EPFL

Thèse n° 10 477

Probing the nature of Dark Energy through the study of large scale structures using spectroscopic surveys and their simulations

Présentée le 14 décembre 2023

Faculté des sciences de base
Laboratoire d'astrophysique
Programme doctoral en physique

pour l'obtention du grade de Docteur ès Sciences

par

Andrei VARIU

Acceptée sur proposition du jury

Prof. H. Brune, président du jury
Prof. J.-P. R. Kneib, directeur de thèse
Dr D. Schlegel, rapporteur
Prof. C. Bonvin, rapporteuse
Prof. M. Hirschmann, rapporteuse

École polytechnique fédérale de Lausanne

2023

"And God said, Let there be light: and there was light.
And God saw the light, that it was good: and God divided the light from the darkness."
— Genesis 1:3-4

To my parents...

Acknowledgements

First and foremost, I would like to thank my parents – without whom nothing would have been possible – Floare-Veronica and Vasile, not only for their love, support and trust during my studies, but also for their guidance since I was a child. I remember dismantling toys with my father's tools trying to understand how they work and instead of being mad, they would attribute this behaviour to curiosity and development of skills. Back then, I did not know what physics was, I only had curiosity. It was my physics teacher, Viorel Solschi, who made me realise – with his funny accent and amazing story-telling capabilities – that if I want to understand how the Universe works, physics represented the perfect tool. During this journey, I was accompanied by my dear old brother-like friend Alexandru-Ioan Pop with whom I had the most interesting and long discussions starting from the most banal subjects ("Cum fac oile în Ardeal?", a Romanian joke) to the deepest questions about Life and the Universe. At the same time, I thank my dear friends Lorena Copil, Diana Bogdan and Ioan Buftea who laid down fundamental bricks in the construction of today's version of me. We all split ways for our bachelor's studies, but they remain important pillars in my life on whom I can always count. Furthermore, I would like to thank my lovely family: grandmothers, uncles, aunts and cousins, who have always been close to me.

During my bachelor's studies, I met Silvia Georgescu and Sebastian Micluța-Câmpeanu that were my "partners in crime" in terms of asking an incommensurable number of questions to the incredible professors that were patient enough to answer all of them. I want to thank all of them for motivating me to push my limits during those years. Out of all the professors, I would like to mention Prof. Ionel Lazanu, who supervised my bachelor's thesis. He knew how to challenge and advise me so that I could take the best decisions forward. As it happens, sometimes you need to have some luck and meet the right person at the right time: while I was barely starting my first year of bachelor's studies, Tudor Pahomi was already applying for master's studies, therefore he represented an example for me that I could follow and encouraged me to apply to EPFL. I, sometimes, think that without Tudor I would not have studied at EPFL nor pursued PhD studies in astrophysics. Thank you, my friend!

At EPFL, I met Prof. Jean-Paul Kneib, who became my PhD supervisor, for which I am very grateful. Apart from helping me develop independence and connections, I have been strongly influenced by Jean-Paul's calmness and positivity. While the road towards the PhD was

sometimes difficult, his "I think it's fine, you can do this", in a calm tone, gave me confidence, motivated me and decreased – at least partially – the stress. On the same page, I could not be more thankful to Cheng Zhao, who closely helped me on my PhD projects. Part of the same 3D-COSMO group, I would like to thank my colleagues and friends Daniel Felipe Forero Sánchez, Jiaxi Yu, Aurélien Verdier, Amelie Tamone and Tianyue Chen for the nice memories and their support!
Furthermore, I could not be more grateful for the LASTRO, where I had the chance of meeting Austin Peel, Cameron Lemon, Eric Paic, Aymeric Galan, Aris Tritsis, James Chan, Yoan Rappaz, i.e. outstanding lab colleagues who became close friends. I also want to thank Jennifer Schober, Favio Diaz, Utsav Akhaury, Frédéric Dux, Estrella Briant and Olivier Genevay for creating a familiar atmosphere in LASTRO.
Lastly, but not least importantly, dancing Salsa in Lausanne, I was able to meet important people that managed to bring balance in my life outside astrophysics: Luis Gutierrez, a close friend and mentor that made me laugh in three years more than in the rest of my life; Mădălina Melzer that has been like a sister, ready to support me during the more difficult periods; Kevin Jamolli, Mélanie Lafond, Mélanie Sautaux, Pascal Esquilat, who are all part of my Salsa family, always ready to rock the dance floor.
Infinitely many thanks to you all!

Lausanne, October 2023 A. V.

Abstract

Measurements of large-scale structure (LSS), as performed on the largest 3D map of over two million extragalactic sources from the Sloan Digital Sky Survey, together with measurements of the cosmic microwave background (CMB) anisotropies, are in complete agreement with a flat ΛCDM Universe. In this model, the accelerating expansion of the Universe is driven by dark energy (Λ), and galaxies are formed under the gravitational pull of cold dark matter. The precise nature of these two dark components remains unknown. The Dark Energy Spectroscopic Instrument (DESI) aims to unravel the mystery of the former by probing the Universe at different epochs through measurements of LSS. This thesis presents an overview of the necessary steps for studying LSS with spectroscopic surveys, along with my contributions toward building realistic galaxy simulations to estimate covariance matrices, as well as improving and developing models to constrain cosmological parameters from real data.

Using the Baryon Acoustic Oscillations (BAO) as a standard ruler, DESI aims to measure the distances of 40 million galaxies and quasars with a sub-percent precision. Achieving such level of precision requires a careful analysis of the systematic effects. Therefore, DESI has initiated a mock challenge to test different methods to construct covariance matrices, which are needed for estimating the precision of the measurements. Chapter 2 presents some techniques to build realistic galaxy simulations starting from simulated dark matter haloes, and how these simulations can be used to compute a covariance matrix. The last section shows that using a Halo Occupation Distribution (HOD) model to assign galaxies to the FastPM dark matter haloes, the resulting galaxy two-point clustering is consistent with the one of the reference N-body simulation. Moreover, the estimated sample covariance matrices are robust against the details of the HOD fitting at the scales of interest for LSS studies.

Chapter 3 is dedicated to the study of cosmic voids as tracers of underdense regions. Voids and galaxies have been part of multi-tracer BAO studies that have provided stronger constraints on cosmological parameters than galaxy studies alone. Nevertheless, voids require careful modelling due to the exclusion effect that affects their clustering. Therefore, the last section introduces two new numerical models of the void clustering that yield unbiased BAO measurements when subjected to a series of robustness tests. Moreover, they are preferred over the previous models, according to the Bayesian analysis.

Key words: Cosmology; Spectroscopic surveys; Numerical simulations; Large-scale structures; Baryon acoustic oscillations; Dark energy

Résumé

Les mesures de la structure à grande échelle (SGE), effectuées sur la plus grande carte 3D de plus de deux millions de sources extragalactiques du Sloan Digital Sky Survey, ainsi que les mesures des anisotropies du fond diffus cosmologique, sont en parfait accord avec un Univers plat ΛCDM. Dans ce modèle, l'expansion accélérée de l'Univers est régie par l'énergie sombre (Λ), et les galaxies se forment sous l'attraction gravitationnelle de matière sombre froide (en anglais Cold Dark Matter – CDM). La nature précise de ces deux composantes obscures reste inconnue. La collaboration Dark Energy Spectroscopic Instrument (DESI) vise à percer le mystère de la première en sondant l'Univers à différentes époques grâce à des mesures du SGE. Cette thèse présente une vue d'ensemble des étapes nécessaires à l'étude des SGE avec des relevés spectroscopiques, ainsi que mes contributions à la construction de simulations réalistes de galaxies pour estimer les matrices de covariance, ainsi qu'à l'amélioration et au développement de modèles pour contraindre les paramètres cosmologiques à partir de données réelles.

En utilisant les oscillations acoustiques baryoniques (OAB) comme règle standard, DESI vise à mesurer les distances de 40 millions de galaxies et de quasars avec une précision inférieure à un pour cent. Pour atteindre un tel niveau de précision, il faut analyser soigneusement les effets systématiques. C'est pourquoi DESI a lancé un défi scientifique pour tester différentes méthodes de construction de matrices de covariance, nécessaires à l'estimation de la précision des mesures. Le chapitre 2 présente quelques techniques pour construire des simulations réalistes de galaxies à partir de halos de matière noire simulés, et comment ces simulations peuvent être utilisées pour calculer une matrice de covariance. La dernière section montre qu'en utilisant un modèle de distribution d'occupation du halo (DOH) pour assigner les galaxies aux halos de matière noire – obtenus par le programme FastPM – la répartition statistique des galaxies qui en résulte est cohérente avec celle de la simulation de référence à N-corps. De plus, les matrices de covariance résultantes sont robustes par rapport aux modifications des paramètres du modèle DOH aux échelles d'intérêt pour les études SGE.

Le chapitre 3 est consacré à l'étude des vides cosmiques en tant que traceurs des régions sous-denses. Les études les plus récentes de OAB, qui incluent les galaxies et les vides cosmiques ont fourni des contraintes plus fortes sur les paramètres cosmologiques que les études de galaxies seules. Néanmoins, les vides nécessitent une modélisation soigneuse en raison de l'effet

d'exclusion qui affecte leur répartition statistique. C'est pourquoi la dernière section présente deux nouveaux modèles numériques décrivant statistiquement des vides qui donnent des mesures OAB non biaisées lorsqu'ils sont soumis à une série de tests de robustesse. En outre, ils sont préférés aux modèles précédents, selon l'analyse bayésienne.

Mots clefs: Cosmologie ; Enquêtes spectroscopiques ; Simulations numériques ; Structures à grande échelle ; Oscillations acoustiques baryoniques ; Énergie sombre

Rezumat

Măsurătorile marii structuri ale Universului (MSU), realizate cu ajutorul celei mai mari hărți 3D, care cuprinde peste două milioane de galaxii și care a fost creată de Sloan Digital Sky Survey, precum și măsurătorile anizotropiei fondului cosmic de microunde sunt în deplin acord cu un Univers ΛCDM plat. În acest model, expansiunea accelerată a Universului este cauzată de energia întunecată (Λ), iar galaxiile se formează sub atracția gravitațională a materiei întunecate (în engleză Cold Dark Matter – CDM). Natura exactă a acestor două componente întunecate rămâne, însă, necunoscută. Colaborația "Dark Energy Spectroscopic Instrument (DESI)" își propune să dezlege misterul energiei întunecate studiind istoria Universului prin măsurători ale MSU. Această teză oferă o prezentare de ansamblu a etapelor necesare pentru studierea MSU cu ajutorul măsurătorilor spectroscopice. Ea cuprinde contribuțiile mele la construirea unor simulări realiste de galaxii în vederea estimării matricilor de covarianță și contribuie la îmbunătățirea și dezvoltarea de modele pentru a constrânge parametrii cosmologici din măsurători.

Utilizând oscilațiile acustice ale barionilor (OAB) pe post de riglă standard, DESI își propune să măsoare distanțele a 40 de milioane de galaxii și quasari cu o precizie de sub un procent. Atingerea unui astfel de nivel de precizie necesită o analiză atentă a efectelor sistematice. Prin urmare, DESI a inițiat o serie de proiecte pentru a testa diferite metode de construire a matricelor de covarianță necesare pentru estimarea preciziei măsurătorilor. Capitolul 2 prezintă câteva tehnici de simulări realiste ale galaxiilor pornind de la simulări ale halo-urilor de materie întunecată și, de asemenea, explică modul în care aceste simulări pot fi utilizate pentru a calcula o matrice de covarianță. Ultima secțiune arată că, utilizând un model de tipul Halo Occupation Distribution (HOD) pentru a repartiza galaxii la halo-urile de materie întunecată – simulate utilizând aproximațiile programului FastPM – distribuția statistică a galaxiilor este în concordanță cu cea din simularea de referință. Mai mult, matricile de covarianță rezultante sunt robuste față de anumite detalii ale modelului HOD la scările de interes pentru studiile MSU.

Capitolul 3 este dedicat studiului vidurilor cosmice. Cele mai recente studii ale OAB, care integrează nu doar galaxiile, ci și vidurile în procesele de măsurare, au demonstrat îmbunătățiri ale măsurătorilor parametrilor cosmologici. Cu toate acestea, vidurile necesită o modelare atentă din cauza efectului de excludere care afectează distribuția lor spațială. Prin urmare,

ultima secțiune prezintă două noi modele numerice pentru a descrie statistic vidurile. Aceste modele permit studierea MSU utilizând OAB cu o mare acuratețe chiar și atunci când sunt supuse unei serii de teste de robustețe. În plus, analiza bayesiană arată o preferință pentru aceste două modele față de modelele anterioare.

Cuvinte cheie: Cosmologie; Măsurători spectroscopice; Simulări numerice; Marea structură a Universului; Oscilațiile acustice ale barionilor; Energie întunecată

Contents

List of Figures

List of Tables

1 Introduction

The main goal of this thesis is to show how the study of the large-scale structure (LSS) using the Baryon Acoustic Oscillations (BAO) as standard ruler allows us to understand the Dark Energy (DE) – for a schematic overview see Figure 1.1. To this end, I have contributed toward building realistic galaxy simulations to estimate covariance matrices, as well as improving and developing models to constrain cosmological parameters from real data. This chapter introduces fundamental notions about the Universe and the LSS. The second chapter explains how one can build galaxy simulations and how they can be used to compute covariance matrices that are needed to estimate the uncertainties of the final measurements. Lastly, the third chapter shows the cosmological results achieved using galaxies and cosmic voids and introduces two new robust numerical models for cosmic voids.

The current chapter begins with the presentation of the notions of space, time and gravity and how they evolved in time, reaching the modern's understanding through the General Theory of Relativity (GR). The space and time form the 4D space-time that adapts such that the maximum speed in vacuum is the speed of light c and it curves under the effect of massive objects, hence explaining the gravitational interaction (Section 1.1). GR allows for a natural description of the Universe at large scales through the Friedmann equations, where it can be seen as statistical homogeneous and isotropic (Section 1.2.1). In addition, GR gives the possibility that a cosmological constant Λ drives the current accelerated expansion of the Universe. Generalising the state equation of Λ, one obtains the more general concept of DE. Given that different kinds of DE can have distinct effects on the expansion of the Universe – hence on the measured cosmological distances – one can gain insight in the nature of DE by measuring distances (Section 1.2.2).

Sections 1.2.4 and 1.2.3 present a brief history of the early Universe starting with a phenomenological description of the inflation. The Big-Bang Nucleosynthesis is introduced as the mechanism that allowed for the creation of light elements such as deuterium, beryllium and lithium. In addition, the study of the primordial abundances of light elements can be used to measure the baryon mass density. Furthermore, the BAO are described as the propagation of the initial fluctuations – that could be explained by inflationary models – in the primordial plasma of

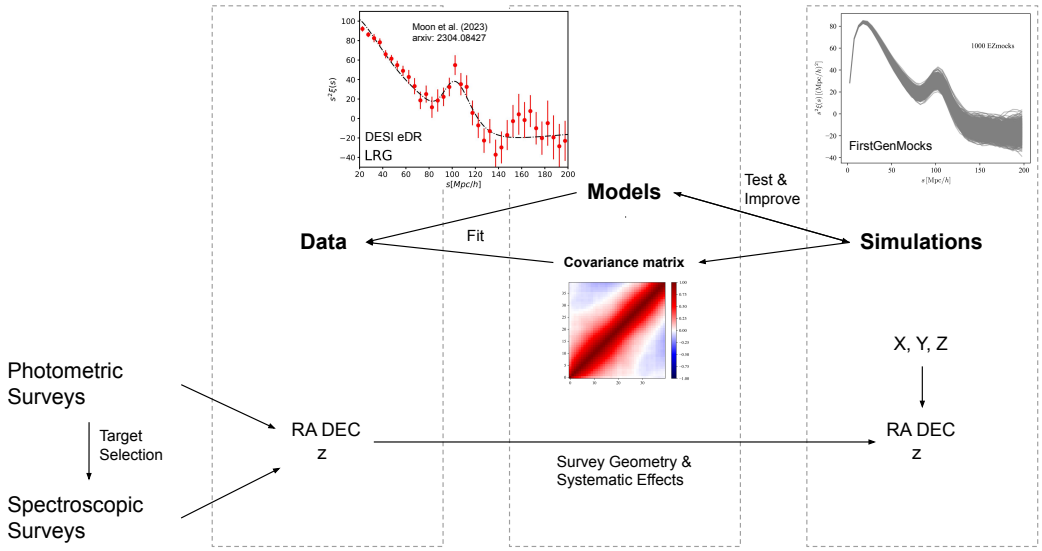


Figure 1.1: A scheme of the large-scale structure analysis. The two upper subplots represent the two-point correlation functions measured from the early Data Release of DESI (in red) and from 1000 galaxy simulations (in grey). The lower plot shows the covariance matrix obtained from the 1000 realisations of the two-point correlation function. See the text for more details.

baryons and photons that stops when electrons and ions recombine and hence imprints an overdense region at around 100 Mpc/h in comoving coordinates. This distance is a standard ruler that can be used to determine cosmological distances and can be measured in the galaxy clustering as shown in Section 1.4.4. After recombination, the photons are free to travel inside the Universe and today we detect them as the Cosmic Microwave Background (CMB), whose temperature fluctuations reveal the effect of the BAO and provide constraints on cosmological parameters.

The matter distribution at the epoch of recombination set the seeds for the LSS formation that is theoretically modelled as described in Section 1.3. Given the stochastic nature of the matter distribution, one must study the LSS statistically through the two-point or higher order correlation functions (and their Fourier counterparts). The first step is to include the physical phenomena that occurred before the recombination through the transfer function computed using rather sophisticated Boltzmann codes (Section 1.3.2).

In order to evolve the matter density field one can solve either perturbatively (Lagrangian or Eulerian perspective, i.e. Sections 1.3.4 or 1.3.5, respectively.) or numerically (e.g. N-body simulations, Section 1.3.6) the Vlasov equation (Section 1.3.3). An important limit of the perturbation theory is the shell crossing, i.e. when multiple streams of matter intersect. Nevertheless, models such as the spherical collapse can improve the LSS study beyond shell crossing since it can provide insight into the formation of dark matter haloes.

In N-body simulations, the dark matter haloes can be detected from the evolved dark matter

field using a variety of algorithms (e.g. Friends-of-Friends, spherical overdensity). The dark matter haloes are required to create realistic galaxy simulations using galaxy-halo connection models, as described in Section 2.2 part of Chapter 2. From the statistical point of view, galaxies and haloes as biased tracers of the dark matter field as described in Section 1.4.3.

Given that in practice, haloes are not directly observable, one creates 3D maps of galaxies (quasars, neutral hydrogen clouds as well) by performing photometric (Section 1.4.1) and spectroscopic surveys (Section 1.4.2). The clustering statistics of these 3D maps is computed as shown in Section 1.4.3 in order to detect the BAO signal and measure the cosmological parameters as explained in Section 1.4.4. To this end, one requires a covariance matrix to estimate the noise in the measurements (Section 2.1), where one method is to create many realistic galaxy simulations and compute their clustering statistics, as described in Section 2.1.1. In terms of realism, in addition to the matching clustering statistics, one has to apply the survey geometry on the cubic simulations, as shown in Section 2.3. Lastly, Section 2.4 – i.e. a submitted first-author paper – describes the galaxy catalogues and the corresponding covariance matrices obtained by applying a Halo Occupation Distribution model on the FASTPM dark matter haloes.

In addition to the BAO signal, the galaxy clustering is affected by the Redshift-Space Distortions (RSD, Section 1.4.5) that are induced by the galaxy peculiar velocities on the redshift measurements. Apart from being a probe of gravity models, RSD dilute the BAO signal. Nevertheless, the BAO reconstruction – introduced in Section 1.4.6 – can partially remove this effect and hence improve the BAO signal. Therefore, BAO reconstruction is standard process applied on galaxy catalogues for the BAO studies.

In the end, one requires a clustering model to fit the measurements together with an estimated covariance matrix. For this, Section 3.2 – i.e. a published first-author paper – introduces two new numerical models for the clustering of cosmic voids that are robust against different systematic effects and against the fitting interval. Section 3.1.1 introduces the concept of cosmic voids and then focuses on Delaunay Triangulation (DT) voids. Lastly, it presents the results of the latest multi-tracer BAO analysis of galaxies and DT voids based on BOSS and eBOSS data, where the combined sample yields better constraints on cosmological parameters than the galaxy sample alone.

1.1 Space, Time and Gravity

Along the history of physics the concepts of time, space and gravity have changed, depending strongly on the available experimental evidences and mathematical tools. Newton has introduced the laws of motion in a Euclidean 3D space in which information can travel instantaneously. In this context, the gravity was an attractive force between massive objects. The Special Theory Relativity (SR) has connected the space and time and provided a universal constant, i.e. the speed of light. Finally, GR explains how gravity and space–time are linked to the presence of mass.

The books: The Feynman Lectures on Physics[1] and the Mechanics Berkeley Course have been a great source of inspiration for the first two subsections (Feynman et al., 2006; Kittel, 1973). Schutz (2009); Weinberg (1972); Rich (2010) are the main references used for the last subsection.

1.1.1 The Newtonian perspective

At the time when Isaac Newton was alive, there were some important observations regarding the moving objects such as Galileo Galilei's law of inertia for horizontal motion and Kepler's laws of planetary motion. Additionally, the highest experimented velocities were of the order of hundreds of meters per second (e.g. cannon balls) which are thousands times smaller than the speed of light. The space was seen as a theatre stage where all motions and interactions occurred and time was the same for all observers.

Newton thus introduced the concept of absolute space in which his laws of motion would be true and all frames that are in relative uniform motion to absolute space would be an inertial frame. Moreover, all these inertial frames would share the so-called universal time. With this view of the world, he formalised what we now call the Newton's laws of motion:

1. A moving body at constant speed or at rest will continue moving at constant speed or remain at rest as long as it does not interact with other bodies.

2. The total force $\boldsymbol{F}$ that acts on a body changes the linear momentum $\boldsymbol{P}$ of the body: $\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}$.

3. When two bodies interact, they apply forces to one another that are equal in magnitude and opposite in direction.

It is important to emphasise that in a non-inertial reference frame these laws can fail, as for example on Earth, one needs to take into account the effect of the rotation by introducing the Coriolis pseudo-force – responsible for the direction of rotation of cyclones.

