk}}}\left(\theta_{\mathrm{c},k} - \bar{\theta}_{\mathrm{c}}\right)^2 \tag{4.34}$$

$$\mathrm{cov}(CC, AA) = \frac{(N_{\mathrm{jk}} - 2)}{\sqrt{2}N_{\mathrm{jk}}}\sum_{k=1}^{N_{\mathrm{jk}}}\left(\theta_{\mathrm{c},k} - \bar{\theta}_{\mathrm{c}}\right)\left(\theta_{\mathrm{a},k} - \bar{\theta}_{\mathrm{a}}\right) \tag{4.35}$$

This gives all the components needed to compute the covariance of TT, using its definition in Eq. 4.25:

$$\begin{aligned}\mathrm{cov}(TT, TT) &= \mathrm{cov}(AA, AA) + \mathrm{cov}(CC, CC) + 2\mathrm{cov}(AA, CC) \\ &= \frac{N_{\mathrm{jk}} - 1}{N_{\mathrm{jk}}} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{\mathrm{a},k} - \bar{\theta}_{\mathrm{a}}\right)^2 + \frac{(N_{\mathrm{jk}} - 2)^2}{2N_{\mathrm{jk}}(N_{\mathrm{jk}} - 1)} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{\mathrm{c},k} - \bar{\theta}_{\mathrm{c}}\right)^2 + \\ &\quad + \frac{\sqrt{2}(N_{\mathrm{jk}} - 2)}{N_{\mathrm{jk}}} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{\mathrm{c},k} - \bar{\theta}_{\mathrm{c}}\right)\left(\theta_{\mathrm{a},k} - \bar{\theta}_{\mathrm{a}}\right)\end{aligned} \tag{4.36}$$

Note how the terms scale differently with the number of the jackknife regions. [7] argue that this inconsistent scaling is the source of the bias that arises with the standard jackknife approach. In the next sections, we will see how adjusting this scaling can enable one to recover an unbiased covariance estimator and demonstrate the need for going beyond the Mohammad-Percival correction to get unbiased covariance estimators in all regimes of galaxy number density.

4.3.2 Mohammad-Percival correction

[7] proposed to weight the cross-pairs CC in order to fix the mismatch in the scaling, as seen in Eq. 4.36. With this weight α multiplying all the CC pair counts, the expression for TT_k becomes

$$TT_k = TT_{\mathrm{tot}} - AA_k - \alpha CC_k. \tag{4.37}$$

The definition of $\theta_{c,k}$ is then generalised to:

$$\theta_{c,k}(\alpha) = \frac{2\alpha}{N - 2\alpha}\left(\frac{N}{2}\overline{CC} - \alpha CC_k\right), \tag{4.38}$$

which also changes slightly the mean of this quantity as $\bar{\theta}_c(\alpha) = \alpha\overline{CC}$.

Following the steps from equations (4.29), (4.31) and (4.34), the modified expression for the covariance of the CC paircounts weighted by α is

$$\mathrm{cov}(\alpha CC, \alpha CC) = \frac{(N_{\mathrm{jk}} - 2\alpha)^2}{2\alpha^2 N_{\mathrm{jk}}(N_{\mathrm{jk}} - 1)} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{c,k} - \bar{\theta}_c\right)^2 \tag{4.39}$$

We see that for $\alpha = 1$ we recover the ordinary jackknife, as it will remove the cross-pairs in the same way as it removes the auto-pairs. Alternatively, by choosing $\alpha = N_{\mathrm{jk}}/\left[2 + \sqrt{2}(N_{\mathrm{jk}} - 1)\right]$ we can achieve equal scaling for the first two terms. Therefore, under the assumption of $\mathrm{cov}(CC, AA) = 0$ we indeed have all the terms scaling with N_{jk} in same manner, which can be seen by rewriting the expression for $\mathrm{cov}(TT(\alpha), TT(\alpha))$ as

$$\begin{aligned}\mathrm{cov}(TT(\alpha), TT(\alpha)) &= \mathrm{cov}(AA, AA) + \mathrm{cov}(\alpha CC, \alpha CC) = \\ &= \frac{N_{\mathrm{jk}} - 1}{N_{\mathrm{jk}}} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{\mathrm{a},k} - \bar{\theta}_{\mathrm{a}}\right)^2 + \frac{(N_{\mathrm{jk}} - 2\alpha)^2}{2\alpha^2 N_{\mathrm{jk}}(N_{\mathrm{jk}} - 1)} \sum_{k=1}^{N_{\mathrm{jk}}} \left(\theta_{c,k} - \bar{\theta}_c\right)^2\end{aligned} \tag{4.40}$$

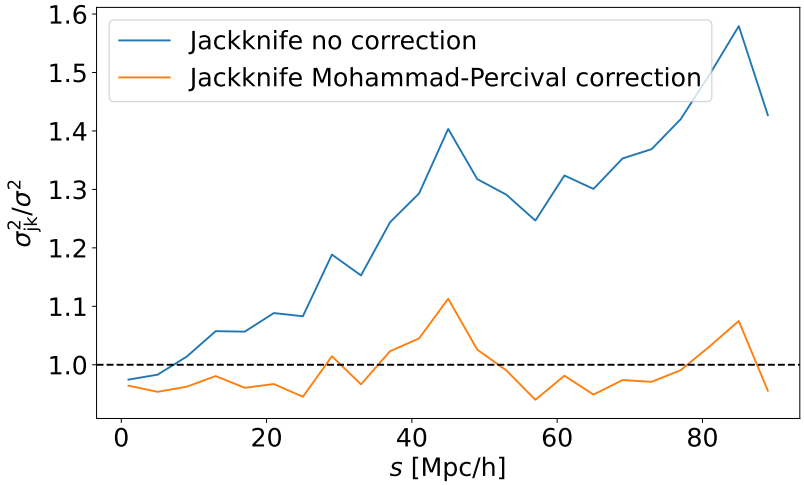

Figure 4.2: Comparison of the accuracy in the estimate of the diagonal elements of the covariance matrix for the real-space correlation functions as a function of scale obtained from 1000 cubic box independent mock catalogues. The ratio is the mean of the diagonal elements obtained using different jackknife approaches to those obtained directly from the ensemble of mocks. The noticeable scale-dependent bias that is visible for the standard jackknife estimate is absent when the Mohammad-Percival correction is employed.

In order to illustrate the effect of introducing the α weighting of [7], we create 1000 Poisson random catalogues in a box with a size of 1 Gpc/h, divide them into 125 cubic regions and then compute the covariance matrices of the real-space correlation function. We do this for both the standard jackknife and jackknife with the Mohammad-Percival correction. The results are presented in Fig. 4.2. We show the ratio of the mean of the diagonal elements, $\sigma^2 \equiv C_{ii}$, of the covariance matrix between jackknife-based $\sigma_{\rm jk}$ and mock-based σ (estimated directly using Eq. 4.22), where the blue curve uses the standard jackknife and the orange one includes the Mohammad-Percival correction. The standard jackknife is over-estimating the covariance with respect to that from the mocks, while introducing the α weighting of [7] for the cross-pairs removes this bias.

4.4 Fitted covariance

The real galaxy density has physical correlations and so galaxy distributions are not Poisson distributions. Therefore, the assumption of $\mathrm{cov}(CC, AA) = 0$ is not valid. With the α weighting of the cross-pairs that was introduced in Section 4.3.2, Eq. 4.35 becomes

$$\mathrm{cov}(\alpha CC, AA) = \frac{(N_{jk} - 2\alpha)}{\sqrt{2}\alpha^2 N_{jk}} \sum_{k=1}^{N_{jk}} \left(\theta_{c,k} - \bar{\theta}_c\right)\left(\theta_{a,k} - \bar{\theta}_a\right). \tag{4.41}$$

We can see that adopting any general fixed value of α unfortunately leaves the scaling

of $\mathrm{cov}(CC, AA)$ different from those of $\mathrm{cov}(AA, AA)$ and $\mathrm{cov}(CC, CC)$, so, in order to try to recover the benefits of the Mohammad-Percival approach, we treat α as a free parameter. We propose therefore to augment the jackknife method with α weighting where the value of α is tuned by fitting the covariance estimate from a limited number of mocks. A scheme that represents the approach is shown in Fig. 4.3. First, let us assume we have a set of N_m mocks $S = \{S_1 ... S_{N_m}\}$. Then, S/S_k denotes the set of mocks with the k^{th} mock removed. Then, we refer to the mock covariance from such a set S/S_k as $C_{ij}[S/S_k]$. We also introduce the α-dependent jackknife covariance obtained from a mock S_k with a chosen α weighting as $C_{ij}[S_k](\alpha)$, from correlation functions constructed with counts following eq. (4.37).

Having that in our possession, we are able to estimate the uncertainty on the diagonal elements of the covariance $\Xi_{ij}(\mathrm{diag}(C))$. First, we resample the given set of mocks and produce N_m covariances $C_{ij}[S/S_k]$. Then we compute the covariance matrix of the diagonals $\Xi_{ij}(\mathrm{diag}(C))$, where we limit ourselves to the diagonal elements as there are not enough degrees of freedom to build a covariance of matrices [8]:

$$\Xi_{ij}(\mathrm{diag}(C)) = \mathrm{cov}(C_{ii}, C_{jj}) = \\ = \frac{N_\mathrm{m} - 1}{N_\mathrm{m}} \sum_{k=1}^{N_\mathrm{m}} (C_{ii}[S/S_k] - C_{ii}[S])(C_{jj}[S/S_k] - C_{jj}[S]) \quad (4.42)$$

In general N_m should be greater than the number of elements in the fitted part of the covariance. However, in the case of a small N_m, one can restrict this to just the diagonal elements of Ξ_{ij}, to ensure that covariance matrix stays non-singular. The next step consists of finding which specific α is needed to obtain a realisation of the covariance matrix to describe $C_{ij}[S]$. First, we can write the α dependent estimator of the covariance $C_{ij}(\alpha)$ based on the mean of N_m α dependent jackknife covariances:

$$C_{ij}(\alpha) = \frac{1}{N_\mathrm{m}} \sum_{k=1}^{N_\mathrm{m}} C_{ij}[S_k](\alpha) \quad (4.43)$$

Then, the χ^2 of the $C_{ii}(\alpha)$ describing the $C_{ii}[S]$ can be written as:

$$\chi^2_C(\alpha) = \sum_{ij} (C_{ii}(\alpha) - C_{ii}[S]) \left(\Xi^{-1}\right)_{ij} (C_{jj}(\alpha) - C_{jj}[S]) \quad (4.44)$$

Following that, we minimise χ^2_C by varying α, such that we obtain $\chi^2_C(\alpha_{\mathrm{min}}) = \mathrm{min}(\chi^2_C(\alpha))$. To justify using the Gaussian likelihood in this procedure, we first notice that we are using only the diagonals of the covariance matrix. That allows us, with sufficiently large N_m, to approximate the distribution of the separate bins of the diagonals C_{ii} with a Gaussian.

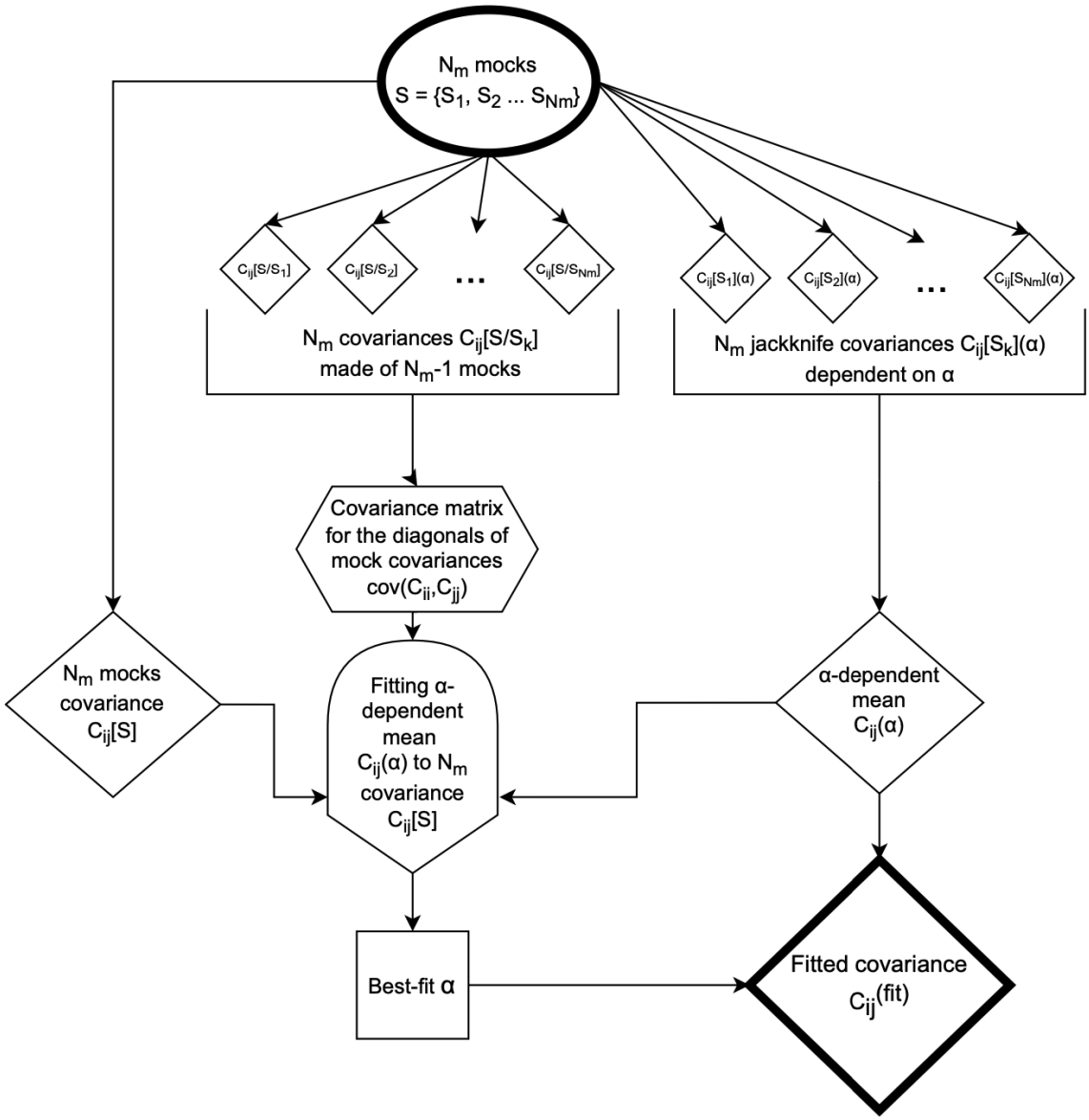

Figure 4.3: Schematic describing the procedure to obtain the fitted covariance $C^{ij}_{\rm fit}$ as defined in Eq. (4.45) and discussed in Section 4.4.

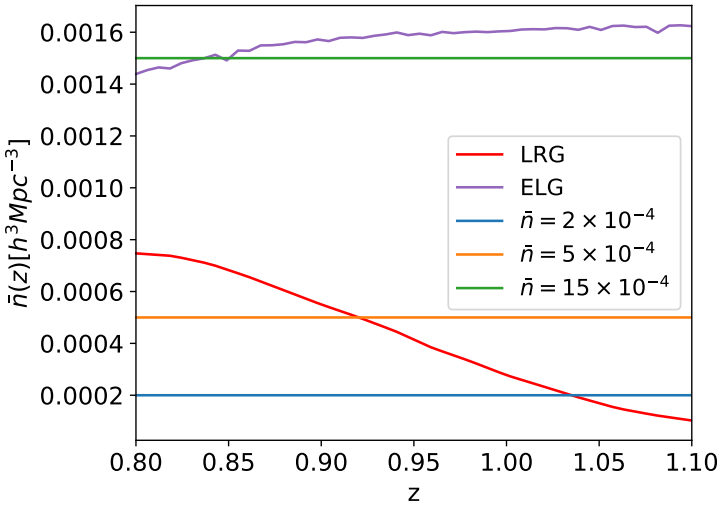

Figure 4.4: Number density dependence on redshift for different datasets used. The lognormal mock samples were chosen to have a constant density selection function, to simplify the matters, while LRG and ELG mock samples follow the expected values from the corresponding DESI survey subsets.

Therefore, our proposed estimator of the α dependent covariance matrix $C_{ij}^{(\mathrm{fit})}$ can be defined as:

$$C_{ij}^{(\mathrm{fit})} = C_{ij}(\alpha_{\mathrm{min}}) = \frac{1}{N_{\mathrm{m}}} \sum_{k=1}^{N_m} C_{ij}[S_k](\alpha_{\mathrm{min}}) \tag{4.45}$$

While only the diagonal of $C_{ij}^{(\mathrm{fit})}$ are used when fitting for α, all the elements of $C_{ij}^{(\mathrm{fit})}$ are consistently adjusted with the value of α that is found. In the original Mohammad-Percival approach, the contribution of the cross-pairs to the covariance is adjusted to match that of the auto-pairs. Our hybrid approach allows us to adjust the cross-pair contribution on the α weighted covariance so that the covariance matches the one obtained from the limited set of mocks. We will show in the next section that by doing so, we can greatly reduce the bias that can appear for dense samples when using the fixed α weighting of [7]. However, the hybrid approach does require more than a single mock to create a covariance estimate, but in the next section we will also show that the number of mocks needed is significantly reduced compared to a purely mock-based approach.

We test the performance of the fitted jackknife method with respect to other covariance matrix estimation methods on different sets of mocks that include RSD and some geometrical effects that we will describe in subsequent sections. For each specific set of mocks we also generate a set of matching random synthetic catalogues.

In section 5.4.3 we present the methodology of the tests that we perform on our mocks. In section 4.4.1, a set of tests is performed on lognormal mocks produced by

the `MockFactory` code[1] with three number densities to explore shot noise-dominated and sample variance-dominated regimes, but also to mimic the DESI LRG and ELG samples. In section 5.4.3 approximate EZmocks mimicking the DESI LRG and ELG samples are used to provide a mock-based covariance matrix which has the level of statistical precision of expected from the DESI Year-5 data. The corresponding number densities can be seen in Fig. 4.4 for LRG EZmocks in red, ELG EZmocks in purple and the different lognormal mocks at $\bar{n} = (2,\ 5,\ 15) \times 10^{-4}$ [Mpc/h]$^{-3}$ in blue, orange and green respectively. We use 1500 lognormal mocks for each space density, and 1000 ELG and LRG EZ mocks respectively.

Both the random and data samples are divided into $N_{\rm jk} = 196$ jackknife regions (the results, shown in Sec. 4.4.1, are not sensitive to $N_{\rm jk}$). The FKP weights are also applied.

4.4.1 Dependence on number density

We create 3 sets of lognormal mocks, each set containing 1500 realisations, for number densities $\bar{n} = 2 \times 10^{-4}$, 5×10^{-4} and $15 \times 10^{-4} h^3 \mathrm{Mpc}^{-3}$ at $z = 1$. Each of the realisations is made from a cubic box with a volume of $(2\mathrm{Gpc}/h)^3$ with grid of size 384^3 and fiducial cosmology with $h = 0.674$, $\sigma_8 = 0.816$ and $\Omega_{\rm m}^{(0)} = 0.31$. The CLASS code [9] is used to generate the initial power spectrum. Redshift space distortions are then added, and each box is cut to have a footprint that covers 15% of the full sky. Each mock is then analysed in the redshift range from 0.8 to 1.2, and the corresponding randoms are generated, which are about 4 times denser than the data mocks. The procedure to obtain the fitted jackknife covariance is summarised in Fig. 4.3 and explained in the previous section. Here, we use $N_{\rm m} = 50$ mocks. We measure correlation functions from the mocks in bins of $5h^{-1}$Mpc. Fig. 4.6 presents the α parameter value distribution, obtained from the fits of the covariances.

Fig. 4.5 shows a measure of the relative bias $\Delta\sigma^2(\xi_\ell)/\sigma(\sigma^2_{\rm Mock})$ between a jackknife-based covariance matrix and the mock-based covariance as a function of pair separation s. For simplicity we only consider the diagonal elements of each covariance matrix estimate. This relative bias is defined as

$$\frac{\Delta\sigma^2(\xi_\ell)}{\sigma(\sigma^2_{\rm Mock})} = \frac{\sigma^2(\xi_\ell) - \sigma^2_{\rm Mock}(\xi_\ell)}{\sigma(\sigma^2_{\rm Mock}(\xi_\ell))}, \tag{4.46}$$

where $\sigma(\xi_\ell)$ is the variance on a given multipole l obtained from the jackknife method, $\sigma_{\rm Mock}(\xi_\ell)$ is the variance on the same multipole obtained from the 1500 lognormal mocks and $\sigma(\sigma^2_{\rm Mock})$ is the uncertainty on the mock-based error bar, determined by applying the classical jackknife delete-one mock estimator to the set of mocks from which the covariance is estimated.

[1] `https://github.com/cosmodesi/mockfactory`

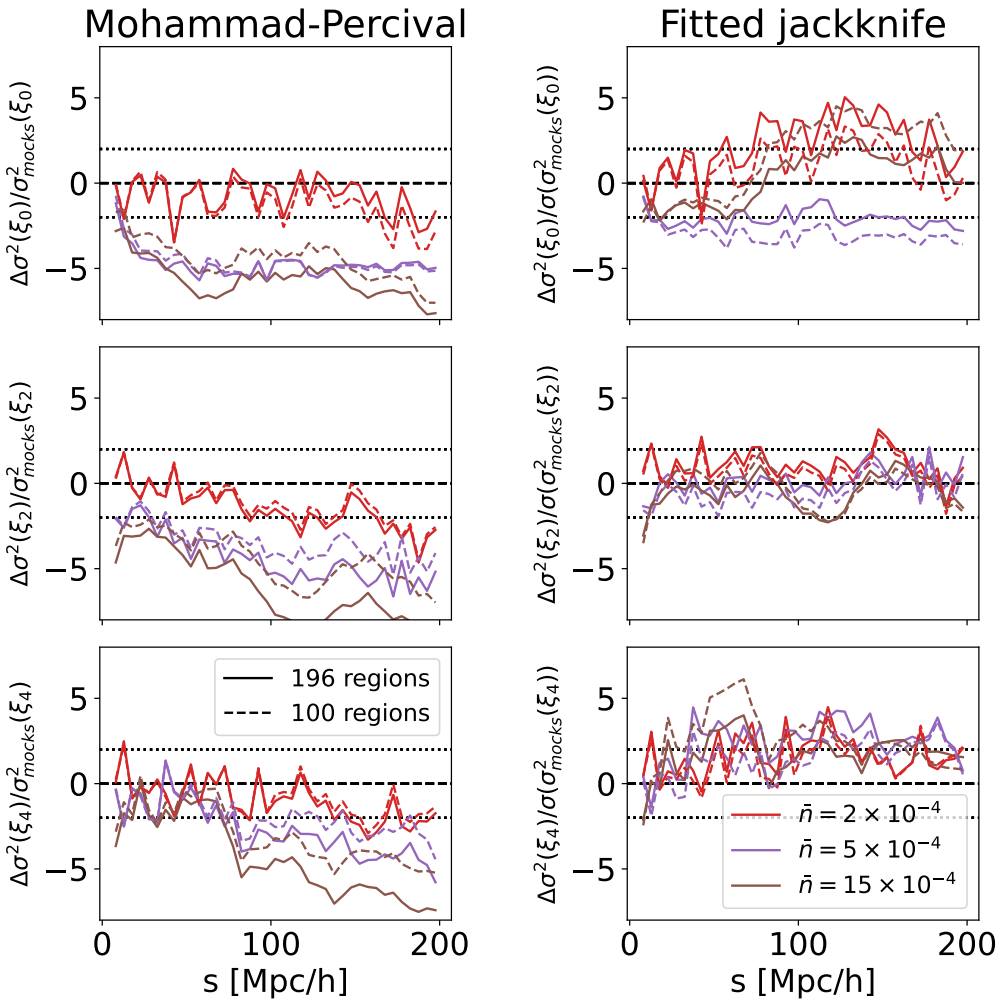

Figure 4.5: The average of the quantity defined in Eq. (4.46) representing the bias of the specific covariance estimation approach plotted as a function of separation, s, for various number densities.

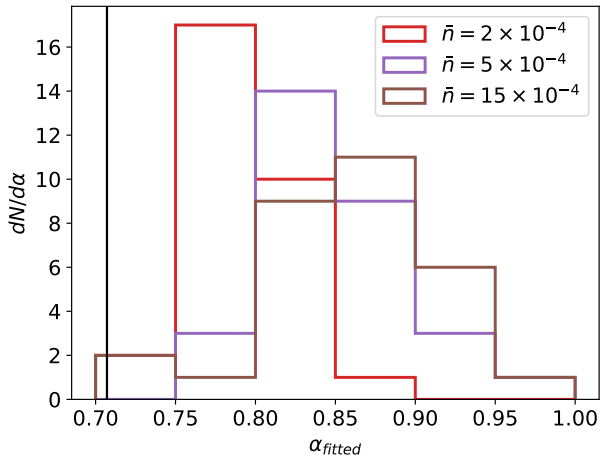

Figure 4.6: Histogram of the α parameter fitted from 50 mocks for lognormal mocks with $\overline{n} = 2 \times 10^{-4}, 5 \times 10^{-4}$and $15 \times 10^{-4}\ h^3\text{Mpc}^{-3}$. The vertical black line shows the value of $\alpha = N_{\text{jk}}/(2 + \sqrt{2}(N_{\text{jk}} - 1))$

The left panel of Fig. 4.5 shows this relative bias of the jackknife method with the Mohammad-Percival correction while the right panel shows the result for our fitted jackknife method. In both cases, the monopole, ξ_0, is displayed in the top panel, the quadrupole, ξ_2, in the middle and the hexadecapole, ξ_4, in the bottom. The coloured lines show different number densities and the solid lines are the baseline configuration of 196 jackknife regions while the dashed lines show the test of using 100 jackknife regions instead. As expected, the underestimation slightly worsens with the increase in the number of jackknife regions, as predicted by eq. (4.35).

However, as the number density $\bar{n}$ increases, the underestimation of the jackknife method with the Mohammad-Percival correction becomes more and more significant, especially for $\bar{n} = 15 \times 10^{-4}\ h^3\text{Mpc}^{-3}$. This underestimation is not visible on the jackknife covariance matrix estimates produced from the random catalogues as shown in Fig. 4.2. As explained in the previous section, the clustering of the data leads to higher covariance due to additional covariance coming from cross-correlations between CC and AA pair counts.

Additionally, there is no strong dependence on the number density for the fitted jackknife method which makes it more robust whatever the density regime of the galaxy sample of interest. It should be noted that for low-density regimes optimal α seems to be closer to the default value of Mohammad-Percival approach, and its fitted estimation in our method introduces additional uncertainty, which makes our method more imprecise as $\overline{n}(z)$ decreases.

We also use a set of 1000 EZmocks generated from N-body simulations with 6 Gpc/h box size. The fiducial cosmology employed is Planck 2018 [10], and the boxes are

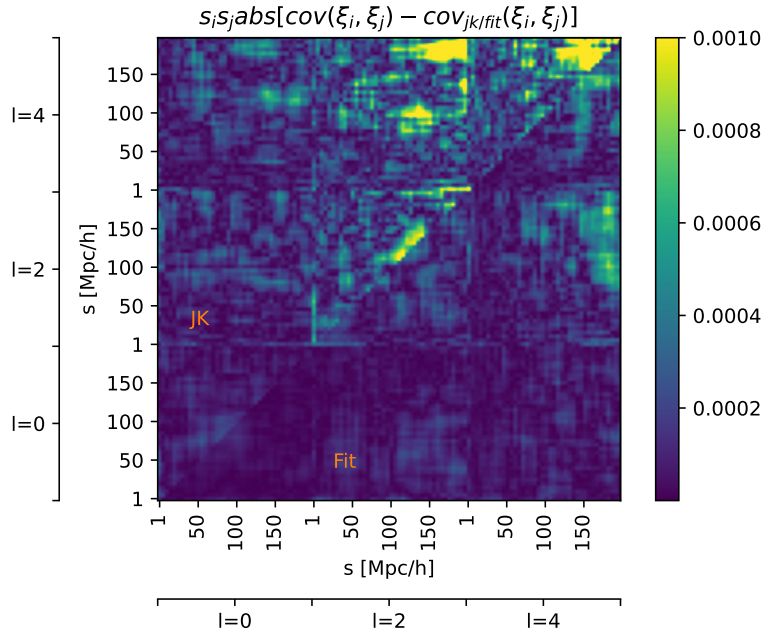

Figure 4.7: Comparison of the deviation of jackknife and fit covariances from the mock covariance multiplied by a square of separation for multipoles $\ell = 0, 2, 4$ for the EZ LRG mocks.

generated at $z = 0.8$ for the LRGs and $z = 1.1$ for the ELGs. We use the redshift range of $z = [0.8, 1.1]$ and the mocks are cut to a footprint that reproduces that planned for the 5-year DESI data in order to match the expected final precision of the mock-based covariance matrix. The comparison of the difference with the mock covariance for the single realisation of the jackknife covariance and the fitted covariance is presented in Fig. 4.7.

On Fig. 4.8 the relative bias of the diagonals of jackknife-based vs mock-based covariances as defined by eq. 4.46 are shown for the LRG sample on the left and for the ELG sample on the right, in a similar way to Fig. 4.5. First, The same trend is seen for the Mohammad-Percival jackknife as we found with the lognormal mocks: the bias of the jackknife method with the Mohammad-Percival correction tends to increase with number density, so from LRG to ELG, and the fitted jackknife is still able to mitigate it. However, we can also notice that the differences are less pronounced in the case of the EZmocks which is due to a bigger volume being probed by the same number density.

Given that for the same number density we see a bigger discrepancy between the jackknife method with Mohammad-Percival correction and the mock-based covariance matrix in the case of the lognormal mocks than with the approximate mocks, we also explore the effect of varying the size of the footprint. We consider the LRG EZmocks and compute the covariance matrix for the three methods (jackknife with Mohammad-Percival, fitted jackknife and mock-based) for the Southern Galactic Cap separately and compare

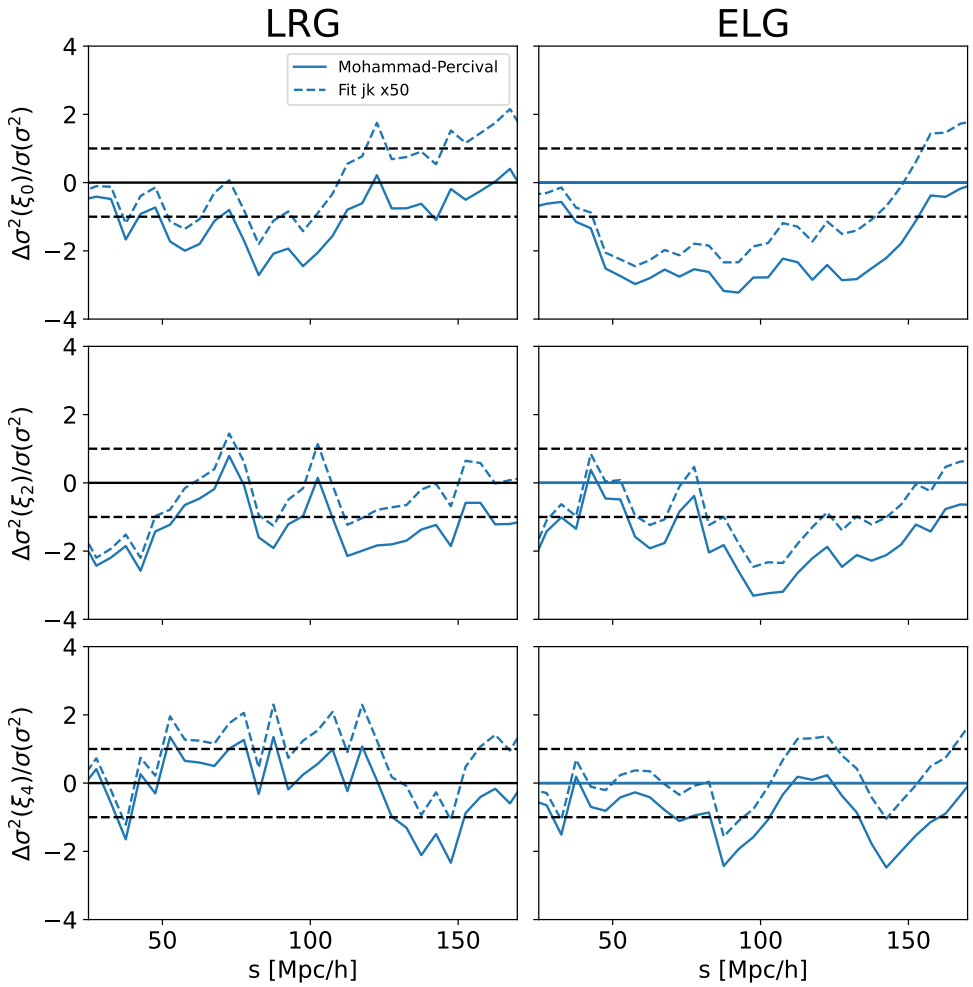

Figure 4.8: The quantity defined in Eq. (4.46) representing the bias of the specific covariance estimation approach plotted for three multipoles of LRG and ELG EZmocks (left and right panels respectively). Solid lines are with Mohammad-Percival correction and dashed lines for the fitted jackknife.

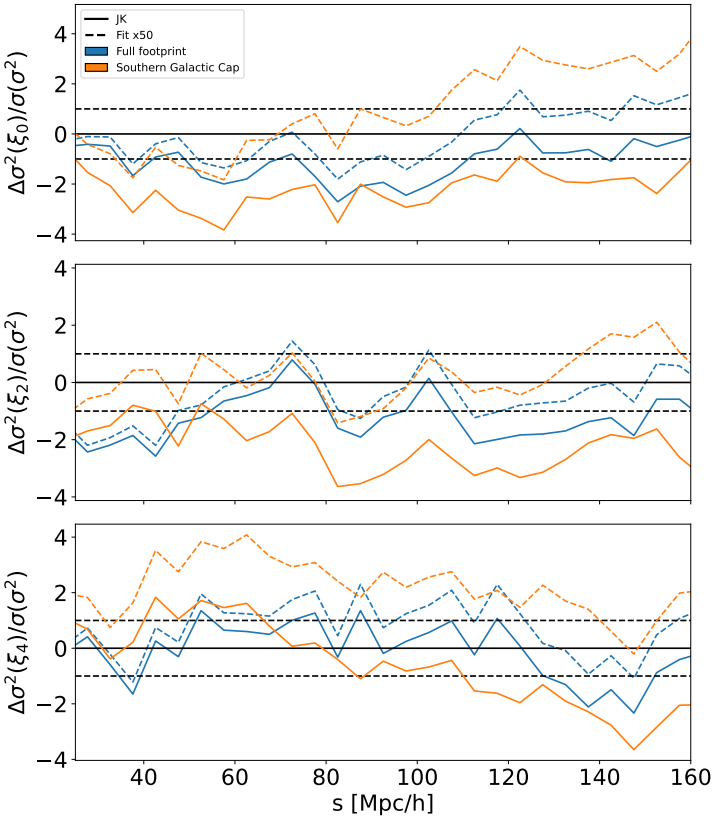

Figure 4.9: Relative bias in the estimate of the variance as defined in equation (4.46) for the fitted jackknife method (dashed) and for the Mohammad-Percival approach (solid) as a function of pair separation for LRG EZ mocks. The orange curves are the results for the SGC while the blue curves are for the larger full footprint.

the results with the ones for the full Y5 footprint. We keep the same number of jackknife regions in all cases.

The results are displayed in Fig. 4.9 where we plot the relative bias as defined by equation (4.46) between a jackknife method and the mock-based covariance as a function of pair separation for the monopole (top), quadrupole (middle) and hexadecapole (bottom). The results for the SGC are shown in orange and the ones for the full footprint in blue. Indeed, we see that the bias associated with the Mohammad-Percival approach is higher when a smaller footprint is used. We also confirm the same effect for the ELG dataset. Therefore, it makes the fitted jackknife covariance method even more useful when the footprint considered is relatively small. We believe that this effect might be related to super-sampling covariance, and if it is indeed the case, it would mean that our approach can successfully take it into account.

Overall, throughout all of the tests for varying number densities, different types of

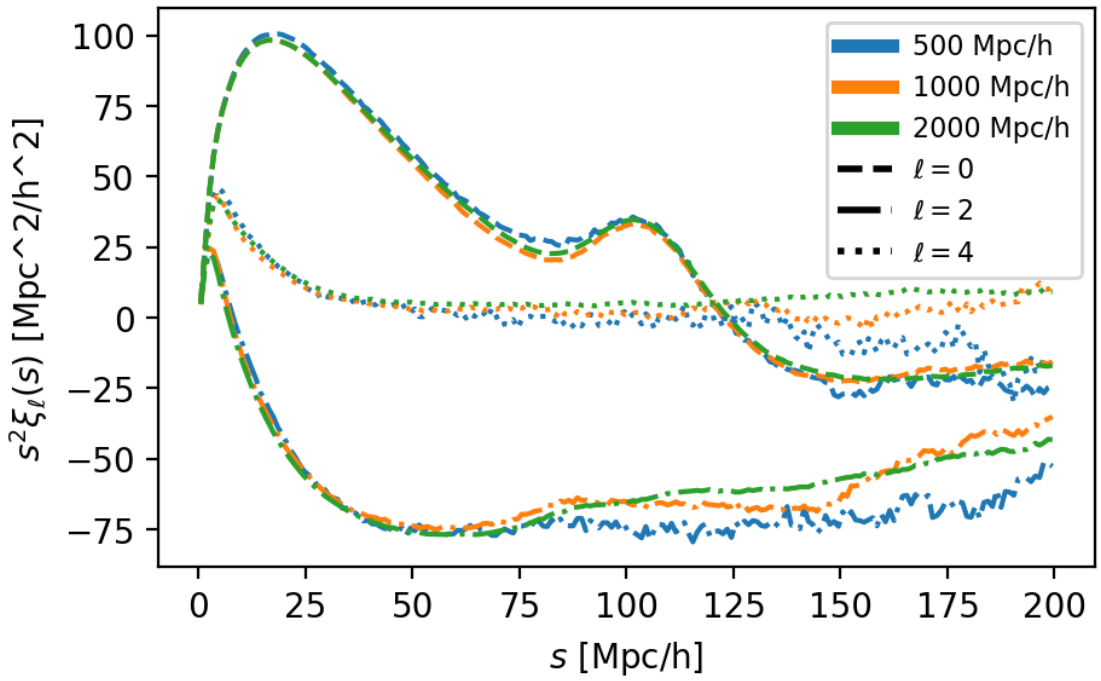

Figure 4.10: The correlation function multipoles multiplied by separation squared s^2 are plotted for three different mock species produced from different GLAM simulation boxes with periodic boundary conditions. We see that the clustering of all three is almost identical, with certain differences appearing for example on the quadrupole only after a scale of 100 Mpc/h

mocks and number of fitted mocks, the fitted jackknife approach shows a considerable improvement over the correction for standard jackknife proposed by [7]. The fitted jackknife approach can achieve an unbiased estimate of the covariance matrix with similar precision to a mock-based covariance but with the major advantage of requiring a much smaller number of mocks.

4.5 Replications

Sometimes producing even a limited number of mocks with both the volume and the resolution required can be a problem. In this case, what is done is replication of the simulation boxes: thanks to the presence of periodic boundary conditions, we can stack into a larger box until the point where we reach the required target volume. The clustering of such mocks will be equivalent to those produced with a sufficient box size for scales smaller than the size of the replicated box and the larger-scale coupling to non-linear scales is negligible.This is illustrated in Figure 4.10, where the averaged multipoles from GLAM boxes with different box sizes.

When looking at the covariances with the boxes with periodic boundary conditions,

the ratio of covariances will be inversely proportional to the ratio of the volumes. That is illustrated in Figure 4.11, where the standard deviations of the multipoles $\sigma(\xi_\ell)$ for different GLAM boxes with varying sizes are presented, with some of them scaled by $\sigma(\xi_\ell^{V_1})/\sigma(\xi_\ell^{V_2}) = \frac{V_2}{V_1}$.

Testing the effect of replications on two different mock sizes (2 Gpc/h and 1 Gpc/h) such that one of them needs replications to mimick BGS, we find that (Figure 4.12) the scaling of the uncertainty is non-trivial, as only the monopole $\ell = 0$ is affected, while as the quadrupole and hexadecupole are negligiably biased, at least for the mock in our posession. The FitCov method described earlier (Section 4.4), as well as jackknife in general, seems to stay untouched by the replication, as long as the replication does not affect one of the multipoles, as can also be seen in Figure 4.12, which might make the production of accurate covariances even cheaper than thought previously, as simulated volumes can be made smaller. The only modification which was done to the original version of the algorithm is not including the monopole part into the fitting procedure, such that only quadrupole and hexadecupole were used in fitting α. Having discovered this effect, we plan to further investigate it with more details in the future.

4.6 Covariance for the DR1 BGS sample

Having described all of the different approaches to producing the covariance matrices, we can try to apply them to the production of the DESI DR1 BGS covariance matrix.

Analytic covariances for the DESI BGS DR1 were created by Misha Rashkovetskiy in configuration space ([3]) and by Otavio Alves in Fourier space (Alves et al. in prep). As the computation of the analytic covariances required a detailed computation of the window function kernels in order to take into account the geometrical effects, the analytic covariances were restricted to the sample with $Mr < -21.5$ for computational reasons. The codes used are public, the one for the configuration space being `RascalC` [2, 11] and the one for the Fourier space - `thecov` [12–14][2]. Both analytic covariances use the 2-point statistics from the observed DR1 BGS dataset as an input, while the window kernels are computed using the corresponding randoms. Additionally, `RascalC` calibrates a shot noise term using the jackknife estimator[2]. Two types of BGS mocks are used to produce mock-based covariances, which I described in Chapter 3: EZmock and GLAM mocks, where we have used 1000 of each. Lastly, I computed the fitted covariance (FitCov) using 25 Abacus mocks.

The comparison of standard deviation σ from different BGS DR1 covariance matrix estimations is shown in Figure 4.13 for the monopole, quadrupole and hexadecapole correlation function, and for the power spectrum in Figure 4.14, where in the lower panels

[2] `https://github.com/cosmodesi/thecov`

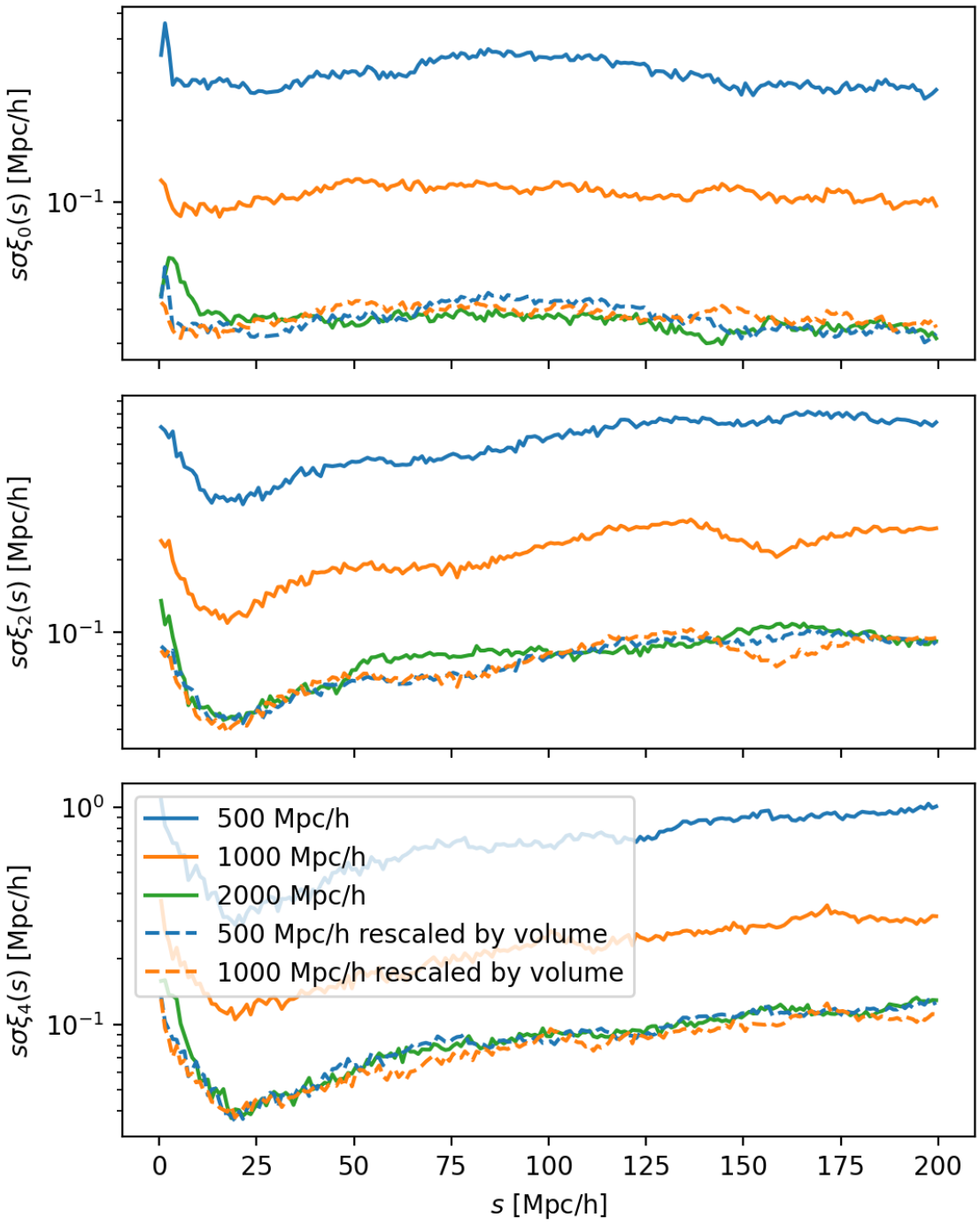

Figure 4.11: Standard deviation σ multiplied by separation s is plotted for three different mock species produced from different GLAM simulation boxes with periodic boundary conditions. The dashed lines represent those from smaller boxes, but rescaled with respect to the simulation volume. We can see, that this simple scaling is able recover σ pretty consistently.

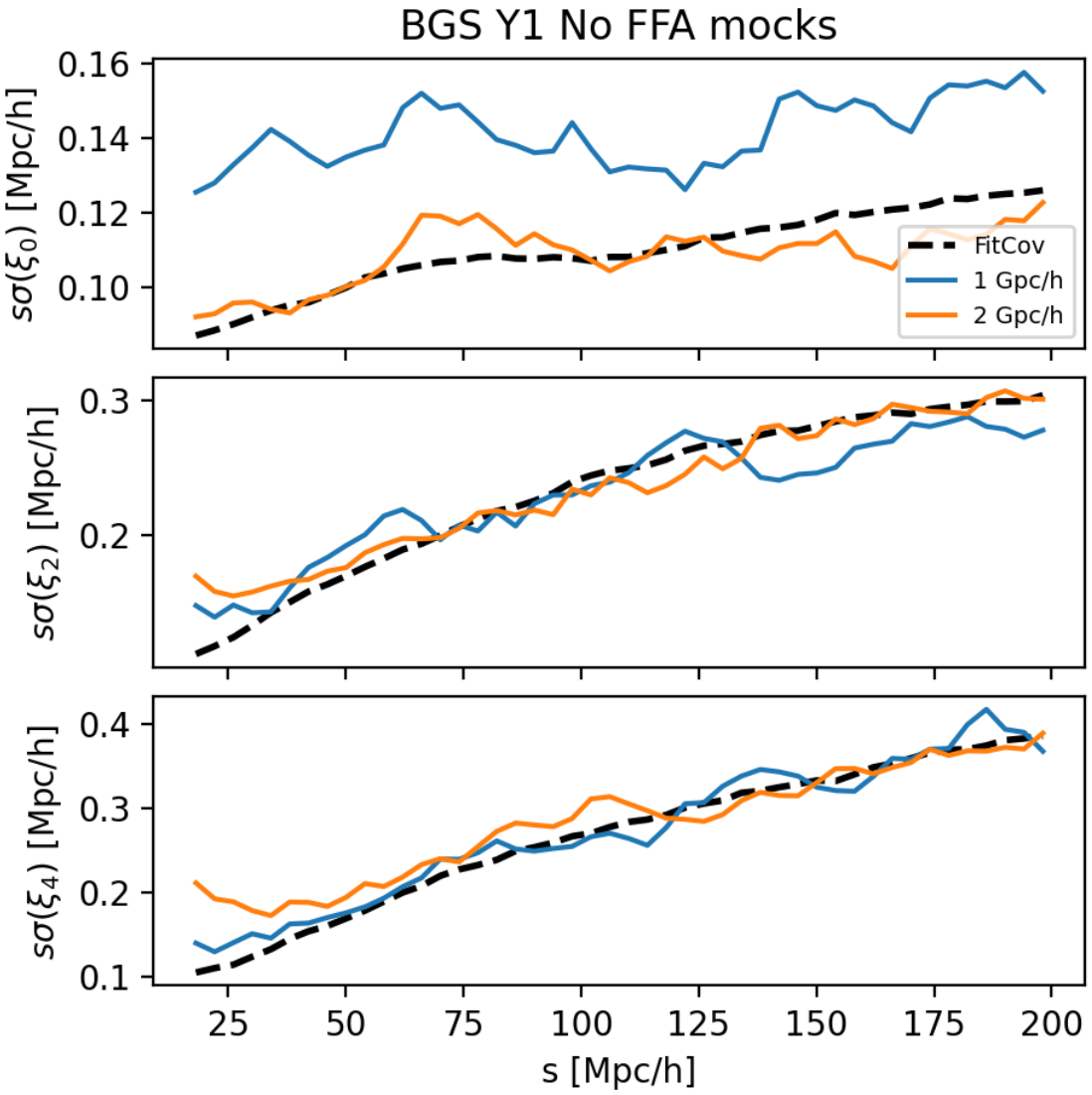

Figure 4.12: Standard deviation σ multiplied by separation s is plotted for three different mock species produced from different GLAM simulation boxes, replicated and cut to the footprint of BGS DR1. We see that the monopole is affected , however quadrupole and hexadecupole stay untouched. The dashed black line represents the FitCov results, produced from 50 mocks with simulated volume of 1 Gpc/h, and fitted in the separation range of $[40, 160]$ Mpc/h.

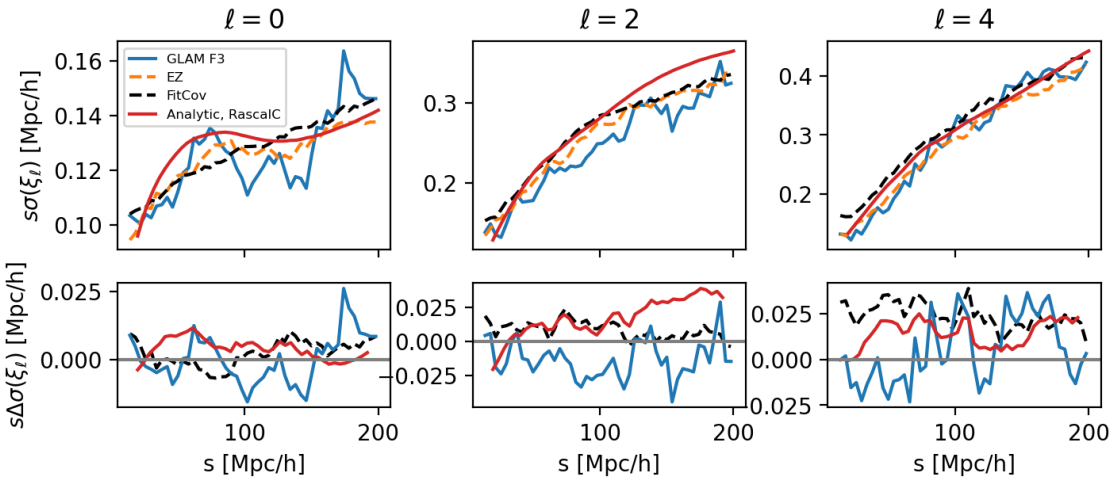

Figure 4.13: The standard deviation computed with different approaches in configuration space.

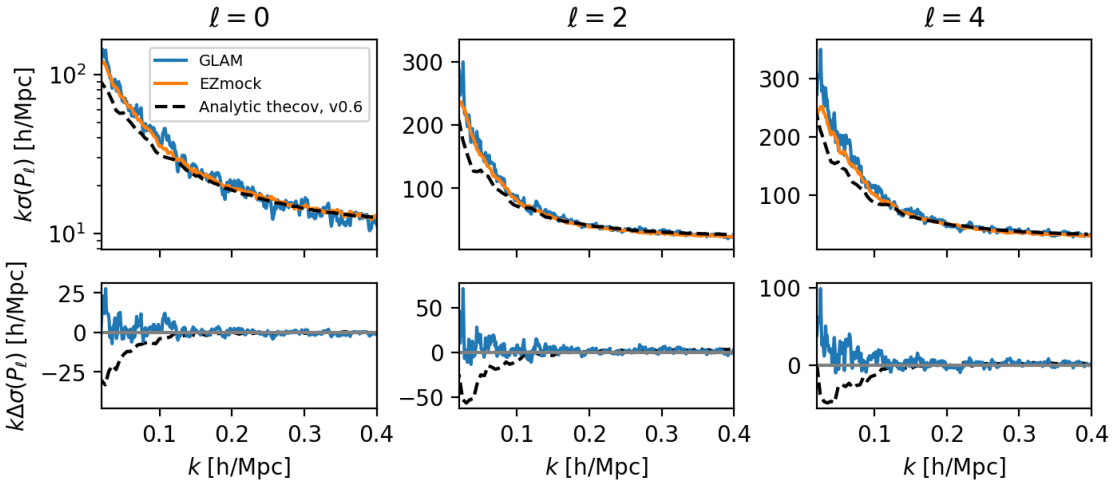

Figure 4.14: The standard deviation computed with different approaches in Fourier space.

we present residiuals with respect to the reference EZmock covariance. We can see that the analytic covariance for the power spectrum tends to underestimate the uncertainty at large k, while the one for the correlation function seems to overestimate the monopole at intermediate scales, and the quadrupole at large scales. For both statistics, EZmock and GLAM covariances are fully consistent. FitCov, developed for the correlation function, is also consistent with the mock covariances.

This gives us a set of covariances we use extensively in Chapter 5 for cosmological inference. Having several of them allows us to cross-verify them, thus ensuring the validity of our approaches to error estimation. Additional tests of the cosmological parameters inferred with the different covariance matrix estimates are presented in Chapter 5.

References

[1] Digvijay Wadekar and Román Scoccimarro. "Galaxy power spectrum multipoles covariance in perturbation theory". In: *Phys. Rev. D* 102 (12 Dec. 2020), p. 123517. DOI: 10.1103/PhysRevD.102.123517. URL: https://link.aps.org/doi/10.1103/PhysRevD.102.123517.

[2] Oliver H E Philcox et al. "rascalc: a jackknife approach to estimating single- and multi-tracer galaxy covariance matrices". In: *MNRAS* 491.3 (Nov. 2019), pp. 3290–3317. ISSN: 0035-8711. DOI: 10.1093/mnras/stz3218. eprint: https://academic.oup.com/mnras/article-pdf/491/3/3290/31490300/stz3218.pdf. URL: https://doi.org/10.1093/mnras/stz3218.

[3] Michael Rashkovetskyi et al. "Validation of semi-analytical, semi-empirical covariance matrices for two-point correlation function for early DESI data". In: *Mon. Not. Roy. Astron. Soc.* 524.3 (2023), pp. 3894–3911. DOI: 10.1093/mnras/stad2078. arXiv: 2306.06320 [astro-ph.CO].

[4] Hartlap, J. et al. "Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix". In: *A&A* 464.1 (2007), pp. 399–404. DOI: 10.1051/0004-6361:20066170. URL: https://doi.org/10.1051/0004-6361:20066170.

[5] Cheng Zhao et al. "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: 1000 multi-tracer mock catalogues with redshift evolution and systematics for galaxies and quasars of the final data release". In: *MNRAS* 503.1 (Feb. 2021), pp. 1149–1173. ISSN: 0035-8711. DOI: 10.1093/mnras/stab510. eprint: https://academic.oup.com/mnras/article-pdf/503/1/1149/36626380/stab510.pdf. URL: https://doi.org/10.1093/mnras/stab510.

[6] P. Norberg et al. "Statistical analysis of galaxy surveys – I. Robust error estimation for two-point clustering statistics". In: *MNRAS* 396.1 (June 2009), pp. 19–38. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2009.14389.x. eprint: https://academic.oup.com/mnras/article-pdf/396/1/19/4066956/mnras0396-0019.pdf. URL: https://doi.org/10.1111/j.1365-2966.2009.14389.x.

[7] Faizan G Mohammad and Will J Percival. "Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function". In: *MNRAS* 514.1 (May 2022), pp. 1289–1301. ISSN: 0035-8711. DOI: 10.1093/mnras/stac1458. eprint: https://academic.oup.com/mnras/article-pdf/514/1/1289/43986041/stac1458.pdf. URL: https://doi.org/10.1093/mnras/stac1458.

[8] John Wishart. "THE GENERALISED PRODUCT MOMENT DISTRIBUTION IN SAMPLES FROM A NORMAL MULTIVARIATE POPULATION". In: *Biometrika* 20A.1-2 (Dec. 1928), pp. 32–52. ISSN: 0006-3444. DOI: 10.1093/biomet/20A.1-2.32. eprint: https://academic.oup.com/biomet/article-pdf/20A/1-2/32/530655/20A-1-2-32.pdf. URL: https://doi.org/10.1093/biomet/20A.1-2.32.

[9] Diego Blas et al. "The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes". In: *JCAP* 2011.07 (July 2011), p. 034. DOI: 10.1088/1475-7516/2011/07/034. URL: https://dx.doi.org/10.1088/1475-7516/2011/07/034.

[10] N. Aghanim et al. "iPlanck/i2018 results". In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. URL: https://doi.org/10.1051/0004-6361/201833910.