Mathematically, the inertial frames are linked by Galilean transformations. Supposing that the frame $R'(x', y', z', t')$ moves along the x axis of the reference $R(x, y, z, t)$ with velocity v_x with respect to R, as in Figure 1.2, the transformations are:

$$t' = t \tag{1.1}$$
$$x' = x - v_x t \tag{1.2}$$
$$y' = y \tag{1.3}$$
$$z' = z, \tag{1.4}$$

[1] https://www.feynmanlectures.caltech.edu/

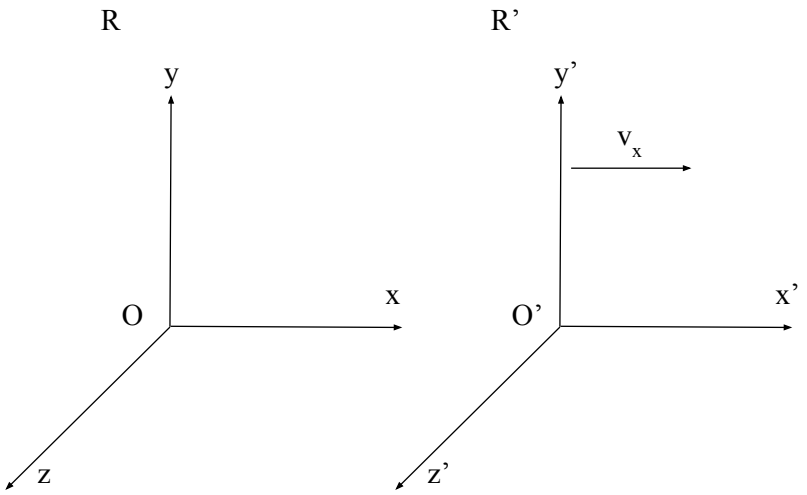


Figure 1.2: The two reference frames have relative movement. If the observer is found in reference R, the reference R' moves farther along the x axis with a velocity v_x.

where (x, y, z) and (x', y', z') are the 3D spatial coordinates of R and R', in which the time is described by t and t', respectively. In the case of a 3D displacement, one can trivially generalise the transformations. This is also known as the Galilean relativity.

Lastly, based on Kepler's laws of planetary motion and the previously mentioned laws, Newton was able to describe the gravity as an attractive force that acts between bodies with masses. Assuming two point sources with masses M and m separated radially by $\boldsymbol{r}$, the gravitational force $\boldsymbol{F}(\boldsymbol{r})$ is:

$$\boldsymbol{F}(\boldsymbol{r}) = -G\frac{Mm}{r^2}\frac{\boldsymbol{r}}{r}, \tag{1.5}$$

where $G = 6.67408 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ is the gravitational constant. Generally, if one wants to determine the gravitational pull exerted by a mass distribution $\rho(\boldsymbol{r})$, on a particle of mass m, one has to solve the Poisson equation for the gravitational potential ϕ:

$$\nabla^2\phi = 4\pi G\rho(\boldsymbol{r}) \tag{1.6}$$

and then obtain the force through the gradient of the potential: $\boldsymbol{F}(\boldsymbol{r}) = -m\nabla\phi$. This is mathematically possible because equation (1.5) has a zero curl.

1.1.2 The Special Theory of Relativity

In the second half of the 19th century, Maxwell has developed the dynamical theory of the electromagnetic field, unifying the electricity and magnetism. Moreover, these equations predicted the existence of electromagnetic waves. Later, Hertz's discovery of transverse elec-

tromagnetic waves which propagated at the same speed as light was a strong experimental support of Maxwell's theory and of the connection between electromagnetism and optics.

Given the fact that the propagation of waves had always involved a medium, it was natural to assume that light would also need a medium through which to propagate in vacuum. This was called the ether. On short, experiments such as Michelson-Morely have not observed evidence of this medium – more details can be found in Jackson (1999) and references therein. Additionally, Maxwell's equations were not invariant under Galilean transformations. These arguments have lead Albert Einstein to develop the SR, based on two postulates:

1. The laws of physics are invariant in all inertial frames of reference.
2. There is a finite universal limiting speed for physical entities in every inertial frame, which is equal to the speed of light c in vacuum.

Using the two postulates, one can derive the Lorentz transformation of coordinates between two inertial reference frames, that have a relative movement as the one illustrated in Figure 1.2:

$$t' = \gamma\left(t - \frac{v_x x}{c^2}\right) \tag{1.7}$$

$$x' = \gamma\left(x - v_x t\right) \tag{1.8}$$

$$y' = y \tag{1.9}$$

$$z' = z, \tag{1.10}$$

where

$$\gamma = \frac{1}{\sqrt{1 - \frac{v_x^2}{c^2}}} \tag{1.11}$$

and c is the speed of light in vacuum. Under these transformations, Maxwell's equations remain invariant to the change of inertial reference frame. Analysing them, one can observe that for a $v_x << c$, $\gamma \approx 1$ and the Lorentz transformations become the previously mentioned Galilean transformations, equations (1.1)-(1.4). This shows, that Newton's laws and vision about the laws of physics are not incorrect, but rather a good approximation when the relative speeds are much lower than the speed of light.

The second postulate breaks the concepts of an absolute space and a universal time. It implies that the space and time have to adapt so that no velocity in an inertial frame becomes larger than c. In other words, space and time become connected and thus, one can define the space-time infinitesimal invariant ds:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - |\mathrm{d}\boldsymbol{x}|^2. \tag{1.12}$$

If we imagine a particle with a velocity $\boldsymbol{u}$ with respect to an inertial reference R, the infinitesi-

mal change in position $\mathrm{d}\boldsymbol{x} = \boldsymbol{u}\mathrm{d}t$ and thus:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2(1-\beta^2), \tag{1.13}$$

where $\beta = u/c$. Analysing the particle in its own reference frame, then $\mathrm{d}\boldsymbol{x}' = 0$ and $\mathrm{d}t' \equiv \mathrm{d}\tau$, where τ is called the proper time of the particle. Thus $\mathrm{d}s = c\mathrm{d}\tau$, with $\mathrm{d}\tau = \mathrm{d}t\sqrt{1-\beta^2}$. The space-time invariant $\mathrm{d}s$ shows whether two events can or cannot be causally connected – timelike separation with $\mathrm{d}s > 0$ and spacelike separation with $\mathrm{d}s < 0$, respectively – since the physical interactions cannot propagate from one point to another with velocities greater than c. In addition, for $\mathrm{d}s = 0$, two events can be connected only by light signals.

Given the connection between space and time, one can define a 4D position contravariant vector, whose elements ξ^α are (ct, x, y, z). Using the summation convention for repeated indices, one can rewrite equation (1.12):

$$\mathrm{d}s^2 = \eta_{\alpha\beta}\mathrm{d}\xi^\alpha\mathrm{d}\xi^\beta, \tag{1.14}$$

where $\eta_{\alpha\beta}$ is the Minkowski metric tensor[2]:

$$\eta_{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{1.15}$$

Furthermore one can differentiate the 4D position to obtain the 4D velocity and acceleration:

$$u^\alpha = \frac{\mathrm{d}\xi^\alpha}{\mathrm{d}\tau} \qquad\qquad a^\alpha = \frac{\mathrm{d}^2\xi^\alpha}{\mathrm{d}\tau^2}. \tag{1.16}$$

Finally, the SR is the standard framework with which all new theories have to be consistent. Precise atomic phenomena, nuclear physics and high-energy physics use and depend on the formalism of SR. In this case, it is obvious to ask for the theory of gravity to be consistent with SR.

1.1.3 The General Theory of Relativity

The SR has been an great step forwards in understanding the fundamental physical laws, however it put Newton's law of gravity under great scrutiny. Newton's gravitational force does not depend on time and it implies that changes in the matter distribution would be instantaneously felt in the gravitational potential. The latter observation is in contradiction to the postulate of SR.

The solution was inspired by a fact observed even by Galileo Galilei: objects of different masses

[2] Note that there are also other sign conventions for the metric.

fall with the same acceleration in a gravitational field. This implies that the inertial mass, m_{I} that appears in the second Newton's law, and the gravitational mass m_{G} part of equation (1.5) are equal. Hundreds of years later, the Baron Eötvös de Vásárosnamény has reconfirmed the $m_{\mathrm{I}} = m_{\mathrm{G}}$ equality using pendulums and a torsion balance to high precision (see Weinberg (1972) for more details). This experimental equality has lead Einstein to postulate what we now call Weak Equivalence Principle: Trajectories of particles in the gravitational field are locally indistinguishable from the trajectories of free particles as viewed from an accelerated reference frame. In other words, in a gravitational field there is an accelerated reference system in which the effect of gravity is cancelled. This is called the free-falling reference frame.

Consequently, if one describes mathematically the SR in an arbitrary coordinate system, the resulting mathematical tools can be used to describe gravity.

Metric

Given a gravitational field, the equivalence principle implies that mathematically, in the free-falling frame, the suitable metric is the Minkowski one, equation (1.15). This means that in a general coordinate system (random reference frame) such as $x^\mu = x^\mu(\xi^\alpha)$, the space-time infinitesimal invariant becomes:

$$\mathrm{d}s^2 = \eta_{\alpha\beta}\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu. \tag{1.17}$$

Therefore, in this frame, the distances are not computed using the Minkowski metric, but rather using the metric:

$$g_{\mu\nu} = \eta_{\alpha\beta}\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}, \tag{1.18}$$

where $g^{\lambda\nu}g_{\nu\mu} = \delta^\lambda_{\ \mu}$ is Kronecker delta, which is 0 for $\mu \neq \lambda$ and 1 otherwise. There are three important observations:

1. $g_{\nu\mu}$ is a description of the effect of the gravity in x^μ coordinates.
2. x^μ are not some special coordinates, as one can mathematically always describe the same physical system in a new set of coordinates.
3. In the presence of gravity, the metric is not globally Minkowskian. One can choose a coordinate system where the metric becomes $\eta_{\alpha\beta}$ on a line, but not on the whole space.

Einstein's equations

In Newtonian framework, the mass distribution $\rho(\boldsymbol{r})$ is determining the potential ϕ through the Poisson equation. In a similar way, the energy-momentum tensor (or stress energy tensor, the relativistic version of the matter distribution) $T_{\mu\nu}$, affects the metric $g_{\mu\nu}$ through the

Einstein's equations:

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{1.19}$$

where $G_{\mu\nu}$ is the Einstein's tensor, $R_{\mu\nu}$ is the Ricci tensor and $R = g^{\mu\nu} R_{\mu\nu}$ is the Ricci scalar.

These equations can be obtained by minimising the Einstein-Hilbert action, which is defined using the simplest (non-trivial) available scalar that includes up to second order derivatives of the metric $g_{\mu\nu}$, i.e the Ricci scalar. In addition, it turns out that one can add a constant Λ in the previous action resulting into

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} - \Lambda g_{\mu\nu}. \tag{1.20}$$

Here, Λ is the cosmological constant that was initially introduced by Einstein to theoretically describe a static and eternal Universe (according to the available observations at that time). However, currently, Λ is used to explain the accelerated expansion of the Universe.

Einstein's equations are set of ten coupled equations that require both initial and boundary conditions. However, the ten coupled equations are reduced to only six independent equations using the Bianchi identities. These equations determine the six independent components of the metric tensor.

Finally, in the case of $\Lambda = 0$ and the weak field approximation ($|\phi| \ll c^2$, $|\boldsymbol{v}| \ll c$, i.e. the gravitational field cannot impose velocities close to the speed of light c), GR reproduces the Newtonian gravity. Mathematically, Einstein's equations become the Poisson equation (1.6) for the gravitational potential ϕ generated by a mass density ρ.

Geodesic equation

In GR, the movement of a free-particle in a curved space-time – i.e. in the presence of gravity – is described by the geodesic equation (the analogue of the second Newton's law):

$$\frac{d^2 x^\mu}{d\lambda^2} + \Gamma^\mu_{\alpha\beta} \frac{dx^\alpha}{d\lambda} \frac{dx^\beta}{d\lambda} = 0, \tag{1.21}$$

where λ represents a general parameter for both massive and massless particles. On one hand, $\lambda = s$ for massive particles. On the other hand, for massless particles $\mathrm{d}s = 0$, thus λ represents a random possible parametrization. For a massive particle, the geodesic equation provides the minimum proper time trajectory of that particle in a gravitational field.

Additionally, $\Gamma^\mu_{\alpha\beta}$ are called the Christoffel symbols and have the following form:

$$\Gamma^\mu_{\alpha\beta} = \frac{1}{2} g^{\mu\rho} \left(\frac{\partial g_{\rho\beta}}{\partial x^\alpha} + \frac{\partial g_{\rho\alpha}}{\partial x^\beta} - \frac{\partial g_{\alpha\beta}}{\partial x^\rho} \right). \tag{1.22}$$

They include information about the local gravitational interaction and the fictitious forces (e.g centrifugal, Coriolis) that arise when using non-inertial reference frames. Thus Γ becomes zero in the absence of gravity and in an inertial frame.

As a short intuitive summary, let us imagine a particle of mass m in the gravitational field provided by the Earth. In practice, one should obtain the metric in the presence of the Earth using the Einstein's equations. Furthermore, the resulting metric should be introduced in the geodesic equation to predict the movement of the particle under the effect of gravity. In case the particle is under the influence of other forces (e.g. electromagnetic ones), they have to be mathematically included in the geodesic equation.

1.2 Modern Cosmology

The modern understanding of the Universe and its evolution is based on GR and statistical mechanics. Boltzmann equations of statistical mechanics describe the collective behaviour of matter and radiation – as there is no interest in the evolution of an individual particle at the scale of the Universe – in a perturbed space-time defined by Einstein's equations.

The theoretical description of the Universe has been guided and complemented by important observational discoveries over the last century: the expansion of the Universe by Hubble (1929); the requirement of the Dark Matter (DM) to explain the Galaxy rotation curve (Zwicky, 1933; Rubin & Ford, 1970), the observation of a strong gravitational lens by Walsh et al. (1979) and the anisotropies in the Cosmic Microwave Background (CMB) by Smoot et al. (1992); the accelerating expansion of the Universe (e.g. Perlmutter et al., 1999), that can be explained by the presence of the Cosmological Constant Λ as a DE. These observations have led to the development of the standard cosmological model Λ Cold Dark Matter (ΛCDM) of the Universe: the Universe is in accelerated expansion due to the dominating DE and contains DM, baryons and a cold sea of photons (CMB).

Given its expansion, Gamow (1946); Gamow (1948) have suggested that the Universe has started as a very dense and hot point-like region called big bang. Moreover, the expansion of the Universe provides a mechanism that can explain the observed cosmological abundance of the chemical elements. Today, it is called the Big-Bang Nucleosynthesis. Lastly, by including an inflationary period, one can explain the fluctuations in the CMB and thus the current structure in the matter distribution. Figure 1.3 illustrates the history of the Universe as function of time and temperature, from the hypothetical period of inflation until today.

1.2.1 The Friedmann Equations

The set of Einstein's equations can be used to understand very different scales, starting from the precession of the perihelion of Mercury, to black holes and even the Universe, (e.g. Schutz, 2009). By making assumptions about the studied problem, one may simplify the set of equa-

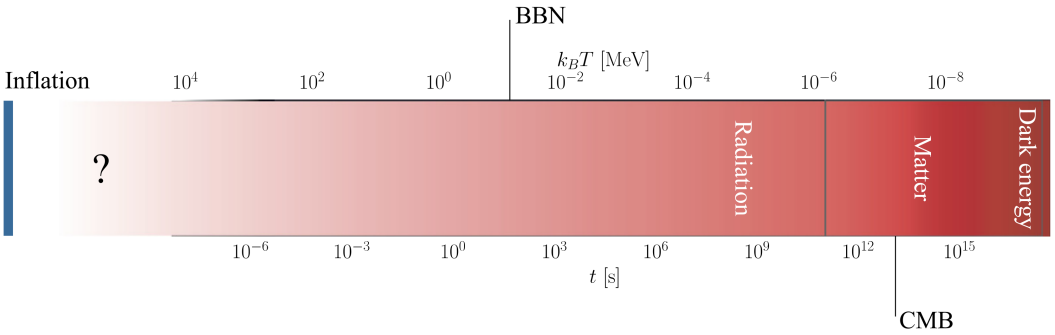


Figure 1.3: The history of the Universe as function of time (in seconds) and temperature (in MeV). Electron-Volt (eV) is a standard way to measure energies (e.g. kinetic, binding) and temperatures in particle physics. Figure 1.11 from Dodelson & Schmidt (2020).

tions and identify exact or near-exact solutions. At large scales – i.e. hundreds of megaparsecs[3] – the Cosmological Principle assumes that the Universe is statistically homogeneous and isotropic. This means that the Universe looks statistically the same from any point in space and in all directions.

The Cosmological Principle transforms the generic $g_{\mu\nu}$ metric – equation (1.18) – into the Friedmann–Lemaître–Robertson–Walker (FLRW) metric:

$$(\mathrm{d}s)^2 = (c\mathrm{d}t)^2 - a(t)^2\left[\frac{(\mathrm{d}\mathbb{X})^2}{1-k\cdot\mathbb{X}^2} + \mathbb{X}^2\left((\mathrm{d}\theta)^2 + \sin^2\theta(\mathrm{d}\varphi^2)\right)\right], \tag{1.23}$$

where c is the speed of light, t is the cosmic time, $(\mathbb{X},\theta,\phi)$ are the spherical comoving coordinates and $a(t)$[4] is the dimensionless scale factor related to the expansion of the Universe. Lastly, k is the curvature of the Universe that can be negative, 0 or positive, meaning an open, flat or closed Universe, respectively. One can perform a change of variables such as

$$\mathrm{d}\mathbb{R} = \frac{\mathrm{d}\mathbb{X}}{\sqrt{1-k\cdot\mathbb{X}^2}} \tag{1.24}$$

and obtain an equivalent representation

$$(\mathrm{d}s)^2 = (c\mathrm{d}t)^2 - a(t)^2\left[(\mathrm{d}\mathbb{R})^2 + S_k^2(\mathbb{R})\left((\mathrm{d}\theta)^2 + \sin^2\theta(\mathrm{d}\varphi^2)\right)\right], \tag{1.25}$$

where:

$$\mathbb{X} = S_k(\mathbb{R}) \equiv \begin{cases} \frac{1}{\sqrt{k}}\sin\sqrt{k}\mathbb{R} & k>0 \\ \mathbb{R} & k=0. \\ \frac{1}{\sqrt{|k|}}\sinh\sqrt{|k|}\mathbb{R} & k<0 \end{cases} \tag{1.26}$$

[3] one parsec, i.e. 1 pc ≈ 3.26 light – years

[4] Some conventions such as the one in Carroll (1997), perform the following substitutions $k \to \frac{k}{|k|}$; $\mathbb{X} \to \sqrt{|k|}\mathbb{X}$; $a \to \frac{a}{\sqrt{|k|}}$, leaving the form of the metric unchanged, but $k = -1, 0, 1$ and $a(t)$ has dimension of distance.

In order to determine the scale factor $a(t)$, one has to include the FLRW metric into Einstein's equations and solve them for a given stress-energy tensor $T_{\mu\nu}$. To this end, let us consider a perfect fluid[5] that fills the homogeneous and isotropic Universe. Therefore, the stress-energy tensor becomes:

$$T_{\mu\nu} = \left(\frac{p}{c^2} + \rho\right) u_\mu u_\nu - p g_{\mu\nu}, \tag{1.27}$$

where p is the pressure, ρ is the mass (energy) density, and u_μ is the velocity quadri-vector. Assuming that the fluid is at rest $u_\mu = (c, 0, 0, 0)$, the remaining components of $T_{\mu\nu}$ are $T_{00} = \rho$ and $T_{ij} = -p g_{ij}$. Finally, imposing the FLRW metric (1.23) and the stress-energy tensor of a perfect fluid at rest into the Einstein's equations (1.20), one obtains the two Friedmann's equations:

$$\frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2} - \frac{\Lambda c^2}{3} = \frac{8\pi \mathrm{G}}{3}\rho, \tag{1.28}$$

$$2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2} - \Lambda c^2 = -\frac{8\pi \mathrm{G}}{c^2} p, \tag{1.29}$$

with the three unknown functions of time: $a(t)$, $\rho(t)$ and $p(t)$. Due to the fact that a cosmological model (e.g. ΛCDM) imposes the components of the cosmological fluid, each with its specific equation of state $p = p(\rho)$, one can solve the resulting set of differential equations and find the time evolution of $a(t)$, $\rho(t)$ and $p(t)$, given the values of k and Λ.

ΛCDM standard model

The current measurements (see Table 1.1) are consistent with an expanding Universe described by a ΛCDM model. In this model, the cosmological constant Λ – that has been initially introduced as a property of space (i.e. a scaling of the metric $g_{\mu\nu}$) – is imagined as a dark energy (or fluid) with a constant density in time and a state equation:

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi \mathrm{G}} \qquad\qquad p_\Lambda = -\rho_\Lambda c^2. \tag{1.30}$$

Therefore, one can define a total energy density $\rho_{\mathrm{tot}} = \rho_\Lambda + \rho$ and a total pressure $p_{\mathrm{tot}} = p_\Lambda + p$.

In this case, the DE explains approximately 70 per cent of the total energy density in the Universe. The remaining 30 per cent is represented by the rest mass of the non-relativistic matter (CDM and baryons) with an equation of state $p_{\mathrm{m}} = 0$. Lastly, the relativistic components (e.g. photons or relativistic neutrinos) with an equation of state:

$$p_{\mathrm{rel}} = \frac{\rho_{\mathrm{rel}} c^2}{3} \tag{1.31}$$

constitute a negligible part of the total energy content of the Universe today. Nevertheless,

[5] A perfect fluid is entirely described by its rest frame mass density and isotropic pressure, i.e. it has no viscosity.

they have played an important role in the evolution of the Universe.

Other representations of the Friedmann's equations

On one hand, working with a total energy density and a total pressure $\rho_{\text{tot}}(t)$ and $p_{\text{tot}}(t)$, one can combine the two Friedmann equations[6] and obtain an equation equivalent to the first law of thermodynamics ($\mathrm{d}E + p\mathrm{d}V = 0$):

$$\frac{\mathrm{d}(\rho_{\text{tot}} c^2 a^3)}{\mathrm{d}t} + p_{\text{tot}} \frac{\mathrm{d}a^3}{\mathrm{d}t} = 0. \tag{1.32}$$

This equation can be obtained from the conservation of the stress-energy tensor $T_{\mu\nu}$, i.e. the $\nu = 0$ component of $\nabla_\mu T^{\mu\nu} = 0$. Moreover, it implies that the rate of change in the energy compensates the work done by the expansion of the Universe.

On the other hand, by subtracting equation (1.29) from equation (1.28) and using $\rho_{\text{tot}}(t)$ and $p_{\text{tot}}(t)$, one obtains:

$$\frac{\ddot{a}}{a} = -\frac{4\pi \mathrm{G}}{3}\left[\frac{3}{c^2} p_{\text{tot}} + \rho_{\text{tot}}\right]. \tag{1.33}$$

It is important to mention that only two out of the four equations (1.28), (1.29), (1.32) and (1.33) are mathematically independent. Nevertheless, depending on the nature of the studied problem, one might find useful to use one of them instead of another. For example, equation (1.32) helps us determine $\rho(a)$, as shown in the next paragraphs.