[11] Ross O'Connell et al. "Large covariance matrices: smooth models from the two-point correlation function". In: *MNRAS* 462.3 (July 2016), pp. 2681–2694. ISSN: 0035-8711. DOI: 10.1093/mnras/stw1821. eprint: https://academic.oup.com/mnras/article-pdf/462/3/2681/8010260/stw1821.pdf. URL: https://doi.org/10.1093/mnras/stw1821.

[12] Otavio Alves and DESI Collaboration. "Analytical covariance matrices of DESI galaxy power spectrum multipoles". (in prep.) 2024.

[13] Digvijay Wadekar and Roman Scoccimarro. "Galaxy power spectrum multipoles covariance in perturbation theory". In: *Phys. Rev. D* 102.12 (2020), p. 123517. DOI: 10.1103/PhysRevD.102.123517. arXiv: 1910.02914 [astro-ph.CO].

[14] Yosuke Kobayashi. "Fast computation of the non-Gaussian covariance of redshift-space galaxy power spectrum multipoles". In: *Phys. Rev. D* 108.10 (2023), p. 103512. DOI: 10.1103/PhysRevD.108.103512. arXiv: 2308.08593 [astro-ph.CO].

Chapter 5

Clustering analysis and its application to DESI BGS

Тому, кто зряч, но светом дня ослеп, —
Смысл голосов, звук слов, событий звенья,
И запах тел, и шорохи растенья, —
Весь тайный строй сплетений, швов и скреп
Раскрыт во тьме.

M. Voloshin, 126

Introduction

During the course of the previous chapters we have discussed the different components needed for obtaining the data from the observed Universe: how to obtain the data, how to model it and how to estimate the uncertainties on it. Now, there is only one last step left: getting the cosmological parameters themselves. In this chapter, we will discuss the various ways to obtain the cosmological parameters from the 2-point clustering statistics. We will start with discussing the most classic approaches: BAO-only and RSD analysis, before moving on to the full-shape analysis employed in DESI instead of the classic RSD and we will discuss the question of compression. Then, we will test the tools we have created for the full-shape analysis using realistic mocks. Eventually, we will present the results of the analysis of the DESI BGS DR1 sample, both single- and multi-tracer.

5.1 Clustering analysis

5.1.1 BAO analysis

The BAO analysis starts with the choice of fiducial cosmology, and generation of the linear power spectrum, which is usually separated into the broadband part without the characteristic BAO wiggle, $P^{\rm fid}_{nw} = P^{\rm fid}_{\rm lin} - P^{\rm fid}_{\rm wig}$ and the wiggle $P^{\rm fid}_{\rm wig}$ itself. The following model is then often adopted [1–3]:

$$P_{\rm BAO}(k,\mu) = B^2 P^{\rm fid}_n w(k) \left[1 + \left(\frac{P^{\rm fid}_{\rm lin}}{P^{\rm fid}_{nw}} - 1\right) e^{-\frac{1}{2}\left(\mu^2 k^2 \Sigma^2_{\parallel} + (1-\mu^2) k^2 \Sigma^2_{\perp}\right)}\right] \tag{5.1}$$

where B is responsible for the general amplitude, while $\Sigma_{\parallel}$ and $\Sigma_{\perp}$ are responsible for the non-linear BAO dumbping, and are usually estimated from simulations.

Additionally to that, one usually marginalises broadband part by adding to each multipole of the power spectrum a following term $g^{\ell}(k)$:

$$g^{\ell}(k) = \sum_{i=1}^{N} A^{\ell}_i k^{2-i} \tag{5.2}$$

where A^{ℓ} are marginalised over. The last thing to do will be to modify the power spectrum to account for the Alcock-Paczynski effect, as for example in eq.69. The configuration space multipoles are produced by Hankel transformations, as for example, was done in [4]. Various prescriptions exist, however, including isotropic BAO fitting, such as used here [5], for example.

Non-linear gravitational evolution, however, broadens the BAO peak in the correlation function, or damps the corresponding oscillations in the power spectrum. Plus, the location of the peak can also be shifted due to galaxy biasing [6, 7]. These effects in total degrade the BAO measurements. There is however a technique to reverse those effects, thus improving the constraints on the BAO, called reconstruction [8]. This modifies the equation 5.1 to be[3]:

$$P_{\rm BAO}(k,\mu) = B^2 (1 + \frac{b}{f}\mu^2 R)^2 P^{\rm fid}_n w(k) \left[1 + \left(\frac{P^{\rm fid}_{\rm lin}}{P^{\rm fid}_{nw}} - 1\right) e^{-\frac{1}{2}\left(\mu^2 k^2 \Sigma^2_{\parallel} + (1-\mu^2) k^2 \Sigma^2_{\perp}\right)}\right] \tag{5.3}$$

where R is the reconstruction smoothing scale, which is set to 0 for the prereconstructed catalogues. For DESI BAO analysis the model was modified further however, for additional details see [9]. Furthermore, reconstruction has been shown to be effective in removing the shifts of the acoustic peak, and reducing the systematic errors in BAO measurements [10–13]. An example of the effect on the BAO measurement is shown on Figure 5.1.

Thus, reconstruction has become a standard technique to use in the galaxy surveys when making BAO measurements, recent DESI results not excluded [4, 9]. However, in

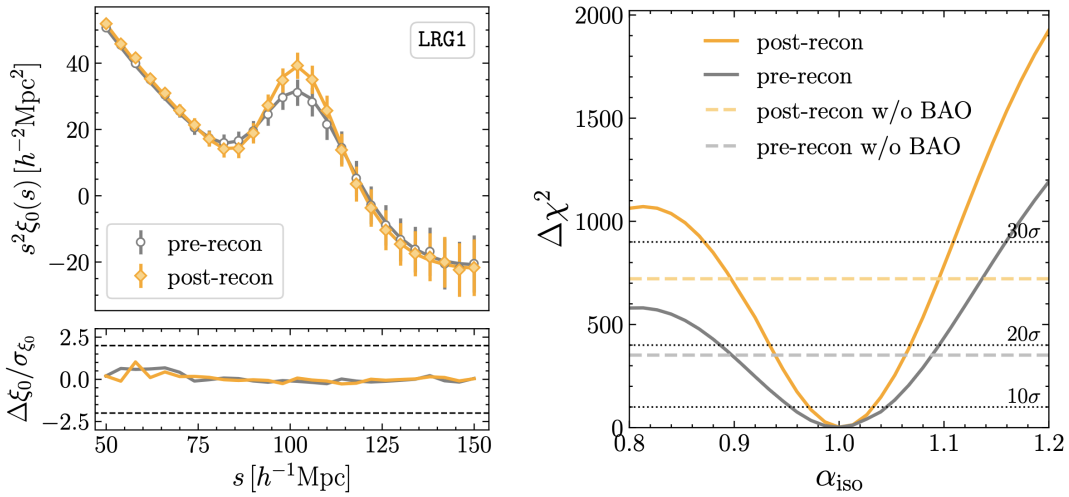

Figure 5.1: *Left*: The dots shown the monopole of the two-point correlation function before and after reconstruction averaged over 25 realisations of Abacus LRG mocks, measured in $0.4 < z < 0.6$. The error bars are taken from a semi-analytic covariance matrix [14]. The solid lines correspond to the best-fit model. *Right*: $\Delta\chi^2$ as a function of the isotropic BAO scaling parameter α_{iso}. Solid and dashed lines show results from the models with and without BAO included. Taken from [15].

this thesis we will not perform the BAO-only measurements, so the interested reader is advised to learn more, for example, from [15].

5.1.2 RSD and full-shape analysis

As in the previous section, we start by the choice of the reference cosmology. In this approach, the cosmology is not varied throughout the posterior exploration. Thus the linear power spectrum is created only once. The generation of the non-linear power spectrum is then performed in two steps in case of the clustering analysis, which were explained in detail in Chapter 2: 1) computation of the perturbation theory kernels, 2) multiplication of those with the nuisance parameters such as biases. Then, using the statistical machinery explained in Introduction, we obtain the values of f and fiducial parameters such as biases. However, we should note that the difference between the fiducial and the observed cosmology is not taken into account at this stage. We can take a look at smaller scales around 30 – 50 h/Mpc , where the RSD effect happens. Let us leave aside the difference between observed and fiducial cosmologies for now. We can note that one of the parameters, usually‘ computed from theory, but which can be easily varied in a fashion similar to biases is the growth rate of structures f. Given its direct connection to the equation of state of dark energy (equation 64), we get a powerful probe. We can add a modification to the modelling of the non-linear power spectrum by accounting

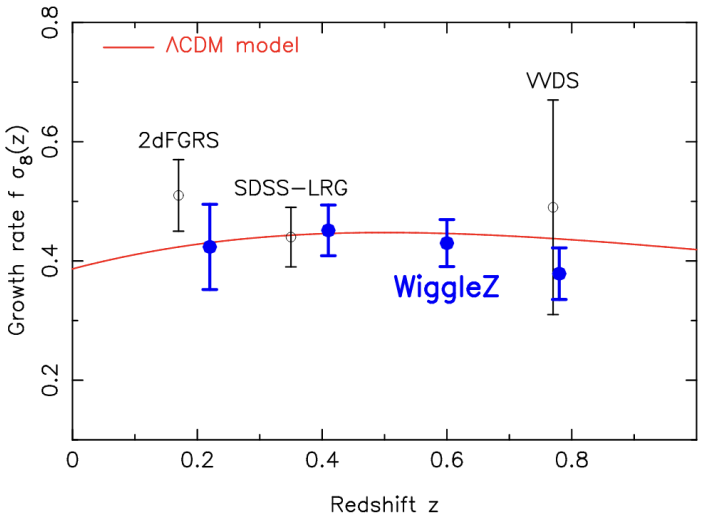

Figure 5.2: Growth rate from various outdated surveys with RSD part inferred only, taken from [19].

for the difference in cosmologies by scaling the resulting statistics as $k_\perp = \tilde{k}_\perp/\alpha_\perp$ and $k = \tilde{k}_\parallel/\alpha_\parallel$, following [16], which will account for the difference between the fiducial and real cosmology. This type of analysis is called the RSD analysis.

There is caveat, though. f is highly correlated with σ_8, to the point where one can write $f = \frac{d\sigma_8(a)}{d\ln a}$. Thus, what is usually reported is $f\sigma_8$ measurement, keeping in mind the unresolved differences between the cosmologies, which is usually valid for smaller redshifts [17]. This is for example one of the methodologies explored in [18].

Once the growth rate is measured, usually at several redshifts (for which the data of different surveys can be used), we can explore the evolution of the growth of over time and compare it with different cosmological models. A classical example of such a plot can be seen in Figure 5.2.

We can now see that by measuring the evolution of $f\sigma_8$, we can later fit the curve resulting from the equation 64, thus, getting an estimate of Ω_m if we assume ΛCDM, or even γ.

Both the RSD and BAO analysis provide information on cosmological parameters. It should be kept in mind though, that the two use the different parts of the target clustering statistics to obtain information: RSD is usually more present on the smaller scales, while BAO dominates large scales. Thus, it is possible to infer information from them simultaneously, thus accounting for the difference in cosmology during the RSD measurement [20]. It should be noted that the AP effect and the RSD effect are degenerate. However, with precise enough measurements, they can be disentangled, and thus this is the reason why it was not used earlier. This technique is called the full-shape clustering analysis, and is the one used throughout this thesis.

5.2 The question of compression

The methods described earlier are examples of methods employing the so-called template compression, meaning that the template linear power spectrum is kept fixed during the inference and the cosmological parameters are not inferred directly, but rather through compressed parameters, such as $f\sigma_8, \alpha_\perp, \alpha_\parallel$, that control the late-time dynamics of the Universe. However, what would happen if we would attempt to use the cosmological parameters directly?

As was detailed in Chapter 2, the computation of the perturbation theory clustering statistics from scratch takes considerable time, if emulators are not used. But, it was shown that a considerable amount of information is lost when performing the fixed-template compressed analysis[21], if not combined with the CMB data and when compared with a direct fitting of the cosmological parameters from the 2-point statistics. This is illustrated on Figure 5.3 where the grey curve shows the constraints from RSD alone, orange from BAO+RSD and blue from Full-Modelling (or direct fitting).

It should be noted, that the template compression can be extended in order to include the information otherwise missing. This is done by introducing the parameter m which controls the slope of the linear power spectrum. This method is called ShapeFit[21] and the linear power spectrum is modified such that:

$$P'_{\rm lin}(k) = P_{\rm lin}(k)\exp\left\{\frac{m}{a}\tanh\left[\ln\left(\frac{k}{k_p}\right)\right] + n\ln\left(\frac{k}{k_p}\right)\right\} \tag{5.4}$$

where a is a fixed parameter that controls the large and small scale limits and how they are reached, while m is variable and controls the slope itself, representing the maximum slope at scale k_p. k_p is usually chosen to correspond to the location where the baryon impression is the strongest and kept fixed. More on this approach can be found in [21, 23], and its application to DESI in [**maus2024**] for the theoretical model we are using in this work, `velocileptors`[24, 25]

In preparation of the DESI DR1 Full-Shape analysis, the three methods (classic RSD, ShapeFit and Full-Modelling) have been compared, and, as can be seen in Figure 5.4, the ShapeFit and Full-Modelling approaches indeed yielded much tighter constraints on the cosmological parameters of ΛCDM, when tested on the AbacusSummit cubic boxes[26].

5.3 Projection effects

One of the problems appearing with posteriors from high-dimensional likelihoods in Bayesian inference is often our inability to present and imagine the results in the ways other than projections with 1D and 2D projections. Thus, for these types of likelihoods the same posteriors might be very misleading. A good example is shown in Figure

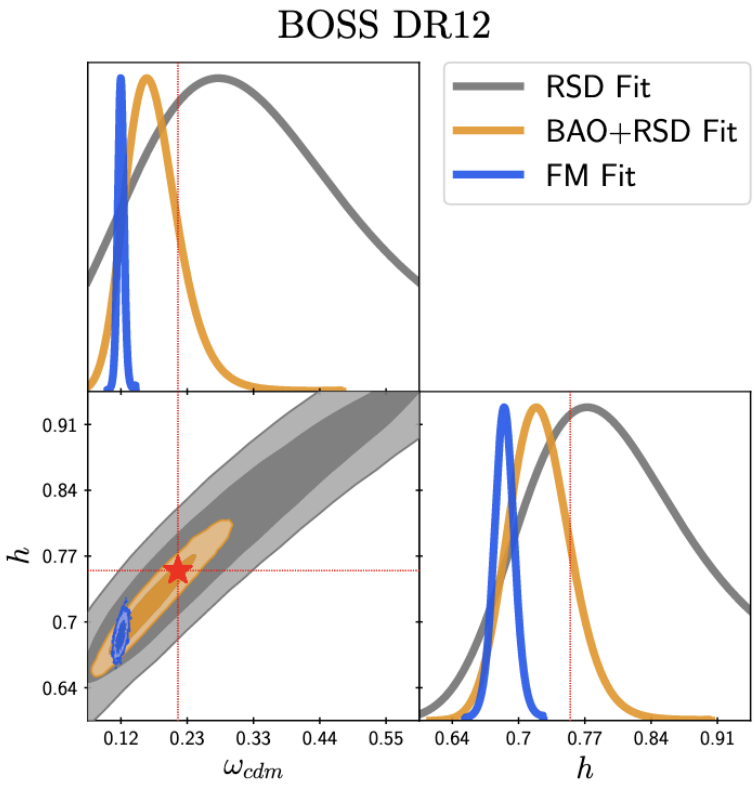

Figure 5.3: Posteriors inferred from various setups for BOSS DR12 data[22], where the compressed parameters have been converted to corresponding cosmological parameters, for 68% and 95% confidence intervals. We see that the full modelling fit considerably improves the constraints. Taken from [23].

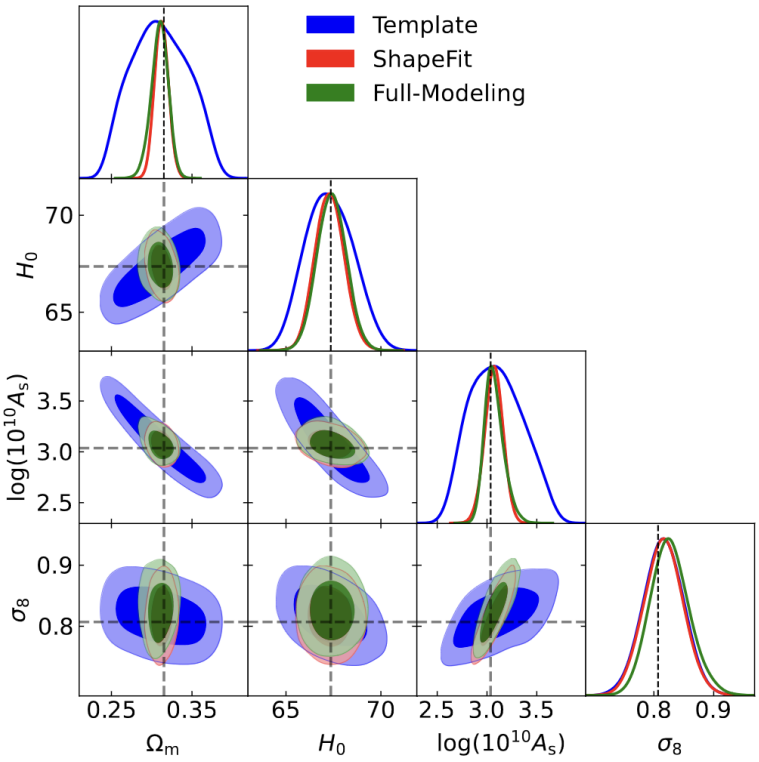

Figure 5.4: A comparison of different analysis approaches (classic RSD, ShapeFit and Full-Modelling) using AbacusSummit cubic boxes[26] in Fourier space with $k_{\max} = 0.2$ with `velocileptors` for inference of standard ΛCDM parameters. Taken from [**maus2024**].

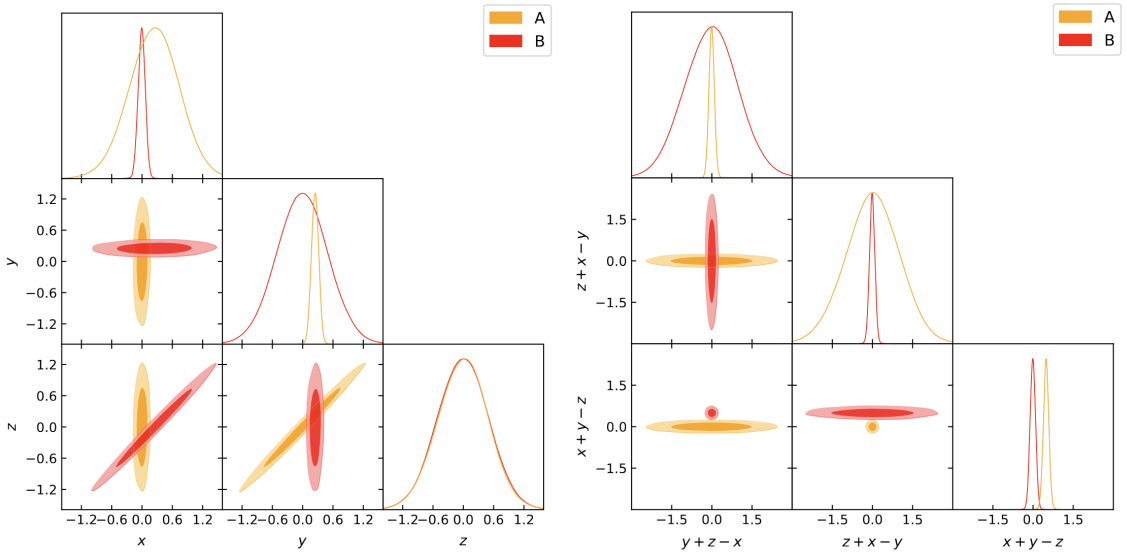

Figure 5.5: An example of the posterior from the likelihood of three parameters x, y and z. We can see that under a specific linear transformation, the posteriors fully compatible on the left panel, become completely incompatible (right panel). Taken from [27].

5.5, where two posteriors are plotted with and without a linear transformation applied, showing how two seemingly compatible distributions become fully incompatible under a linear transformation.

It was shown that the analysis of DR1 data can be susceptible to these effects (e.g. [**maus2024**]), usually called projection effects. It is also shown there that the projection effects might present themselves in the increased sensitivity of the resulting posteriors to the starting values and priors used. That is the reason why only the monopole and quadrupole will be used in the official DESI DR1 full-shape analysis, as additional degrees of freedom coming from counterterms responsible for accurate modelling of the hexadecapole increase the sensitivity on priors.

For DR1 we observe two kinds of projection effects: 1) prior weight effect, which manifests itself when the center of the prior of the nuisance parameter is shifted with respect to its true value, which can potentially shift the most likely value of the cosmological parameters and 2) prior volume effect, which arises from the marginalisation over the nuisance parameter prior volumes, thus potentially shifting the mean of the cosmological parameters' posterior away from the most likely value. Both effects depend on how constraining the dataset is. The effects are illustrated in Figure 5.6, where one can see the prior volume effect manifesting in the shifts of the maximal likelihood points with respect to the means, while the prior weight effect can be seen in the configuration-dependent shift of the maximal likelihood point.

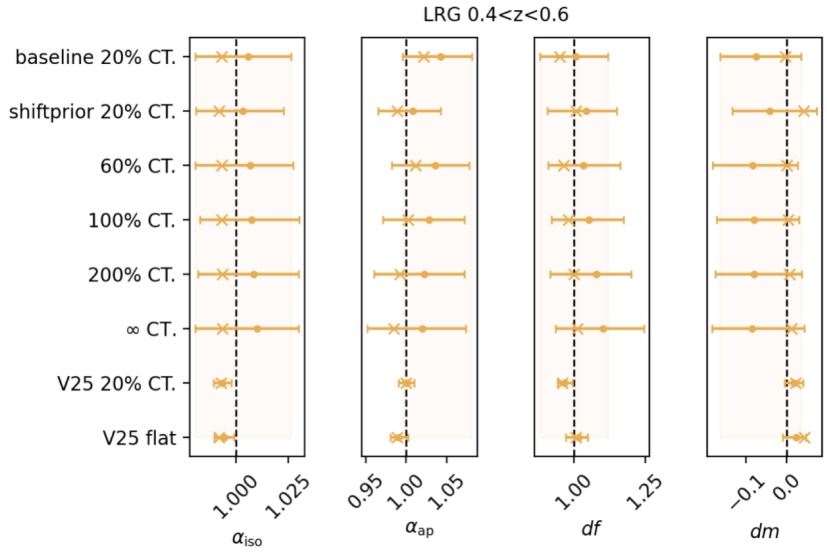

Figure 5.6: Tests on the mean of the LRG cutsky mocks with $0.4 < z < 0.6$ using different priors on the nuisance parameters, with solid lines and points representing the intervals and means obtained my MCMC sampling, while crosses represented the maximal likelihood points. The percentage corresponds to the uncertainty in the Gaussian prior of the nuisance parameter as a proportion of the nuisance parameter. V25 corresponds to the covariance matrix with the precision of 25 mocks. Credit: Ruiyang Zhao.

5.4 Tests of tools and methodology

In this section, we present the various tests performed on different sets of simulations tailored to assess the robustness of the tools being tested and their impact on the cosmological parameters of interest. We will start with testing the classic compressed full-shape pipeline we have developed, then we will make extensive tests of the neural network-based emulator for the 2-point statistics from Chapter 2. After that we will present the tests performed to validate the Fitcov approach for covariance matrix from Chapter 4 and its comparison with mock-based and analytic covariance matrices from Chapter 3. We will finish the section with the tests of the single- and multi-tracer approaches on BGS DR1 mocks.

5.4.1 General full-shape pipeline

We have started our tests with the classic compressed full-shape inference pipeline in order to constrain $\alpha_{\parallel}$, $\alpha_{\perp}$ and $f\sigma_8$. The results were published in [5], and here we briefly summarise the methodology and the results.

We fit our RSD model to the correlation function multipoles from the three datasets: SDSSbao the dataset obtained from the SDSS survey[28] with redshifts below $z = 0.3$,

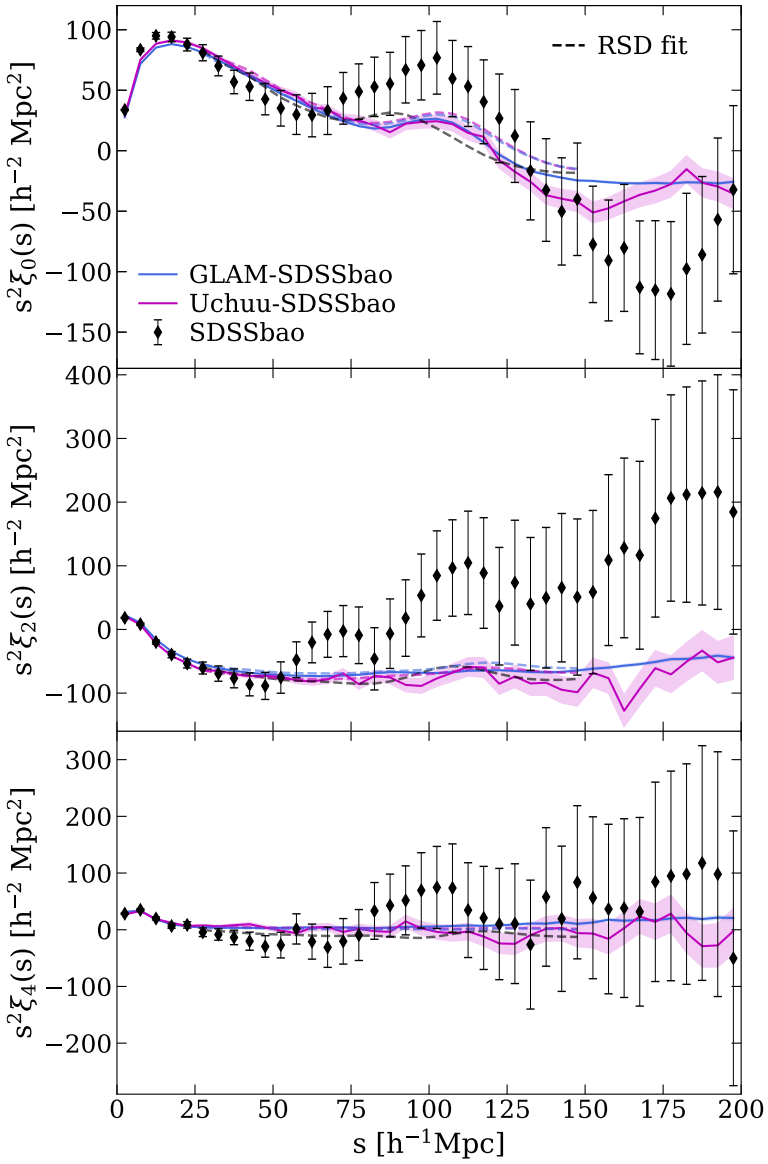

Figure 5.7: The monopole, quadrupole, and hexadecapole of the two-point correlation function measured in our three SDSS BAO data set: SDSSbao (filled symbols), GLAM-SDSSbao (blue lines) and Uchuu-SDSSbao (pink lines). Errors have been estimated from the covariance matrix of the 5100 GLAM-SDSSbao lightcones. For Uchuu (GLAM) the errors correspond to the error on the mean of the 32 (5100) mocks. Following the same colour code, we also plot the best-fit RSD model (dashed lines) for each data set.

described in more detail in [29], and two sets of simulations created to mimick this catalogue, Uchuu-SDSSbao, produced using the Uchuu simulation[30] and GLAM-SDSSbao produced with GLAM[31]. Both are described in more detail in [5]. in the separation range [25, 145] h/Mpc, with bins of width 5 h/Mpc using `velocileptors`package [24, 25], the publicly available EFT model described in Chapter 2. In addition to the cosmological parameters $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$, we also estimate the Lagrangian biases b_1, b_2, as defined in Chapter 2 and the Fingers-of-God parameter $\sigma_{\rm FOG}$. Unphysical values of parameters are avoided by setting the priors to $f\sigma_8 > 0$, $b_1 > -1$, and $\sigma_{\rm FOG} > 0$. The first Lagrangian bias is related to the Eulerian bias by $b_{1,\rm Eulerian} = 1 + b_1$. The effective redshift of the SDSS sample is computed to be $z_{\rm eff} = 0.15$, and is given by the following formula [32, 33]:

$$z_{\rm eff} = \frac{\sum_{i,j} w_i w_j (z_i + z_j)/2}{\sum_{i,j} w_i w_j} \tag{5.5}$$

, where z_i are the redshifts of the galaxies in the survey and w_i are the attributed weights responsible for targeting and observing systematic effects.

The correlation function multipoles corresponding to our best-fit models can be seen in Fig 5.7. We observe good agreement with Uchuu-SDSS and GLAM-SDSS measurements. The RSD model predicts a position of the BAO peak in the monopole that is too low compared with the observed position of the peak. However, this disagreement is within the noise limits.

The minimum of χ^2 is found using `iminuit` [34], which achieves convergence near the minimum using the first and approximate second derivatives. Errors are estimated from the region of $\Delta\chi^2 = 1$ of the marginalized χ^2 distribution, and they are allowed to be asymmetric. We also run Monte Carlo Markov chains (MCMC) with the `emcee` package [**emcee**] in order to compute the likelihood surface of our set of fitted parameters. Their convergence is checked with the Gelman-Rubin convergence test [35, 36].

We first test our RSD pipeline on the GLAM-SDSS and Uchuu-SDSS lightcones. The best-fit results of the parameters $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$ are summarised in Fig. 5.8. We confirm that the theoretical model recovers the fiducial Planck values within 1σ.

We then apply the pipeline to the SDSS sample. Our results are listed in Table 5.1. Both χ^2 minimization and MCMC sampling methods provide consistent results, and the small difference in the errors and parameter values are attributed to more accurate treatment of the posteriors with MCMC chains. The parameter distributions for $\sigma_{\rm FOG}$ are non-Gaussian, as it is restricted to positive values, but the best-fit value is consistent with 0. Additionally, we compare our obtained $f\sigma_8$ and b_1 with Howlett et al. [17], finding a good agreement, but we obtain a $\gtrsim 30\%$ increase in precision on $f\sigma_8$ which can be attributed to our better estimate of the covariance matrix (see Table 5.1).

The distribution of parameter values and their uncertainties can be seen in Fig. 5.8,

Parameter	χ^2 minimization	MCMC	Reference
$f\sigma_8$	$0.65^{+0.14}_{-0.15}$	$0.60^{+0.15}_{-0.16}$	$0.63^{+0.24}_{-0.27}$
$\alpha_\parallel$	$1.00^{+0.08}_{-0.06}$	$1.04^{+0.14}_{-0.09}$	N/A
$\alpha_\perp$	$1.18^{+0.07}_{-0.09}$	$1.16^{+0.08}_{-0.09}$	N/A
$b_{1,\mathrm{Eulerian}}$	$1.53^{+0.19}_{-0.19}$	$1.59^{+0.22}_{-0.20}$	$1.36^{+0.29}_{-0.26}$
b_2	$-1.8^{+2.7}_{-1.1}$	$-0.8^{+1.7}_{-1.5}$	N/A
σ_{FOG}[h/Mpc]	$0^{+7.4}_{-0.0}$	$4.0^{+2.8}_{-2.7}$	N/A

Table 5.1: RSD fitted parameters from the SDSS data, obtained by χ^2 minimization and using Bayesian MCMC inference. The first two columns correspond to the results from this paper, while the last column shows the values from Howlett et al. [17] for comparison. Only $f\sigma_8$ and $b_{1,\mathrm{Eulerian}}$ estimated values from Howlett et al. [17] are found in the literature.

Parameter	Uchuu	GLAM	Expected value
$f\sigma_8$	$0.44^{+0.02}_{-0.02}$	$0.420^{+0.002}_{-0.002}$	0.46
$\alpha_\parallel$	$0.98^{+0.03}_{-0.03}$	$0.991^{+0.002}_{-0.002}$	1.00
$\alpha_\perp$	$0.98^{+0.02}_{-0.01}$	$0.985^{+0.001}_{-0.001}$	1.00
$b_{1,\mathrm{Eulerian}}$	$1.44^{+0.03}_{-0.03}$	$1.430^{+0.003}_{-0.003}$	N/A
b_2	$-0.8^{+0.2}_{-0.2}$	$-0.641^{+0.03}_{-0.03}$	N/A
σ_{FOG}[h/Mpc]	$0.0^{+1.5}_{-0.0}$	$3.73^{+0.08}_{-0.08}$	N/A

Table 5.2: RSD fitted cosmological parameters from the means of Uchuu and GLAM correlation functions, obtained using χ^2 minimization in comparison with the values predicted by the fiducial cosmology.

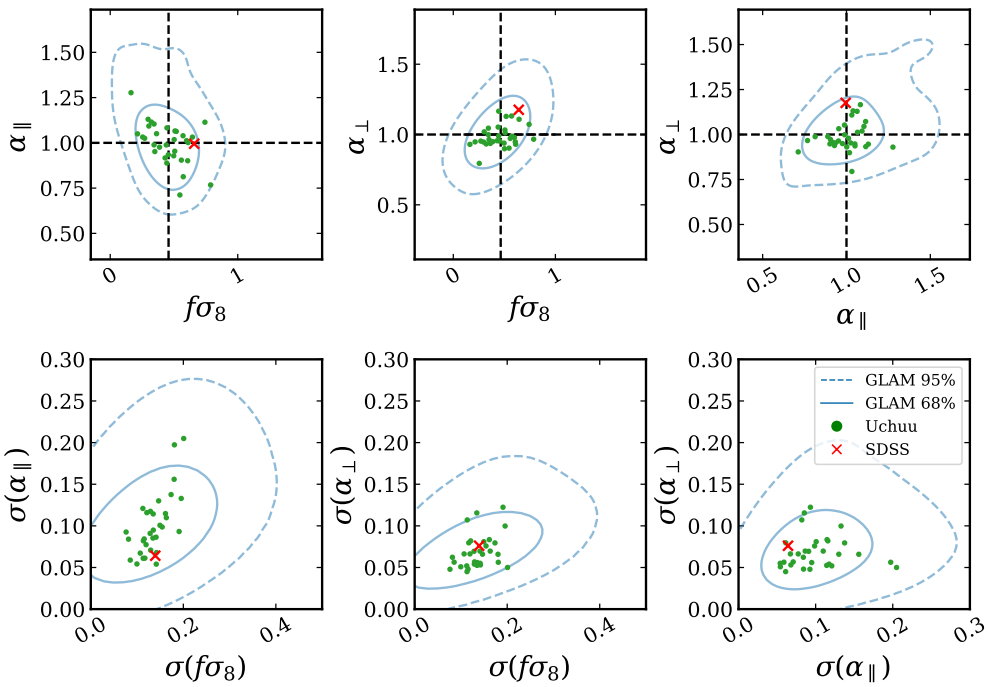

Figure 5.8: *Top row*: Parameter values of $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$ measured from the GLAM-SDSS (blue contours) and Uchuu-SDSS (green points) lightcones, together with SDSS data (red cross). Dashed black lines represent the expected values for the fiducial cosmology. *Bottom row*: The corresponding parameter errors.

where the results from the SDSS sample together with those from GLAM and Uchuu are shown in blue, orange, and red respectively, and the black lines show the expected values of the parameters for the Planck fiducial cosmology. The results from GLAM- and Uchuu-SDSS are consistent with each other and with the SDSS data, meaning that the mocks are a fair statistical representation of the data. We confirm that the fits we have performed are valid and have an acceptable χ^2/dof, for GLAM mocks being $\chi^2/\mathrm{dof} = 1.00 \pm 0.18$, for Uchuu mocks being $\chi^2/\mathrm{dof} = 1.07 \pm 0.18$ and for the SDSS sample being $\chi^2/\mathrm{dof} = 0.98$. It should be noted, however, that due to a very high number of GLAM mocks, the results from the fits of the mean were expected to be slightly biased due to the precision becoming larger than the systematic uncertainty from our model, which is designed for the analysis of only one realisation.

Finally we use our best-fit values of the BAO-inferred $\alpha_{\parallel}$ and $\alpha_{\perp}$ parameters to provide a measurement of the Hubble distance and the (comoving) angular diameter distance at the effective redshift of the SDSSbao sample, i.e. $D_{\mathrm{H}}(z = 0.15)/r_d = 27.9^{+2.2}_{-1.8}$, $D_{\mathrm{M}}(z = 0.15)/r_d = 5.1^{+0.3}_{-0.4}$. These results are valuable since they constitute the first two-dimensional distance measurements from SDSS data.

5.4.2 Theoretical modelling

We assess the performance of the neural network emulator in inferring the cosmological parameters with respect to the original code by fitting the mean of the 25 Abacus mocks for the three datasets: DESI LRG-like at $z = 0.5$, DESI LRG-like at $z = 0.8$ and DESI ELG-like at $z = 0.8$. Fig. 5.9 shows the cosmological constraints obtained from full modelling using either the neural network emulator (red) or `MomentExpansion` module of `velocileptors` (green). The dashed curves represent the 1σ error from each model. Both models yield very consistent results, both for the best-fit values and the uncertainty.

We summarise the results of the comparison between the neural network emulator (triangles) and the analytic code (circles) in Fig. 5.10 where we show the best-fit values and 1σ error for the three main cosmological parameters that are well constrained with Full-Shape analysis. We can see that both methods yield very consistent and similar results with less than 0.3σ for the largest difference seen on h.

We also want to assess the performance of the emulator with respect to the expected values given the cosmological model that was used for the simulations. In order to do that, we fit the 25 individual realisations for each case and in Fig. 5.11, we show the results for the cosmological parameters by plotting the difference between measured and truth divided by the error on the measured parameter as a function of mock number. Blue triangles represent the results for LRG at $z = 0.5$, green ones for LRG $z = 0.8$ and orange ones for ELG at $z = 0.8$. All the results are consistent with the expected truth values within 1-2 σ, which further validates the ability of the emulator to recover precise and

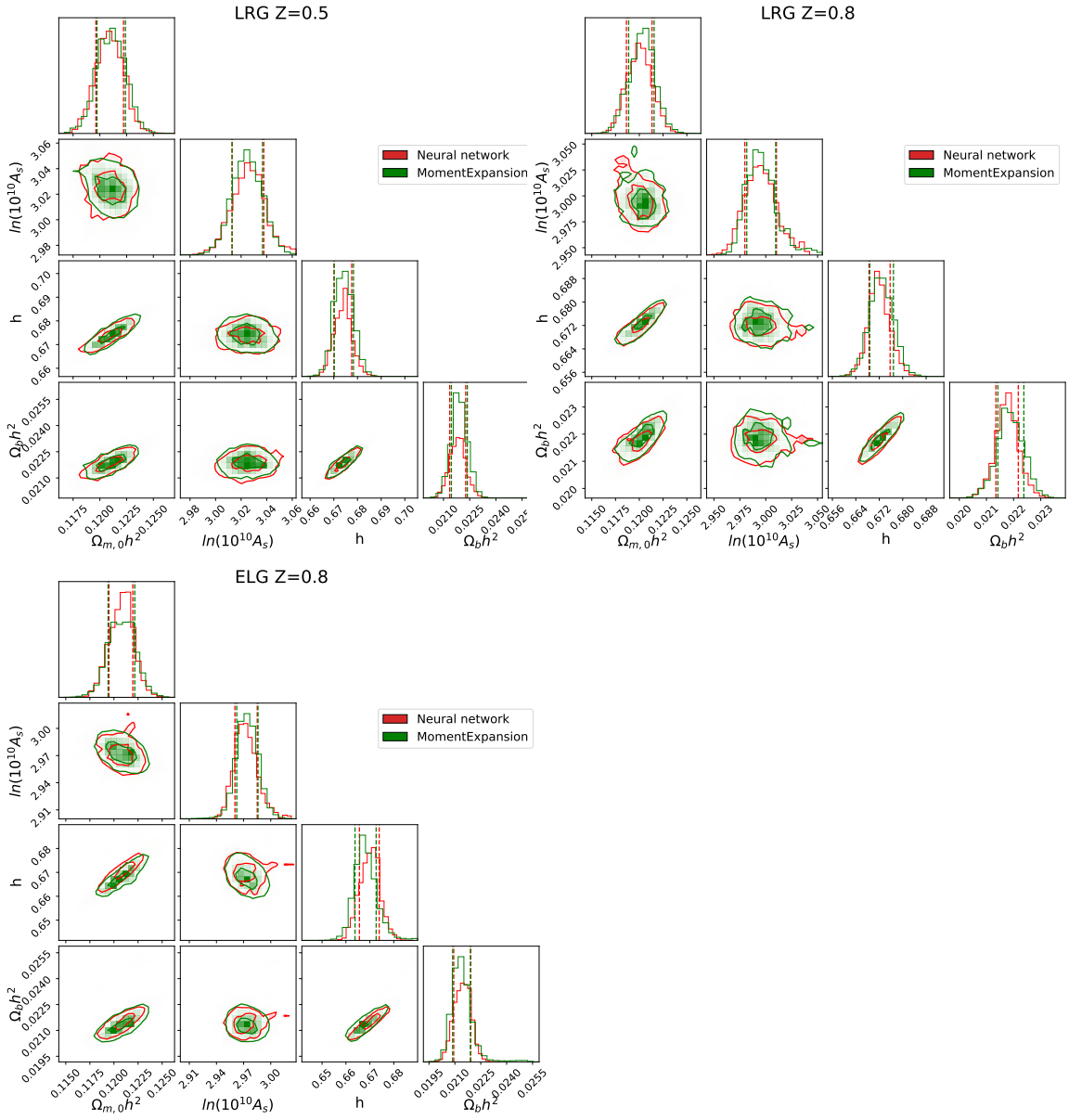

Figure 5.9: Comparison of cosmological constraints obtained with the neural network emulator (red) and with `MomentExpannsion` (green) when fitting the mean of the 25 Abacus mocks for the three configurations.

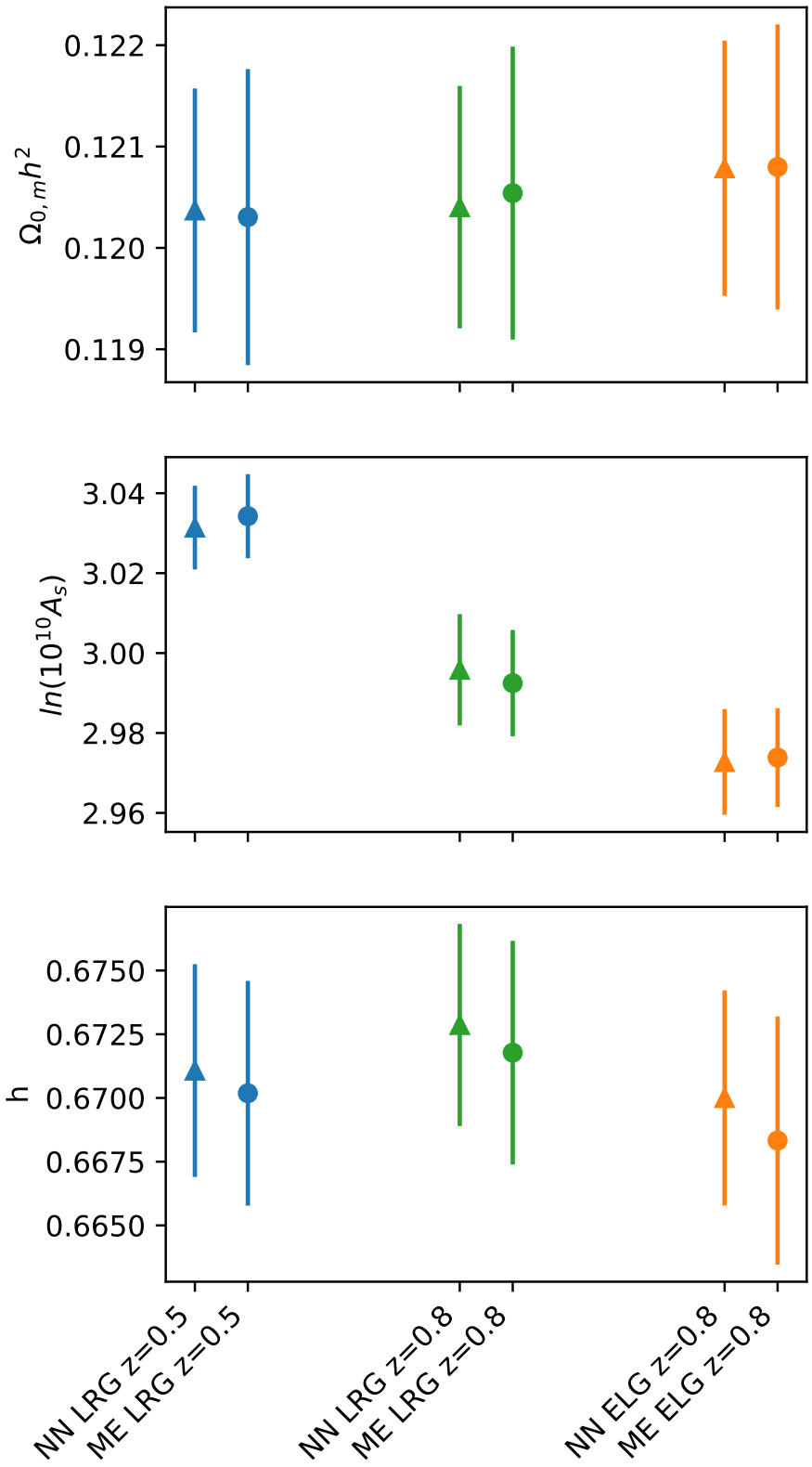

Figure 5.10: The cosmological parameters obtained from Full-Modelling fits with the neural network emulator and with the original code obtained from the mean of different mock types with rescaled covariance matrix.

$\bar{n}(z)(h^3\mathrm{Mpc}^{-3})$	Mock	Mohammad-Percival	Fit
2×10^{-4}	1.03	1.40	1.05
5×10^{-4}	0.99	1.42	1.05
15×10^{-4}	1.00	1.56	1.08

Table 5.3: For each of the estimation methods we tabulate the standard deviation σ of $(f\sigma_{8i} - \overline{f\sigma_8})/\sigma_i(f\sigma_8)$, over independent fits, i. For the mock covariance method $\sigma \approx 1$ (as expected when all the fits are performed consistently with the same covariance), for the fitted covariance method it is also quite close to unity, but for the jackknife method $\sigma > 1.4$, which shows a much higher degree of deviation from the truth.

unbiased cosmological constraints.

5.4.3 FitCov

To test the performance of different covariance estimation techniques we infer $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$ by fitting the theoretical predictions of the multipoles to the ones from the mocks using covariances from the estimators reviewed earlier. The theoretical power spectrum $P_{\rm FOG}$ is obtained using the `MomentumExpansion` module of the `velocileptors` package [for more details, see 24, 25], which is generated using a 6-parameter model and we follow the classic full-shape approach by fitting the linear growth rate $f\sigma_8$ and the Alcock-Paczynski parameters [16] $\alpha_\parallel$ and $\alpha_\perp$. The model also includes first- and second-order biases b_1, b_2 and the effective Fingers Of God parameter (FOG) $\sigma_{\rm FOG}$.

We use a likelihood maximisation method to find the χ^2 minima using `iminuit` [34]. Errors are estimated from the region of $\Delta\chi^2 = 1$ of the marginalized χ^2 distribution, and they are allowed to be asymmetric. The choice of a frequentist method of analysis is motivated by its low computational cost.

In Fig. 5.12, the first row shows the distributions of reduced χ^2 for different choices of $\bar{n}$, and the other rows show the marginalised 2D-distributions of parameters and their uncertainties for respectively $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$. The distributions of reduced χ^2 show the goodness of the individual fits for the three methods. The contours in the bottom panel show how, for all the parameters, the spread from the Mohammad-Percival jackknife in green is in general much wider than the one from the mock covariance in red both in terms of uncertainty and parameter values. While in case of the fitted jackknife covariance, the blue contours are very similar to the mock covariance ones. Presumably, this improvement comes from using 50 realisations rather than one. In Fig. 5.13 we also show in the same form the performance from the standard jackknife in comparison with the Mohammad-Percival corrected jackknife and mock-based covariance. As expected,

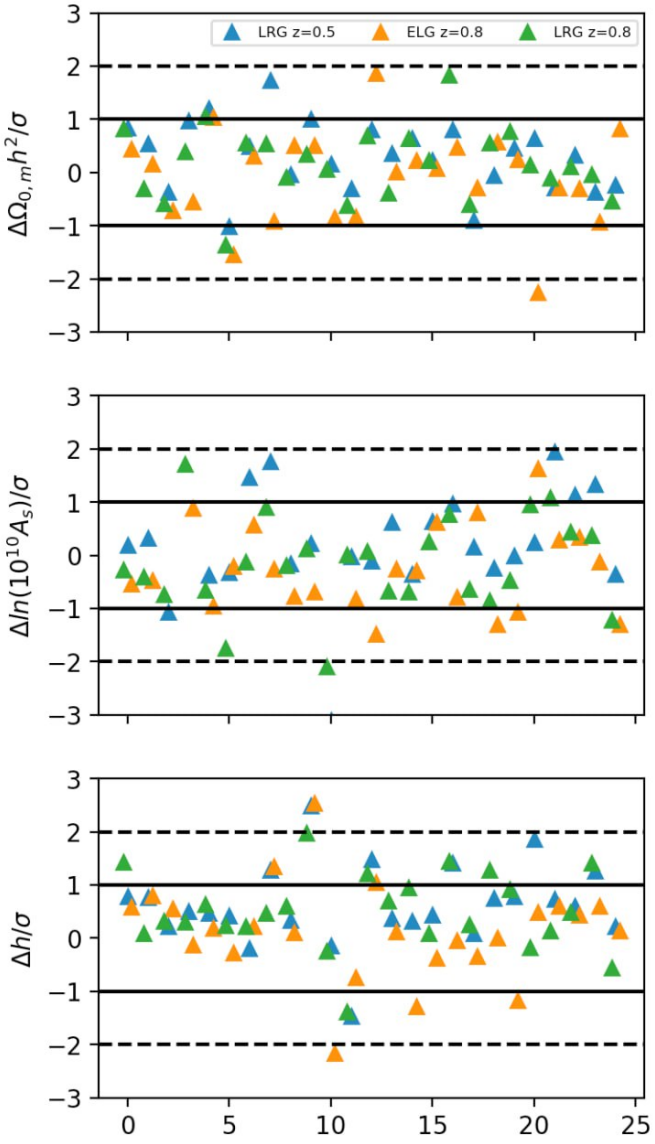

Figure 5.11: The deviation of different cosmological parameters in terms of the error σ from the expected theoretical value obtained from the individual mock fits for three mock types: blue for LRGs with $z = 0.5$, green for LRGs with $z = 0.8$ and orange for ELGs with $z = 0.8$.

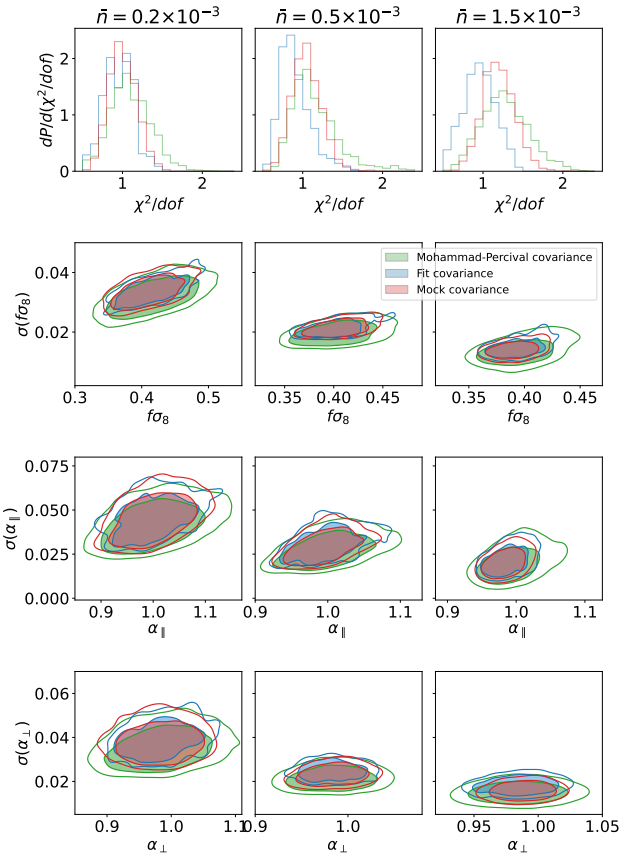

Figure 5.12: The summary of the results from the cosmological fits from the lognormal mocks with varying density (one for each column and with density in $(\mathrm{Mpc/h})^{-3}$ indicated at the top) for the three covariance matrix estimation methods: jackknife covariance with Mohammad-Percival correction in green, fitted jackknife covariance in blue and mock covariance in red. The top panels show the histograms of the reduced χ^2, while the three bottom ones show the marginalised 2D-distributions of parameters and their uncertainties for $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$, obtained from the set of fits.

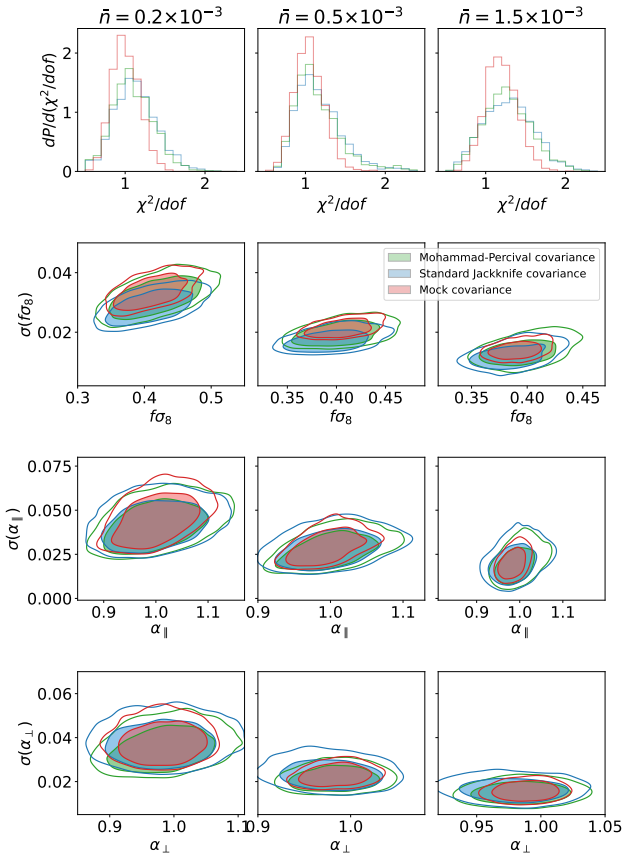

Figure 5.13: The summary of the cosmological fits from the lognormal mocks with a varying density. Similar to Fig. 5.12 but withdifferent methods of estimating the covariance matrix: jackknife covariance with the Mohammad-Percival correction in green, mock covariance in red (the same contours as on Fig. 5.12) and standard jackknife covariance in blue.

the standard jackknife produces slightly larger contours, which are noticeably shifted with respect to the mock covariance, especially for $f\sigma_8$.

To additionally test the validity of our inference approaches, we will define the quantity

$$x = \frac{\eta - \bar{\eta}}{\sigma(\eta)}, \tag{5.6}$$

where η is an inferred parameter from a specific fit, $\bar{\eta}$ is the mean from all the fits, and $\sigma(\eta)$ is the error estimation from a specific fit. The distribution of quantity x is called a pull distribution. If η follows a Gaussian distribution, the distribution of x will form a normal distribution with $\bar{x} = 0$ and $\sigma(x) = 1$.

For the mock covariance, we fit the 1500 available samples, while for the Mohammad-Percival jackknife and for the fitted jackknife 50 random mocks are fitted using 30 realisations of the covariance, under the assumption that all of the covariance estimators are probing the same underlying likelihood.

Pull distributions for $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$ are presented in Fig. 5.14 for each number density of the lognormal mocks. The fitted jackknife and mock covariance pull distributions have Gaussian-shape with $\sigma = 1$ normal distributions as expected, while the pull distributions obtained when using Mohammad-Percival jackknife are slightly wider, which is due the covariance being less precise. We can see it quantitatively in the Table 5.3, where the standard deviations of the distributions from Fig. 5.14 are presented. That is due to various shifts of the distributions obtained from fitting to different jackknife covariances. This is not the case for the fitted approach, however.

Overall, for all the number densities, the performance of the fitted jackknife method using 50 mocks is much better than that of the standard jackknife with the Mohammad-Percival correction, and, most importantly, it gives similar performance as the covariance matrix created from 1500 mocks.

We also test the performance of the approach when varying the number of mocks used for producing the fitted covariance. We test using 10, 25 and 50 mocks and report the results on the cosmological fits in Fig. 5.15, following the same methodology as explained before for 50 mocks. The precision on the marginalised 2D contours of the cosmological parameters of interest starts to drop noticeably when 10 mocks are used, while it remains stable between 25 and 50 mocks.