Solutions of the Friedmann equations

In what follows, we describe the evolution of a flat Universe (i.e. $k = 0$) using Friedmann equations throughout the radiation, matter, and dark energy dominated epochs, successively.

Radiation dominated Universe: In a Universe dominated by radiation $\rho_{\text{tot}} = \rho_{\text{rad}}$. As radiation is relativistic $p_{\text{tot}} = p_{\text{rad}} = \rho_{\text{rad}} c^2/3$. Thus, using equations (1.32) and (1.28) one obtains, respectively:

$$\rho_{\text{rad}} = \rho_{0\text{rad}} \left(\frac{a_0}{a}\right)^4, \qquad a(t) = a_0 \left(\frac{t}{t_0}\right)^{1/2} \tag{1.34}$$

and finally $\rho_{\text{rad}} \propto t^{-2}$, where the subscript 0 denotes the values measured today. The universe is expanding but given the fact that $\ddot{a} < 0$, the expansion is decelerating.

[6] After computing the time derivative of equation (1.28) one must subtract from it the equation (1.29).

Matter dominated Universe: In a Universe dominated by pressureless matter (baryonic matter and DM) $\rho_{\rm tot} = \rho_{\rm m}$ and $p_{\rm tot} = 0$. Therefore, equations (1.32) and (1.28) provide respectively:

$$\rho_{\rm m} = \rho_{0\rm m}\left(\frac{a_0}{a}\right)^3, \qquad a(t) = a_0\left(\frac{t}{t_0}\right)^{2/3} \tag{1.35}$$

Consequently, $\rho_{\rm m} \propto t^{-2}$, i.e. the density has the same time dependence as in a radiation dominated Universe. Moreover, the Universe is expanding, but the deceleration is larger than in the previous case.

Dark Energy dominated Universe: Lastly, in this case, $\rho_{\rm tot} = \rho_\Lambda$ is a constant. Using equation (1.28), it follows that:

$$a(t) = a_0 \exp\left(\sqrt{\frac{\Lambda}{3}}c(t-t_0)\right). \tag{1.36}$$

This means that the expansion of the Universe is exponentially accelerating.

Figure 1.4 shows the energy density of the three main components (matter, radiation and DE) as function of the age of the Universe. Additionally, it shows the evolution of the scale factor as function of the age of the Universe. One can observe that the radiation density has been the highest in the first 40 thousands years indicating a radiation dominated era. As a result, the scale factor increases as a power law function of time that is consistent with the one found in equations (1.34). Furthermore, for ≈ 10 billions years matter has dominated in the entire Universe. This imposes a scale factor that depends on time as a power law with an index consistent with 2/3 (see equation (1.35)). Lastly, in the last three billions years, the DE has driven an exponential expansion of the Universe.

General solution: Let us assume a ΛCDM Universe with a curvature k, matter, radiation and DE as Λ. One can treat the evolution of the matter and radiation as independent – i.e. matter does not turn into radiation and radiation does not transform into matter. As a consequence $\rho_{\rm tot}$ and $p_{\rm tot}$ are solutions of equation (1.32), given the equation of state of each component, where:

$$\rho_{\rm tot}(a) = \rho_{\rm m}(a) + \rho_{\rm rad}(a) + \rho_\Lambda \qquad p_{\rm tot} = p_{\rm m} + p_{\rm rad}(a) + p_\Lambda \tag{1.37}$$

and $\rho(a)$ for the three components are the previously found solutions. To summarise, equation (1.32) provides $\rho_{\rm tot}(a)$ and the state equations connect p and ρ. Therefore, there is only one remaining Friedmann equation that can be used from the system of all independent equations, in order to obtain $a(t)$. Introducing $\rho_{\rm tot}(a)$ in equation (1.28), the resulting differential

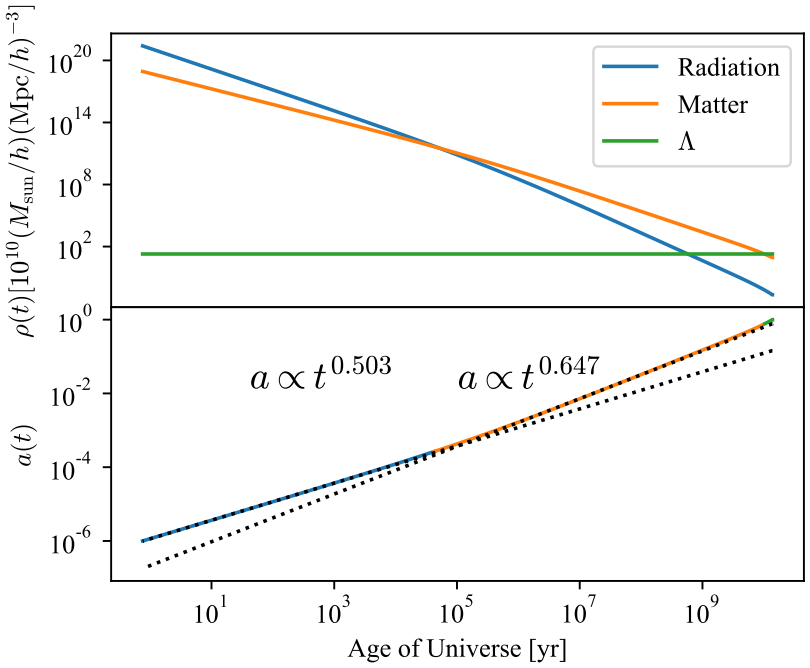


Figure 1.4: The energy density and the scale factor as function of the age of the Universe. The curves are computed using Planck 2015 (Planck Collaboration et al., 2016) flat ΛCDM cosmological parameters by employing ASTROPY (Astropy Collaboration et al., 2022) and NBODYKIT (Hand et al., 2018) PYTHON packages. The dotted curves represent the best-fitting curves for $[10, 10^3]$ years and $[10^6, 10^8]$ years intervals, in order to capture the radiation and matter dominated epochs, respectively.

equation of $a(t)$ can be solved given k, Λ, $\rho_{0\rm m}$ and $\rho_{0\rm rad}$[7]. In practice, one uses the redshift representation of equation (1.28), given $\rho_{\rm tot}(a)$.

Redshift representation of the first Friedmann equation

The redshift z shows how much the observed light wavelength ($\lambda_{\rm obs}$) is shifted towards the red end of the spectrum compared to the emitted light ($\lambda_{\rm em}$) when there is a relative motion between the observer and emitter. Mathematically, it is defined as follows:

$$1 + z = \frac{\lambda_{\rm obs}}{\lambda_{\rm em}}. \tag{1.38}$$

Due to the expansion of the Universe, the light from distant galaxies is redshifted. Thus, the scale factor is related to the redshift:

$$a = \frac{a_0}{1+z}, \tag{1.39}$$

[7]The values of k, Λ, $\rho_{0\rm m}$ are provided by cosmological measurements

where a_0 is the value of the scale factor today and $a_0 = 1$ by convention. Due to the fact that the redshift is a direct outcome of the Spectroscopic Surveys (see Section 1.4.2), equation (1.28) is usually presented under a different form. In the following paragraphs, we show how to transform equation (1.28), given $\rho_{\text{tot}}(a)$.

Firstly, the Hubble expansion rate $H(t)$ and the critical density $\rho_{\text{c}}(t)$ transform equation (1.28) into:

$$\frac{kc^2}{a^2H^2} = \frac{\rho_{\text{tot}}}{\rho_{\text{c}}} - 1, \tag{1.40}$$

where

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)}, \qquad \rho_{\text{c}}(t) \equiv \frac{3H^2(t)}{8\pi G}. \tag{1.41}$$

This intermediate representation suggests a relationship between the curvature k of an expanding Universe, ρ_{tot} and ρ_{c}:

$$\begin{cases} k > 0, \text{ closed universe if } \rho_{\text{tot}} > \rho_{\text{c}} \\ k = 0, \text{ flat universe if } \rho_{\text{tot}} = \rho_{\text{c}} \\ k < 0, \text{ open universe if } \rho_{\text{tot}} < \rho_{\text{c}}. \end{cases}$$

Furthermore, let us define:

$$\Omega(t) \equiv \frac{\rho(t)}{\rho_{\text{c}}(t)} \qquad \Omega_k \equiv -\frac{kc^2}{a^2H^2}. \tag{1.42}$$

Given ρ_{tot} from equation (1.37) and the previous definitions, equation (1.40) becomes:

$$H^2(t) = H_0^2 \left[\Omega_{0\Lambda} + \Omega_{0\text{m}} \left(\frac{a_0}{a}\right)^3 + \Omega_{0\text{rad}} \left(\frac{a_0}{a}\right)^4 + \Omega_{0k} \left(\frac{a_0}{a}\right)^2 \right], \tag{1.43}$$

where Ω_0 (for the curvature k and all components: radiation, matter, Λ) and H_0 represent the values measured today at $t = t_0$, H_0 is known as the Hubble's constant and $a_0 = 1$ is the value of the scale factor today by convention.

Finally, transforming the scale factor into redshift using equation (1.39), one obtains:

$$H^2(z) = H_0^2 \left[\Omega_{0\Lambda} + \Omega_{0\text{m}} (1+z)^3 + \Omega_{0\text{rad}} (1+z)^4 + \Omega_{0k} (1+z)^2 \right]. \tag{1.44}$$

In practice, one measures the values of the Ω_0 parameters and H_0, in order to determine $H(z)$ and thus the scale factor $a(t)$. However, $\Omega_{0\text{rad}} \approx 10^{-5}$, thus it is usually neglected in many cosmological calculations at low enough redshifts.

Figure 1.5 shows the evolution of Ω and Hubble parameters as function of the redshift for a flat (i.e. $k = 0$) ΛCDM model. One can observe that the sum of the three components is equal

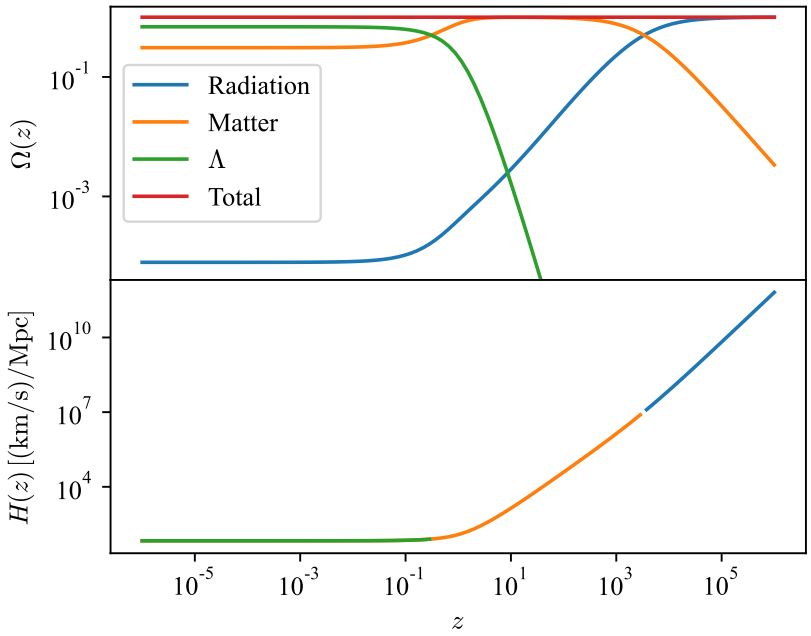


Figure 1.5: Ω and Hubble parameters redshift evolution. The curves are computed using Planck 2015 (Planck Collaboration et al., 2016) flat ΛCDM cosmological parameters by employing ASTROPY (Astropy Collaboration et al., 2022) and NBODYKIT (Hand et al., 2018) PYTHON packages.

to one, representing equation (1.40). Additionally, one can observe again the two transitions: radiation – matter at $z \approx 3400$ and matter – dark energy at $z \approx 0.3$. Lastly, the Hubble parameter has been decreasing with time, asymptotically approaching a minimum in the DE dominated era.

wCDM model

This model generalises the state equation for Λ – i.e. equation (1.30) – and thus the time dependency of the energy density to:

$$p = w\rho c^2 \qquad \rho(t) = \rho_0 \left(\frac{a_0}{a}\right)^{3(1+w)}. \tag{1.45}$$

For $w = -1$, one obtains the standard negative pressure and constant energy density of Λ.

The previous general state equation of DE reflects into the Friedmann's equation as follows:

$$H^2(z) = H_0^2 \left[\Omega_{0\mathrm{DE}}(1+z)^{3(1+w)} + \Omega_{0\mathrm{m}}(1+z)^3 + \Omega_{0\mathrm{rad}}(1+z)^4 + \Omega_{0k}(1+z)^2\right]. \tag{1.46}$$

This model is called $o w$CDM, while for $\Omega_k = 0$, the model is simply wCDM.

Figure 1.6 shows the effect of different w and $\Omega_{0\Lambda}$ values on the Hubble parameter. One can

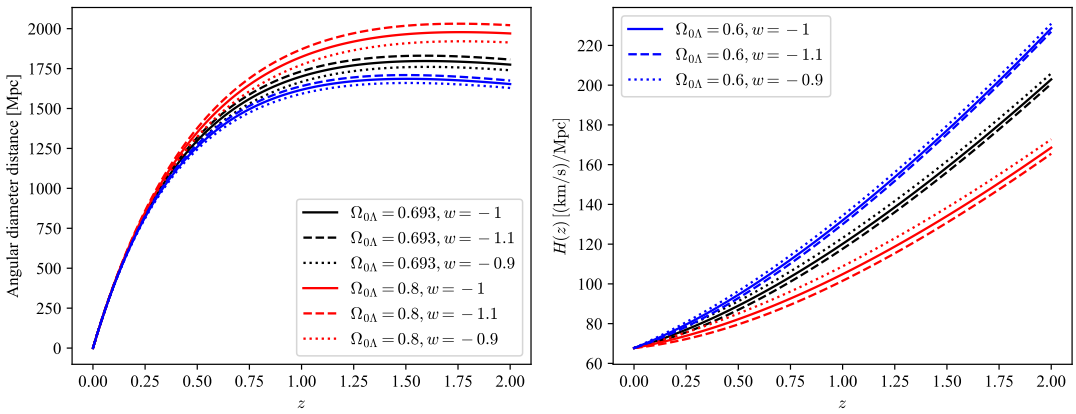


Figure 1.6: The redshift dependence of the angular diameter distance, equation (1.59), and the Hubble parameter equation (1.46). The curves are computed for a flat wCDM Universe, with the cosmological parameters shown in the legend and $H_0 = 67.7\,(\mathrm{km/s})/\mathrm{Mpc}$.

observe that the Hubble parameter is the most sensitive on the changes in the values at large redshift. Therefore, cosmological measurements that can put constraints on $H(z)$ at large redshifts can more easily differentiate between models of DE, see Section 1.4.2.

1.2.2 Distances in the Universe

An important issue in astrophysics and cosmology is the estimation of distances to far galaxies. The fact that the Universe is expanding provides us the intuitive connection between redshift and distance, however the actual conversion is not trivial and depends on the used technique and on cosmological parameters. This section is mainly inspired from Hogg (1999); Davis & Lineweaver (2004). Figure 1.7 shows all the discussed distances in this section.

Comoving distance

The expansion of the Universe is encoded in the scale factor $a(t)$ from FLRW metric – equation (1.23) – thus using $(\mathbb{X}, \theta, \phi)$ comoving coordinates that expand with the Universe, one can compute a differential comoving distance that is independent on the expansion:

$$\mathrm{d}x^2 \equiv \frac{(\mathrm{d}\mathbb{X})^2}{1 - k \cdot \mathbb{X}^2} + \mathbb{X}^2 \left((\mathrm{d}\theta)^2 + \sin^2\theta(\mathrm{d}\varphi^2)\right) = (\mathrm{d}\mathbb{R})^2 + S_k^2(\mathbb{R}) \left((\mathrm{d}\theta)^2 + \sin^2\theta(\mathrm{d}\varphi^2)\right). \tag{1.47}$$

The interpretation of this distance comes naturally when one looks at the trajectory of a photon. In this case, $\mathrm{d}s = 0$ and thus from FLRW one obtains:

$$\mathrm{d}x = \frac{c\,\mathrm{d}t}{a(t)}. \tag{1.48}$$

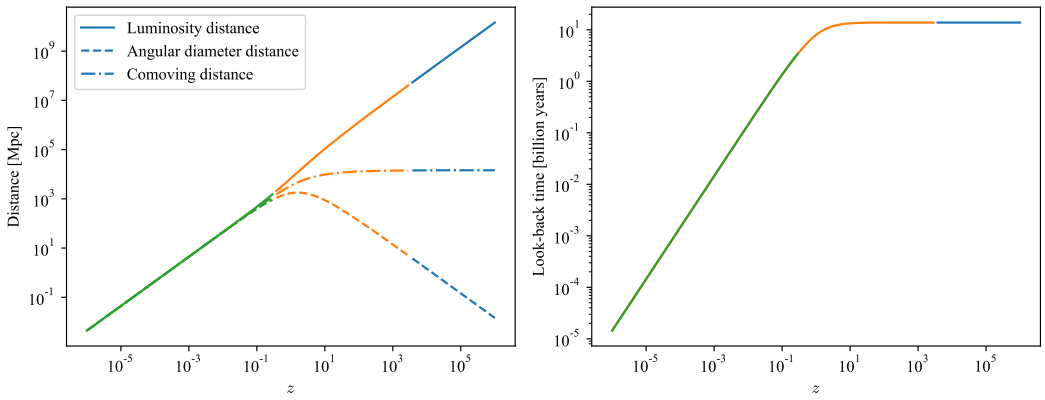


Figure 1.7: Cosmological distances and the look-back time. The curves are computed using Planck 2015 (Planck Collaboration et al., 2016) flat ΛCDM cosmological parameters by employing ASTROPY (Astropy Collaboration et al., 2022) and NBODYKIT (Hand et al., 2018) PYTHON packages. The colour convention is the same as in Figure 1.5: blue – radiation dominated epoch, orange – matter dominated epoch, green – Λ dominated epoch.

Practically, the comoving distance is the distance travelled by a photon in a given time, when the expansion is taken into account trough the scale factor. Consequently, by integrating this equation from $t_{\rm em}$, the time when a source emits a photon, until $t_{\rm obs}$, the time the observer detects the photon, one obtains the comoving distance between the observer and that emitter:

$$x = \int_{t_{\rm em}}^{t_{\rm obs}} \frac{c\,\mathrm{d}t}{a(t)}. \tag{1.49}$$

Particle horizon. The particle horizon is a specific case of equation (1.49):

$$\eta(t) = \int_0^t \frac{c\,\mathrm{d}t'}{a(t')}. \tag{1.50}$$

It is the distance light could have travelled since the beginning of the Universe until a time t. Practically, it denotes a region of causal contact at time t.

Radial comoving distance. In this scenario, the observer and the emitter are displaced only along the radial coordinate, thus $\mathrm{d}\theta = \mathrm{d}\phi = 0$. Therefore, $x = \mathbb{R}$. Additionally, given the fact that there is a one to one correspondence between the cosmological redshift and the time, one can change the time variable t to redshift z. Let us consider today ($a_0 = 1$) an observer on Earth receiving light from galaxies at different redshifts z. One can thus change the time variable t to z in equation (1.49):

$$x = \mathbb{R} = \int_0^z \frac{c\,\mathrm{d}z'}{H(z')} (= D_{\rm C}), \tag{1.51}$$

where D_C is the notation in Hogg (1999).

Transverse comoving distance. Considering two events at the same redshift (i.e. same radial distance, thus $d\mathbb{X} = d\mathbb{R} = 0$), but separated by an angle $\delta\theta$ on the sky, the comoving distance between the two is $(D_M \delta\theta \equiv)\mathbb{X}\delta\theta = S_k(\mathbb{R})\delta\theta$, where $\mathbb{X}$ is the transverse comoving distance and is shown explicitly in equation (1.26). In addition, D_M is the notation in Hogg (1999).

Proper distance

Let us imagine two cars travelling with a relative velocity between them, thus the distance between them changes in time. Consequently, if one wants to measure the separation between the two cars with a ruler, one can do it only at a fixed time. In a similar way, the proper distance between two far away galaxies – in an expanding Universe – is defined at a fixed time, so that one can "place" a ruler between the two galaxies and measure the distance. In this case, at a given time t, we have $dt = 0$, thus the $(ds)^2 = -a^2(t)(dx)^2$. The fact that the invariant is negative means that the two points separated by the proper distance r and comoving distance x:

$$r(t) \equiv a(t)x \tag{1.52}$$

are not in causal contact at time t. By convention, today at t_0, $a(t_0) = 1$, thus the proper distances are numerically equal to the comoving distances.

As previously mentioned, the comoving distance x does not change with the expanding Universe, however, galaxies have a peculiar motion due to the gravitational interaction. As a consequence x has also a time dependence, thus differentiating equation (1.52) with respect time t leads to

$$\dot{r}(t) = \dot{a}(t)x + a(t)\dot{x}(t), \tag{1.53}$$

where

$$v_{\rm rec}(t) \equiv \dot{a}(t)x(t) = H(t)r(t) \tag{1.54}$$

is the recession velocity (also called the Hubble's law) and $u(t) \equiv a(t)\dot{x}(t)$ is the peculiar velocity.