For LRG- and ELG-like EZMocks we also infer the values of the cosmological parameters $f\sigma_8$, $\alpha_\parallel$ and $\alpha_\perp$, using the same methodology as for the lognormal mocks. The results of the fits are shown in Fig. 5.16 where the first row shows the χ^2/dof distribution and the other rows show the marginalised 2D contours for best-fit values and uncertainties on the cosmological parameters. We confirm the findings with the lognormal mocks that the fitted jackknife method provides results which are in much better agreement with the mock-based method while the jackknife method with the Mohammad-Percival correction clearly over-estimates the uncertainties on all the cosmological parameters. The effect is

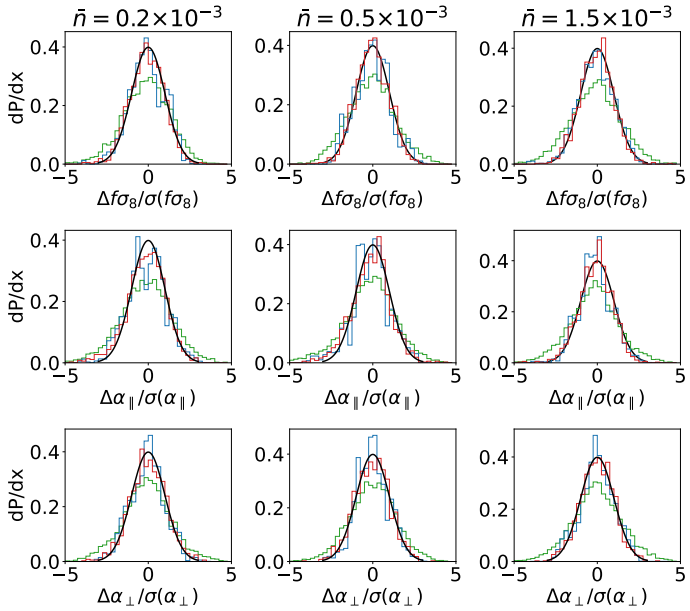

Figure 5.14: Pull distributions for different covariance estimation techniques with results from fits on various lognormal mocks, shown for 3 different number densities indicated at the top in $(\mathrm{Mpc/h})^3$. Line colors follow those in Fig. 5.12

also stronger as the number density of the galaxy sample increases. Moreover, as we have fewer mocks than for the tests with the lognormal mocks, we can notice that the fitted covariance based on 50 mocks actually produces smaller contours overall than the mock covariance which uses 1000 EZmocks.

In Fig. 5.17, we show the pull distributions as defined by eq. 5.6 for various cosmological parameters. The standard deviations of the $f\sigma_8$ pull distributions are presented on Table 5.4. The results are also similar to the ones obtained with the lognormal mocks: both the fitted jackknife and mock covariances produce a Gaussian shape with $\sigma = 1$, while the standard deviation of the pull distribution obtained using the Mohammad-Percival correction for the jackknife method is larger (σ=1.5, 1.8 for LRG and ELG respectively).

Survey	Mock	Mohammad-Percival	Fit
LRG	1.04	1.49	1.07
ELG	1.08	1.80	1.07

Table 5.4: Standard deviation σ of $(f\sigma_{8,i} - \overline{f\sigma_8})/\sigma_i(f\sigma_8)$, where i is a separate fit for each of the methods. We can see, that for the mock covariance, it is close to 1 (as it is supposed to be when all of the fits share the same covariance.), for fitted covariance it is quite close to 1, but for the jackknife estimate it usually takes values > 1.4, which shows a much higher degree of deviation from what we assume to be the truth.

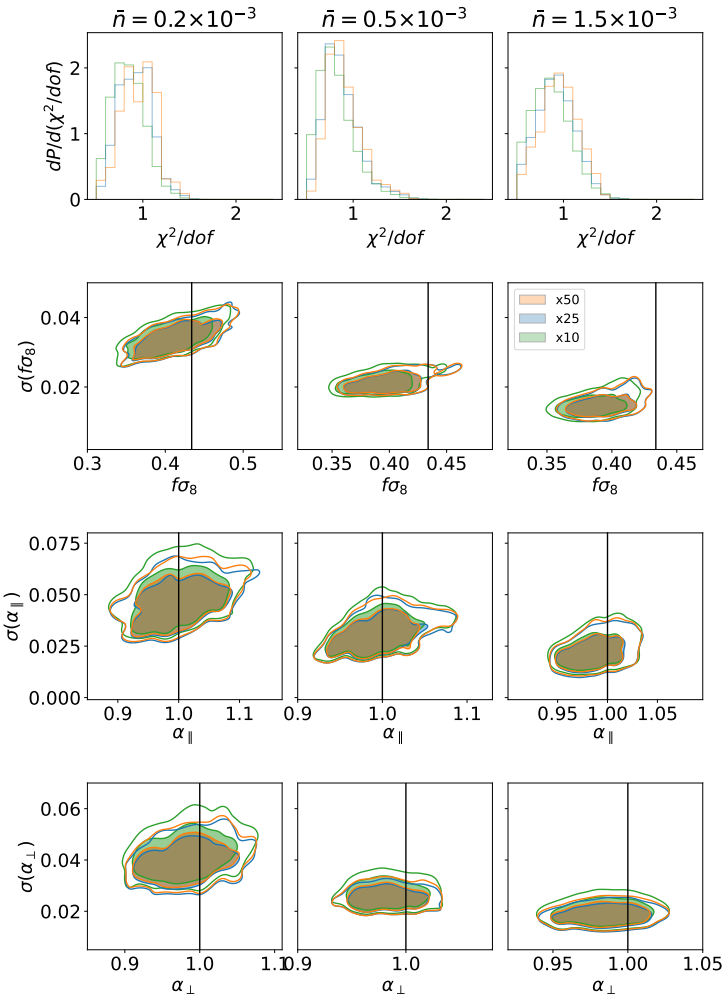

Figure 5.15: The summary of the cosmological fits when using different numbers of mocks to obtain the fitted jackknife covariance: the default number of 50 mocks in red, 25 mocks in blue and 10 mocks in green. The figure is organised like Fig. 5.12.

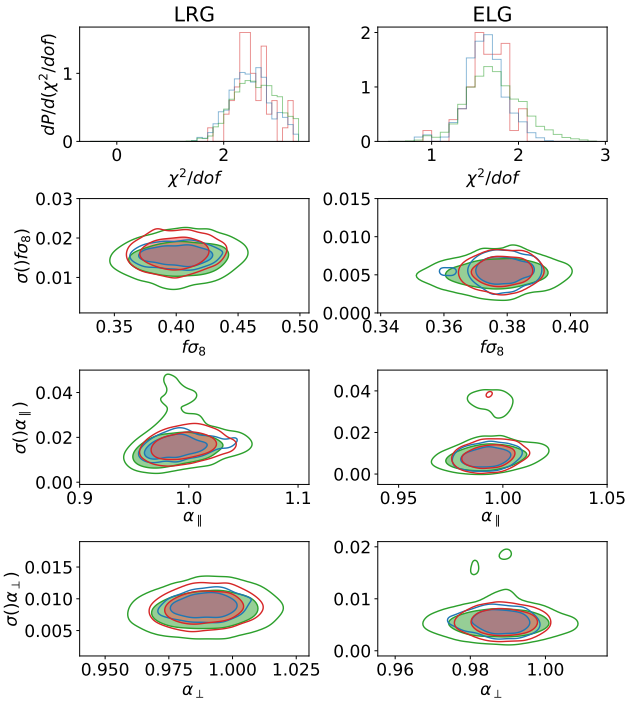

Figure 5.16: The summary of the cosmological fits for the EZ mocks for LRGs and ELGs (left and right column respectively), similar to layout of Fig. 5.12.

This quantitative test thus demonstrates that the fitted jackknife method performs better in estimating an unbiased and accurate covariance matrix for the two-point correlation function.

Overall, throughout all of the tests for varying number densities, different types of mocks and number of fitted mocks, the fitted jackknife approach shows a considerable improvement over the correction for standard jackknife proposed by [37]. The fitted jackknife approach can achieve an unbiased estimate of the covariance matrix with similar precision to a mock-based covariance but with the major advantage of requiring a much smaller number of mocks.

5.4.4 Covariances for DR1 standard analysis

Most of the mocks created in the context of this PhD work aim at modelling the covariance. As FitCov is only applicable to the configuration space, and the official DESI DR1 Full-Shape analysis is performed in Fourier space, mock-based covariance estimates remain a necessity.

I have participated in the production of the EZmocks, and developed a pipeline for the production of the GLAM mocks, thus these two varieties will be investigated. More details can be found in Chapter 3. Here we present the final versions of the covariances for DR1 analysis in Figure 5.18, for analytic, EZmock and GLAM covariances, where

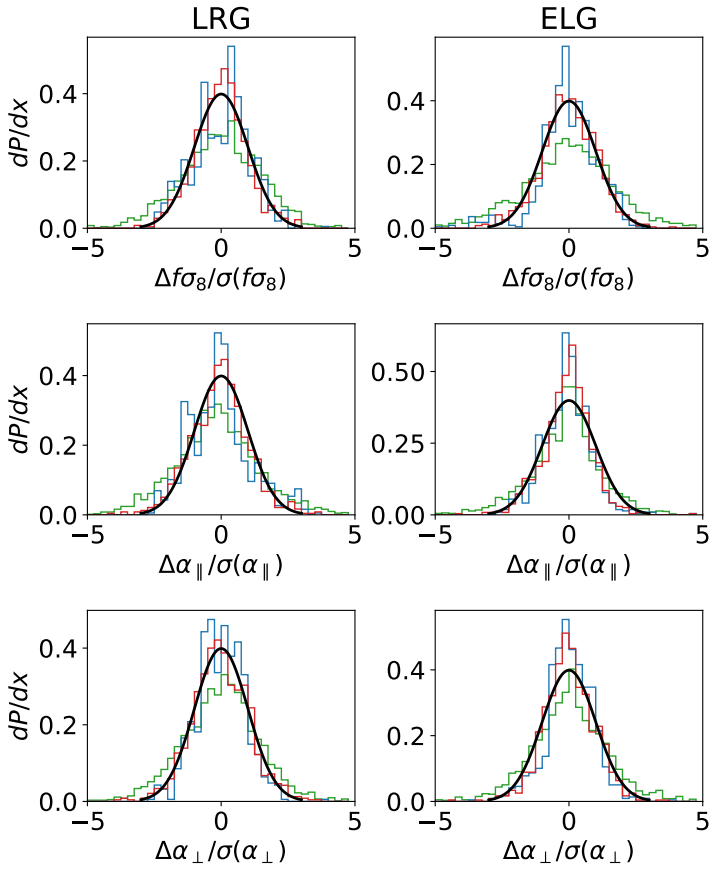

Figure 5.17: Pull distributions for different covariance estimation techniques with results from fits on LRG and ELG mocks with line colours as in Fig. 5.14

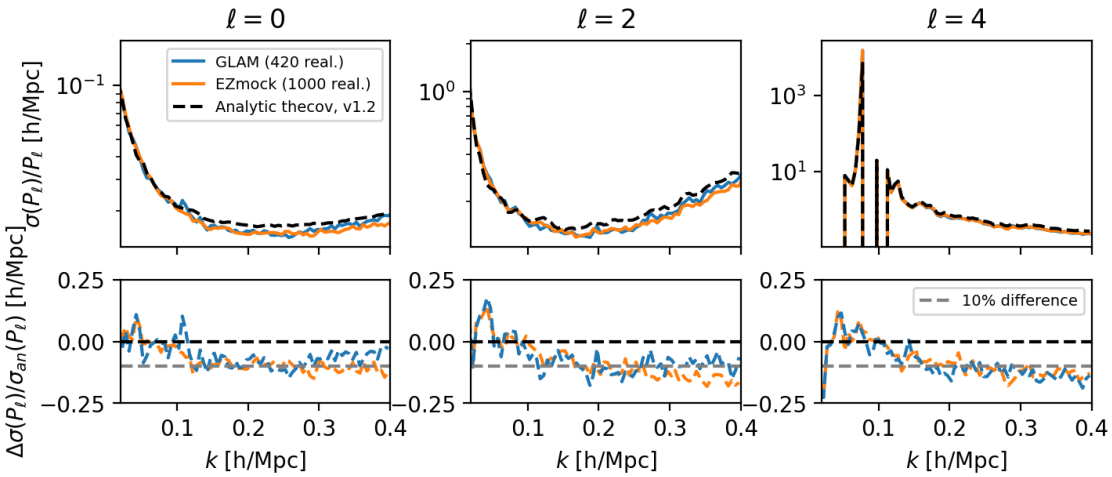

Figure 5.18: *Upper:* Ratio of standard deviation obtained by various methods(GLAM mocks, EZmocks, analytic covariance `thecov`[38–40]) to the power spectrum of the corresponding multipoles averaged over 25 Abacus realisations plotted together. *Lower:* Difference of standard devsiation obtained from mocks with the analytic covariance divided by the analytic standard deviation. We can see that the analytic covariance overestimates the variance by $\sim 10\%$.

the top panels show the signal to noise ratio, while the lower one is a deviation of mock covariances from the analytic one, for $\ell = 0, 2, 4$. The analytic covariance was created using `thecov`[38–40], and was calibrated on the DR1 sample.

In order to test them for BGS DR1 standard analysis, we have performed two tests. In the first one, we fit the 25 Abacus mocks to three covariance matrices in our possession. The results of the comparison are seen in Figures 5.19 and Figure 5.20, where each point on the figure has as coordinates values and errors from fits with different covariances. It should be noted, that GLAM mocks are still in production as of writing this thesis, thus the results here can be considered as preliminary. The fits are performed using MCMC sampling with `emcee`[41]. The parameter df is defined as $df = f_{\text{best-fit}}/f_{\text{fid}}$. We can see on these plots that the results of GLAM and EZmock are very consistent with each other, however analytic covariance tends to slightly overestimate the errorbars for $q_\perp$.

Then, we perform the same test but using 500 GLAM mocks and using likelihood minimisation with `iminuit`[34] instead of Monte-Carlo sampling. The results are presented in Figures 5.21 and 5.22. The results are consistent with those obtained when fitting the Abacus mocks. Meanwhile GLAM and EZmock covariances for both tests are fully consistent.

To additionally test the performance of Fitcov with respect to the traditional mock-based covariance (EZmock), we perform the fit on the mean of 25 Abacus BGS DR1 mocks with $M_r < -21.5$ in configuration space for the ShapeFit approach. The results on the compressed parameters showing the consistency between two methods are presented

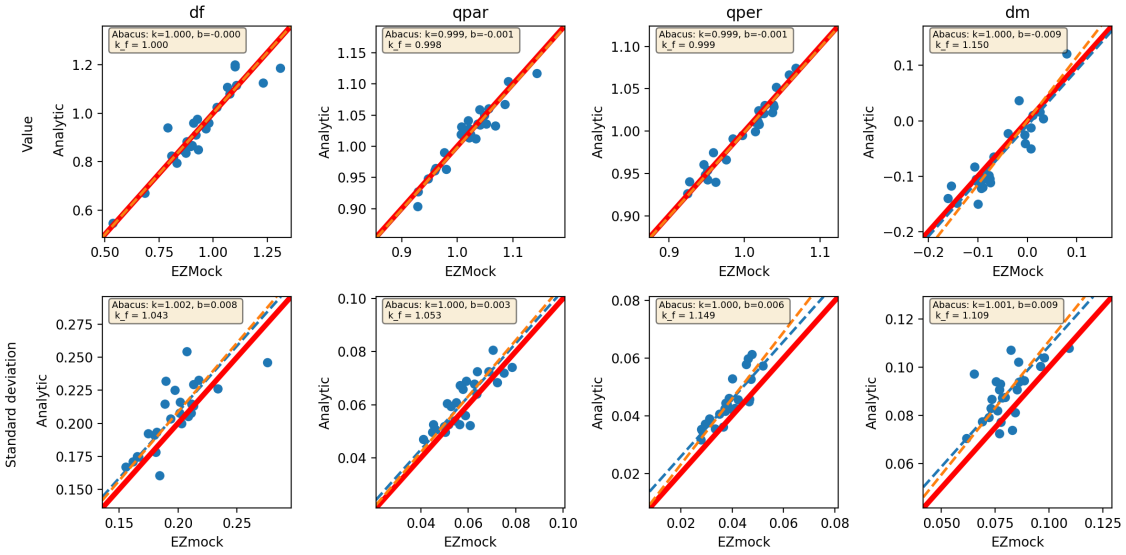

Figure 5.19: Comparison of the fits performed with analytic and EZmock covariance matrices on 25 Abacus DR1 realisations. On the upper panel the best-fit values of the parameters are shown, where the y-axis corresponds to values obtained using analytic covariance, while on x-axis the same values but obtained with EZmock covariance. The lower panel compares the errors from the fits. On each of the plots the thick red line corresponds to the line $y = x$, blue line is $y = kx + b$ and orange line is $y = k_f x$, where k, b and k_f are obtained via least-squares method. These values are presented in yellow boxes on each plot.

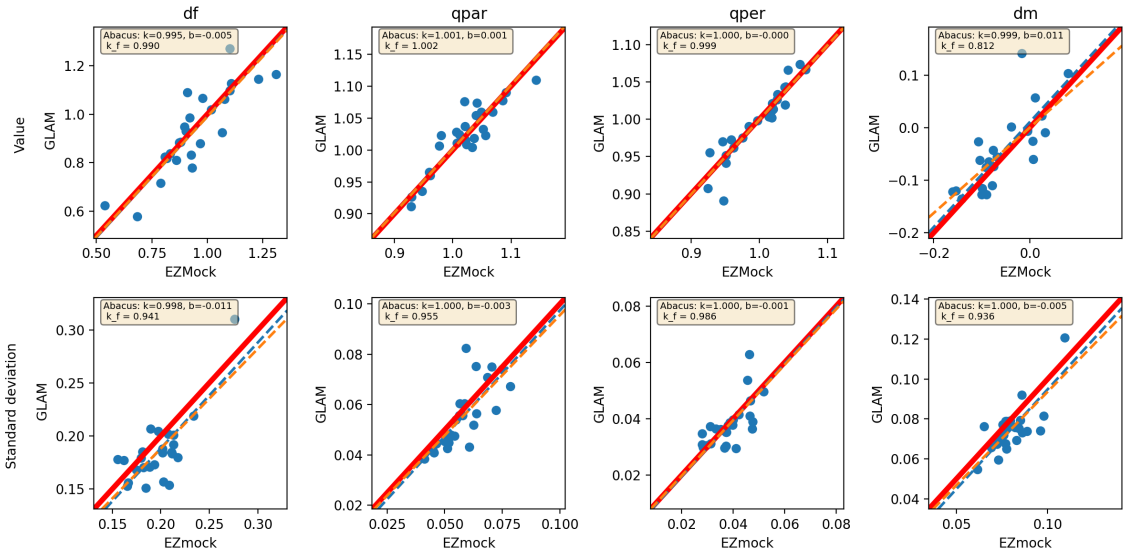

Figure 5.20: Same as in Figure5.19, but for the GLAM mock covariance instead of analytic.

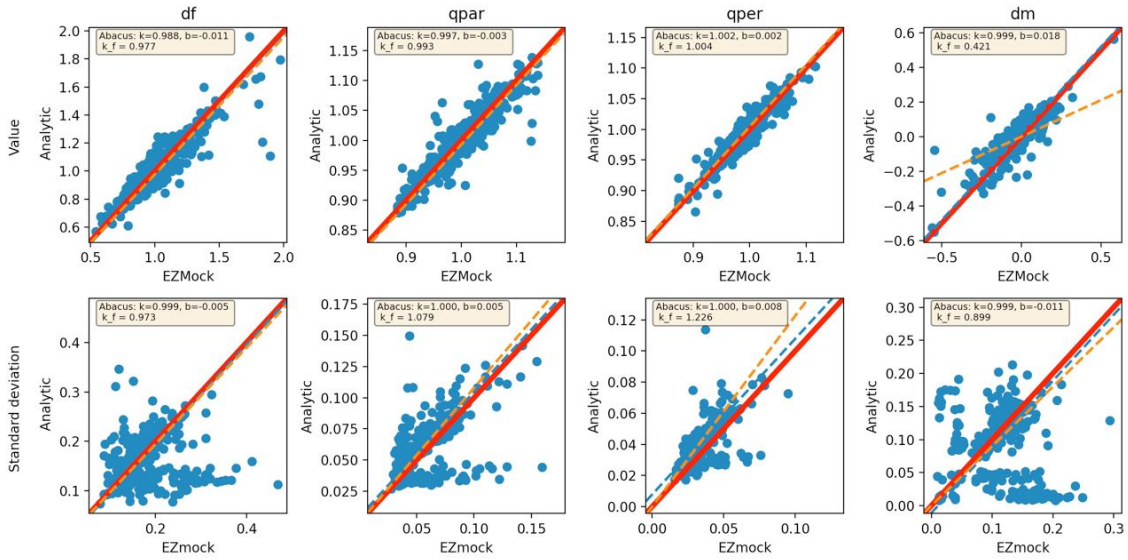

Figure 5.21: Same as in Figure5.19, but instead of 25 Abacus mocks, 500 GLAM mocks are being fitted, where the inference is done by likelihood minimization with `iminuit`[34].

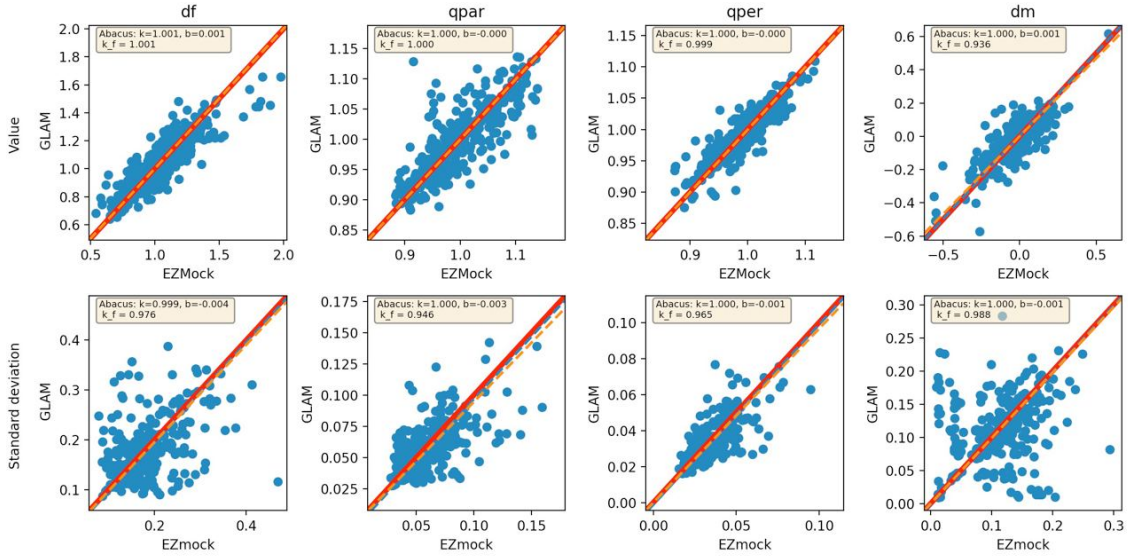

Figure 5.22: Same as in Figure5.21, but for the GLAM mock covariance instead of analytic.

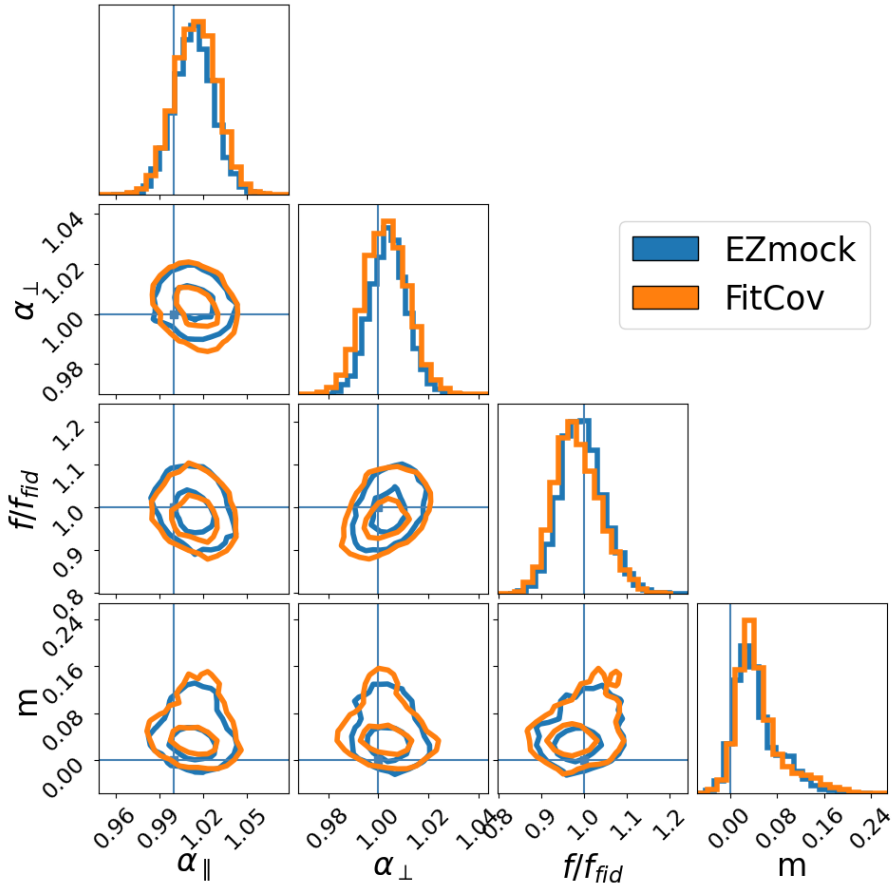

Figure 5.23: The compressed parameters obtained from the mean of 25 Abacus BGS DR1 mocks with $M_r < -21.5$ with covariance matrix scaled accordingly.

in Figure 5.23.

We also perform the same test but using our full-modelling pipeline, and the results ar also very consistent between using FitCov and the traditional EZmock covariance. The results are presented in Figure 5.24.

5.4.5 BGS Bright single tracer tests

Having everything ready for the offical DESI DR1 analysis of BGS with $M_r < -21.5$, we can attempt to use the techniques developed earlier to tackle the full DESI BGS Bright sample. We recall that the main reason for not analysing the full BGS Bright sample is the lack of enough simulations to build a mock-based covariance. Therefore, we use the Fitcov method to obtain a covariance matrix that we calibrated on the 25 Abacus mocks. We then fit the mean of 25 Abacus mocks using FitCov.

We will now use the NN emulator model described in Chapter 2 in order to perform a Full Modelling analysis. We infer 4 cosmological parameters ($\omega_{cdm}, h, \log 10^{10} A_s, \omega_b$), with ω_b restricted by a BBN prior[42], as well as well as additional nuisance parameters with second-order bias and corresponding counter-terms as well as a FOG parameter:

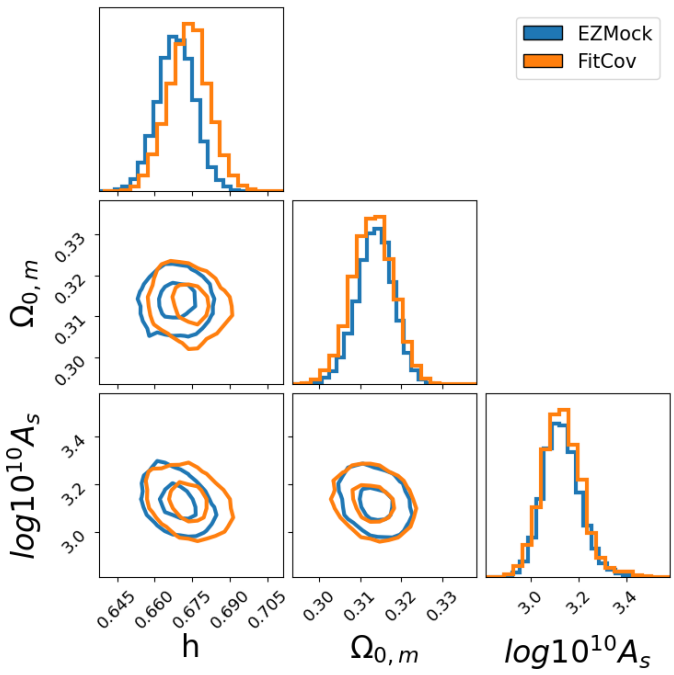

Figure 5.24: The cosmological parameters obtained from the mean of 25 Abacus BGS DR1 mocks with $M_r < -21.5$ with covariance matrix scaled accordingly with full-modelling

$(b_1, b_2, \alpha_0, \alpha_2, \sigma_v)$. We use the monopole and quadrupole in our setup in the separation range of $[26, 150]$Mpc/h. The priors for these parameters are described in table 5.5 (note that some of them are defined by the ranges of validity of our emulator).

We can see in Figure 5.25, where the averaged over 25 realisations multipoles from Abacus mocks have been sucessfuly modelled by the emulator, with almost all of the residuals between the modelled and the measured ones remaining within 1σ uncertainty on the mean.

As seen in Figure 5.26, our pipeline manages to yield the correct values of the cosmological parameters, with full sample being slightly less biased than the one with $M_r < -21.5$. The results between the two also seem to be much more compatible than the ones obtained using ShapeFit.

5.4.6 BGS Bright Multitracer tests

As for the single-tracer analysis, we test the multi-tracer pipeline using the 25 Abacus mocks. If we look at the rest frame $g - r$ colour distribution in DESI BGS DR1 shown in Figure 5.27, we see two peaks that correspond to two separable populations of galaxie. The vertical line indicates the separation between red and blue galaxies that we define for the multi-tracer analysis.