Luminosity distance

Given a light source of known luminosity L, the flux F measured by an observer is:

$$F = \frac{L}{4\pi D_L^2}, \tag{1.55}$$

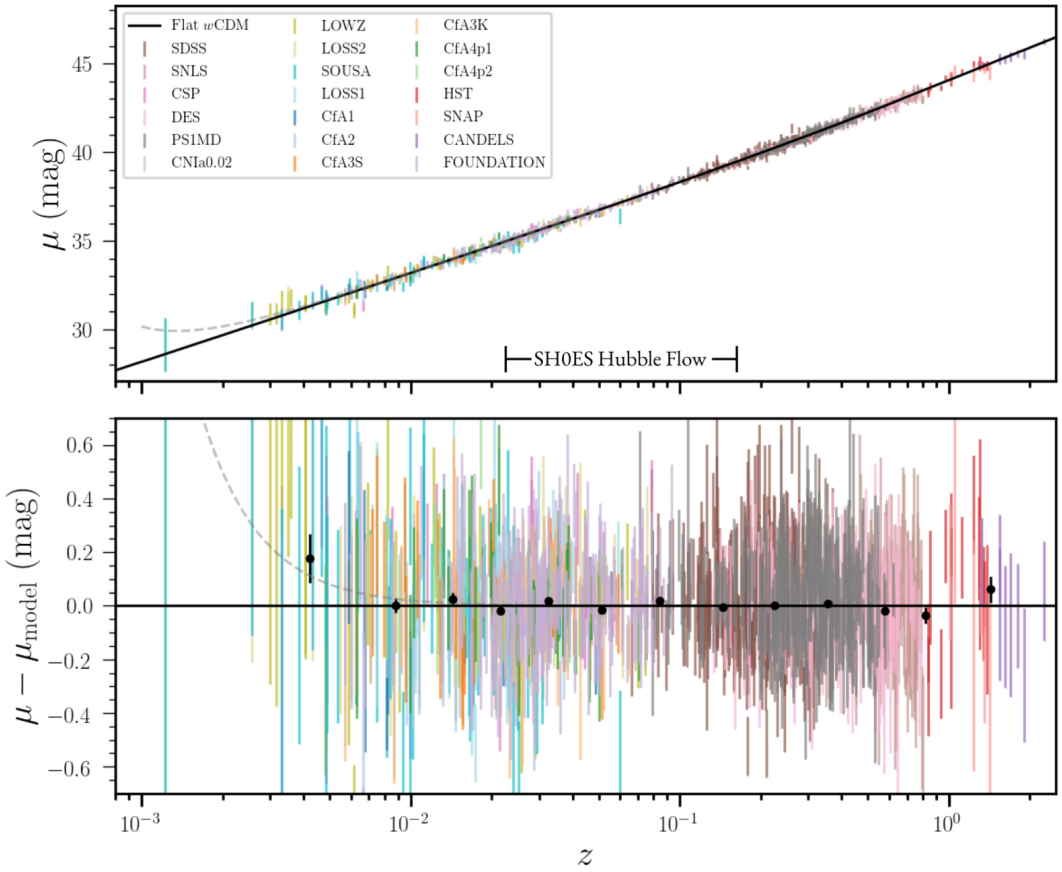


Figure 1.8: (a) The distance modulus for 1550 Ia supernovae part of 18 different surveys (different colors) as function of their redshift. (b) The best-fitting magnitude residuals. The fitting has been performed for a flat wCDM model, resulting in $\Omega_{0\rm m} = 0.309^{+0.063}_{-0.069}$, $\Omega_{0\Lambda} = 0.691^{+0.069}_{-0.063}$ and $w = -0.9 \pm 0.14$. Figure 4 of Brout et al. (2022)

where $D_{\rm L}$ is the luminosity distance between the observer and the light source

$$D_{\rm L}(z) = (1+z)\, a_0 \mathbb{X}(= (1+z)\, a_0 D_{\rm M}) \tag{1.56}$$

and $\mathbb{X}$ is the transverse comoving distance between the source and the observer – $a_0 = 1$ by convention.

The luminosity distance is used in cosmological measurements based on Type Ia Supernovae (SNIa). The SNIa are thermonuclear explosions of carbon-oxygen white dwarfs that surpass the Chandrasekhar mass limit by acquiring matter from a binary partner. This limit makes SNIa nearly photon "Standard Candles" and thus useful in estimating distances, see Rich (2010) for more details. Starting from the definition of the distance modulus μ:

$$\mu \equiv 5 \log \frac{D_{\rm L}}{10{\rm pc}} = 2.5 \log F - 2.5 \log \frac{L}{4\pi (10{\rm pc})^2}, \tag{1.57}$$

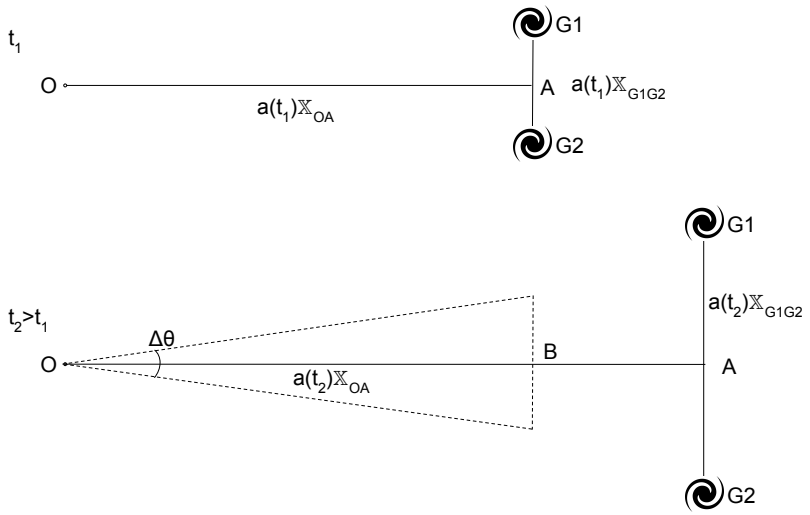


Figure 1.9: Illustration of the angular diameter distance. The light emitted by the galaxies G1 and G2 at time t_1 is detected by the observer O at time $t_2 > t_1$. Thus, the observer sees the two galaxies as they were at t_1. See text for details. Galaxy pictogram from https://www.iconspng.com/image/35332/spiral-galaxy

the luminosity and flux are experimentally calibrated and measured, and D_L is modelled theoretically as in equation (1.56), see Figure 1.8. Schmidt et al. (1998); Riess et al. (1998); Perlmutter et al. (1999) have performed measurements up to $z \approx 0.83$ using 60 SNIa leading to the first observations of and accelerated expansion of the Universe due to the $\Omega_{0\Lambda} \neq 0$. More recently, Brout et al. (2022) have measured cosmological parameters using 1550 SNIa in the redshift interval $z \in [0.001, 2.26]$.

Angular diameter distance

Figure 1.9 illustrates two moments in the evolution of the Universe. At time t_1, the two galaxies are situated at a very large proper distance $a(t_1)\mathbb{X}_{\rm OA}$ from the observer O, compared to the separation $a(t_1)\mathbb{X}_{\rm G1G2}$ between them. The photons emitted at t_1 arrive to the observer at time $t_2 > t_1$. Thus, the observer sees the galaxies as they were at t_1. When measuring the angular separation $\Delta\theta$, the observer obtains:

$$\Delta\theta = \frac{a(t_1)\mathbb{X}_{\rm G1G2}}{a(t_1)\mathbb{X}_{\rm OA}} = \frac{\mathbb{X}_{\rm G1G2}}{\mathbb{X}_{\rm OA}}, \tag{1.58}$$

where the angular diameter distance between the observer and the galaxies is $D_{\rm A} \equiv a(t_1)\mathbb{X}_{\rm OA}$. Consequently, one can express the angular diameter distance as function of the redshift z of a light-source:

$$D_{\rm A}(z) = \frac{a_0\mathbb{X}}{1+z}\left(= \frac{a_0 D_{\rm M}}{1+z}\right), \tag{1.59}$$

where $\mathbb{X}$ is the transverse comoving distance – $a_0 = 1$ by convention.

Let us assume that there exists a Standard Ruler, whose size $\mathbb{X}_{\text{standard ruler}}$ in comoving coordinates is known and does not change in time. In this case, the measured angular size of the standard ruler found at a redshift z provides a constrain for the $D_{\text{A}}(z)$ that depends on cosmological parameters:

$$\Delta\theta^{\text{m}} = \frac{\mathbb{X}_{\text{standard ruler}}}{(1+z)D_{\text{A}}(z)}. \tag{1.60}$$

We show in Section 1.2.3 a practical example of a standard ruler. Figure 1.6 illustrates the effect of different w and $\Omega_{0\Lambda}$ values on the D_{A} and suggests that the constraints on the D_{A} at large redshifts $z > 1$ can distinguish between different models of DE. We detail in Section 1.4.4 how D_{A} can be measured at different redshifts.

Without entering in details, another example of cosmological measurements which involves measuring angular diameter distances is the time-delay technique that can be used to measure H_0 (Refsdal, 1964; Rhee, 1991; Shajib et al., 2020). These measurements are using the effect of strong gravitational lensing, in which the rays of light from a far away varying light-source (e.g. a quasar or a supernova) are deviated due to a massive object (e.g. galaxy, cluster of galaxies) found between the Earth and the source, leading to multiple images of the source. This means that the light of the source reaches the observer along different paths that could lead to different detection times of the light. The time difference (also called time-delay, Δt) induced by the different paths is

$$\Delta t \propto \frac{D_{\text{d}} D_{\text{s}}}{D_{\text{ds}}} \propto H_0^{-1}, \tag{1.61}$$

where D_{d} is the angular diameter distance to the lens (the massive object between the source and the observer), D_{s} is the angular diameter distance to the source and D_{ds} is the angular diameter distance between the source and the lens.

Look-back time

Due to the finite speed of light, an observer on Earth receives today the photons emitted by a galaxy in the past. The farther the galaxy is, the longer the travelling time of the light becomes, therefore earlier periods are observed. Given the fact that farther galaxies have higher redshifts, there is a direct one-to-one mapping between the time and the redshift:

$$\int_t^{t_0} \mathrm{d}t = t_0 - t = \int_0^z \frac{\mathrm{d}z}{(1+z)H(z)}, \tag{1.62}$$

where $t_0 \approx 13.8$ billion years is the age of the Universe today, t is the age of the Universe when the galaxy at redshift z has emitted the light that the observer detects today and $t_0 - t$ is the look-back time. Observing a higher redshift galaxy means that the observer looks even further

back in time. Figure 1.7 illustrates the look-back time as function of redshift. One can notice that the majority of the history of the Universe is captured up to $z \approx 1.5$, as $t(z \approx 1.5) \approx 4$ billion years.

1.2.3 The Big-Bang

As already unveiled in previous sections, the Universe is expanding, therefore the matter and radiation densities decrease with time. Moreover, analysing the spectrum of the CMB today, one observes that it matches with the one of a black body whose temperature is $T_0 = 2.72548 \pm 0.00057\,\mathrm{K}$ (Fixsen, 2009). Due to the fact that the wavelengths of the emitted radiation depend on the temperature and the fact that the expansion decreases the energy of photons with a, the temperature of the cosmic plasma scales as (Dodelson, 2003; Dodelson & Schmidt, 2020):

$$T(t) = \frac{T_0}{a(t)}. \tag{1.63}$$

According to the Big-Bang theory, the Universe started 13.8 billion years ago as an extremely dense and hot plasma. Non-controversial physics – as described by Rich (2010) – allows the understanding of the Universe since it had a temperature of $T \approx 1\,\mathrm{GeV}$ and the matter was a homogeneous soup of quarks gluons and leptons. The expansion of the Universe decreased the temperature, allowing for hadrons, nuclei and finally atoms to form. The formation of neutral atoms allowed for the photons to freely travel inside the Universe, giving birth to the CMB.

Furthermore, the gravitational collapse of atoms lead to the formation of stars and galaxies, which "reionized" the neutral gas inside the Universe.

Big-Bang Nucleosynthesis

This subsection is mainly based on Rich (2010); Dodelson (2003); Dodelson & Schmidt (2020); Workman et al. (2022). Starting from $t \approx 10^{-6}\,\mathrm{s}$ ($T \approx 400$ MeV) after the Big-Bang, the temperature was low enough that hadrons such as neutrons and protons could be produced. Moreover, during the first second after the Big-Bang, the weak interactions have been in thermal equilibrium, meaning that neutrinos ν could easily interact with the surrounding particles, keeping a fixed ratio of neutrons and protons $n/p = \mathrm{e}^{-\Delta \mathrm{m}/T}$, where $\Delta \mathrm{m} = \mathrm{m}_n - \mathrm{m}_p = 1.293\,\mathrm{MeV}$ is the neutron-proton mass difference.

When the Universe celebrated its first second, the temperature cooled down to $T \approx 1\,\mathrm{MeV} \approx 10^{10}\,\mathrm{K}$[8]. At that moment, the cosmic plasma was composed out of:

- Photons, electrons and positrons as relativistic particles in equilibrium through electro-

[8] 1 Kelvin: $1\,\mathrm{K} = -272.15\,^{\circ}\mathrm{C}$

magnetic interactions (e.g. $e^+ + e^- \leftrightarrow \gamma + \gamma$);

- Neutrions ν as decoupled relativistic particles: ν could travel freely and thus the weak interactions were no longer in thermal equilibrium. This moment is also called "freeze-out".

- Baryons as nonrelativistic particles: due to the initial asymmetry of baryons and anti-baryons, at $T \approx 1$ MeV, most anti-baryons have annihilated, thus:

$$\eta \equiv \frac{n_b}{n_\gamma} \approx 6 \times 10^{-10}, \tag{1.64}$$

where n_b is the number density of baryons and n_γ is the photon number density. Additionally, the ratio of neutrons to protons is $n/p = e^{-\Delta_{m/T_{fr}}} \approx 1/6$ (Workman et al., 2022).

The formation chain of complex nuclei starts with the production of deuterium (D or ^{2}H): $p + n \rightarrow \gamma + D$. Even though the nuclear binding energy for deuterium is $\Delta_D = 2.23$ MeV and the $T \approx 1$ MeV, deuterium cannot be formed until temperature drops to $T \approx 100$ keV. The reason is that photons follow a black-body distribution of energies at a given temperature, thus the number of photons per baryon that have the energy larger than Δ_D is larger or equal than unity. Therefore the photo-dissociation prevents the formation of the deuterium and further nucleosynthesis.

As a consequence, while the temperature was 1 MeV > T > 100 keV, i.e. for approximately three minutes, neutrons[9] are free to β-decay ($n \rightarrow p + e^-(\beta^-) + \overline{\nu}_e$), reaching a neutron fraction of $n/p \approx 1/7$. At $T \approx 100$ keV, most neutrons got glued into ^{4}He – mostly through $D + p \rightarrow \gamma + {}^3He$ and $^3He + D \rightarrow p + {}^4He$ – determining the quantity of helium, see Figure 1.10:

$$Y_p = \frac{2(n/p)}{1 + n/p} \approx 0.25. \tag{1.65}$$

Given the lack of metastable or stable elements with mass number $A = 5$ or $A = 8$ and the increasing efficiency of the coulomb barrier between charged nuclei, it became challenging to form heavier elements than ^{4}He based on the two primary species ^{4}He and ^{1}H. Consequently, nuclear reactions froze-out at $T \approx 30$ keV, resulting in a stable abundance of ^{7}Li, ^{4}He, ^{3}He and D, see Figure 1.10.

One can predict the abundances of these elements, however the observations are performed at later epochs, after star formation has taken place. One has to search for regions with low metal abundance in order to measure light element abundances that are more similar to the primordial values, given the fact that stars produce heavier elements ("metals") such as C, N, O and Fe. After measuring the deuterium abundance, Cooke et al. (2018) have estimated

[9] Free neutrons have a mean lifetime of 878.4 ± 0.5 s (Workman et al., 2022).

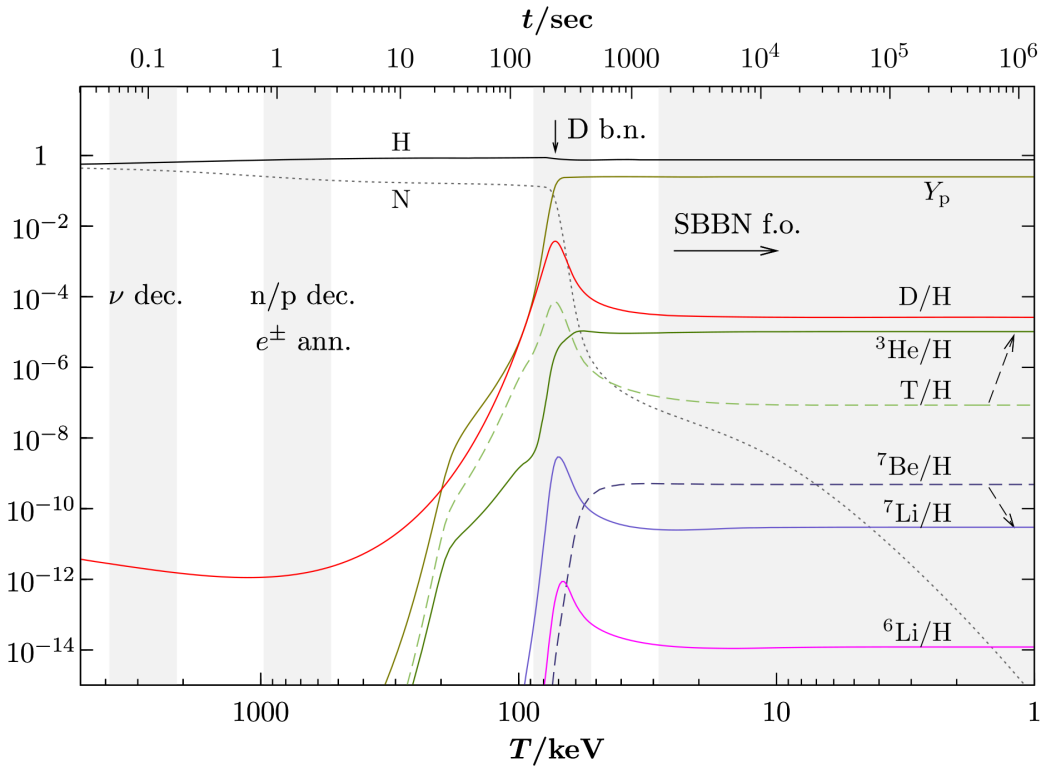


Figure 1.10: Temperature and time dependent abundances of light elements. Figure 4.3 from Dodelson & Schmidt (2020) and Figure 1 from Pospelov & Pradler (2010)

today's density of baryons:

$$100\Omega_{0\mathrm{b}}h^2 = 2.166 \pm 0.015 \pm 0.011, \tag{1.66}$$

where the first error comes from uncertainty in the measurement of deuterium abundance and the second term is the error introduced by the BBN calculations.

Baryonic Acoustic Oscillations

In this section, we describe in a phenomenological manner the Baryonic Acoustic Oscillations (BAO; Peebles & Yu, 1970) based on the Eisenstein et al. (2007a) description in configuration space. This oscillations propagated in the first 400 thousand years of the Universe until the ions and electrons recombined and the baryonic matter decoupled from photons. A schematic description is shown in Figure 1.11.

The initial quantum fluctuations[10] occur in the CDM, neutrinos and primordial plasma of baryons and photons. As discussed in the previous section about BBN, neutrinos decoupled

[10] Inflationary models can provide mechanisms for the initial fluctuations, as mentioned in Section 1.2.4. Here, we skip the inflationary period and start directly with fluctuations in the primordial plasma.

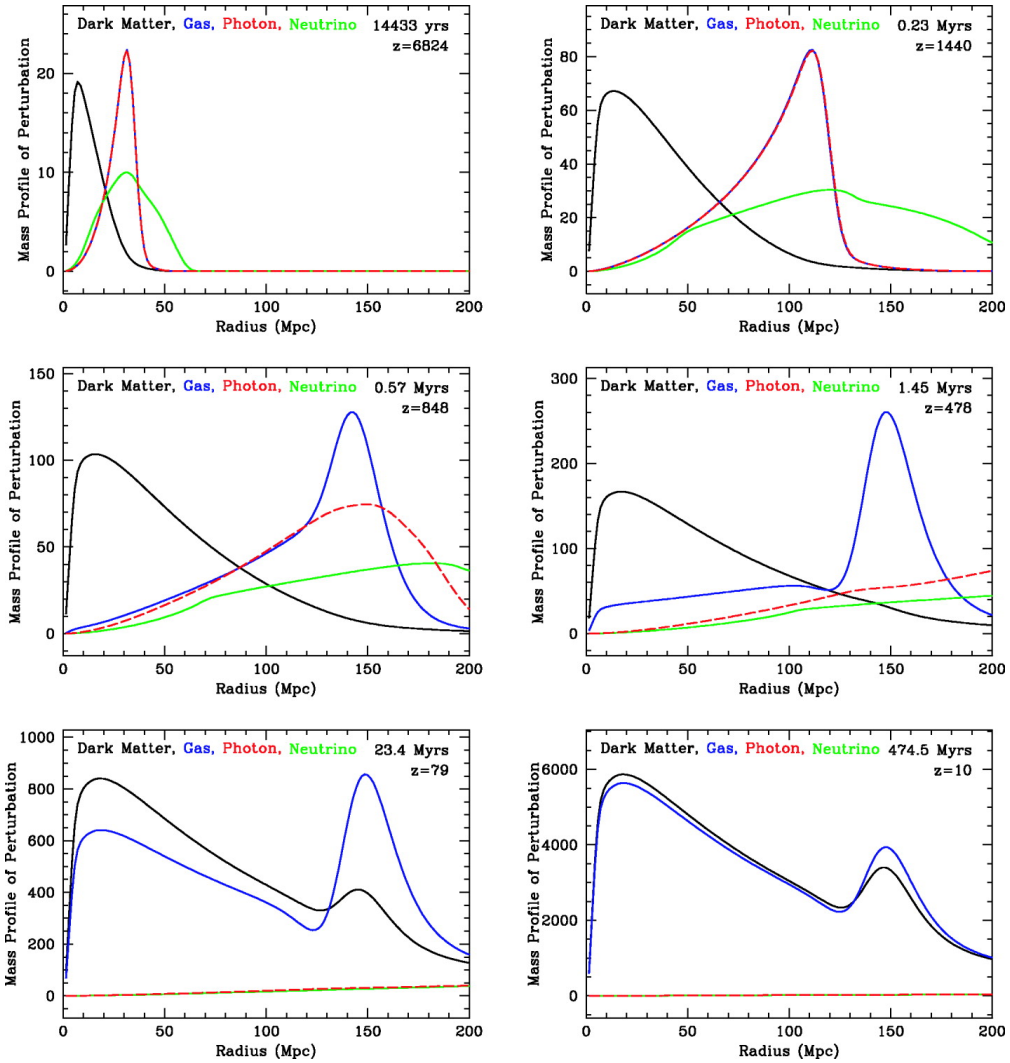


Figure 1.11: The propagation of Baryonic Acoustic Oscillations in the primordial plasma of photons, baryons and neutrinos. Figure 1 of Eisenstein et al. (2007a).

from the baryon-photon plasma in the first second after the Big-Bang, thus the evolution of neutrino fluctuations was less correlated to the fluctuations in the plasma. Moreover, CDM interacts only through gravitational forces with the plasma so they did not follow the same evolution.

Before decoupling, the thermal equilibrium between the baryons, electrons and photons is kept through the photon-electron scattering[11] and electron-proton Coulomb scattering. Practically, baryons are indirectly coupled to the photons, through electrons. Therefore, the term baryonic matter includes both baryon and electron species.