In order to check whether the clustering of those two populations of galaxies is indeed different, we compute the corresponding correlation functions with weights from

Table 5.5: Parameters used for the full-modelling inference and their priors

Parameter	Range/Prior
$\omega_{\rm cdm}$	[0.10, 0.2]
$\omega_{\rm b}$	$\mathcal{N}(0.02237, 0.00037)$
$\log\left(10^{10}A_s\right)$	[2.5, 3.5]
h	[0.6, 0.8]
b_1	[-1, 5]
b_2	$\mathcal{N}(0, 5)$
b_s	$\mathcal{N}(0, 10)$
α_0	$\mathcal{N}(0, 30)$
α_2	$\mathcal{N}(0, 50)$
σ_v	[0,50000]

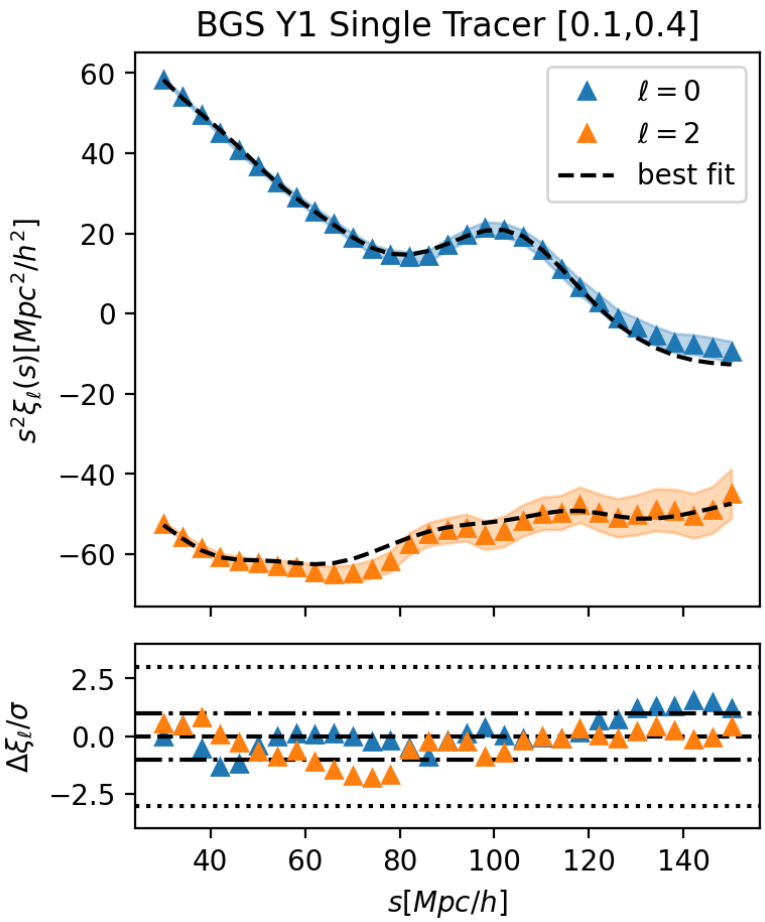

Figure 5.25: *Upper*: Averaged over 25 realisations correlation function multipoles with shaded values representing the error on the mean, and the dashed line being the modelled multipoles for the best-fit values of parameters. *Lower*: The deviation of the given multipoles from the best-fit ones divided by the uncertainty. We can see, that for all of the scales, the deviation does not exceed 2.5σ.

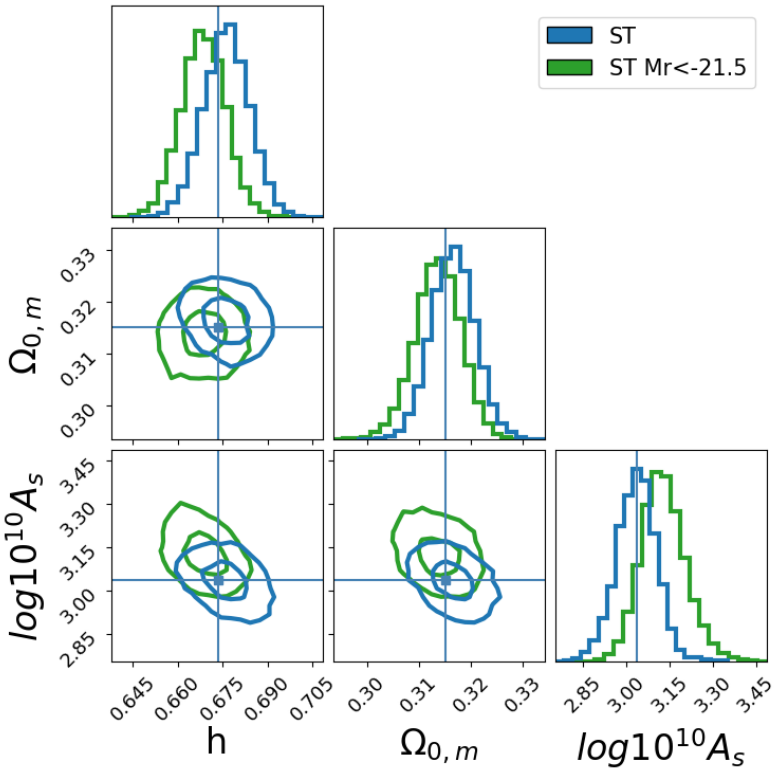

Figure 5.26: The posteriors obtained from the parameter inference on the mean of 25 Abacus mocks with full sample and with the magnitude limited one. The blue line represents the expected values of parameters.

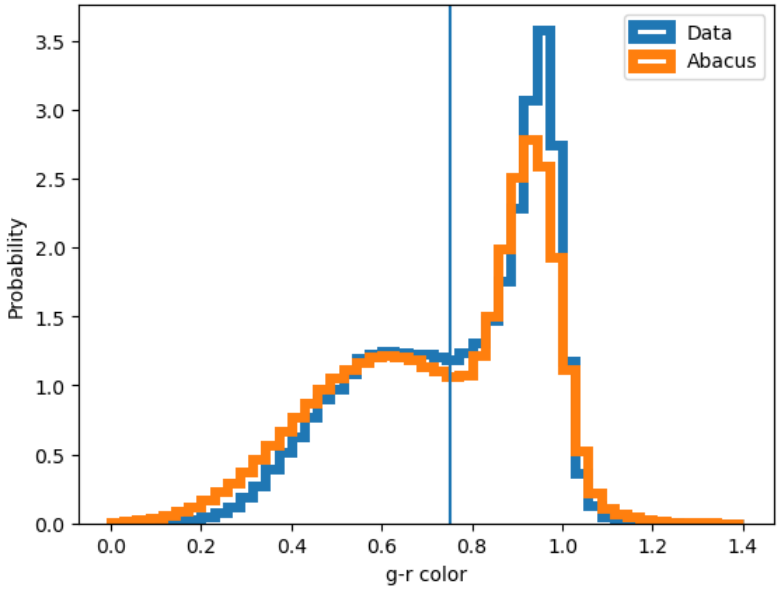

Figure 5.27: Distributions of rest frame $g-r$ colour for Abacus mocks and DR1 data. We can see a well defined bi-variate Gaussian distribution. The thin vertical line indicates the cut chosen throughout this thesis to separate the galaxies into red and blue populations.

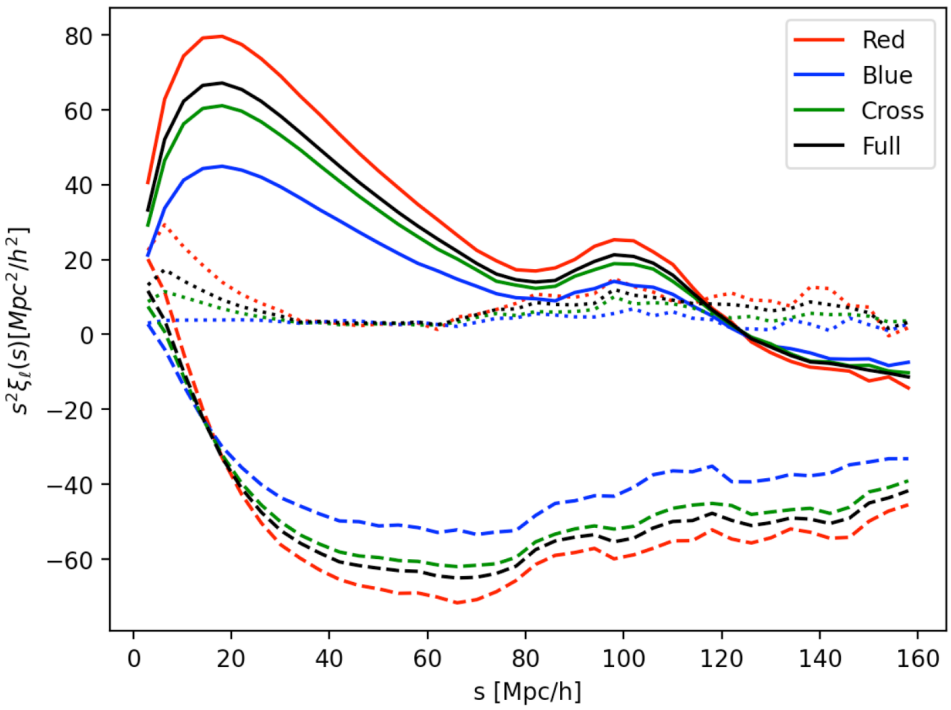

Figure 5.28: The monopoles (solid), quadrupoles (dashed) and hexadecupoles (dotted) of the correlation function averaged over 25 realisations for various tracers taken from the BGS DR1 Abacus mocks: red, blue, cross-correlations between the two, and the two tracer populations combined (full).

Chapter 1. The mean multipoles from 25 Abacus mocks are presented in Figure 5.28 for the auto-correlation of the full sample, the red galaxies, the blue galaxies and their cross-correlation. We clearly see the difference in the galaxy bias by noticing the difference in the correlation functions amplitudes between red and blue galaxies. Red galaxies are older, more massive galaxies, thus they are clustered more strongly than blue ones, which are usually younger and reside in less massive halos [43].

As for the single-tracer analysis of the BGS Bright Full sample, the only mocks available are the 25 Abacus mocks which are not enough to build an accurate mock-based covariance. Therefore, for multi-tracer, we again rely on FitCov to estimate the covariance matrix. The algorithm stays the same, with the exception that α-parameter is estimated for each correlation function (blue and red autocorrelation ones, as well as the crosscorrelation one) separately. Figure 5.29 shows the variance of the three multipoles of three afforementioned correlation functions from 25 Abacus mocks and from FitCov. We see that Fitcov managed to get compatible estimations of the variances for all three.

In order to validate the multi-tracer Full-Shape pipeline, we first perform a test of consistency between red and blue galaxies to ensure that the two populations indeed trace the same underlying matter density field. We fit the mean of the 25 Abacus mocks following the same methodology as in section 5.4.5 with the same cosmological and

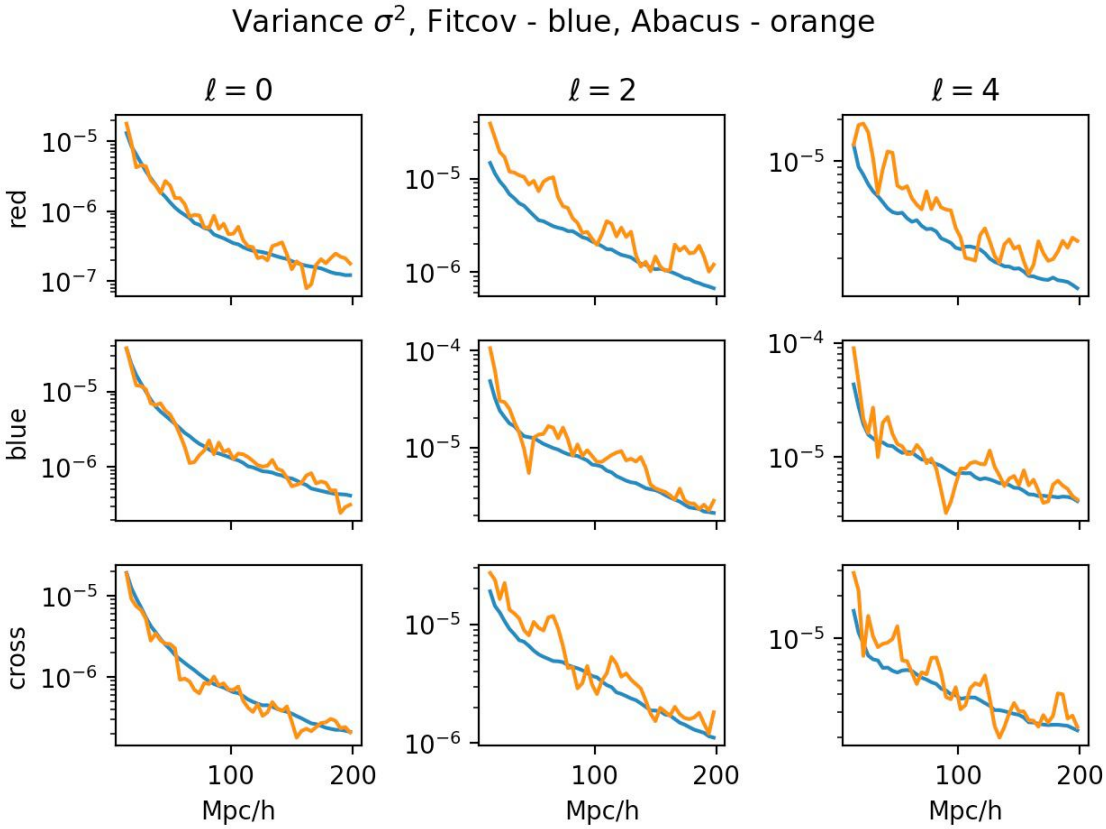

Figure 5.29: Variances on monopole, quadrupole and hexadecupole of the correlation function for various tracers for DR1 estimated from 25 Abacus mocks, with Fitcov in blue, and without in orange.

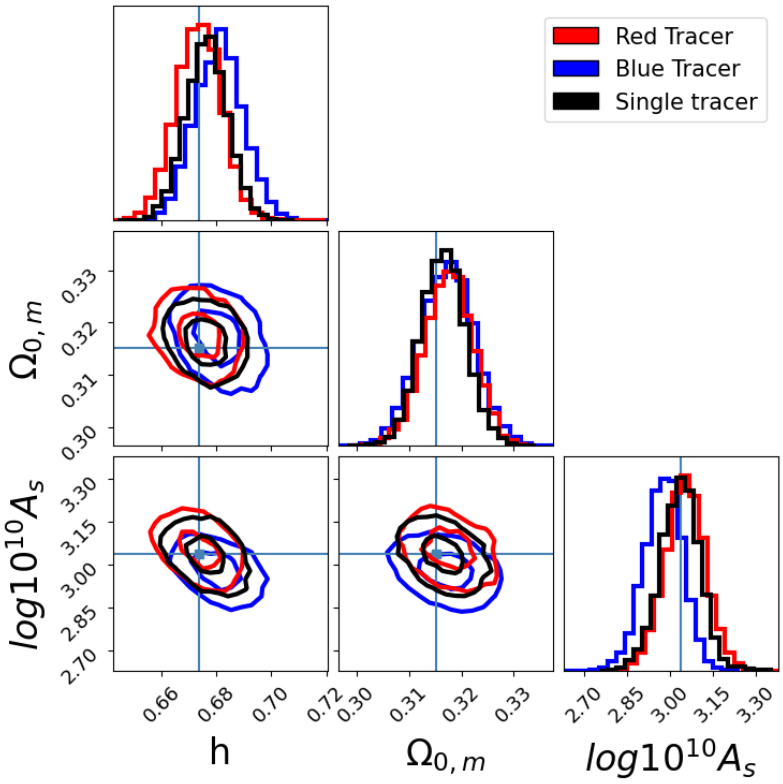

Figure 5.30: The posteriors obtained from the parameter inference on the mean of 25 Abacus mocks for red, blue and single tracer separately. Lines represent the expected values.

nuisance parameters. The results on the ΛCDM parameters can be seen in Figure 5.30 for the blue and red galaxies, together with the full BGS Bright sample. They are all consistent with each other and with respect to the expected values. The fact that the uncertainties between them are comparable as well, is a good representation of the cosmic variance, we introduced in the Introduction: combining the subsets in a traditional way does not improve the constraints

Then, we proceed to test the multitracer inference. The same methodology was used, with the nuisance parameters being fitted separately for each of the tracer in the multitracer case, while the cosmological parameters are common. We should note that the cross correlation function is treated as a separated tracer and thus has the same set of biases and counterterms marginalised over as for the blue and red tracers.

The resulting posteriors are shown in Figure 5.31.

The results between single and multi-tracer are consistent with each other and with respect to the expected values. Moreover, we can see that the multitracer allows us to obtain a $\sim 15\%$ improvement in the precision of $\log 10^{10} A_s$ compared to single-tracer. This validates the full set of the tools necessary for performing a multitracer Full-Modelling analysis of the DESI BGS DR1 data. The summary of the full-modellin tests in con-

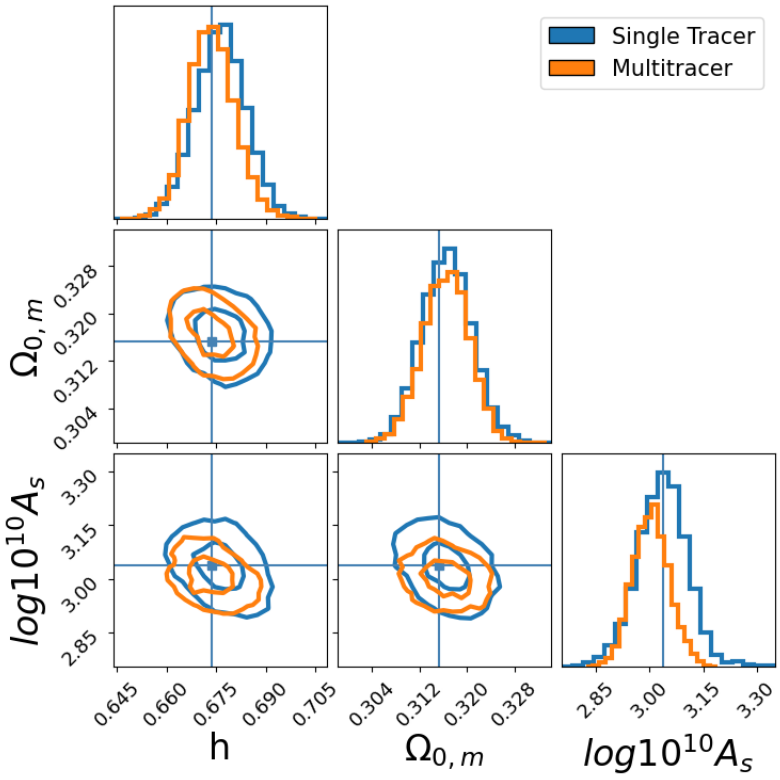

Figure 5.31: The posteriors obtained from the parameter inference on the mean of 25 Abacus mocks for multi- and single tracer analysis.

figuration space using our pipeline on different versions of DESI BGS Abacus mocks is presented in Figure 5.32.

5.4.7 Systematic error budget

One of the important test of the pipeline for the clustering analysis is the estimation of the systematic error budget. For DR1 2-point clustering analysis, the collaboration has looked at 7 different potential sources of systematic effects, all summarised in **??** with the corresponding error estimates obtained from the tests on the realistic mocks from Chapter 3.

Theoretical systematics come from the deficiencies in the chosen theoretical modelling. Three different EFT models in Fourier space have been tested against Abacus simulations:`velocileptors`[**maus24b**, 24, 25], `folps`[44] and `pybird`[45, 46] and they also have been compared with each other in [47]. A fourth EFT model based in configuration space (`EFT-GSM` has also been tested and compared against the other models which can be used for correlation function as well (except `FOLPS`) in [48],Throughout these tests, we decided to use $k_{\rm max}$ = 0.2h/Mpc and $s_{\rm min}$ = 26Mpc/h such that no significant systematics for DR1 are detected. Credit: Mark Maus, Hernan Noriega, Yan Lai, Sadi Ramirez.

Additionally, one needs to estimate the effect of the HOD prescription on the cosmo-

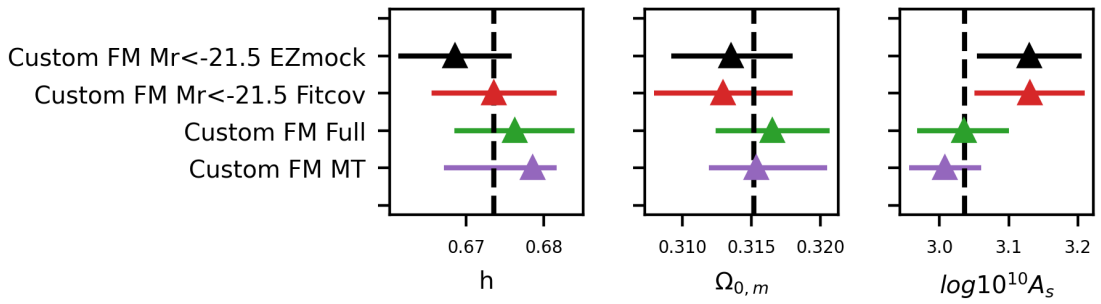

Figure 5.32: Constraints on different cosmological parameters obtained using mean of 25 Abacus mocks with rescaled covariance with full-modelling.

logical parameters of the analysis. For that purpose, mocks with varying HOD have been created for each tracer. Before unblinding, the HOD systematics for LRG and ELG have been studied leading to up to 0.3σ contribution to the systematic error budget. However, we realise that the HOD effect can be easily distinguished from prior weight effect, therefore both effects are currently studied together in order to estimate a systematic error budget that reflects both. The study for BGS and QS0 is also ongoing. Credit: Nathan Findlay and the mock team.

The choice of the fiducial cosmology can also affect our fits, as the computation of comoving distances and fiducial linear power spectra for approaches like ShapeFit would rely on it. It was found out that the choice of cosmology for conversion of redshifts is not producing any significant biases. However, that was not the case for the fiducial cosmology for the initial linear power spectrum generation, yielding a 0.45σ contribution on $f\sigma_8$ for ShapeFit analysis into the error budget. The impact of varying the fiducial cosmology is still ongoing for Full-Modelling. Credit: Rafaela Gsponer, Fanny Arlin Rodriguez, Hernan Noriega and Sadi Ramirez.

The systematics related to the fibre-assignment and the collisions it can lead to are tested as well. To mitigate for that effect, we employ a θ-cut, a technique described in [49] which consists in removing angular scales below 0.05 deg. The tests on the realistic mocks with simulated fibre-assignment and the comparison of their performance to the ones without, yielded an additional 0.2σ contribution to the systematic budget. Credit: Ruiyang Zhao, Mathilde Pinon and the KP3 team.

Imaging systematics due to the inhomogeneities in the target selection can also affect the cosmological measurements. The BGS corresponding to bright objects which do not hint the limit depth of the imaging surveys, it doesn't suffer strong variations of its galaxy density with observing conditions, contrary to other faint targets such as the ELG. Therefore, a linear combination to determine weights that correct for any remaining trend in the target density is enough to mitigate the effect and lead to no significant contribution to the systematic error budget. For ELG and QSO, additional correction beyond imaging

weights (which are estimated using deep learning techniques) are needed to account for remaining systematics. Credit: Ruiyang Zhao and the KP3 team.

The spectroscopic systematics can come from three effects: catastrophic failures of redshift measurements, redshift uncertainty and redshift success rate dependence on observing conditions. These effects have been studied using realistic Abacus mocks, yielded up to 0.2σ contribution to the systematic error budget for ELG which are the most sensitive to spectroscopic systematics. For the other tracers, including BGS, the effect is less and thus considered as negligible. Credit: Jiaxi Yu, Alex Krolewski and the KP3 team [50, 51].

The systematics related to the covariance estimation are tested by comparing two types of covariance matrix estimates: analytic and mock-based. The details of the tests for BGS were presented in the previous subsection. We estimate the contribution to the systematic budget to be up to 0.2σ. Credit: Daniel Forero-Sanchez, Otavio Alvez, Misha Rashkovetskyi and Svyatoslav Trusov.

Table 5.6: Summary of the individual systematic errors obtained when running the pipeline using ShapeFit or Full-Modelling for various realistic mocks and blinded data. Note that this table provides a non-detailed estimate and that in some cases, the recession of our mocks is not sufficient to quote any statistically significant detection of a systematic. Credit: DESI Collaboration

Systematic	Error estimate (DR1 error)	Comment
Theoretical	Not detected for DR1	
HOD	Not estimated yet for BGS	Up to 0.3σ for DR1 ELG
Fiducial cosmology	Not estimated yet for BGS	Up to 0.45σ on $f\sigma_8$ for DR1 LRG
Fibre-collisions	Up to 0.2σ of DR1 error	
Imaging	Not detected for DR1 BGS	Up to 0.25σ for ELG and QSO
Spectroscopic	Not detected for DR1 BGS	Up to 0.2σ on $f\sigma_8$ for DR1 ELG
Covariances	Up to 0.2σ for DR1 BGS	Similar for the other tracers

It should be noted, that **the final systematic error budget has not been determined yet as some systematic effects are still investigated. Moreover, the systematic error propagation has not yet been performed on the results presented further**, thus only statistical errors are considered for now.

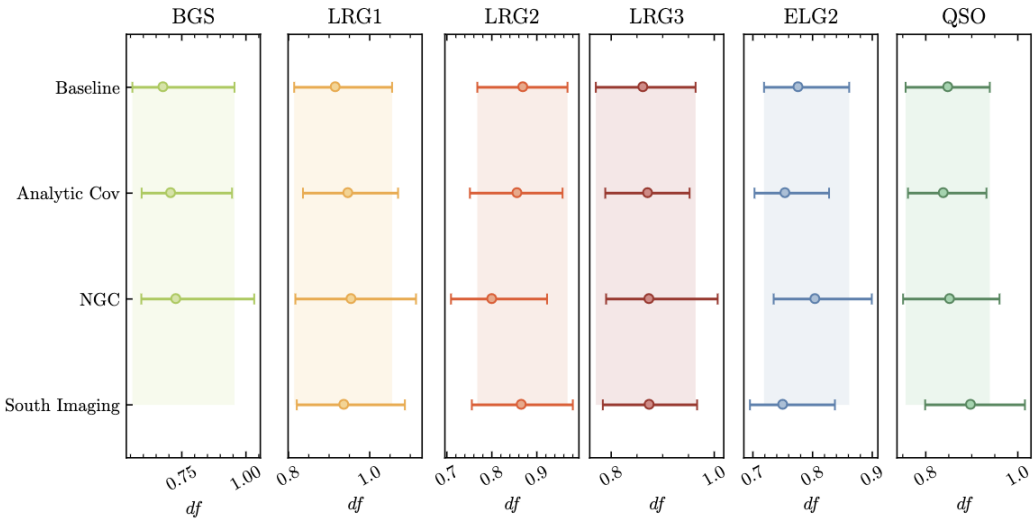

Figure 5.33: Measurements of the ratio of the growth rate to the fiducial ones with different setups performed on the blinded DR1 data for all of the clustering catalogues of DESI. Credit: Ruiyang Zhao

5.5 DESI BGS DR1 analysis

5.5.1 DESI DR1 official Full-Shape analysis: blinded tests

In order to validate the analysis on the data, and avoid confirmation bias the data is first blinded. The redshifts of the galaxies in the catalogues are modified with two distinct shifts, first obtaining comoving coordinates with one fiducial cosmology, and then converting them back to redshifts with a different one, thus creating an effect similar to the AP effect, thus called an AP-shift, and biasing the measurements of AP parameters, and the shift based on the gradient of the reconstructed galaxy density field, thus mimicking the RSD effect, hence blinding the growth of structures and which is referred to as RSD-shift. More on the procedure can be found in [52]. The blinding procedure allows us to do intensive tests on the data while avoiding the possibility of confirmation bias. An example of some consistency tests performed on the DR1 blinded data can be seen on Figure 5.33.

5.5.2 DESI BGS $M_r < -21.5$

As was mentioned previously, the official BGS DR1 analysis is based on Full-Modelling in Fourier space [53] in the range of $k = [0.02, 0.2]$ h Mpc^{-1} in bins of $\Delta k = 0.005$h Mpc^{-1}. The BGS sample has also a cut in absolute magnitude $M_r < -21.5$ and an effective redshift $z_{\rm eff} = 0.295$.

We perform the same analysis in configuration space using the official DESI tools (aka 'desilike') and with our pipeline that relies on Fitcov and on the neural-network emulator.

Table 5.7: Parameters used for the ShapeFit inference and their priors

Parameter	Range/Prior
f/f_{fid}	[0, 2]
$\alpha_{\parallel}$	[0.8,1.2]
$\alpha_{\perp}$	[0.8, 1.2]
m	[-3, 3]
b_1	[-1, 5]
b_2	$\mathcal{N}(0, 5)$
b_s	$\mathcal{N}(0, 10)$
α_0	$\mathcal{N}(0, 30)$
α_2	$\mathcal{N}(0, 50)$
σ_v	[0,50000]

We use the separation range $s = [26, 150]$h Mpc^{-1} with bin of width 4 Mpch^{-1}. The multipoles of the data with uncertainties and the best-fit theoretical ones are shown in Figure 5.35. In all 3 cases the EZmock covariance is used. We have also performed the ShapeFit analysis with the parameters fitted and their priors specified in Table 5.7. The summary of the results are shown in Figure 5.34.