The mean free path of the photon-electron scattering was much less than the Hubble distance,

[11] Compton and Thomson scattering, see Dodelson & Schmidt (2020)

therefore the overdense regions had slightly higher temperatures and thus larger pressures. This created a gradient in pressure that drove a spherical acoustic wave into the plasma. The sound waves propagated at the speed of sound c_s (Eisenstein & Hu, 1998):

$$c_s(z) = \frac{c}{\sqrt{3}} \left[1 + \frac{3\rho_b(z)}{4\rho_\gamma(z)} \right]^{-1/2}, \tag{1.67}$$

where $\rho_b(z)$ and $\rho_\gamma(z)$ are the time-dependent (redshift-dependent) energy densities of baryons and photons, respectively and c is the speed of light. After the recombination, the pressure-supplying photons evaded the plasma and the speed of sound decreased sharply. However, the baryonic fluctuations propagated until the moment when the baryonic matter decoupled from the photons. Consequently, the comoving distance travelled by the oscillation, also called sound horizon r_s is:

$$r_s = \int_0^{t_d} \frac{c_s(t)\mathrm{d}t}{a(t)} = \int_{z_d}^{\infty} \frac{c_s(z)\mathrm{d}z}{H(z)}, \tag{1.68}$$

where t_d and z_d are the time and the redshift of the decoupling[12] of the baryonic matter from the photons and $H(z)$ is the Hubble parameter, see Table 1.1 for numerical values of r_s. This comoving distance can be considered a standard ruler and it can be used to estimate distances as we show in Section 1.4.4.

After decoupling, there were a CDM overdensity at the initial position and a spherical shell of plasma around it. Furthermore, both attracted gas and CDM, being seeds of the gravitational instability and thus structure formation. Finally, the photons became free to travel throughout the Universe, with an energy density decreasing with a^4. Today, we see them as the CMB.

Cosmic Microwave Background

In 1960s, Arno Penzias and Robert Wilson were experimenting with a microwave antenna for telecommunications and astronomy when they observed an isotropic flux of microwaves across the sky. This was the discovery of the CMB (Penzias & Wilson, 1965a,b). Twenty five years later, NASA's Cosmic Background Explorer (COBE) satellite was into Earth's orbit to further study the CMB detecting for the first time tiny fluctuations in the temperature (Smoot et al., 1992). COBE's successor, Wilkinson Microwave Anisotropy Probe (WMAP; Bennett et al., 2003), launched in 2001, has further improved the measurements of these fluctuations and thus provided increased precision on cosmological parameters (Hinshaw et al., 2013). The state-of-the-art CMB anisotropy measurements are provided by the ESA's Planck mission (2009-2013; Tauber et al., 2010).

Figure 1.12 shows the CMB spectrum from WMAP measurements and a black body spectrum that follows well the measurements. Combining multiple measurements from the literature,

[12] Also known as the redshift of the drag epoch, $z_d \approx 1059$. This redshift is slightly smaller than the redshift at recombination $z_* \approx 1089$, see Planck Collaboration et al. (2020b) for more details

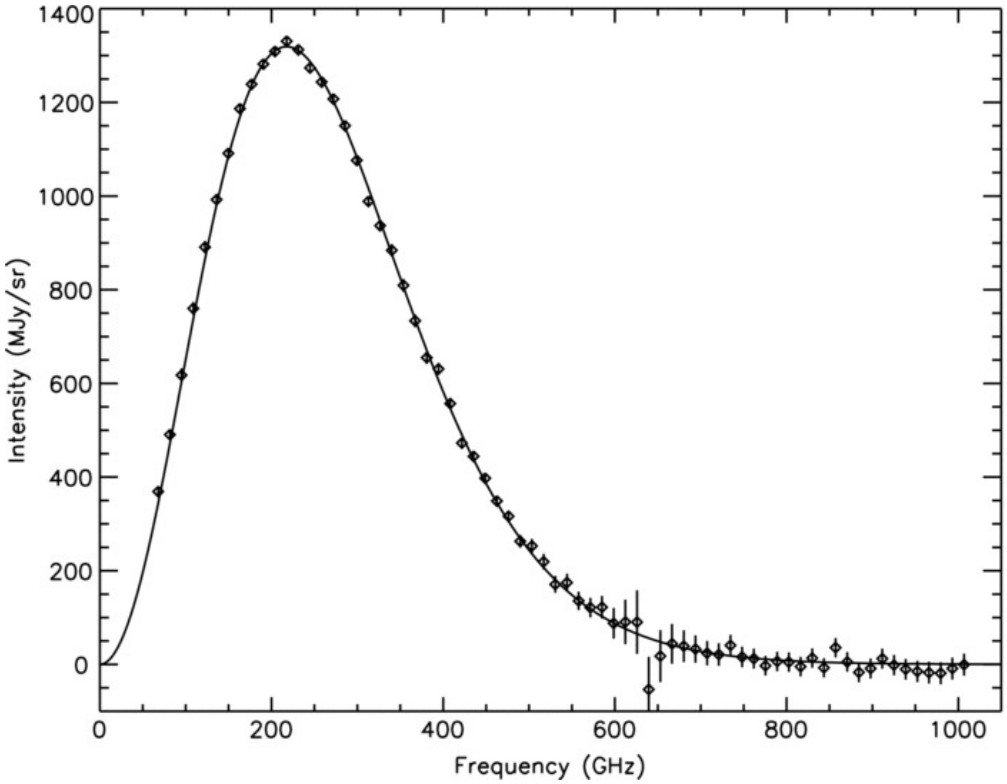


Figure 1.12: The black body Spectrum of CMB with a temperature of $T_0 = 2.72548 \pm 0.00057\,\mathrm{K}$. Figure 3 of Fixsen (2009)

Fixsen (2009) has obtained a CMB black body temperature of $T_0 = 2.72548 \pm 0.00057\,\mathrm{K}$.

Nevertheless, looking at Figure 1.13, one can observe temperature anisotropies of the order of:

$$\frac{\Delta T}{T} \approx 10^{-5}. \tag{1.69}$$

They are directly connected to the fluctuations in the primordial plasma of baryons and photons that also caused the BAO.

Given the fact that the CMB provides 2D measurements on the surface of a sphere, one can expand the temperature anisotropies into spherical harmonics $Y_{\ell m}(\theta,\phi)$ (Dodelson & Schmidt, 2020):

$$\frac{\Delta T}{T} = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta,\phi), \tag{1.70}$$

where $a_{\ell m}$ contain all the information found in the temperature field and ℓ and m are conjugate to the real space θ and ϕ angles on the sky. For density perturbations, one can only make predictions about the distribution from which they are drawn. Given that $\langle \frac{\Delta T}{T} \rangle = 0$, $\langle a_{\ell m} \rangle = 0$,

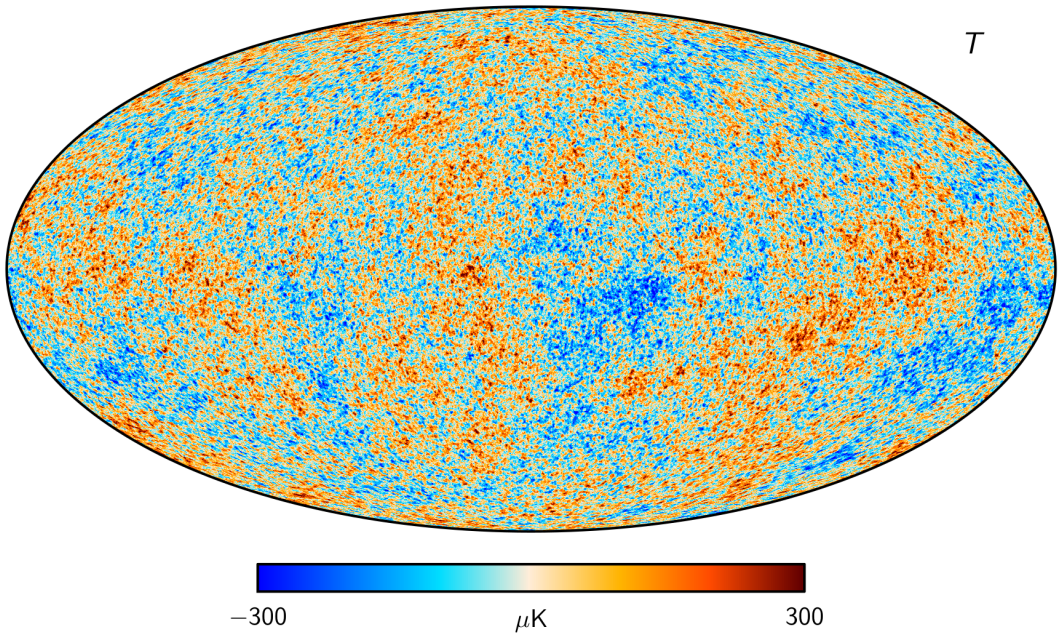


Figure 1.13: CMB temperature anisotropy map from https://wiki.cosmos.esa.int/planck-legacy-archive/index.php/CMB_maps.

thus one can get information from the variance C_ℓ:

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = \delta_{\ell\ell'}\delta_{mm'}C_\ell. \tag{1.71}$$

Due to the mathematical properties of the spherical harmonics, for a given ℓ, one can measure $2\ell+1$ independent m modes, which means that there is a fundamental uncertainty that affects the knowledge one can extract from C_ℓ. This uncertainty ΔC_ℓ is called the cosmic variance[13] and affects mostly the low ℓ (Dodelson & Schmidt, 2020):

$$\Delta C_\ell = C_\ell \sqrt{\frac{2}{2\ell+1}}. \tag{1.72}$$

In practice, one computes the amplitude of the temperature fluctuations $D_\ell^{TT} = (2\pi)^{-1}\ell(\ell+1)C_\ell$, that is shown in Figure 1.14. There are three main features that can be observed:

1. Late-time Integrated Sachs-Wolfe effect at large scales, i.e. $\ell < 30$ [14]. The fluctuations at these scales are strongly affected by the gravitational potential of galaxy clusters or voids found in the paths of CMB photons from the last scattering surface (the epoch of decoupling) to the observers on Earth (today). For example, a photon entering in a potential well of a cluster increases its energy due to the gravitational blueshift[15]. Given

[13] Cosmic variance affects the matter clustering as well, see Section 1.3.1 for a discussion.
[14] (Sachs & Wolfe, 1967; White & Hu, 1997; Rich, 2010; Dodelson & Schmidt, 2020)
[15] The opposite of redshift

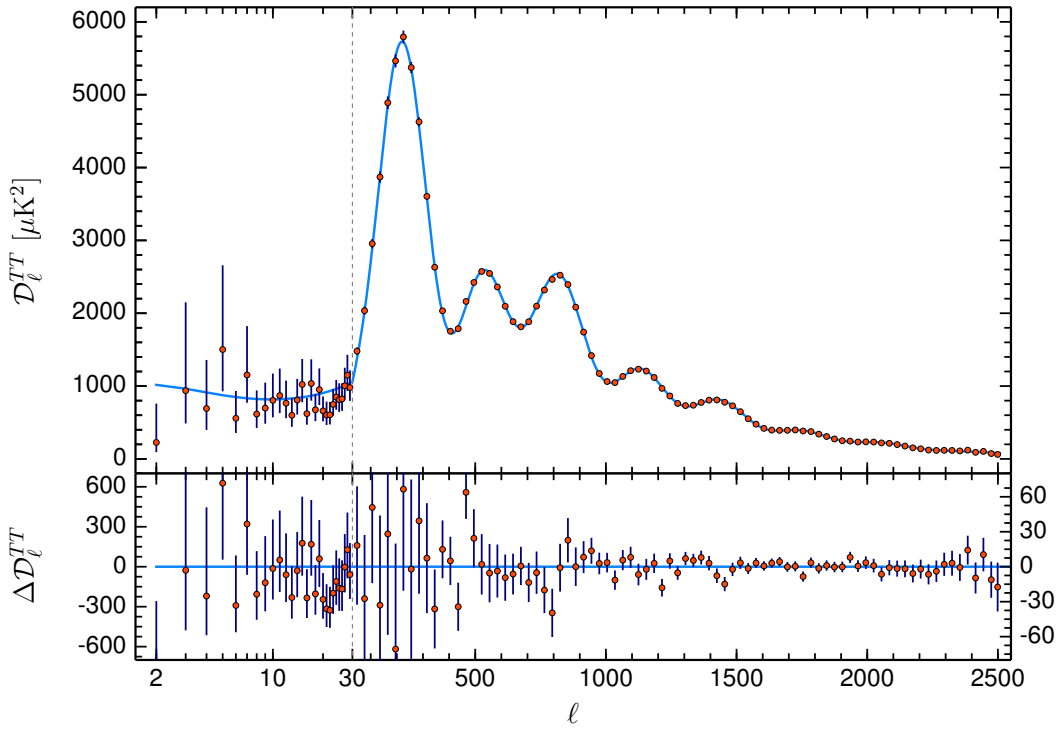


Figure 1.14: The amplitude of the temperature fluctuations D_ℓ^{TT} (red points with error bars) and its best-fitting flat ΛCDM model (blue). The residuals are shown in the lower panel. Figure 1 of Planck Collaboration et al. (2020b)

the faster expansion of the Universe in the DE dominated epoch, the potential well of the cluster gets shallower during the travel of the photon, thus the loss of energy to get out of the well is lower than the initial gain. Consequently, a photon slightly increases its frequency when it passes through a cluster. The opposite is true for a photon entering a void.

2. The BAO are visible in the temperature power spectrum as wiggles, given that in the configuration space it should be a spherical shell, i.e. a radial peak[16].

3. Photon diffusion at small scales that induces the Silk damping (Silk, 1968). Photons scatter off on electrons and thus have a random walk of a given distance $\lambda_D \approx \lambda_{\rm MFP}\sqrt{n_e\sigma_{\rm T}H^{-1}}$ (H–Hubble parameter, n_e–number density of electrons, $\sigma_{\rm T}$ – Thomson[17] cross section , $\lambda_{\rm MFP}$ – mean free path). This random walk washes out fluctuations smaller than λ_D (Dodelson & Schmidt, 2020).

In order to model such high precision CMB measurements, one must employ sophisticated codes that solve the Boltzmann equations that describe the physics before recombination, as described in Section 1.3.2.

[16]The Fourier transform of a sharp peak is a sine wave
[17]Thomson scattering is the elastic interaction between electromagnetic radiation and charged particles.

Parameter	H_0	$\Omega_{0\Lambda}$	$\Omega_{0\mathrm{m}}$	-
Unit	(km/s)/Mpc			-
ΛCDM				-
BAO (eB)	-	0.7010 ± 0.0160	0.2990 ± 0.0160	-
CMB (P18)	67.27 ± 0.60	0.6834 ± 0.0084	0.3166 ± 0.0084	-
CMB + BAO (eB)	67.61 ± 0.44	0.6881 ± 0.0059	0.3119 ± 0.0059	-
CMB + mtBAO	67.96 ± 0.39	0.6930 ± 0.0051	0.3070 ± 0.0051	-
BBN + BAO	67.35 ± 0.98	0.7010 ± 0.0160	0.2990 ± 0.0160	-
BBN + mtBAO	67.58 ± 0.91	0.7100 ± 0.0150	0.2900 ± 0.0150	-
oΛCDM				$\Omega_{0\mathrm{k}}$
BAO	-	$0.637^{0.084}_{-0.074}$		$0.078^{+0.086}_{-0.099}$
CMB	$54.5^{+3.3}_{-3.9}$	$0.561^{+0.050}_{-0.041}$		$-0.044^{+0.019}_{-0.014}$
CMB + BAO	67.59 ± 0.61	0.6882 ± 0.0060		-0.0001 ± 0.0018
wCDM				w
BAO	-	$0.729^{+0.017}_{-0.038}$		-0.69 ± 0.15
CMB	-	$0.801^{+0.057}_{-0.022}$		$-1.58^{+0.16}_{-0.35}$
CMB + BAO	$68.4^{+1.4}_{-1.5}$	0.694 ± 0.012		$-1.034^{+0.061}_{-0.053}$

Parameter	$100\Omega_{0\mathrm{b}}h^2$	$\Omega_{0\mathrm{CDM}}h^2$	r_{drag}	n_{s}
Unit			Mpc	
ΛCDM				-
CMB (P18)	2.236 ± 0.015	0.1202 ± 0.0014	147.05 ± 0.30	0.9649 ± 0.0044
BAO + BBN (eB)	-	-	149.3 ± 2.8	-

Table 1.1: The measurements of cosmological parameters using different probes. The values of the parameters for oΛCDM and wCDM are from eBOSS(eB; Alam et al., 2021). The Planck18 measurements (in addition, $\ln\left(10^{10}A_{\mathrm{s}}\right) = 3.045 \pm 0.016$, at $k_{\mathrm{p}} = 0.05\,\mathrm{Mpc}^{-1}$) (P18; Planck Collaboration et al., 2020b) include the temperature and polarisation power spectra. The "mt" – in CMB + mtBAO and BBN + mtBAO – stands for multi-tracer BAO analysis that includes voids (see Zhao et al., 2022, for more details and for BBN + BAO results). We discuss the different measurements in Section 1.4.4 and Section 3.1.1.

Table 1.1 shows the constraints on cosmological parameters obtained using Planck 2018 data (Planck Collaboration et al., 2020b). The value of the curvature parameters Ω_k is very close to zero. This means that Ω_{tot} – i.e. equations (1.33) and (1.42) – had to be tuned to the value of one to a level of $\approx 10^{-60}$ (see Rich (2010)) at very early times in the history of the Universe. The necessity of this kind of fine-tuning is called the Flatness Problem.

Moreover, the fact that the temperature fluctuations are of the order of 10^{-5} means that the primordial plasma was in an almost perfect thermal equilibrium. However, today we can observe the CMB photons from patches in the plasma that were not in causal contact at the time of recombination[18]. This means that plasma could not have thermalized in that period of the history of the Universe. This is called the Horizon Problem.

Lastly, the Big-Bang theory does not provide a mechanism to create the fluctuations seen in the CMB temperature maps. However, we know that these fluctuations are the seeds for the large-scale structure formation.

1.2.4 Inflation

In order to solve the Horizon and Flatness problems, Guth (1981); Linde (1982); Albrecht & Steinhardt (1982) have introduced the concept of inflation.

The inflation has been introduced as an exponential expansion of the very early Universe even before the radiation dominated epoch, during which the scale factor $a(t)$:

$$a(t) = a_e \mathrm{e}^{H_{\text{inf}}(t-t_e)}, \tag{1.73}$$

for $t_b < t < t_e$ and where H_{inf} is a constant Hubble parameter during the inflation, t_b and t_e represent the beginning and end time of inflation. Lastly, a_e is the scale factor at the end of inflation (Dodelson & Schmidt, 2020). This exponential expansion allowed the communication over larger distances before it occurred, i.e. one can approximate that for $t < t_b$, the scale factor $a \approx a(t_b)$, allowing for large enough patches to be in thermal equilibrium. Finally, the exponential inflation would just extend the space and spread the thermalized patch.

In order to fulfill the CMB observations (and thus solve the Horizon problem), the scale factor must increase by at least a factor of 10^{26} at the end of inflation, i.e. $a_e/a_b > 10^{26}$, if one approximates the beginning of the inflation at a temperature $T_b \approx 10^{15}$ GeV. The whole process should last for $\approx 10^{-34}$ s (Rich, 2010). Additionally, Rich (2010) shows that such an expansion would smooth out a possible curvature k of the initial Universe so that it allows for the current observed values of Ω_k without the requirement of fine tuning Ω_{tot} during very early epochs.

There are many attempts to theoretically define and describe the field that drives the inflation (see review of Workman et al. (2022)), however it is certain that at the end of it, the Universe

[18] On the sky, the $\approx 2^\circ$ angular separation denotes the limit above which patches were not in causal contact.

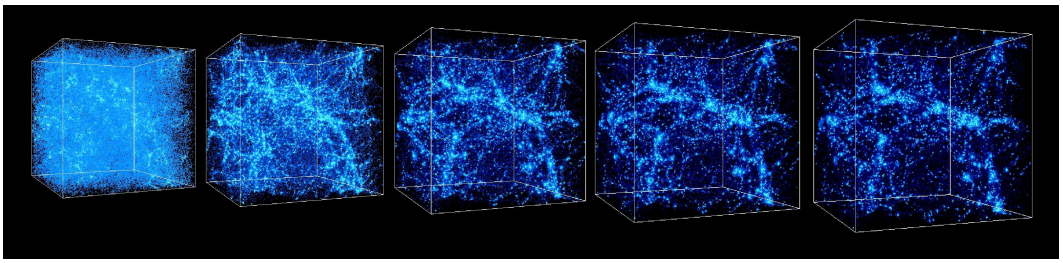


Figure 1.15: Large-Scale Structure formation using a N-body simulation of cold dark matter from $z = 10$ to the present epoch. Figure from http://cosmicweb.uchicago.edu/filaments.html

was filled with radiation. Thus, there should be a conversion mechanism between the two states. Today, these mechanisms and the nature of inflation are still unknown, therefore it is beyond the scope of the thesis to get into more details.

Finally, the inflation field can give birth to Gaussian quantum fluctuations – with mean zero – that can translate into fluctuations of the matter and radiation fields observed in the CMB, see Dodelson & Schmidt (2020)

1.3 Large-Scale Structure

Observations of the distribution of galaxies and matter in the Universe have revealed that the matter is not randomly and uniformly distributed inside the Universe. At scales of the order of hundreds of megaparsecs, structures such as filaments, sheets and super clusters (nodes) in the matter distribution can be observed. The structures have been most likely formed from the primordial fluctuations in the collisionless cold dark matter (CDM) that have evolved under the gravitational interaction.

It is important to notice that the primordial fluctuations occurred on a large range of scales and they have been imprinted in all components of the Universe – i.e. photons, baryons, neutrinos and CDM – except DE. Consequently, their evolution – except the CDM that interacts only gravitationally – has been influenced by multiple physical phenomena, before the recombination of ions and electrons, see Section 1.3.2. The distribution of the fluctuations in the photon density field at the recombination can be observed as temperature fluctuations in CMB observations such as WMAP and Planck (e.g. Planck Collaboration et al., 2020a), see Section 1.2.3. Lastly, the matter (CDM and baryonic matter) distribution, at the moment of recombination, can be regarded as the seed of the Large-Scale Structure (LSS) formation.

Figure 1.15 shows the time evolution of the CDM distribution in a simulated box of side length 43 Mpc, from a redshift $z = 10$ to $z = 0$ (present time). One can observe that at very early times, the distribution of matter seems to be closer to uniformity. Nevertheless due to the gravitational collapse, the small initial seeds grow, giving birth to the nodes and filaments.

In this section, we provide a brief phenomenological and analytical description of LSS and

its evolution. Due to the stochastic aspect of the density fields, we introduce in Section 1.3.1 some important statistical tools used to study the LSS (inspired by the review of Bernardeau et al. (2002)). In Section 1.3.2, we explain the impact of CDM, baryons and neutrinos on the linear matter power spectrum (i.e. the matter power spectrum after decoupling, but still at high enough redshift). Section 1.3.3 presents the Vlasov equation for collisionless CDM that embodies its gravitational evolution. Sections 1.3.4 and 1.3.5 introduce Perturbation Theory (PT) as a technique to solve the Vlasov equation and study the LSS formation and evolution (mostly based on Bernardeau et al. (2002); Peebles (1980)). The last section, Section 1.3.6, presents the N-body simulations as discrete numerical solutions of the Vlasov equation.