We observe that, in configuration space for BGS, the uncertainty on h and $\Omega_{0,m}$ are much tighter, which could be a representation of projection effects propagating into the cosmological parameter estimation. We notice that for $\Omega_{0,m}$, Full-Modelling seems to give tighter constraints. We see as well that the Fourier space Full-Modelling tends to underestimate $\Omega_{0,m}$ compared to other two techniques, while configuration space does the same with $\log 10^{10} A_s$. We note that the ShapeFit results remain consistent but the error bars are significantly larger in Fourier space for some parameters. That can be due to the projection effects, and thus the reason the baseline DESI analysis using Full-Modelling, which was also our focus as well. When comparing to their Full-Modelling counterparts in Fourier space, the Full-Modelling constraint acquires a $\sim 1.\sigma$ tension on the h. For the configuration space, the tension of the same magnitude appears for the $log 10^{10} A_s$. Comparing Full-Modelling in configuration space (red) to the one in Fourier space (brown), we see that they acquire $\sim 1\sigma$ tension on both $\Omega_{0,m}$ and $\log 10^{10} A_s$.

5.5.3 DESI BGS Bright

As mentioned in the previous section on mocks, thanks to the development of a hybrid covariance with FitCov, we can analyse the full BGS Bright, trying to bring back as much precision as possible on the cosmological parameters. We perform a Full-Modelling inference from the full BGS Bright dataset using the same s-binning as in the previous

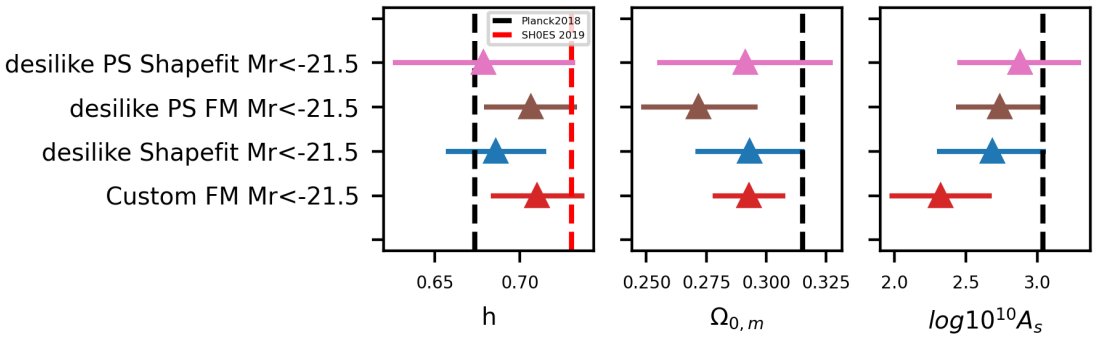

Figure 5.34: Constraints on different cosmological parameters obtained using DR1 with $M_r < -21.5$ with different approaches.The pink point represents the analysis with ShapeFit in Fourier space, the brown point is the same but using full-modelling instead of ShapeFit, the blue point represents the ShapeFit analysis in configuration space and the red point is the full-modelling results obtained with our custom pipeline.

subsection and with an effective redshift $z_{\text{eff}=0.275}$. The BGS Bright DR1 multipoles with the best-fit multipoles as well as the residuals are shown in Figure 5.36.

Figure 5.37 shows the constraints on the cosmological parameters for BGS Full (orange) compared to the volume-limited BGS sample with $M_r < -21.5$. We can see that the results are consistent, with Full BGS Bright bringing noticeably smaller contours. For $\Omega_{0,m}$ the improvement is 24%, and for the $\log 10^{10} A_s$ we see an improvement of 22% in precision. The errorbars for h are only 4% smaller for Full BGS Bright.

The last step is to perform the multitracer analysis of the BGS DR1 Full. First we compute the red, blue and cross- correlation functions. The monopole and quadrupole of the auto correlation functions are plotted in Figure 5.38. We note a strange feature in the quadrupole of the blue tracer that we are still investigating at the time of writing this manuscript. It can be related to untreated imaging systematics, thus the cosmological results presented starting from here **should not be treated as final and must be taken with a grain of salt.**

Using the pipeline we covered in more detail in the previous section, we can finally perform the multitracer analysis of DESI BGS. The results are presented in Figure 5.39.

We can see a shift, for example, with respect to h, which is driven by the afforementioned feature in the quadrupole of the correlation function of the blue tracer. This is especially evident looking at the results obtained from fitting the individual tracers (Figure 5.40).

With respect to other analysis approaches presented earlier, we can notice (Figure 5.41) that in configuration space the constraints tend to be in general more stringent, which might be attributed to the absence of need to marginalise over the shot noise. We also note that the constraints from various approaches tend to be consistent with each

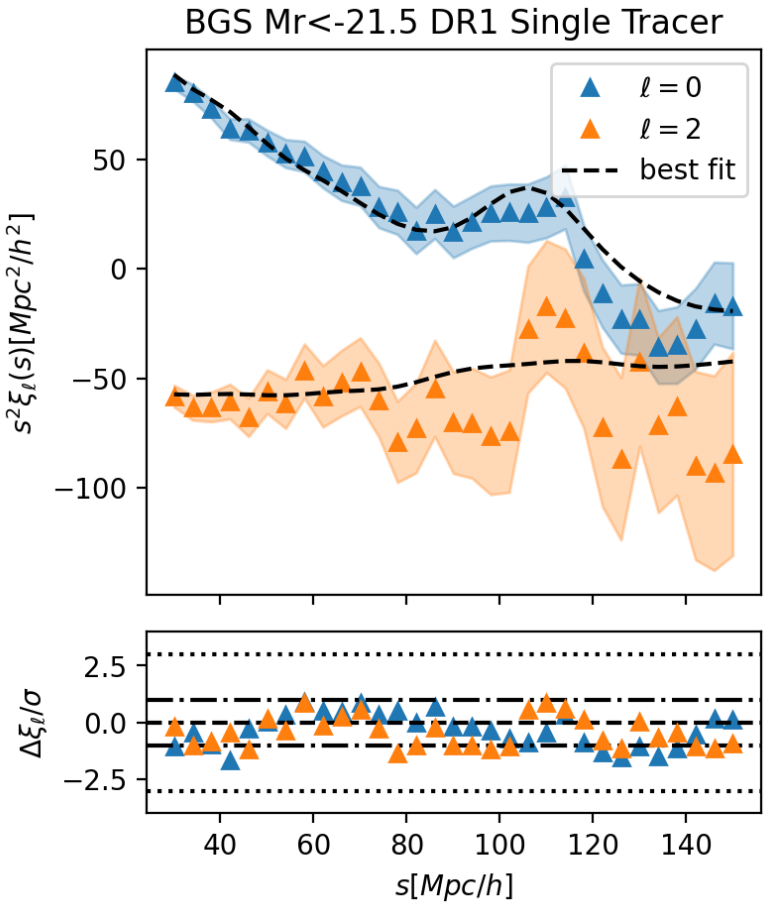

Figure 5.35: *Upper*: Correlation function multipoles of the BGS $M_r < -21.5$ DR1 with shaded values representing the errorbar, and the dashed line being the modelled multipoles for the best-fit values of parameters. *Lower*: The deviation of the given multipoles from the best-fit ones divided by the uncertainty. We can see, that for most of the scales, the deviation does not exceed 1σ.

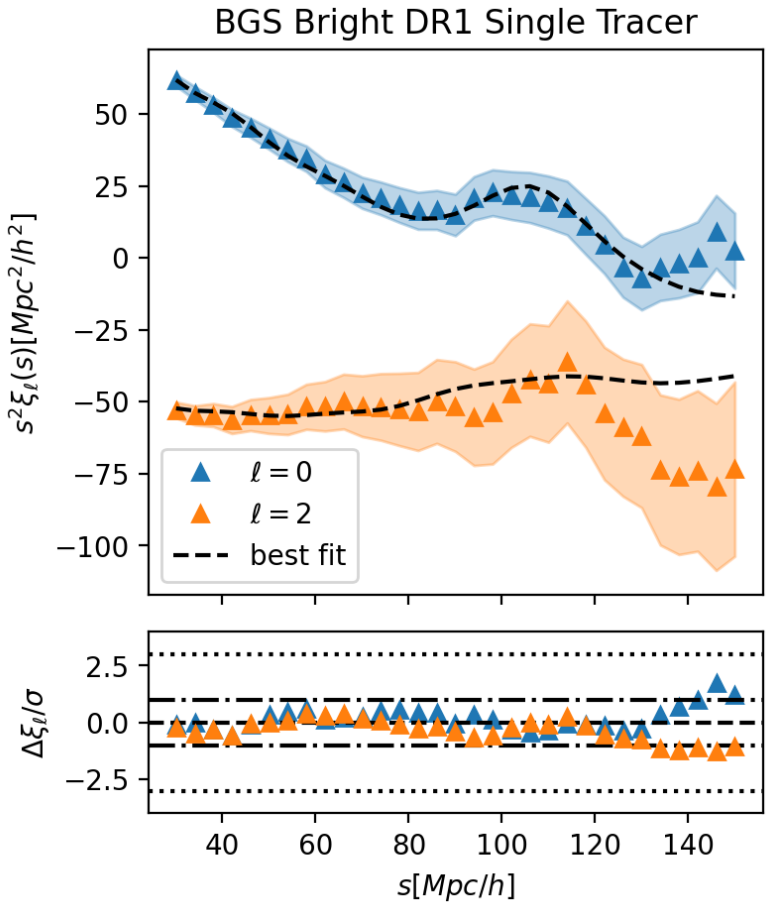

Figure 5.36: *Upper*: Correlation function multipoles of the BGS Bright DR1 with shaded values representing the errorbar, and the dashed line being the modelled multipoles for the best-fit values of parameters. *Lower*: The deviation of the given multipoles from the best-fit ones divided by the uncertainty. We can see, that for most of the scales, the deviation does not exceed 1σ.

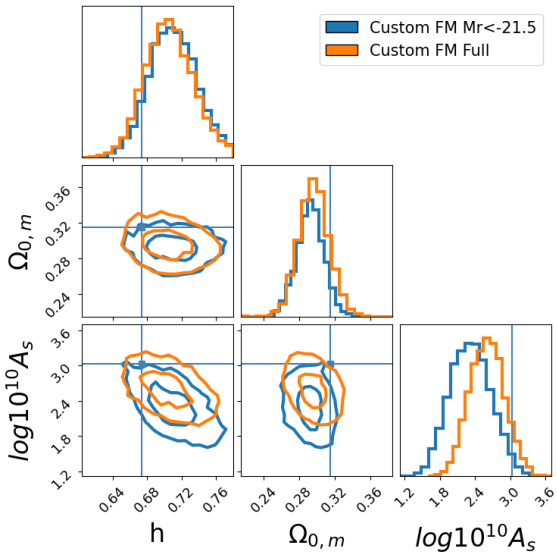

Figure 5.37: The posterior distributions of cosmological parameters obtained using BGS DR1 with a cut on $M_r < 21.5$ and the full BGS Bright obtained with Full-Modelling.

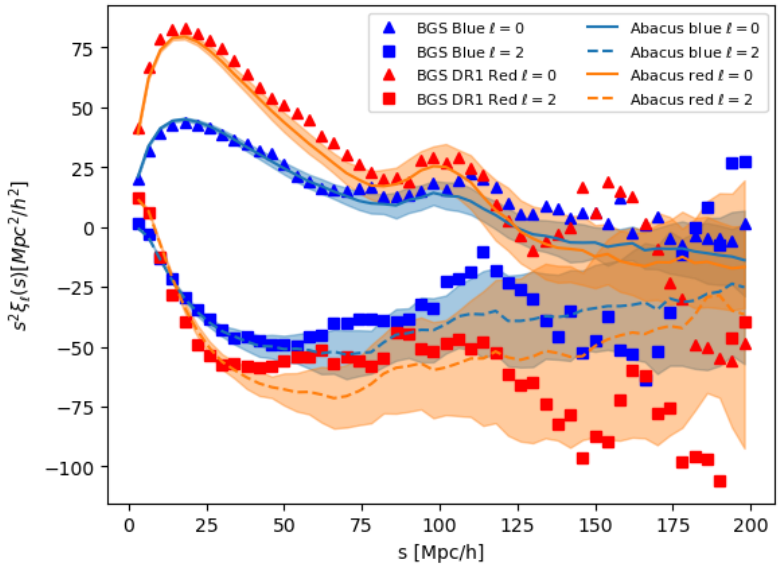

Figure 5.38: Monopoles and quadrupoles of the correlation functions of red and blue tracers of DESI BGS DR1 (markers) and of the Abacus BGS DR1 mocks (solid and dashed lines).

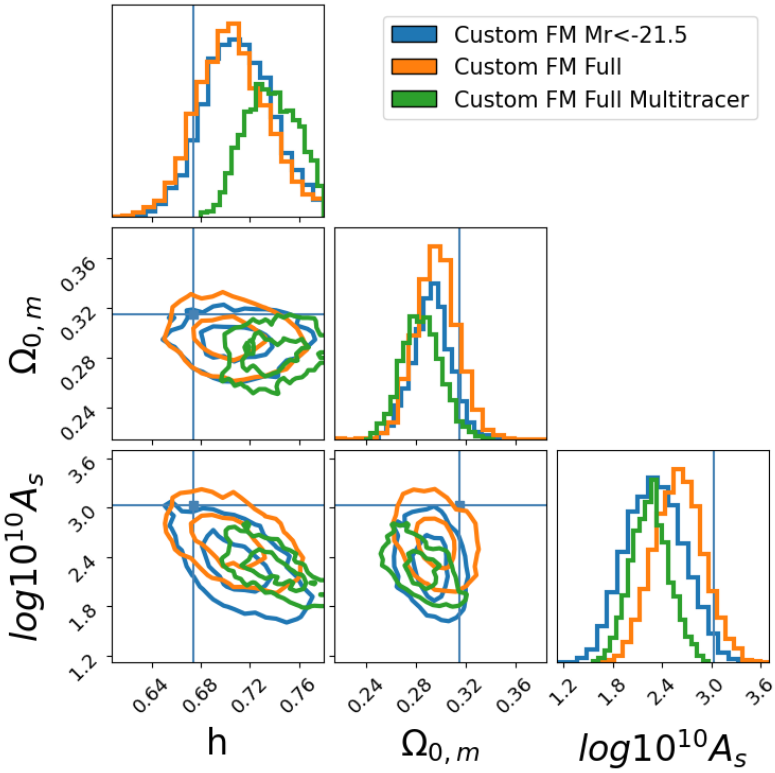

Figure 5.39: The posterior distributions of cosmological parameters obtained using DESI BGS DR1 with various analysis configurations, where blue lines correspond to Planck2018 best-fit values[54].

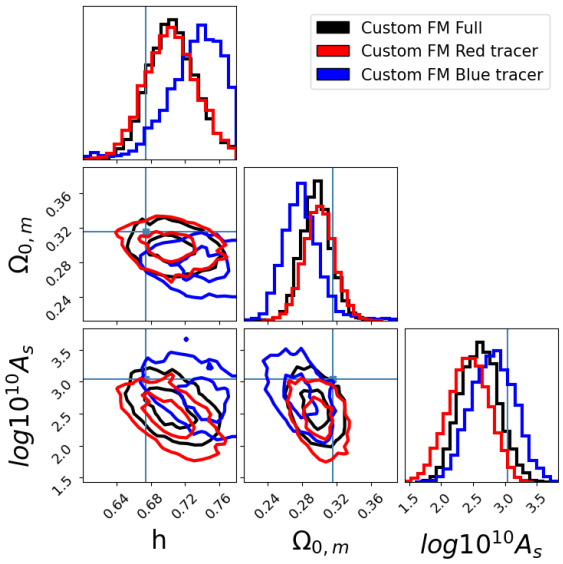

Figure 5.40: The posterior distributions of cosmological parameters obtained using DESI BGS DR1 with various tracers.

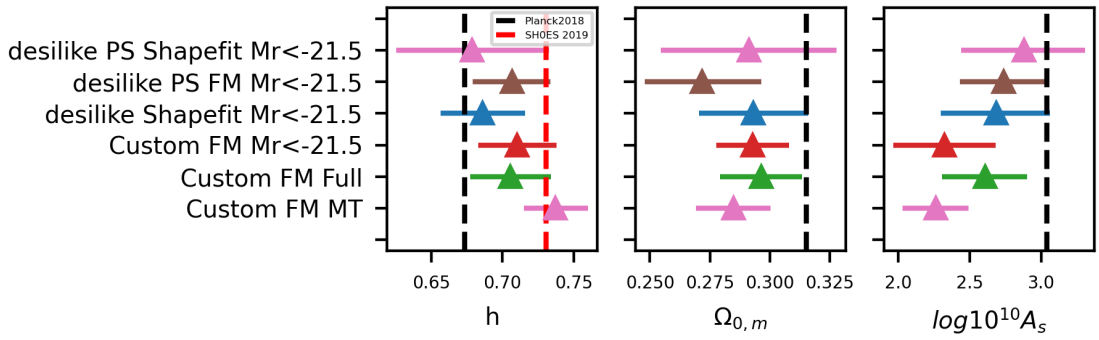

Figure 5.41: Constraints on cosmological parameters h, $\Omega_{0,m}$ and A_s from various ways of analysing DESI BGS DR1

other, with multitracer ones moving away from Planck[54] the most. The constraints are also summarised in table 5.8.

Table 5.8: Summary of the cosmological parameters obtained using full-modelling on DESI BGS in various configuration

	h	$\Omega_{0,m}$	$log10^{10}A_s$
Custom Mr-21.5	$0.710^{+0.028}_{-0.027}$	$0.293^{+0.015}_{-0.015}$	$2.322^{+0.359}_{-0.355}$
desilike PS Mr-21.5	$0.707^{+0.027}_{-0.027}$	$0.272^{+0.025}_{-0.024}$	$2.734^{+0.308}_{-0.303}$
Custom Full	$0.706^{+0.028}_{-0.028}$	$0.296^{+0.017}_{-0.017}$	$2.606^{+0.295}_{-0.302}$
Custom MT	$0.737^{+0.023}_{-0.022}$	$0.285^{+0.015}_{-0.016}$	$2.261^{+0.229}_{-0.230}$
Custom Blue	$0.730^{+0.032}_{-0.031}$	$0.280^{+0.019}_{-0.021}$	$2.814^{+0.345}_{-0.341}$
Custom Red	$0.704^{+0.031}_{-0.030}$	$0.300^{+0.017}_{-0.018}$	$2.435^{+0.325}_{-0.325}$
Custom Cross	$0.704^{+0.026}_{-0.026}$	$0.298^{+0.017}_{-0.017}$	$2.511^{+0.320}_{-0.311}$

We can further compare our results to the rest of the DESI tracers, and these are shown in Figure 5.42, for h, $\Omega_{0,m}$ and A_s. All of the results other than the multitracer one have been obtained using Full-Modelling in Fourier space. We see that DESI BGS DR1 with multitracer becomes mostly an outlier when it comes to the A_s, however, the nature of this deviation, as mentioned earlier, is in investigation, as of writing this section.

References

[1] Florian Beutler et al. “The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Fourier space”. In: *Monthly Notices of the Royal Astronomical Society* 464.3 (Sept. 2016), pp. 3409–3430. ISSN: 0035-8711. DOI: 10.1093/mnras/stw2373. eprint: https:

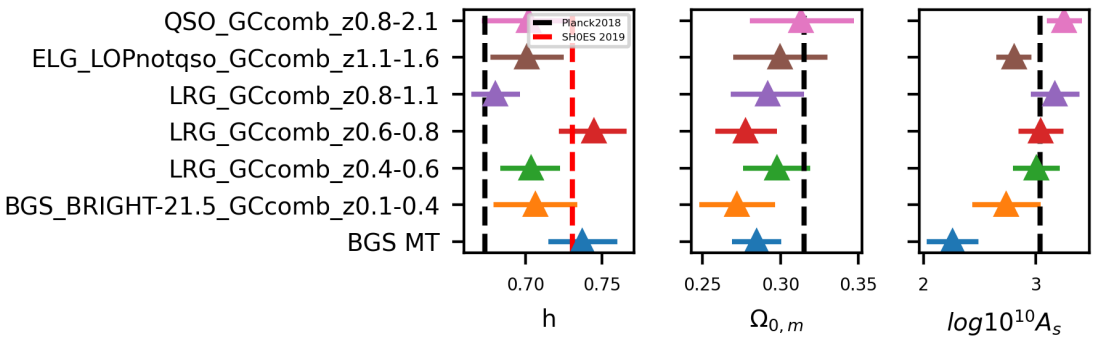

Figure 5.42: Constraints on cosmological parameters h, $\Omega_{0,m}$ and A_s from various DESI tracers.

//academic.oup.com/mnras/article-pdf/464/3/3409/17703479/stw2373.pdf. URL: https://doi.org/10.1093/mnras/stw2373.

[2] Héctor Gil-Marín et al. "The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic power spectrum between redshifts 0.6 and 1.0". In: *Monthly Notices of the Royal Astronomical Society* 498.2 (Aug. 2020), pp. 2492–2531. ISSN: 0035-8711. DOI: 10.1093/mnras/staa2455. eprint: https://academic.oup.com/mnras/article-pdf/498/2/2492/33776452/staa2455.pdf. URL: https://doi.org/10.1093/mnras/staa2455.

[3] Samuel Brieden et al. "Model-agnostic interpretation of 10 billion years of cosmic evolution traced by BOSS and eBOSS data". In: *Journal of Cosmology and Astroparticle Physics* 2022.08 (Aug. 2022), p. 024. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2022/08/024. URL: http://dx.doi.org/10.1088/1475-7516/2022/08/024.

[4] DESI Collaboration et al. 2024. arXiv: 2404.03002 [astro-ph.CO].

[5] C. A. Dong-Páez et al. "The Uchuu–SDSS galaxy light-cones: a clustering, redshift space distortion and baryonic acoustic oscillation study". In: *Mon. Not. Roy. Astron. Soc.* 528.4 (2024), pp. 7236–7255. DOI: 10.1093/mnras/stae062. arXiv: 2208.00540 [astro-ph.CO].

[6] Daniel J. Eisenstein et al. "On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter". In: *The Astrophysical Journal* 664.2 (Aug. 2007), pp. 660–674. ISSN: 1538-4357. DOI: 10.1086/518755. URL: http://dx.doi.org/10.1086/518755.

[7] Shi-Fan Chen et al. 2024. arXiv: 2402.14070 [astro-ph.CO].

[8] Daniel J. Eisenstein et al. "Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak". In: 664.2 (Aug. 2007), pp. 675–679. DOI: 10.1086/518712. arXiv: astro-ph/0604362 [astro-ph].

[9] DESI Collaboration et al. *DESI 2024 III: Baryon Acoustic Oscillations from Galaxies and Quasars*. 2024. arXiv: 2404.03000 [astro-ph.CO]. URL: https://arxiv.org/abs/2404.03000.

[10] Hee-Jong Seo et al. "Nonlinear Structure Formation and the Acoustic Scale". In: *The Astrophysical Journal* 686.1 (Oct. 2008), pp. 13–24. ISSN: 1538-4357. DOI: 10.1086/589921. URL: http://dx.doi.org/10.1086/589921.

[11] Hee-Jong Seo et al. "HIGH-PRECISION PREDICTIONS FOR THE ACOUSTIC SCALE IN THE NONLINEAR REGIME". In: *The Astrophysical Journal* 720.2 (Aug. 2010), pp. 1650–1667. ISSN: 1538-4357. DOI: 10.1088/0004-637x/720/2/1650. URL: http://dx.doi.org/10.1088/0004-637X/720/2/1650.

[12] Marcel Schmittfull et al. "Eulerian BAO reconstructions and N-point statistics". In: *Phys. Rev. D* 92 (12 Dec. 2015), p. 123522. DOI: 10.1103/PhysRevD.92.123522. URL: https://link.aps.org/doi/10.1103/PhysRevD.92.123522.

[13] Zhejie Ding et al. "Theoretical systematics of Future Baryon Acoustic Oscillation Surveys". In: *MNRAS* 479.1 (May 2018), pp. 1021–1054. ISSN: 0035-8711. DOI: 10.1093/mnras/sty1413. eprint: https://academic.oup.com/mnras/article-pdf/479/1/1021/25129090/sty1413.pdf. URL: https://doi.org/10.1093/mnras/sty1413.

[14] Michael Rashkovetskyi et al. "Validation of semi-analytical, semi-empirical covariance matrices for two-point correlation function for early DESI data". In: *Mon. Not. Roy. Astron. Soc.* 524.3 (2023), pp. 3894–3911. DOI: 10.1093/mnras/stad2078. arXiv: 2306.06320 [astro-ph.CO].

[15] E. Paillas et al. *Optimal Reconstruction of Baryon Acoustic Oscillations for DESI 2024*. 2024. arXiv: 2404.03005 [astro-ph.CO].

[16] C. Alcock and B. Paczynski. "An evolution free test for non-zero cosmological constant". In: 281 (Oct. 1979), p. 358. DOI: 10.1038/281358a0.

[17] Cullan Howlett et al. "The clustering of the SDSS main galaxy sample – II. Mock galaxy catalogues and a measurement of the growth of structure from redshift space distortions at z=0.15". In: *MNRAS* 449.1 (Mar. 2015), pp. 848–866. ISSN: 0035-8711. DOI: 10.1093/mnras/stu2693. eprint: https://academic.oup.com/mnras/article-pdf/449/1/848/17335801/stu2693.pdf. URL: https://doi.org/10.1093/mnras/stu2693.

[18] A. Pezzotta et al. “The VIMOS Public Extragalactic Redshift Survey (VIPERS): The growth of structure at 0.5 lt;zlt; 1.2 from redshift-space distortions in the clustering of the PDR-2 final sample ”. In: *Astronomy amp; Astrophysics* 604 (July 2017), A33. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201630295. URL: http://dx.doi.org/10.1051/0004-6361/201630295.

[19] M. Drinkwater et al. “The WiggleZ Galaxy Survey shows that dark energy is real”. In: May 2012, p. 071. DOI: 10.22323/1.134.0071.

[20] Beth A. Reid et al. “The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measurements of the growth of structure and expansion rate at z = 0.57 from anisotropic clustering”. In: *MNRAS* 426.4 (Nov. 2012), pp. 2719–2737. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2012.21779.x. eprint: https://academic.oup.com/mnras/article-pdf/426/4/2719/3295253/426-4-2719.pdf. URL: https://doi.org/10.1111/j.1365-2966.2012.21779.x.

[21] Samuel Brieden et al. “ShapeFit: extracting the power spectrum shape information in galaxy surveys beyond BAO and RSD”. In: *JCAP* 2021.12 (Dec. 2021), p. 054. DOI: 10.1088/1475-7516/2021/12/054. URL: https://dx.doi.org/10.1088/1475-7516/2021/12/054.

[22] Kyle S. Dawson et al. “The Baryon Oscillation Spectroscopic Survey of SDSS-III”. In: 145.1, 10 (Jan. 2013), p. 10. DOI: 10.1088/0004-6256/145/1/10. arXiv: 1208.0022 [astro-ph.CO].

[23] Samuel Brieden et al. “ShapeFit: extracting the power spectrum shape information in galaxy surveys beyond BAO and RSD”. In: *JCAP* 2021.12 (Dec. 2021), p. 054. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2021/12/054. URL: http://dx.doi.org/10.1088/1475-7516/2021/12/054.

[24] Shi-Fan Chen et al. “Consistent modeling of velocity statistics and redshift-space distortions in one-loop perturbation theory”. In: *JCAP* 2020.07 (July 2020), p. 062. DOI: 10.1088/1475-7516/2020/07/062. URL: https://dx.doi.org/10.1088/1475-7516/2020/07/062.

[25] Shi-Fan Chen et al. “Redshift-space distortions in Lagrangian perturbation theory”. In: *JCAP* 2021.03 (Mar. 2021), p. 100. DOI: 10.1088/1475-7516/2021/03/100. URL: https://dx.doi.org/10.1088/1475-7516/2021/03/100.

[26] Nina A Maksimova et al. “AbacusSummit: a massive set of high-accuracy, high-resolution N-body simulations”. In: *MNRAS* 508.3 (Sept. 2021), pp. 4017–4037. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2484. eprint: https://academic.oup.com/mnras/article-pdf/508/3/4017/40811763/stab2484.pdf. URL: https://doi.org/10.1093/mnras/stab2484.

[27] Will Handley and Pablo Lemos. "Quantifying tensions in cosmological parameters: Interpreting the DES evidence ratio". In: *Phys. Rev. D* 100 (4 Aug. 2019), p. 043504. DOI: 10.1103/PhysRevD.100.043504. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.043504.

[28] James E. Gunn et al. "The 2.5 m Telescope of the Sloan Digital Sky Survey". In: *The Astronomical Journal* 131.4 (Apr. 2006), pp. 2332–2359. ISSN: 1538-3881. DOI: 10.1086/500975. URL: http://dx.doi.org/10.1086/500975.

[29] Michael R. Blanton et al. "New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys". In: 129.6 (June 2005), pp. 2562–2578. DOI: 10.1086/429803. arXiv: astro-ph/0410166 [astro-ph].

[30] Tomoaki Ishiyama et al. "The Uchuu simulations: Data Release 1 and dark matter halo concentrations". In: 506.3 (Sept. 2021), pp. 4210–4231. DOI: 10.1093/mnras/stab1755. arXiv: 2007.14720 [astro-ph.CO].