1.3.1 Statistical description

Due to the fact that the LSS formation must have started from the fluctuations in the primordial plasma, it is useful to define the density contrast or the cosmic density field $\delta(\boldsymbol{x})$:

$$\delta(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{\overline{\rho}} - 1, \tag{1.74}$$

where $\overline{\rho}$ is the spatial average of the $\rho(\boldsymbol{x})$ and $\boldsymbol{x}$ is – usually – the comoving 3D position at which the density is evaluated.

The matter density field (together with other fields such as the velocity divergence field or the cosmic gravitational potential, defined in the next subsection) can be described in the early Universe by a Gaussian Random Field (GRF Planck Collaboration et al., 2020c), meaning that its values are randomly sampled from a Gaussian distribution with mean zero. This has two main consequences:

1. the density field must be must be studied statistically;
2. there is an intrinsic uncertainty called cosmic variance (e.g. Somerville et al., 2004) that affects all clustering measurements.

The fundamental problem in cosmology is that we do not have access to other sampled universes. Therefore, we must assume the ergodicity principle, which makes the equivalence between the ensemble average (of multiple Universes) and the volume average (in a single Universe). Consequently:

1. the $\langle R \rangle$ ensemble average operator is used interchangeably with $\overline{R}$, where R is a random field;
2. the cosmic variance can be reduced by probing larger volumes in our Universe.

A GRF with mean zero, i.e. $\langle \delta(\boldsymbol{x}) \rangle = 0$, can be entirely described by its variance. Therefore, we

define the two-point correlation function (2PCF, $\xi(s)$):

$$\xi(s) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{s})\rangle \tag{1.75}$$

and the power spectrum $P(k)$:

$$\langle \delta(\boldsymbol{k})\delta(\boldsymbol{k}_1)\rangle = \delta_D(\boldsymbol{k}+\boldsymbol{k}_1)P(k) \tag{1.76}$$

of the matter density field $\delta(\boldsymbol{x})$, where $\boldsymbol{s}$ is the separation vector between two positions with an absolute value s. The power spectrum is the Fourier Transform[19] (FT) of the 2PCF:

$$\xi(s) = \int \mathrm{d}^3\boldsymbol{k} P(k) \exp(i\boldsymbol{k}\cdot\boldsymbol{s}), \tag{1.77}$$

where δ_D is the three-dimensional Dirac delta distribution and the wave vector $\boldsymbol{k}$ is the Fourier counterpart of the separation $\boldsymbol{s}$. The absolute values of $\boldsymbol{s}$ and $\boldsymbol{k}$ (i.e. s and k) are used assuming that the Universe is statistically homogeneous and isotropic. In this case, the FT becomes the Hankel transform[20]:

$$\xi(s) = 4\pi \int_0^\infty k^2 P(k) \frac{\sin(ks)}{ks} \mathrm{d}k \qquad P(k) = \frac{4\pi}{(2\pi)^3} \int_0^\infty s^2 \xi(s) \frac{\sin(ks)}{ks} \mathrm{d}s, \tag{1.78}$$

where the ratio $j_0(ks) = \sin(ks)/(ks)$ is the spherical Bessel function with index 0. Nonetheless, Redshift Space Distortions (Section 1.4.5) and the Alcock & Paczynski (AP Alcock & Paczynski, 1979) effect (Section 1.4.4) introduce aniosotropies in measurements. As a consequence, one must adapt the previous equations to include these effects (Section 1.4.3).

Another important observation is that the matter density field becomes non-Gaussian due to the gravitational evolution. Therefore, higher order moments (such as bispectrum and the equivalent three-point correlation function) are needed to entirely describe the density field (see e.g. Bernardeau et al., 2002).

1.3.2 Transfer function

In the early epochs of the Universe, the fluctuations can be described as following a Gaussian distribution with mean zero and variance:

$$P_{\mathrm{primordial}}(k) = A_\mathrm{s} k^{n_\mathrm{s}}, \tag{1.79}$$

known as primordial power spectrum, where n_s is the scalar index and A_s is the amplitude of the variations at a certain pivot k_p, see Table 1.1. For $n_\mathrm{s} = 1$, the fluctuations are scale-free (Harrison, 1970; Zeldovich, 1972). Afterwards, the density fields of the components (i.e.

[19] The convention used in this section is the one from Bernardeau et al. (2002), i.e. $\tilde{A}(\boldsymbol{k},\tau) = \int \frac{\mathrm{d}^3\boldsymbol{x}}{(2\pi)^3} \exp(-i\boldsymbol{k}\cdot\boldsymbol{x}) A(\boldsymbol{x},\tau)$, $A(\boldsymbol{x},\tau) = \int \mathrm{d}^3\boldsymbol{k} \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) A(\boldsymbol{k},\tau)$

[20] The Hankel transform is equivalent to a Fourier transform in spherical coordinates along the radial component (e.g. Karamanis & Beutler, 2021).

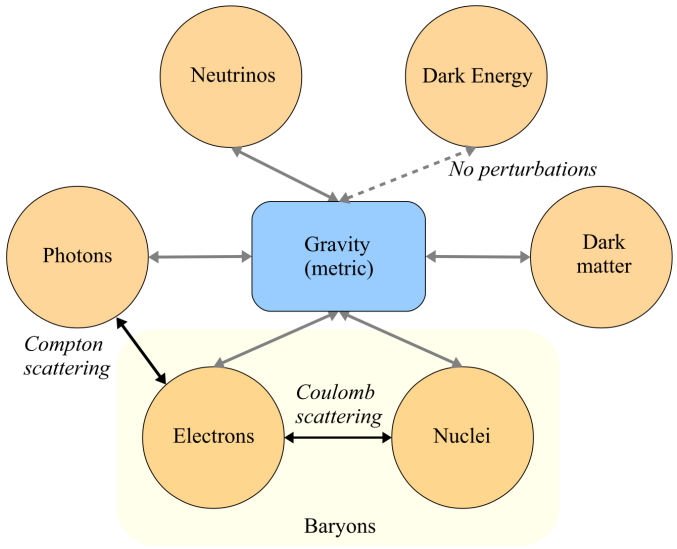


Figure 1.16: The interactions between the components of the Universe are described by Boltzmann-Einstein equations. The Coulomb scattering between nuclei and electrons binds them into a single component that is called baryons. Figure 5.1 from Dodelson & Schmidt (2020)

baryons, CDM, photons, neutrinos) evolve in time, through the radiation, matter and Λ dominated eras.

In order to follow the evolution of the CDM, baryons, photons and neutrinos up to the decoupling, one has to solve the Einstein equation together with one Boltzmann equation for each component. This set of equations is coupled through the physical interactions between all components (see Figure 1.16), therefore it must be solved numerically with codes such as (CAMB; Lewis et al., 2000) and (CLASS; Blas et al., 2011). The solution is the transfer function $T(k)$ that adapts the primordial power spectrum by including the physical effects:

$$P_{\rm m}^{\rm L}(k) = A_{\rm s} k^{n_{\rm s}} |T(k)|^2, \tag{1.80}$$

where $P_{\rm m}^{\rm L}(k)$ is the matter power spectrum (also called linear power spectrum) sometime after decoupling, while the evolution is still linear.

The evolution of perturbations is dependent on the balance between opposing effects. On one hand, the gravitational force pulls matter towards over dense regions. On the other hand, the expansion of the Universe pulls apart the particles of all species such that the perturbations grow more slowly when the Universe is expanding faster. In addition, photons exert a pressure that is proportional to the density, pushing the plasma of baryons and photons towards lower density, hindering the accumulation of baryonic matter.

In what follows, we briefly present the effects of the components on the transfer function based on the analytical solutions of Dodelson & Schmidt (2020) in some limiting cases that are

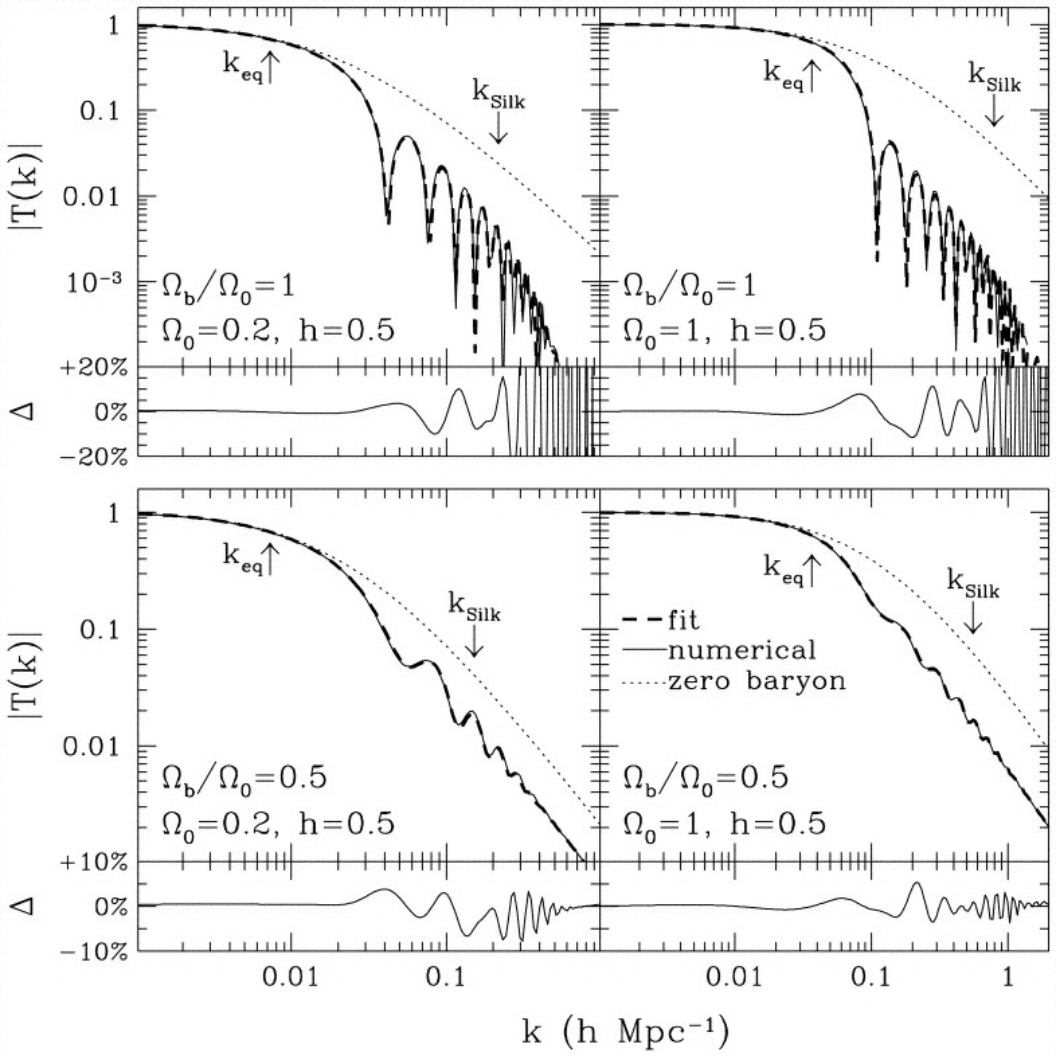


Figure 1.17: Examples of transfer functions (numerical results and best-fitting curves). In this figure, $\Omega_0 = \Omega_b + \Omega_{0CDM}$, where Ω_0, Ω_b Ω_{0CDM} are the total matter, the baryonic and the CDM densities today, respectively. Additionally, $H_0 = 100h$ km/s/Mpc. Figure 3 from Eisenstein & Hu (1998)

defined with respect to the particle horizon $\eta(t)$, equation (1.50). The description is done in Fourier space because the evolution of each k mode can be treated independently. Figure 1.17 shows examples of transfer functions for different matter and baryonic densities.

Cold Dark Matter. CDM interacts only gravitationally with the other components, thus the perturbations in the CDM are mainly dependent on the gravitational forces and expansion of the Universe. The small fluctuations that enter the horizon in the matter dominated era (i.e. $k\eta < 1$ until the matter dominated era) grow as a power-law of time, during both radiation and matter dominated epochs. In contrast, smaller fluctuations that are inside the horizon during the radiation epoch have a logarithmic grow in time until the matter starts to dominate and the growth becomes proportional to a power-law of time. The logarithmic growth is much slower than the power-law, thus for $k > k_{\rm eq}$, ($k_{\rm eq}$ represents the scale of particle horizon at radiation-matter equality epoch, it depends linearly on $\Omega_{0\rm m}$ (Eisenstein & Hu, 1998)) the transfer function drops significantly (Dodelson & Schmidt, 2020)), while for $k < k_{\rm eq}$, the transfer function is close to one.

Baryons. Before the recombination, baryons are strongly coupled to the photons and thus the plasma of baryons and photons experiences acoustic oscillations as described in Section 1.2.3. These waves are observed as oscillatory behaviour in the transfer function. One can observe in Figure 1.17 that without baryons the oscillations vanish. In addition, due to the fact that radiation pressure hinders the gravitational collapse below the sound horizon scale, the growth of the baryonic fluctuations is suppressed. Consequently, one can neglect the effect of the baryons to the gravitational wells at those scales (Eisenstein & Hu, 1998) and thus, the amplitude of the transfer function is lower than in the case without baryons. After the recombination, the baryonic fluid becomes pressureless and its perturbations start following the ones in CDM by falling into the CDM gravitational potential wells.

Photons. As mentioned in Section 1.2.3, the diffusion of photons due to the scattering on electrons, wash out the fluctuations at small scales ($k > k_{\rm Silk}$, Silk damping (Silk, 1968)). Additionally, photons can push the baryons away from overdensities to underdense regions, washing away the fluctuations (i.e. Compton drag). This is seen as an exponential damping of the acoustic oscillations.

Massive neutrinos. A first effect of neutrinos on the growth of perturbations is related to the energy density evolution, which initially decreases with a^{-4} and then slows down to a^{-3}. This changes the Hubble parameter that enters in to the growth factor (see equation (1.104)). Secondly, the neutrino perturbations smaller than the free-streaming scale (i.e. the distance travelled by neutrinos as they escape the high-density regions) are washed out, weakening the gravitational pull on the CDM. This means that the amplitude of CDM perturbations

is lower than it could have been in the absence of massive neutrinos at scales smaller than the free-streaming scale. The more massive the neutrinos are, the more important their contribution is to the gravitational potential well. As a consequence, the CDM fluctuation amplitude decreases with the neutrinos mass (for more details, see e.g. Dodelson & Schmidt, 2020; Agarwal & Feldman, 2011).

1.3.3 Vlasov equation for collisionless Cold Dark Matter

In the previous subsection, we have presented the evolution of fluctuations in a phenomenological way, given the coupled Einstein-Boltzmann equations of all components. This description is useful to estimate the resulting CDM linear density field after baryons-photons decoupling. In what follows, we present the mechanism to describe the evolution of the CDM perturbations in an arbitrary homogeneous and isotropic background Universe filled with matter and dark energy Λ and with a curvature k, that follows the Friedmann equations[21]:

$$kc^2 = (\Omega_{\rm tot}(\tau) - 1)\,\mathscr{H}^2(\tau), \tag{1.81}$$

$$\frac{\partial \mathscr{H}(\tau)}{\partial \tau} = \left(\Omega_\Lambda(\tau) - \frac{\Omega_{\rm m}(\tau)}{2}\right)\mathscr{H}^2(\tau), \tag{1.82}$$

where τ is the conformal time $\mathrm{d}t = a(\tau)\mathrm{d}\tau$.

Even though there is no direct evidence of CDM particles, there are multiple theoretical models attempting to describe the DM. Nevertheless, one can approximate the CDM particles by a non-relativistic collisionless fluid that obeys the Vlasov equation of the phase space $f(\boldsymbol{x}, \boldsymbol{p}, \tau)$:

$$\frac{\mathrm{d}f}{\mathrm{d}\tau} \equiv \frac{\partial f}{\partial \tau} + \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau}\cdot\nabla_x f + \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\tau}\cdot\nabla_p f = 0, \tag{1.83}$$

where $\boldsymbol{x}$ is the comoving coordinate $\boldsymbol{r} = a\boldsymbol{x}$, with $\boldsymbol{r}$ the proper distance (coordinate); $\boldsymbol{p}$ is the linear momentum and τ represents the conformal time. The phase space density function $f(\boldsymbol{x}, \boldsymbol{p}, \tau)$ of the CDM particles is defined as a comoving density:

$$\mathrm{d}N_{\rm particles} = f(\boldsymbol{x}, \boldsymbol{p}, \tau)\mathrm{d}^3\boldsymbol{x}\mathrm{d}^3\boldsymbol{p} = f(\boldsymbol{r}, \boldsymbol{p}, t)\mathrm{d}^3\boldsymbol{r}\mathrm{d}^3\boldsymbol{p}. \tag{1.84}$$

This is in contrast to other functions, such as the matter density $\rho(\boldsymbol{x}, \tau)$[22].

Practically, the Vlasov equation is the continuum limit of the Hamiltonian mechanics with gravitational forces (Angulo & Hahn, 2022). Consequently, the natural linear momentum

[21] The two shown equations are the conformal time representation of equations (1.40) and (1.33), where $\mathscr{H} \equiv \mathrm{d}\ln a/\mathrm{d}\tau = Ha$.

[22] Mathematically, if $\tilde{\mathscr{F}}(\boldsymbol{r}, t)$ is a function of proper (physical) distance and cosmic time and $\mathscr{F}(\boldsymbol{x}, \tau)$ is a function of comoving distance and conformal time, there is a conversion function $\mathscr{C}$ that $(\boldsymbol{r}, t) = \mathscr{C}(\boldsymbol{x}, \tau)$, such that $\tilde{\mathscr{F}}(\boldsymbol{r}, t) = (\tilde{\mathscr{F}} \circ \mathscr{C})(\boldsymbol{x}, \tau)$ and $\mathscr{F} = (\tilde{\mathscr{F}} \circ \mathscr{C})$. Thus, the physical interpretation of a function is the same in both sets of coordinates. Consequently, we simply use the notation $\mathscr{F}(\boldsymbol{r}, t) = \mathscr{F}(\boldsymbol{x}, \tau)$, unless specified otherwise.

degree of freedom of a particle is (see Peebles (1980) for more details):

$$\boldsymbol{p}(\boldsymbol{x},\tau) = am\boldsymbol{u}(\boldsymbol{x},\tau), \tag{1.85}$$

where m is the particle mass and $\boldsymbol{u}(\boldsymbol{x},\tau) = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau}$ is its peculiar velocity defined through the total velocity $\boldsymbol{v} = \dot{\boldsymbol{r}}$[23]:

$$\boldsymbol{v}(\boldsymbol{r},t) = H(t)\boldsymbol{r} + a\dot{\boldsymbol{x}} \qquad \boldsymbol{v}(\boldsymbol{x},\tau) = \mathscr{H}(\tau)\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x},\tau). \tag{1.86}$$

Furthermore, the equations of motion of a particle become:

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \equiv \frac{\ddot{a}}{a}\boldsymbol{r} + \frac{1}{ma}\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla_r\phi(\boldsymbol{r},t) \qquad \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\tau} \equiv \frac{\partial\mathscr{H}(\tau)}{\partial\tau}\boldsymbol{x} + \frac{1}{ma}\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\tau} = -\nabla_x\phi(\boldsymbol{x},\tau), \tag{1.87}$$

where ϕ is sourced by the dark energy component Λ and the matter, through the Poisson equation (see Peebles (1980) for more detailed arguments):

$$\nabla_r^2\phi(\boldsymbol{r},t) = 4\pi G\rho(\boldsymbol{r},t) - \Lambda c^2 \qquad \frac{1}{a^2}\nabla_x^2\phi(\boldsymbol{x},\tau) = 4\pi G\rho(\boldsymbol{x},\tau) - \Lambda c^2, \tag{1.88}$$

and ρ denotes the proper matter density. This works in the approximation that particles interact only through Newtonian gravity – given the non-relativistic aspect (low velocities) at scales smaller than the Hubble radius – but in an expanding Universe.

So far, we have shown formulas using both $(\boldsymbol{r},t)$ and $(\boldsymbol{x},\tau)$, but in the next paragraphs we restrict the description solely to $(\boldsymbol{x},\tau)$. The background model, see e.g. Peebles (1980), neglects any non-linear coupling between the evolution of the Universe through the scale factor $a(t)$ and the inhomogeneities in the matter field. Therefore, one can split ρ and ϕ in a component corresponding to the uniform background related to the evolution of $a(t)$ and a component related to the evolution of the matter fluctuations. Given the total matter density ρ, we define the matter density contrast:

$$\delta(\boldsymbol{x},\tau) = \frac{\rho(\boldsymbol{x},\tau)}{\overline{\rho}(\tau)} - 1, \tag{1.89}$$

where $\overline{\rho}$ is the spatial average of the $\rho(\boldsymbol{x},\tau)$ and is the component that enters in the Friedmann equation (1.28). Furthermore, the cosmological gravitational potential ϕ can be decomposed as well:

$$\phi(\boldsymbol{x},\tau) \equiv \bar{\phi}(\boldsymbol{x},\tau) + \Phi(\boldsymbol{x},\tau), \tag{1.90}$$

where $\Phi(\boldsymbol{x},\tau)$ is sourced by the fluctuations in the matter field ($\delta(\boldsymbol{x},\tau)$) and $\bar{\phi}(\boldsymbol{x},\tau)$ is related to the uniform and homogeneous background that affects the expansion of the Universe. Analysing the equation (1.87), one can observe the "independent" effects of the two compo-

[23]The dot derivative of a function f is defined with respect to t the cosmic time, i.e. $\dot{f} = \frac{\mathrm{d}f}{\mathrm{d}t}$

nents of ϕ.

$$\bar{\phi}(\boldsymbol{x},\tau) = -\frac{1}{2}\frac{\partial\mathscr{H}(\tau)}{\partial\tau}x^2, \tag{1.91}$$

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\tau} = -am\nabla_x\Phi(\boldsymbol{x},\tau), \tag{1.92}$$

i.e. $\bar{\phi}$ imposes the Hubble flow on a particle, given the evolution of the Universe and Φ affects the momentum of a CDM "particle", through Newtonian gravity. Using the definitions of Ω parameters and equations (1.91) and (1.82), one obtains the Poisson equation for Φ:

$$\nabla_x^2\Phi(\boldsymbol{x},\tau) = \frac{3}{2}\Omega_{\mathrm{m}}(\tau)\mathscr{H}^2(\tau)\delta(\boldsymbol{x},\tau). \tag{1.93}$$

Finally, replacing equations (1.85), (1.92) in equation (1.83), one obtains a non-linear partial differential equation with seven variables:

$$\frac{\partial f}{\partial\tau} + \frac{\boldsymbol{p}}{ma}\cdot\nabla_x f - am\nabla_x\Phi(\boldsymbol{x},\tau)\cdot\nabla_p f = 0. \tag{1.94}$$

The purpose of the numerical simulations is to resolve the Vlasov–Poisson set of equations – eqs. (1.93) and (1.94) – for a given number of particles N, that fixes the mass resolution of the numerical simulation, see Section 1.3.6.