[31] Anatoly Klypin and Francisco Prada. "Dark matter statistics for large galaxy catalogues: power spectra and covariance matrices". In: *MNRAS* 478.4 (June 2018), pp. 4602–4621. ISSN: 0035-8711. DOI: 10.1093/mnras/sty1340. eprint: https://academic.oup.com/mnras/article-pdf/478/4/4602/25096258/sty1340.pdf. URL: https://doi.org/10.1093/mnras/sty1340.

[32] Richard Neveux et al. "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from the anisotropic power spectrum of the quasar sample between redshift 0.8 and 2.2". In: *MNRAS* 499.1 (Sept. 2020), pp. 210–229. DOI: 10.1093/mnras/staa2780. eprint: https://academic.oup.com/mnras/article-pdf/499/1/210/33842640/staa2780.pdf. URL: https://doi.org/10.1093/mnras/staa2780.

[33] Jiamin Hou et al. "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2". In: *MNRAS* 500.1 (Oct. 2020), pp. 1201–1221. DOI: 10.1093/mnras/staa3234. eprint: https://academic.oup.com/mnras/article-pdf/500/1/1201/34369971/staa3234.pdf. URL: https://doi.org/10.1093/mnras/staa3234.

[34] Hans Dembinski and Piti Ongmongkolkul et al. "scikit-hep/iminuit". In: (Dec. 2020). DOI: 10.5281/zenodo.3949207. URL: https://doi.org/10.5281/zenodo.3949207.

[35] Andrew Gelman and Donald B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472. ISSN: 08834237. URL: http://www.jstor.org/stable/2246093 (visited on 07/21/2022).

[36] Stephen P. Brooks and Andrew Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. DOI: 10.1080/10618600.1998.10474787. eprint: https://www.tandfonline.com/doi/pdf/10.1080/10618600.1998.10474787. URL: https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474787.

[37] Faizan G Mohammad and Will J Percival. "Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function". In: *MNRAS* 514.1 (May 2022), pp. 1289–1301. ISSN: 0035-8711. DOI: 10.1093/mnras/stac1458. eprint: https://academic.oup.com/mnras/article-pdf/514/1/1289/43986041/stac1458.pdf. URL: https://doi.org/10.1093/mnras/stac1458.

[38] Otavio Alves and DESI Collaboration. "Analytical covariance matrices of DESI galaxy power spectrum multipoles". (in prep.) 2024.

[39] Digvijay Wadekar and Roman Scoccimarro. "Galaxy power spectrum multipoles covariance in perturbation theory". In: *Phys. Rev. D* 102.12 (2020), p. 123517. DOI: 10.1103/PhysRevD.102.123517. arXiv: 1910.02914 [astro-ph.CO].

[40] Yosuke Kobayashi. "Fast computation of the non-Gaussian covariance of redshift-space galaxy power spectrum multipoles". In: *Phys. Rev. D* 108.10 (2023), p. 103512. DOI: 10.1103/PhysRevD.108.103512. arXiv: 2308.08593 [astro-ph.CO].

[41] Daniel Foreman-Mackey et al. "¡tt¿emcee¡/tt¿: The MCMC Hammer". In: *PASP* 125.925 (Mar. 2013), pp. 306–312. ISSN: 1538-3873. DOI: 10.1086/670067. URL: http://dx.doi.org/10.1086/670067.

[42] V. Mossa et al. "The baryon density of the Universe from an improved rate of deuterium burning". In: *Nature* 587.7833 (Nov. 2020), pp. 210–213. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2878-4. URL: http://dx.doi.org/10.1038/s41586-020-2878-4.

[43] W. W. Morgan and N. U. Mayall. "A Spectral Classification of Galaxies". In: 69.409 (Aug. 1957), p. 291. DOI: 10.1086/127075.

[44] Hernán E. Noriega et al. *Fast computation of non-linear power spectrum in cosmologies with massive neutrinos*. Aug. 2022. arXiv: 2208.02791 [astro-ph.CO].

[45] Guido D'Amico et al. "Limits on *w*CDM from the EFTofLSS with the PyBird code". In: *JCAP* 01 (2021), p. 006. DOI: 10.1088/1475-7516/2021/01/006. arXiv: 2003.07956 [astro-ph.CO].

[46] Y. Lai et al. "A comparison between Shapefit compression and Full-Modelling method with PyBird for DESI 2024 and beyond". In: *arXiv e-prints*, arXiv:2404.07283 (Apr. 2024), arXiv:2404.07283. DOI: 10.48550/arXiv.2404.07283. arXiv: 2404.07283 [astro-ph.CO].

[47] M. Maus et al. "A comparison of effective field theory models of redshift space galaxy power spectra for DESI 2024 and future surveys". In: *arXiv e-prints*, arXiv:2404.07272 (Apr. 2024), arXiv:2404.07272. DOI: 10.48550/arXiv.2404.07272. arXiv: 2404.07272 [astro-ph.CO].

[48] S. Ramirez-Solano et al. "Full Modeling and Parameter Compression Methods in configuration space for DESI 2024 and beyond". In: *arXiv e-prints*, arXiv:2404.07268 (Apr. 2024), arXiv:2404.07268. DOI: 10.48550/arXiv.2404.07268. arXiv: 2404.07268 [astro-ph.CO].

[49] M. Pinon et al. *Mitigation of DESI fiber assignment incompleteness effect on two-point clustering with small angular scale truncated estimators*. 2024. arXiv: 2406.04804 [astro-ph.CO]. URL: https://arxiv.org/abs/2406.04804.

[50] Jiaxi Yu et al. "ELG Spectroscopic Systematics Analysis of the DESI Data Release 1". In: *arXiv e-prints*, arXiv:2405.16657 (May 2024), arXiv:2405.16657. DOI: 10.48550/arXiv.2405.16657. arXiv: 2405.16657 [astro-ph.CO].

[51] A. Krolewski et al. "Impact and mitigation of spectroscopic systematics on DESI DR1 clustering measurements". In: *arXiv e-prints*, arXiv:2405.17208 (May 2024), arXiv:2405.17208. DOI: 10.48550/arXiv.2405.17208. arXiv: 2405.17208 [astro-ph.CO].

[52] Samuel Brieden et al. "Blind Observers of the Sky". In: *JCAP* 2020.09 (Sept. 2020), pp. 052–052. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2020/09/052. URL: http://dx.doi.org/10.1088/1475-7516/2020/09/052.

[53] DESI Collaboration. "DESI Full-shape analysis". (in prep.) 2024.

[54] N. Aghanim et al. "iPlanck/i2018 results". In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. URL: https://doi.org/10.1051/0004-6361/201833910.

Conclusions

Все!

Eralash

DESI cosmological constraints

As it was mentioned in the Introduction, usually the results of one survey are combined with other cosmological probes in order to achieve the maximal constraining power. This is the final stage of the cosmological analysis, and also where many effects are taken into account, and should be treated with great care.

However, I would still like to share some, even though very preliminary, constraints, which, however, **should be treated as very preliminary**. Thus, I will restrain from quoting any tensions. The combinations were done under the assumption that all of the likelihoods are Gaussian (which is a far-fetched assumption for some of them), and then sampling with `emcee`[1] from the combined likelihood, using the posterior from the combined fit of all other data (DESI except BGS, Planck[2], DES[3]) as Gaussian prior. Thesystematic error budget has not been added to the covariances matrices from DESI yet. The results in terms of ΛCDM are shown in Figure 5.43. They are mostly consistent with each other, with Planck2018 [2] dominating. Using the multitracer analysis of the BGS, however, reduces the constraints slightly with respect to other tracers and shifts the central value a bit further away from the Planck values, as expected given the differences we highlighted in the previous Chapter and that we are still investigating.

We also look at the constraints for w_0w_aCDM, trying to constrain the dynamic Dark Energy model. w_0w_aCDM is a parametrisation of Dark Energy with an additional term w_a, which is responsible for the dynamic behaviour of Dark Energy, and in the ΛCDM model will be equal to zero. The overall equation of state then reads[4]:

$$w(z) = w_0 + \frac{w_a z}{1+z} \tag{5.7}$$

Using the w_0w_aCDM version of the emulator presented in Chapter 2, we unfortunately were not able to reach enough constraining power to produce reasonable estimations of w_0w_a with BGS DR1 alone. However, combining it with data from Planck[2], DES[3]

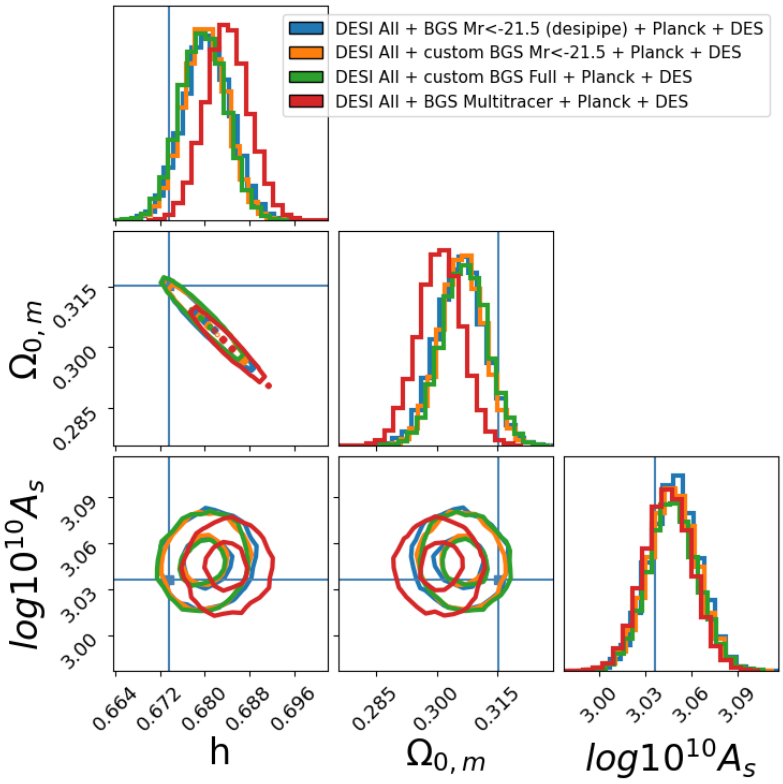

Figure 5.43: Constraints on cosmological parameters provided by combination of DESI data with various types of analysis with Planck2018[2] and DES data[3] .

and other tracers of DESI, we can put constraints on w_0w_aCDM. The preliminary results are shown in Figure 5.44.

We see that the inclusion of the Full-Shape constraints in general improves by almost a half the constraints on the equation of state of dark energy, with multitracer analysis improving them even slightly further with the data favouring the Dark Energy model with the decaying expansion.

That's all, folks!

In this manuscript we started with describing the basics of General Relativity and cosmology in the Introduction, where we also covered various basics of observational cosmology and clustering analysis. In Chapter 1 we have discussed the DESI experiment, the largest spectroscopic survey yet, the technical details related to the data acquisition and the data itself used in this work. In Chapter 2 we have presented the modern approaches to modelling the 2-point clustering statistics, and presented the neural network emulator tool that we developed, which significantly speeds up the computation of the 2-point statistics model. In Chapter 3 we have discussed the techniques of simulating the large-scale galaxy surveys using N-body simulations, and we also presented the effort to create mocks for

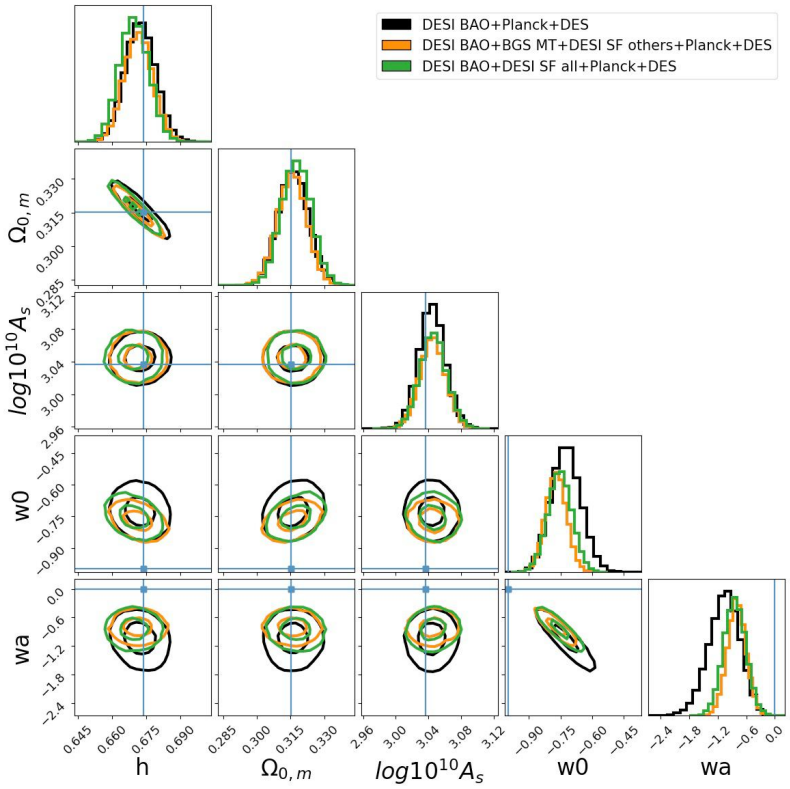

Figure 5.44: Constraints on cosmological parameters provided by combination of DESI data with various types of analysis with Planck2018[2] and DES data[3] in w_0w_aCDM.

the DESI BGS that we led. In Chapter 4 we have discussed the error estimation and covariance production for the DESI BGS, in particular we presented a hybrid approach for accurate covariance matrix that requires about 40 times less simulations than the traditional method. In Chapter 5, we have introduced various techniques of analysing the 2-point clustering statistics, from BAO-only to Full-Shape analysis, both single- and multi-tracer. We presented all the tests we performed on realistic mocks to validate and assess the performance of the tools we have developed in order to analyse the BGS DR1 data, and showing that these techniques allow to squeeze tighter constraints on a dataset plagued by many difficulties, from computational to statistical (cosmic variance), and we presented how those can be circumvented. Eventually, we showed that some cosmological constraints can be improved by almost 40% with respect to the official DESI DR1 BGS analysis. From the preliminary results, we can see that the improvement achieved can also affect the constraints on cosmological parameters when combined with other tracers and probes, thus being valuable assets for the large-scale cosmological analysis that aims at shedding light on the nature of dark energy and gravity.

Future prospects

Many of the techniques developed throughout this thesis can be generalised and improved further.

The emulator modelling approach we showed in Chapter 2 has a potential of going beyond the perturbation theory by employing N-body simulations for the computation of the expansion terms, and not relying on the analytical ones as it is done now. It could also be used for other clustering statistics, following the same expansion ideology. Additionally, the precision can be potentially raised by using alternative types of neural networks, like KANs[5].

The FitCov covariance approach has a potential to be generalised beyond 2-point correlation function to Fourier space, and maybe even to approaches like density split [6]. And of course, it has a potential to be further used for DESI BGS, both for DR2 and the final 5-year DESI BGS catalogue.

The set of GLAM simulations, beside being used for covariance modelling, thanks to its N-body nature, can be used for extensive n-point statistics tests, as well as covariances for those. Being produced with the same methods as the DESI high-fidelity mocks, gives it a possibility to be used for extensive tests of systematic effects. And the volume of the BGS boxes is enough in order to be used even for the final 5-year, DESI BGS catalogue if magnitude cut is taken into accout. Additionally to that, the smaller GLAM boxes, reproducing the full BGS Bright can also be of use, especially if the replication issues are figured out.

Finally, the multitracer analysis showed that it does improve constraints on the cosmological parameters. When it comes to inference beyond ΛCDM, however, more tests are required to get to a conclusion, but the preliminary results are very optimistic. When it comes to the data itself, the peculiar feature of the blue tracer quadrupole has to be studied, in order to eliminate as much as possible the possibility of the unwanted systematics effects affecting the cosmological results. We will focus on that goal and publish the multitracer analysis of the DESI BGS DR1 in coordination with the official DESI DR1 Full-Shape analysis.

In the future, more optimal data splits of the BGS data than a split in optical colour should be investigated, such as the density-split technique based on local density which is expected to provide tighter constraints on the cosmological parameters. In the meantime, DESI continues to take data, and more results are going to arrive from new methods that aim at efficiently and optimally extract robust cosmological information.

References

[1] Daniel Foreman-Mackey et al. "¡tt¿emcee¡/tt¿: The MCMC Hammer". In: *PASP* 125.925 (Mar. 2013), pp. 306–312. ISSN: 1538-3873. DOI: 10.1086/670067. URL: http://dx.doi.org/10.1086/670067.

[2] N. Aghanim et al. "iPlanck/i2018 results". In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. URL: https://doi.org/10.1051/0004-6361/201833910.

[3] DES Collaboration et al. *The Dark Energy Survey: Cosmology Results With 1500 New High-redshift Type Ia Supernovae Using The Full 5-year Dataset*. 2024. arXiv: 2401.02929 [astro-ph.CO]. URL: https://arxiv.org/abs/2401.02929.

[4] V. Motta et al. *Taxonomy of Dark Energy Models*. 2021. arXiv: 2104.04642 [astro-ph.CO]. URL: https://arxiv.org/abs/2104.04642.

[5] Ziming Liu et al. *KAN: Kolmogorov-Arnold Networks*. 2024. arXiv: 2404.19756 [cs.LG]. URL: https://arxiv.org/abs/2404.19756.

[6] E. Paillas et al. *Optimal Reconstruction of Baryon Acoustic Oscillations for DESI 2024*. 2024. arXiv: 2404.03005 [astro-ph.CO].

Acknowledgments

Habe nun, ach! Philosophie,
Juristerey und Medicin,
Und leider auch Theologie!
Durchaus studirt, mit heißem
Bemühn.
Da steh' ich nun, ich armer Thor!

Faust, Johann Wolfgang von
Goethe

First of all, I would like to thank my supervisor Pauline Zarrouk for these wonderful 3.5 years. I am extremely grateful to her, as she has given me an invaluable amount of understanding of physics and science. Without her this work would have been impossible.

I would also address my biggest thanks to Peder Norberg and Shaun Cole, who were following my work for all these years, and provided some of the most valuable input and advice on every step of this journey.

I am also extremely grateful to Alex Smith and Paco Prada, whose help was crucial when working with the simulations, and without whom that part of the thesis wouldn't have been there at all.

And of course DESI, the collaboration which I was really happy to be a part of.

I would like to also thank the LPNHE and cosmology group of it especially for the great atmosphere. Separately, I would like to thank the PhD and postdoc students, for their warm welcome and the sense of community they upheld throughout these years.

And of course, thank you for the jury, for reading and reviewing my manuscript, and for their very pleasant feedback!

Ну и из личного, спасибо всем кто поддерживал меня за пределами рабочего пространства (а некоторые и в его пределах) в течении всего этого пути: моим родителям, сестре, Владе, mes colocs, Leander, Sixtine, Clemence, Саине, Филиппу, Степе, Алисе, Hannah, Anthony, Doriane, Диме, Насте, Гоше и многим многим другим! Без вас ничего этого не было бы!

Publications

[1] Svyatoslav Trusov et al. “The two-point correlation function covariance with fewer mocks”. In: *Monthly Notices of the Royal Astronomical Society* 527.3 (Nov. 2023), pp. 9048–9060. ISSN: 0035-8711. DOI: 10.1093/mnras/stad3710. eprint: https://academic.oup.com/mnras/article-pdf/527/3/9048/54731887/stad3710.pdf. URL: https://doi.org/10.1093/mnras/stad3710.

[2] Svyatoslav Trusov et al. 2024. arXiv: 2403.20093 [astro-ph.CO].

[3] Svyatoslav Trusov et al. “Multitracer analysis of DESI DR1”. (in prep.) 2024.

[4] C. A. Dong-Páez et al. “The Uchuu–SDSS galaxy light-cones: a clustering, redshift space distortion and baryonic acoustic oscillation study”. In: *Mon. Not. Roy. Astron. Soc.* 528.4 (2024), pp. 7236–7255. DOI: 10.1093/mnras/stae062. arXiv: 2208.00540 [astro-ph.CO].

[5] DESI Collaboration et al. 2024. arXiv: 2404.03002 [astro-ph.CO].

[6] DESI Collaboration. “DESI Full-shape analysis”. (in prep.) 2024.

[7] DESI Collaboration et al. *DESI 2024 III: Baryon Acoustic Oscillations from Galaxies and Quasars*. 2024. arXiv: 2404.03000 [astro-ph.CO]. URL: https://arxiv.org/abs/2404.03000.

[8] DESI Collaboration et al. *DESI 2024 IV: Baryon Acoustic Oscillations from the Lyman Alpha Forest*. 2024. arXiv: 2404.03001 [astro-ph.CO]. URL: https://arxiv.org/abs/2404.03001.

Corona Astralis

A poem by Maximillian Voloshin, translated to English by Alex Romanovsky

To realms of love we are delusive comets
Locked out of the trusted orbits' path.
The truth of dreams escapes the earthly wrath,
The voices of the suns of midnight call us.

Evaded holy bath in Lethe's cold course,
Our spirit's poisoned, recollections tough,
We take the extraanimary scuff
Of the expelled, the strangers, the uncommons.

The sighted one, but dazzled by the day,
Who is alive but born for prison stay,
Who praises earth as holy relegation,

Who visions dreams and recollects the names,
Not pleasure of reunion he gains
In love, but secret joys of separation.

1.

To realms of love we are delusive comets,
An axis dashed through crystal sphere keys.
From fire clouds, heavenly unease,
From cosmic storms we bring our glimmer covert -

And spread it further... Let the skies of cobalt
Depict us as a sword to earthly peace, -
To sun we run, like Icarus of Greece,
Unraveling our wind and flare coat,

And touch it - but, surprising, run away,
To night eternal from the light of day

In our one-way parabolic departure.

Our spirit rides us - not a rim's enough -
Through the unsetting sunsets' purple parcher,
Locked out of the trusted orbits' path.

2.

Locked out of the trusted orbits' path,
Unmatched in prayer books of perfect order -
Deprived of earth we'll be at earthly border
By earthly servants of the earthly math.

Insane's our incense, and our ship's a lath,
Like bees gone stray, we seek our swarm by odor.
We passed between our warder and rewarder
And city fire fills our sail with laugh.

For breath of storms' mysterious appeal,
By scrolls of trails, by tangled road turns
We hasten, and our way is hard and stern.

Epigraph for Chapter 3

So let the thunders ring the clouds' peal,
Let doubts swirl embittering and tough!
The truth of dreams escapes the earthly wrath.

3.

The truth of dreams escapes the earthly wrath,
In the brocade of rays, the dawns retire,
The purl of mornings joins the daylight choir
The wane of moon will molder and burn off.

The braids of light, the olive of the dove
Old ripple grinds to beads in gentle gyre,
But Tabor nights we worship and aspire
Will outlive the lower solar craft.

Our eyes resist to noonday aspirations
Of desert stiffness, topaz constellations,
Or resin streams, or rays of golden shine.

The day of night unfading is our compass.
In moonlight silk, like servants of the shrine,
The voices of the suns of midnight call us.

4.

Epigraph for Introduction

The voices of the suns of midnight call us...
Our eyes are lost in telescopes' wells.
The stars' and planets' diamonds constell
In nebulae's and clusters' whirls and corals.

From Alpha Canis to the Capricornus
To Seven Sisters to the Argo's Sail
They cross the heavens telling their tale,
The seekers, perseverers and owners.

Epigraph for Chapter 1

Oh dust of planets! Swarm of holy bees!
I measured, weighed and totaled all of these,
I gave them names, and balances, and contours..

But knowledge made not stellar fear fade.
Our memories of darkest ages stayed,
Evaded holy bath in Lethe's cold course!

5.

Evaded holy bath in Lethe's cold course,
We left the bleach to calmness of the night.

The wellspring of amnesia we denied,
We pledged no vow. We have not been collared.

Epigraph for Chapter 4

The circuit's cut. The binding spells are quartered.
When, to the rest, the day is turquoise bright
And shining creeks in meadows never hide,
We see the lights astray on every corner.

The rustle of cane, the will-o'-wisp of swamps
The useless wind entangles, and stomps, and romps,
And brooms the helpless flock of Kore's kingdom,

Pelides guards, as if his eyeballs starve...
No honey cures us, no scent of linden;
Our spirit's poisoned, recollections tough. 6.

Our spirit's poisoned, recollections tough.
Our spirit grew from darkness, herb-resembling.
It bears venom of the tomb-resenting
And time-resenting womb of undercroft.

But such on earth impenetrable raft
No porphyry, no marble can assemble
That would delay the fury, bound, settle
The lava flows in our vessels pathed.

Oh graves o' worlds! Of suns forever set urn!
The corpse of moon, the lifeless face of Saturn
The brain remembers, heart retains the snuff.

The mind developed in the stellar crashes,
But spirit's buried in the heap of ashes;
We take the extraanimary scuff!

7.

We take the extraanimary scuff,

The weight of dolor, poison of the fire.
The waving flag of all the griefs' empire
Is rustling in the yearning, mourning puff.

But still, despite the wounds, the fire gruff,
The flesh that lets us barely respire,
Laocoön pulls snakes like strings of lyre
Yet not a word he's saying; not a half.

We'll give up not the glory of the pain,
Nor joys of prison, nor pride of bitter chain,
Nor elevation of the doom and jail

For Lethe's peace and all the worldly romance!
We bear to the world the Holy Grail
Of the expelled, the strangers, the uncommons.

8.

Of the expelled, the strangers, the uncommons,
Who longed for being but could not become,
The heritage of song is never calm.
Bird's is the nest, beast's lair, ours scorn is.

Debts never paid, gifts wasted, eggs uncoddled,
Path never stepped surrendered us to harm
From all trails' mists, from waves of any palm -
The honey spilt, confessions unrecorded.

Resurge to strive, to seek, to find yourself,
To love your shame, its humble, bitter scent,
Fall to the ground, search for desert dew,

Come to the strangers' settlements and pray
For broken bread; become a rhapsode new,
The sighted one, but dazzled by the day.

9.

Epigraph for Chapter 5

The sighted one, but dazzled by the day,

Conceives the voices, words and chains of reasons,
The body odors, rustle of trees arisen,
The secret lace beside and far away,

As Phoebes leaves them never in dismay
But serves to them the wisdom of the lizard.
The manger holds the God. The cave of prison
Becomes the Christmas cave, the Holy Tray.

Great Mother Night, in her dark womb a-bearing
The precious Fruit the Father did return,
Bestows treasures on the one in turn

Who was, by jealous Sun, expelled to dreary
And lifeless land, and heartless forces' play,
Who is alive but born for prison stay.

10.

Who is alive but born for prison stay,
Can see the edges of the painted coffin,
The Sun's canoe, Anubis's muzzle scoffing,
And fields of corn in orderly array.

Bulls plough, sickles harvest, flail in play,
Rafts slide along, birds nest, beasts nap as often;
That's what he sees as shroud wrinkles soften
In flips of days and flips of people's way.

Without joy, without grief and tears
He sees the people's vain and restless fears,
No second thought, nor even question 'why'.

Without being, will and aspiration
He savors peace the other would deny
Who praises earth as holy relegation.

11.

Who praises earth as holy relegation,
Rejects the spacious meadows' remit.

He meets in every instant, every beat
The distant planets' sparks of relivation,

As if the dead demand commemoration
And he is scraping glyphs on stone crypt;
The holy letters on the scratched concrete
Of known look can't form a combination.

He raises dust of earthly roadways,
Apostate priest, a deity free from praise,
In all the things old ornaments revealing,

The one who puts decay to knees and tames,
Who sees the Death - and lowers eyes revering,
Who visions dreams and recollects the names.

12.

Who visions dreams and recollects the names,
Who hears grass stalks' interrupting speeches,
He who conceives the knocking future's breaches,
To whom the sea is singing when inflames;

Who bears thought, like coat, on the frames,
Who with the soil his own soul bleaches,
Who lit his candles to the holy teachers,
Who pulled the cover from adored remains,

Who did not turn his feet to earthly pleasure
Of sisters' dance, or lunacy divine,
Who to the wine press wasted no vine,

Who, Orpheus-like, not a single measure
Held off from crossing, but returned in vain,
Not pleasure of reunion he gains.

13.

Not pleasure of reunion he gains
Who scorned the sweet oblivion from passion,
Who never knew the bodily concession,
Who drank no wine that places deadly stains.

From hope and achievement he restrains
His busy shoulders, turning to recession,
And takes no bonds, and quenches the compassion
The Moon ignites to forge our lively chains.

His own grief that he cannot divide,
Like ripple of seascape desolate and wide,
Will not be shared. Vinegar to him

Is dew. He chooses not alleviation
By peaceful waters at the final rim
In love, but secret joys of separation.

14.

In love, but secret joys of separation,
Daydreaming cinder, meeting pain we hold.
We shall not tread the moonlit linen fold,
Nor lock our lips in silent decoration.

Epigraph for Chapter 2

We treasure not our needless revelations,
Escape from precious arms to visions cold,
Face-blind, but in the names our trust is bold
On twisted trails of our peregrination.

From all degrees of darkness we are seen
By eyes of foes full of rage and sin,
With stars and sun no joy we share common,

But strain our path to stretches ever dark
Remembering our inner exile mark;
To realms of love we are delusive comets!