In the next two subsections, we present two frameworks to further understand the dynamics of the CDM fluid: Eulerian and Lagrangian. The first one looks at the fluid as a field and thus works with density and velocity fields that are functions of spatial coordinates. In other words, the CDM fluctuations stay at fixed positions, but their amplitudes grow or decay. On the other hand, the Lagrangian point of view analyses the evolution of a chunk of fluid. In the current situation, the Lagrangian framework solves the equation of motion for each CDM particle.

1.3.4 Eulerian perspective

Given the difficulty to solve the Vlasov–Poisson set of equations and the fact that we are interested in the evolution of the spatial distribution, we can limit ourselves to the study of the linear momentum moments of $f(\boldsymbol{x},\boldsymbol{p},\tau)$. The zeroth, first and second order moments are:

$$\int \mathrm{d}^3\boldsymbol{p}\; f(\boldsymbol{x},\boldsymbol{p},\tau) \equiv \frac{a^3}{m}\rho(\boldsymbol{x},\tau), \tag{1.95}$$

$$\int \mathrm{d}^3\boldsymbol{p}\; \frac{\boldsymbol{p}}{am} f(\boldsymbol{x},\boldsymbol{p},\tau) \equiv \frac{a^3}{m}\rho(\boldsymbol{x},\tau)\boldsymbol{u}(\boldsymbol{x},\tau), \tag{1.96}$$

$$\int \mathrm{d}^3\boldsymbol{p}\; \frac{p_i p_j}{a^2m^2} f(\boldsymbol{x},\boldsymbol{p},\tau) \equiv \frac{a^3}{m}\rho(\boldsymbol{x},\tau)u_i(\boldsymbol{x},\tau)u_j(\boldsymbol{x},\tau) + \sigma_{ij}(\boldsymbol{x},\tau). \tag{1.97}$$

Taking into account the definition of the phase space density function $\mathrm{d}N_{\mathrm{particles}} = f(\boldsymbol{x},\boldsymbol{p},\tau)\mathrm{d}^3\boldsymbol{x}\mathrm{d}^3\boldsymbol{p}$, the a^3/m factors transform the proper mass density $\rho(\boldsymbol{x},\tau)$, as defined above, in a comoving

number density. Furthermore, $\boldsymbol{u}(\boldsymbol{x},\tau)$ is the peculiar velocity flow and $\sigma_{ij}(\boldsymbol{x},\tau)$ denotes the stress tensor. Generally, for a fluid $\sigma_{ij} = -\mathscr{P}\delta^{\mathrm{K}}_{ij} + \eta(\nabla_i u_j + \nabla_j u_i - \frac{2}{3}\delta^{\mathrm{K}}_{ij}\nabla\cdot\boldsymbol{u}) + \zeta\delta^{\mathrm{K}}_{ij}\nabla\cdot\boldsymbol{u}$, where $\mathscr{P}$ is the pressure, η and ζ are viscosity coefficients.

The stress tensor describes how different the particle motions are compared to single coherent flows (i.e. single stream). Therefore, in the early stages of the structure formation – before gravitational collapse and virialization – , one can set $\sigma_{ij} = 0$ for the CDM fluid (ideal fluid with zero pressure). Nevertheless, even during later periods one can meaningfully model the structure formation at sufficiently large scales using the approximation $\sigma_{ij} \approx 0$. Deviations from this value indicate the existence of velocity dispersion induced by the multiple streams, also known as shell crossing.

Computing the zeroth order moment of the Vlasov equation (1.94), one obtains the continuity equation:

$$\frac{\partial\delta(\boldsymbol{x},\tau)}{\partial\tau} + \nabla_x\cdot[(1+\delta(\boldsymbol{x},\tau))\,\boldsymbol{u}(\boldsymbol{x},\tau)] = 0, \tag{1.98}$$

where ρ is replaced from equation (1.89). If one multiplies the continuity equation with $\boldsymbol{u}(\boldsymbol{x},\tau)$ and subtracts the result from the first order of Vlasov equation, one obtains the Euler equation that describes the conservation of momentum:

$$\frac{\partial\boldsymbol{u}(\boldsymbol{x},\tau)}{\partial\tau} + \mathscr{H}(\tau)\boldsymbol{u}(\boldsymbol{x},\tau) + \boldsymbol{u}(\boldsymbol{x},\tau)\cdot\nabla_x\boldsymbol{u}(\boldsymbol{x},\tau) = -\nabla_x\Phi(\boldsymbol{x},\tau) - \frac{1}{\rho}\frac{\partial}{\partial x_j}(\rho\sigma_{ij}). \tag{1.99}$$

Mathematically, a vector field can be entirely described by its divergence and curl. Consequently, this fact is used to simplify the Euler equation by computing its divergence and curl[24]. Therefore, the velocity field $\boldsymbol{u}$ is replaced by $\theta(\boldsymbol{x},\tau) \equiv \nabla\cdot\boldsymbol{u}(\boldsymbol{x},\tau)$ and its vorticity $\boldsymbol{w} = \nabla\times\boldsymbol{u}(\boldsymbol{x},\tau)$.

Eulerian Linear Perturbation Theory

The CDM fluctuations in the Universe at really large scales are very small compared to the uniform background, thus assuming $\delta(\boldsymbol{x},\tau) \ll 1$ and $\nabla_x\boldsymbol{u} \ll \mathscr{H}$, one can keep only linear terms in the equations (1.98), (1.99), i.e. δ^2, $\delta\boldsymbol{u}$ and $\boldsymbol{u}^2$ can be neglected. Moreover, one can set $\sigma_{ij} = 0$. Therefore, one obtains:

$$\frac{\partial\delta(\boldsymbol{x},\tau)}{\partial\tau} + \theta(\boldsymbol{x},\tau) = 0 \tag{1.100}$$

$$\frac{\partial\boldsymbol{u}(\boldsymbol{x},\tau)}{\partial\tau} + \mathscr{H}(\tau)\boldsymbol{u}(\boldsymbol{x},\tau) = -\nabla_x\Phi(\boldsymbol{x},\tau). \tag{1.101}$$

Analysing these two equations in Fourier space, one can notice that different $\boldsymbol{k}$ modes – i.e. Fourier analogues of the position $\boldsymbol{x}$ – have an independent evolution. In other words, modes are not coupled in Eulerian Linear Perturbation Theory.

[24] because $\nabla\times\nabla\Phi = 0$ and $\nabla\cdot\nabla\Phi = \nabla^2\Phi$, see equation (1.93)

If one computes the divergence of the linear Euler equation and the time derivative of the continuity equation, one can replace the $\frac{\partial\theta(\boldsymbol{x},\tau)}{\partial\tau}$ and $\theta(\boldsymbol{x},\tau)$ terms into the linear Euler equation. Furthermore, due to the fact that in the resulting equation only $\delta(\boldsymbol{x},\tau)$ depends on position $\boldsymbol{x}$, one can split the density field in a time-dependent component $D_1(\tau)$ – the linear growth factor – and a position-dependent component: $\delta(\boldsymbol{x},\tau) = D_1(\tau)\delta(\boldsymbol{x},0)$. Consequently, the final second order differential equation is:

$$\frac{\mathrm{d}^2 D_1(\tau)}{\mathrm{d}\tau^2} + \mathscr{H}(\tau)\frac{\mathrm{d}D_1(\tau)}{\mathrm{d}\tau} = \frac{3}{2}\Omega_{\mathrm{m}}(\tau)\mathscr{H}^2(\tau)D_1(\tau). \tag{1.102}$$

The general solution of this type of equation is a sum of two independent solutions:

$$\delta(\boldsymbol{x},\tau) = D_1^{(+)}(\tau)A(\boldsymbol{x}) + D_1^{(-)}(\tau)B(\boldsymbol{x}), \tag{1.103}$$

where $D_1^{(+)}$ is the so-called the fast growing mode, the $D_1^{(-)}$ is the slow growing mode and lastly, $A(\boldsymbol{x})$ and $B(\boldsymbol{x})$ describe the initial conditions. In a Universe filled with matter and dark energy, i.e. $H^2(a) = H_0^2\left[\Omega_{0\Lambda} + \Omega_{0\mathrm{m}}a^{-3} + (1-\Omega_{0\mathrm{m}}-\Omega_{0\Lambda})a^{-2}\right]$:

$$D_1^{(+)} = \frac{H(a)}{H_0}\frac{5\Omega_{0\mathrm{m}}}{2}\int_0^a \frac{\mathrm{d}a}{[aH(a)/H_0]^3} \qquad D_1^{(-)} = \frac{H(a)}{H_0}. \tag{1.104}$$

One can notice that the evolution of δ in this linear approximation is local. In other words, $\delta(\boldsymbol{x},\tau)$ is only influenced by the $\delta(\boldsymbol{x},0)$ and the $\frac{\mathrm{d}\delta}{\mathrm{d}t}(\boldsymbol{x},0)$, i.e. the initial conditions at position $\boldsymbol{x}$, (Peebles, 1980).

Finally, using the solutions for $\delta(\boldsymbol{x},\tau)$ and equation (1.100) one obtains the solution for θ:

$$\theta(\boldsymbol{x},\tau) = -\mathscr{H}(\tau)\left[f(\Omega_{0\mathrm{m}},\Omega_{0\Lambda})D_1^{(+)}A(\boldsymbol{x}) + g(\Omega_{0\mathrm{m}},\Omega_{0\Lambda})D_1^{(-)}B(\boldsymbol{x})\right], \tag{1.105}$$

where one defines g and the linear growth rate of structure f:

$$f(\Omega_{0\mathrm{m}},\Omega_{0\Lambda}) \equiv \frac{1}{\mathscr{H}}\frac{\mathrm{d}\ln D_1^{(+)}}{\mathrm{d}\tau} \qquad g(\Omega_{0\mathrm{m}},\Omega_{0\Lambda}) \equiv \frac{1}{\mathscr{H}}\frac{\mathrm{d}\ln D_1^{(-)}}{\mathrm{d}\tau}. \tag{1.106}$$

In order to completely understand the evolution of the velocity field $\boldsymbol{u}$ one has to know its curl together with its divergence θ. Computing the curl of equation (1.101), one obtains that $\boldsymbol{w} = \nabla\times\boldsymbol{u}(\boldsymbol{x},\tau) \propto a^{-1}$. This means that in the linear regime, any initial curl decays with the expansion of the Universe.

The theoretical matter power spectrum in linear theory (also called, the 0-loop or the tree-level contribution) is thus:

$$P^{(0)}(k,\tau) = \left[D_1^{(+)}\right]^2 P_{\mathrm{m}}^{\mathrm{L}}(k), \tag{1.107}$$

where $P_{\mathrm{m}}^{\mathrm{L}}(k)$ is defined in equation (1.80). Linear theory describes well the CDM fluctuations

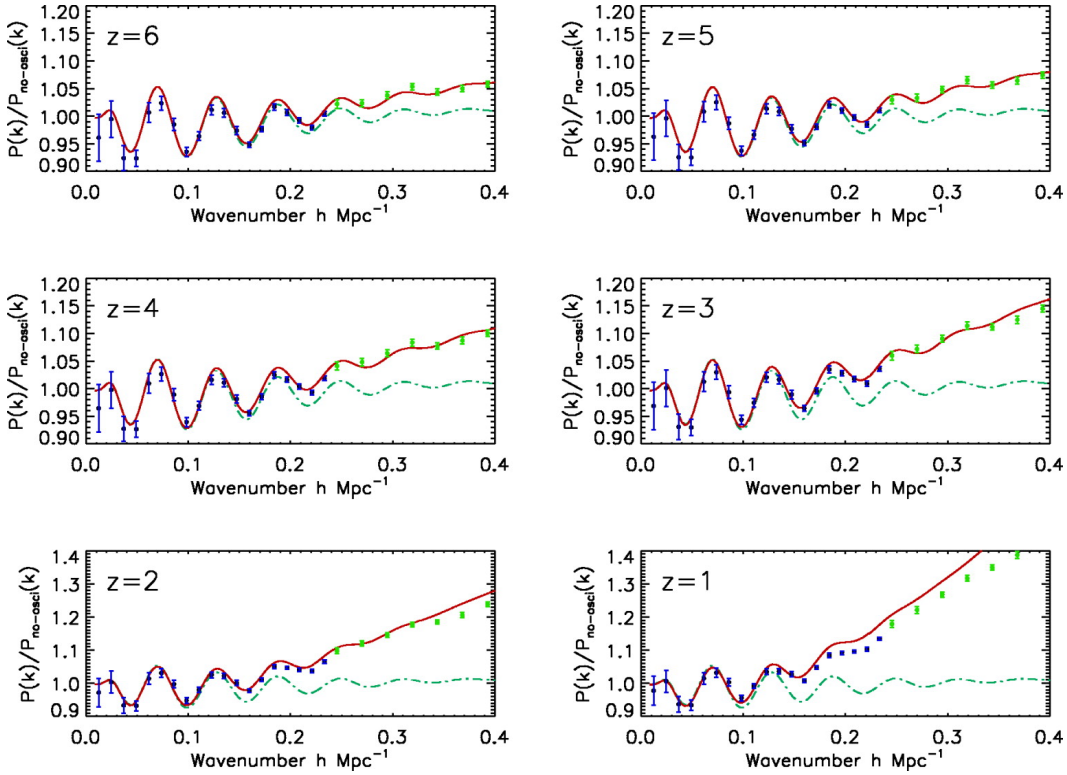


Figure 1.18: The blue and green points denote the power spectra of N-body simulations. The dot-dashed lines illustrate the linear power spectrum. The solid red lines denote the power spectrum up to 1-loop correction (i.e. the density field up to third order in Eulerian perturbation theory). The power spectra are divided by the no-wiggle (no-oscillation, i.e. no BAO wiggles) power spectrum computed as in Eisenstein & Hu (1998). The curves shown at different redshifts. Figure 3 from Jeong & Komatsu (2006)

on scales larger than 0.1 h/Mpc (see Figure 1.18). Theoretical modelling below this threshold requires the addition of non-linear terms into the description of the density field.

Eulerian Non-Linear Perturbation Theory

Perturbation theory is based on the fact that linear fluctuations are small and thus supposes that density and velocity fields can be approximated starting from the linear solutions $\delta^{(1)}$ and $\theta^{(1)}$, on top of which one adds non-linear terms – e.g. $\delta^{(2)}$ and $\theta^{(2)}$ quadratic in the initial density field:

$$\delta(\boldsymbol{x},\tau) = \sum_{n=1}^{\infty} \delta^{(n)}(\boldsymbol{x},\tau), \qquad \theta(\boldsymbol{x},\tau) = \sum_{n=1}^{\infty} \theta^{(n)}(\boldsymbol{x},\tau) \tag{1.108}$$

In this scenario, one must suppose that both the initial vorticity and the σ_{ij} are zero to ensure that the vorticiy degrees of freedom can be neglected. A non-zero initial vorticity would be amplified by non-linear effects at small enough scales.

In contrast to the linear theory, when one computes the Fourier transform of the equations (1.98) and (1.99), one obtains:

$$\frac{\partial\tilde{\delta}(\boldsymbol{k},\tau)}{\partial\tau}+\tilde{\theta}(\boldsymbol{k},\tau)=-\int \mathrm{d}^3\boldsymbol{k}_1\mathrm{d}^3\boldsymbol{k}_2\delta_D(\boldsymbol{k}-\boldsymbol{k}_{12})\alpha(\boldsymbol{k}_1,\boldsymbol{k}_2)\tilde{\theta}(\boldsymbol{k}_1,\tau)\tilde{\delta}(\boldsymbol{k}_2,\tau) \tag{1.109}$$

$$\frac{\partial\tilde{\theta}(\boldsymbol{k},\tau)}{\partial\tau}+\mathscr{H}(\tau)\tilde{\theta}(\boldsymbol{k},\tau)+\frac{3}{2}\Omega_{\mathrm{m}}(\tau)\mathscr{H}^2(\tau)\tilde{\delta}(\boldsymbol{k},\tau)= \tag{1.110}$$

$$=-\int \mathrm{d}^3\boldsymbol{k}_1\mathrm{d}^3\boldsymbol{k}_2\delta_D(\boldsymbol{k}-\boldsymbol{k}_{12})\beta(\boldsymbol{k}_1,\boldsymbol{k}_2)\tilde{\theta}(\boldsymbol{k}_1,\tau)\tilde{\theta}(\boldsymbol{k}_2,\tau), \tag{1.111}$$

where one can observe the mode coupling that occurs due to the non-linear terms through:

$$\alpha(\boldsymbol{k}_1,\boldsymbol{k}_2)\equiv\frac{\boldsymbol{k}_{12}\cdot\boldsymbol{k}_1}{k_1^2},\qquad\qquad \beta(\boldsymbol{k}_1,\boldsymbol{k}_2)\equiv\frac{k_{12}^2(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)}{2k_1^2k_2^2} \tag{1.112}$$

, where $\boldsymbol{k}_{12}=\boldsymbol{k}_1+\boldsymbol{k}_2$ and δ_D is the three-dimensional Dirac delta distribution. The mode coupling implies the non-locality[25] of the density and velocity fields evolution, which occurs already at the second order perturbation theory (Peebles, 1980).

In order to compute higher order density and velocity field terms, there are clear recipes that involve mathematical kernels and recursive relations. In a similar way, one can compute higher order corrections for the power spectrum, by including the non-linear density terms. In this case, we provide the correction up to the first order (also known as 1-loop):

$$P(k,\tau)=P^{(0)}(k,\tau)+P^{(1)}(k,\tau), \tag{1.113}$$

where

$$P^{(1)}(k,\tau)=\left[D_1^{(+)}\right]^4\left(P_{22}(k,\tau)+P_{13}(k,\tau)\right). \tag{1.114}$$

The i,j indices from $P_{ij}(k)$[26] denote the order of the density field correction $\delta^{(i)}$, $\delta^{(j)}$.

Figure 1.18 displays the resulting power spectrum in comparison to the linear case and the reference N-body simulation. One can observe that adding non-linear terms – the 1-loop correction – to the power spectrum, improves the match with the reference. Nevertheless, the more detailed study of Gil-Marín et al. (2012) shows that a precise (i.e. less than on per cent deviation) description of the N-body reference can be achieved only up to $k\approx 0.05\,h/\mathrm{Mpc}$ for $z=0$ and $k\approx 0.1\,h/\mathrm{Mpc}$ for $z=1$, by adding the 1-loop correction. However, higher order corrections can improve considerably the agreement.

[25] The non-locality refers to the fact that the evolution of the density field at a certain position depends on the values of the density field at other positions as well.

[26] Depending on the convention, P_{13} might be multiplied by a factor of two, see Jeong & Komatsu (2006); Gil-Marín et al. (2012)

1.3.5 Lagrangian perspective

In contrast to Eulerian perspective where one follows the evolution of fields, in the Lagrangian framework, one tracks the individual trajectories of particles (or fluid elements). The initial position $\boldsymbol{q}$ of a particle is connected to the final Eulerian position $\boldsymbol{x}$ by the displacement field $\boldsymbol{\Psi}(\boldsymbol{q})$:

$$\boldsymbol{x}(\tau) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, \tau). \tag{1.115}$$

As a consequence, one performs a change of variables from $\boldsymbol{x}$ to $\boldsymbol{q}$ using the fact that the total mass has to be conserved by this change, i.e. $\overline{\rho}\,[1+\delta(\boldsymbol{x},\tau)]\,\mathrm{d}^3\boldsymbol{x} = \overline{\rho}\mathrm{d}^3\boldsymbol{q}$. In this case, the Jacobian $J(\boldsymbol{q},\tau)$ of transformation between Eulerian and Lagrangian spaces is connected to the density field:

$$1+\delta(\boldsymbol{x},\tau) = \frac{1}{\mathrm{Det}\left(\delta^K_{ij} + \dfrac{\partial \Psi_i}{\partial q_j}\right)} \equiv \frac{1}{J(\boldsymbol{q},\tau)}. \tag{1.116}$$

Furthermore, the derivatives change as follows:

$$\frac{\partial}{\partial x_i} = \left(\delta^K_{ij} + \Psi_{i,j}\right)^{-1} \frac{\partial}{\partial q_j}, \tag{1.117}$$

where $\Psi_{i,j} = \dfrac{\partial \Psi_i}{\partial q_j}$ is the tidal tensor.

Following individual particles, one can rewrite the equation of motion, i.e. equation (1.92):

$$\frac{\mathrm{d}\boldsymbol{x}^2}{\mathrm{d}\tau^2} + \mathscr{H}(\tau)\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} = -\nabla_x \Phi \tag{1.118}$$

and further apply a divergence and replace the coordinates equation (1.115), such that one obtains:

$$J(\boldsymbol{q},\tau)\nabla_x \cdot \left[\frac{\mathrm{d}\boldsymbol{\Psi}^2}{\mathrm{d}\tau^2} + \mathscr{H}(\tau)\frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}\tau}\right] = \frac{3}{2}\Omega_{\mathrm{m}}(\tau)\mathscr{H}^2(J-1). \tag{1.119}$$

One can further change the ∇_x using equation (1.117) and then express the equation of motion completely with Lagrangian coordinates.

Interestingly, the regions where shell crossing induced by multi-stream flow occurs[27] make the Jacobian zero. In practice, the Lagrangian Perturbation Theory (LPT) series

$$\Psi(\boldsymbol{q},\tau) = \Psi^{(1)}(\boldsymbol{q},\tau) + \Psi^{(2)}(\boldsymbol{q},\tau) + \ldots \tag{1.120}$$

converges until the first shell crossing takes place. Consequently, the LPT predictions are

[27] the Eulerian final positions $\boldsymbol{x}$ where particles with different initial positions $\boldsymbol{q}$ arrive at the same time

limited by this moment. Nevertheless, Rampf & Hahn (2021) have shown that using LPT one can robustly study the first shell crossing in a ΛCDM Universe.

Equation (1.116) expresses the intrinsic non-linear connection between the density field and the displacement field. This means that a slight change in the displacement field of a particle introduces non-linear information into the Eulerian density and velocity fields. Mathematically, this is observed as non-zero Eulerian PT kernels for all orders, when the LPT is truncated to a given order. In order words, even the first order LPT includes non-zero higher order Eulerian PT terms.

Zel'dovich Approximation and Linear Perturbation Theory

The Zel'dovich Approximation (ZA; Zel'dovich, 1970; White, 2014) provides a solution for the linear displacement field $\Psi^{(1)}(\boldsymbol{q},\tau)$ by imposing the linear Eulerian PT solution at large scales. Expanding equation (1.116) and keeping the first order terms, one obtains:

$$\nabla_q \boldsymbol{\Psi}^{(1)} = -D_1(\tau)\delta(\boldsymbol{q}), \tag{1.121}$$

where ZA is implicitly used. Neglecting vorticity, this equation completely determines the first order displacement, where D_1 is found again with equation (1.102)[28].

The solutions for the displacement fields are curl-free up to the second-order, assuming the initial conditions are in the growing mode. Consequently, it is often useful to define Lagrangian potentials φ such that:

$$\boldsymbol{\Psi}^{(1)} = -D_1^{(+)}\nabla_q\varphi^{(1)} \qquad\qquad \boldsymbol{x} = \boldsymbol{q} - D_1^{(+)}\nabla_q\varphi^{(1)}. \tag{1.122}$$

Furthermore, the velocity field $\boldsymbol{u}$ can be computed as follows:

$$\boldsymbol{u} = -D_1^{(+)} f_1 \mathscr{H}(\tau)\nabla_q\varphi^{(1)}, \tag{1.123}$$

where $f_1 \equiv \left(\mathrm{d}\ln D_1^{(+)}\right) / \left(\mathrm{d}\ln a\right)$.

Further analysis of this solution shows that the particle evolution is local, i.e. it is independent on the other particles. This physically means that when multiple CDM streams cross, they do not interact between themselves and thus the high-density regions are incorrectly too diffuse using ZA.

Nevertheless, ZA is a useful theoretical description of non-linear structure formation. Applying ZA in equation (1.116), the density field becomes:

$$1 + \delta(\boldsymbol{x},\tau) = \frac{1}{[1-\lambda_1 D_1(\tau)]\,[1-\lambda_2 D_1(\tau)]\,[1-\lambda_3 D_1(\tau)]}, \tag{1.124}$$

[28] If one replaces $\boldsymbol{\Psi}^{(1)}$ in equation (1.119), one obtains equation (1.102)

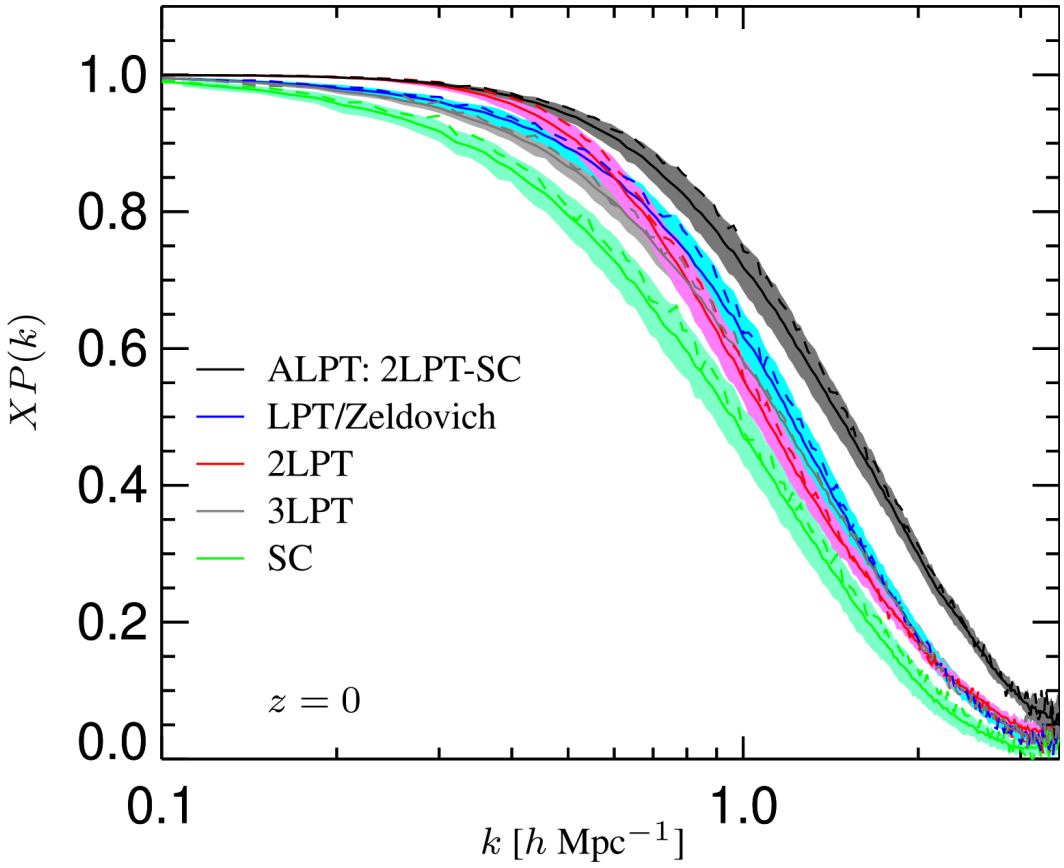


Figure 1.19: $XP(k) \equiv \langle|\delta_{\rm approx}(\boldsymbol{k})\delta^*_{N-{\rm body}}(\boldsymbol{k})|\rangle / \left(\sqrt{P_{\rm approx}(k)}\sqrt{P_{N-{\rm body}}(k)}\right)$ denotes the normalised cross-power spectra between the evolved matter fields with N-body simulation and analytical approximations from Lagrangian Perturbation Theory and Spherical Collapse (SC) model. Figure 4 from Kitaura & Hess (2013)

where $(\lambda_1, \lambda_2, \lambda_3)$ are the three local eigenvalues of the $\Psi_{i,j}$. The values of λ describe four evolution scenarios for the density field $\delta(\boldsymbol{x})$:

1. planar collapse, when one eigenvalue is positive and larger than the rest;
2. filamentary collapse, when two eigenvalues are positive and larger than the third;
3. spherical collapse, when all eigenvalues are positive and equal;
4. evolution of an underdense region, when all eigenvalues are negative.

Figure 1.19 displays in blue the normalised cross-power spectrum between the ZA evolved CDM density field and a N-body simulation. One can observe that up to $k \approx 0.1\,h$/Mpc, ZA density field is almost entirely correlated to the N-body simulation. Nonetheless, a high degree of correlation is maintained up to $k \approx 0.3\,h$/Mpc.

Higher order Perturbation Theory

The second-order LPT (2LPT) correction to the displacement field is:

$$\nabla_q \boldsymbol{\Psi}^{(2)} = \frac{1}{2} D_2(\tau) \sum_{i \neq j} \left(\Psi^{(1)}_{i,i} \Psi^{(1)}_{j,j} - \Psi^{(1)}_{i,j} \Psi^{(1)}_{j,i} \right), \tag{1.125}$$

where $D_2(\tau)$ is the second-order growth factor. This factor can be approximated by $D_2(\tau) \approx -\frac{3}{7} D_1^2(\tau) \Omega_{0\mathrm{m}}^{-1/143}$ up to 0.6 per cent precision for flat ΛCDM with $0.01 \leq \Omega_{0\mathrm{m}} \leq 1$. In a similar way to the ZA, one can express the solutions using the Lagrangian potentials φ:

$$\boldsymbol{\Psi}^{(2)} = D_2^{(+)} \nabla_q \varphi^{(2)} \qquad \boldsymbol{x} = \boldsymbol{q} - D_1^{(+)} \nabla_q \varphi^{(1)} + D_2^{(+)} \nabla_q \varphi^{(2)}. \tag{1.126}$$

Moreover, the velocity field $\boldsymbol{u}$ has the form:

$$\boldsymbol{u} = -D_1^{(+)} f_1 \mathscr{H}(\tau) \nabla_q \varphi^{(1)} + D_2^{(+)} f_2 \mathscr{H}(\tau) \nabla_q \varphi^{(2)}, \tag{1.127}$$

where $f_i \equiv \left(\mathrm{d} \ln D_i^{(+)} \right) / \left(\mathrm{d} \ln a \right)$.

Figure 1.19 shows that the 2LPT substantially improves the evolution of the density field with respect to the ZA at smaller scales. A high correlation with the N-body is maintained up to $k \approx 0.4\, h/\mathrm{Mpc}$ for 2LPT. This significant improvement occurs due to the inclusion of the non-local aspect (also called gravitational tidal effects) of the gravitational instability in the 2LPT.

In contrast to Eulerian PT, LPT does not provide a recursive solution to determine higher order terms. Moreover, it has been shown that in most interesting cases, the third-order LPT (3LPT) is not improving significantly the clustering description compared to the 2LPT, see for example Figure 1.19. Consequently, it is beyond the scope of this thesis to study in more details the higher-order corrections.

As previously mentioned, LPT is limited by the shell crossing, thus it cannot describe well the small scales. Therefore, Kitaura & Hess (2013) have proposed to use a combination of the 2LPT to evolve the large scales and the Spherical Collapse approximation to model the small scales. This is called Augmented-LPT (ALPT).

Spherical Collapse

Let us imagine a ΛCDM Universe, in which the CDM density field fluctuations are described by δ, the average matter density is $\overline{\rho}_{\mathrm{m}}(t)$. If we focus our attention onto a spherical overdense region like the one in Figure 1.20, one can interpret it as a part of the Universe with an average matter density $\tilde{\rho}_{\mathrm{m}} > \overline{\rho}_{\mathrm{m}}$, where the Friedmann equation (1.33):

$$\frac{\ddot{\tilde{a}}}{\tilde{a}} = -\frac{4\pi \mathrm{G}}{3} \left[\frac{3}{c^2} p_\Lambda + \tilde{\rho}_{\mathrm{m}} \right], \tag{1.128}$$

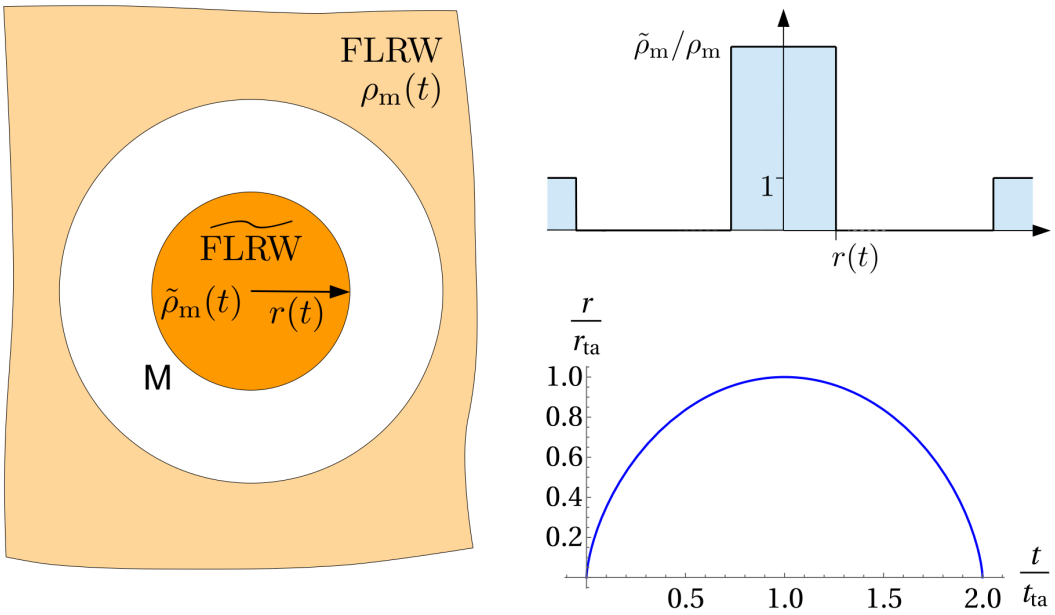


Figure 1.20: A sketch of a Spherical Collapse together with its time evolution in a matter dominated Universe. "ta" denotes the turnaround point after which the overdensity collapses. Figure 12.7 from Dodelson & Schmidt (2020)

with p_Λ being the pressure imposed by the Λ. This results into the Newton equation of motion for a spherical mass of proper radius $r(t)$ and mass M:

$$\ddot{r}(t) = -\frac{GM}{r^2(t)} + \frac{8\pi G}{3}\rho_\lambda r(t). \tag{1.129}$$

This equation can be solved numerically.

Nonetheless, in a matter dominated Universe, one can obtain analytical solutions as function of a parameter θ:

$$r(t) = \frac{r_{\rm ta}}{2}(1-\cos\theta), \tag{1.130}$$

$$t = \frac{t_{\rm ta}}{\pi}(\theta - \sin\theta). \tag{1.131}$$

Figure 1.20 displays the radius $r(t)$ as function of time. Initially, the fluctuation increases in size due to the expansion of the Universe. The maximum achieved size of the spherical overdensity is the turnaround radius $r_{\rm ta}$ and this occurs at time $t_{\rm ta}$. After this moment, the fluctuation starts to collapse. The turnaround point depends on the initial size and δ.

This is the Spherical Collapse (SC) model, an approximation that provides an analytical understanding of the non-linear evolution of spherical perturbations (see e.g. Peebles, 1980; Rich, 2010; Dodelson & Schmidt, 2020).

One can further estimate the critical value of the overdensity $\delta_{\rm cr}$ that constitutes a threshold

above which, fluctuations would collapse. Dodelson & Schmidt (2020) argue that the inclusion of the Λ in the computation of this threshold is minor and provides the $\delta_{\rm cr} \approx 1.686$. Additionally, it has been shown that if the density of the perturbation reaches values that are ≈ 200 times larger than the background density, it collapses into a DM halo. Therefore, this threshold can be used to find haloes in DM simulations, see Section 1.3.6. The SC model can be further used to estimate the number density of halos per halo mass (also called halo mass function).

In addition, the SC model can be used to estimate the time evolution of the fluctuations (Bernardeau, 1994):

$$1 + \delta(\boldsymbol{x}, \tau) \approx \left(1 - \frac{2}{3}\delta^{(1)}\right)^{-3/2}, \tag{1.132}$$

where $\delta^{(1)} = \delta^{\rm L}_{\rm m}(\boldsymbol{x}) D_1(\tau)$ is the linearly evolved initial density field (i.e. after recombination, see Sections 1.3.2 and 1.3.4). Furthermore, Mohayaee et al. (2006) have estimated the divergence of the displacement field based on equation (1.132):

$$\nabla_q \cdot \boldsymbol{\Psi}_{\rm SC} = 3\left[\left(1 - \frac{2}{3}\delta^{(1)}\right)^{1/2} - 1\right]. \tag{1.133}$$

As previously mentioned, Kitaura & Hess (2013) have combined the power of 2LPT to describe the large scales together with the SC approximation for the small scales to develop ALPT. Figure 1.21 shows a comparison of different structure formation models. One can observe that 2LPT exhibits strong shell crossing, while for ALPT this effect is reduced in knots (nodes) and in thick filaments. The additional success of ALPT is that it can capture small filaments that are formed in the N-body simulations, but not present in 2LPT.

Figure 1.19 shows that the SC has lower power even than the ZA, proving that 2LPT is indeed necessary for larger scales. Moreover, ALPT has more power towards smaller scales than 2LPT, having a strong correlation with the N-body simulation up to $k \approx 0.5h/\text{Mpc}$. Additionally, the decrease in power for ALPT is less steep than for 2LPT.

Tosone et al. (2021) have further improved the description of smaller scales using generalisations of the SC model. Therefore, together with the 2LPT at large scales, they have surpassed the ALPT performance.

1.3.6 N-body simulations

A detailed presentation of the N-body simulations is beyond the purpose of this thesis, therefore we refer to the review of Angulo & Hahn (2022) and the references therein. The Eulerian and Lagrangian perspectives together with Perturbation Theory helped at solving the Vlasov equation for the collisionless fluid-like CDM that evolves under the gravitational interaction. Nevertheless, as previously discussed, PT fails to describe strongly non-linear evolution, more specifically, it cannot describe the gravitational evolution after shell crossing (adding spherical

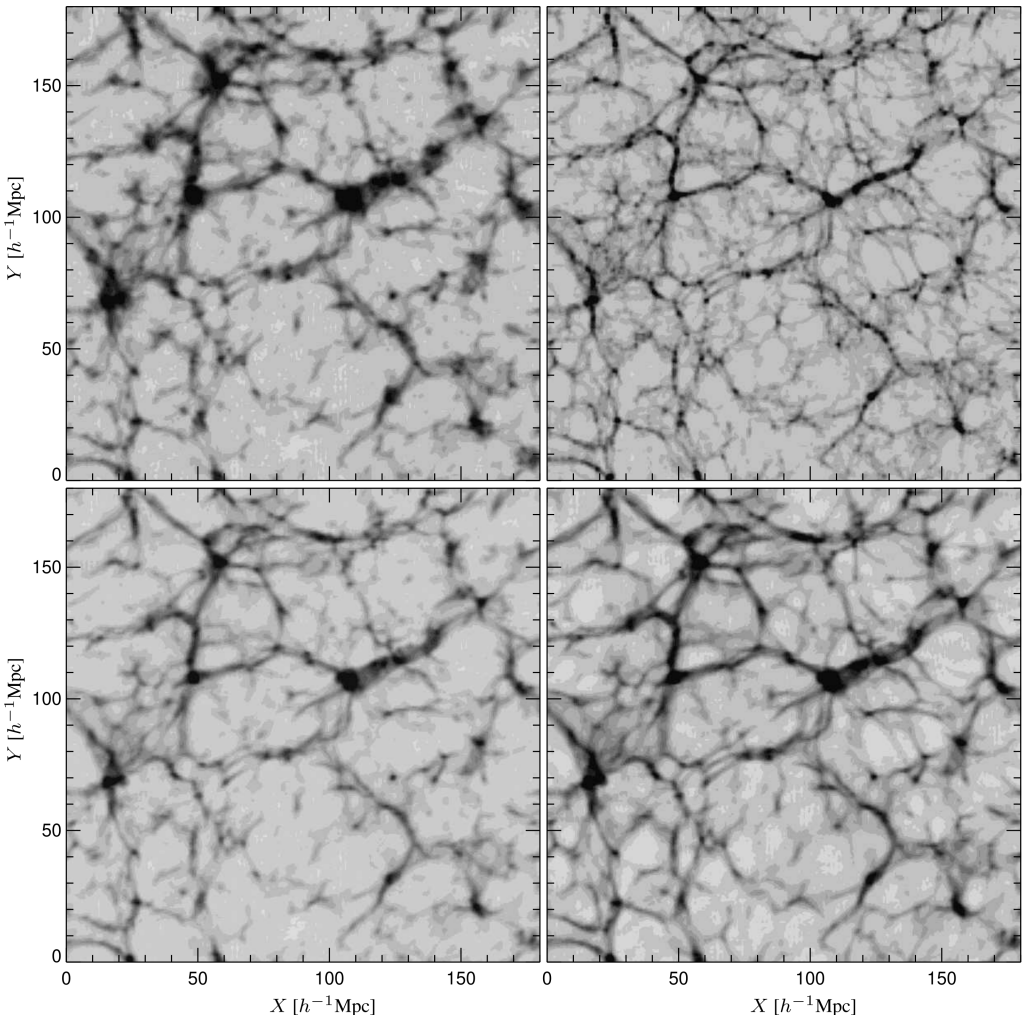


Figure 1.21: Density fields evolved with different structure formation models. Upper-left: 2LPT; upper-right: N-body simulation; lower-left: 2LPT with collapse threshold; lower-right: ALPT (2LPT + SC). Figure 2 from Kitaura & Hess (2013)

collapse can help, as shown for ALPT, however). Figures 1.18 1.19 1.21 show comparisons between N-body simulations and PT.

A solution to the non-linear modelling is to perform N-body simulations, that practically solve a discrete version of the Vlasov equation (i.e. Hamiltonian equations of motion) after sampling the phase space using N particles with $(\boldsymbol{X}_i, \boldsymbol{P}_i), i = 1...N$:

$$f_N(\boldsymbol{x}, \boldsymbol{p}, \tau) = \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \sum_{i=1}^{N} \frac{M_i}{m} \delta_D\left(\boldsymbol{x} - \boldsymbol{X}_i(\tau) - \boldsymbol{n}L\right) \delta_D\left(\boldsymbol{p} - \boldsymbol{P}_i(\tau)\right), \tag{1.134}$$

where L is the side-length of a box, on which one imposes periodic boundary conditions. $M_i = \overline{M} = \Omega_{0\mathrm{m}}\rho_{0\mathrm{c}}V/N$ represents the particle mass in the simulation and m is the mass of the actual CDM "particle". Nonetheless, there are simulations where N particles can have

different masses.

An important aspect of the N-body simulations is setting the initial positions and velocities of the N particles. A traditionally used technique – back-scaling – in N-body simulations is to solve the linear order Einstein-Boltzmann EB) equations up to $a = a_{\rm target}$ (i.e. the target time in the history of the Universe), in order to get a density field that includes the physics before the last scattering, see Section 1.3.2. Practically, a Gaussian random field is sampled to match the output power spectrum of EB numerical codes. Furthermore, the resulting field is scaled linearly to very early times $a \to 0$ and rescaled using the required nLPT approximation to $a_{\rm start}$ (the initial redshift of the N-body simulation, usually $z_{\rm start} \approx 100$). Finally, the particles are allowed to evolve under the gravitational interaction.

Starting from the Vlasov equation, the equations of motion[29] for the N gravitationally interacting bodies are:

$$\frac{\mathrm{d}\boldsymbol{X}_i}{\mathrm{d}\tau} = \frac{\boldsymbol{P}_i}{M_i a} \qquad\qquad \frac{\mathrm{d}\boldsymbol{P}_i}{\mathrm{d}\tau} = -aM_i \nabla_x \Phi|_{\boldsymbol{X}_i}, \tag{1.135}$$

where Φ is sourced by the fluctuations in the matter field and can be computed from the discrete estimation of ρ – obtained from equation (1.134) – using Poisson equation. The discretisation of ρ sets the quality the force calculation and how close the simulation is to the continuous limit.

Apart from the number of particles and their mass, N-body simulations depend on the time evolution (in practice, the number of time steps and the order of the steps to evolve the particles) and on the force calculation.

Time evolution

Considering that the phase-space (conjugate coordinates and momenta) area must be conserved [30], specific numerical integration techniques have been developed to accommodate this demand. One such technique is the second order "leap-frog" integrator, which applies a drift-kick-drift (DKD) scheme or KDK one in one step, where the drift updates the positions of the particles and the scale factor and the kick updates the linear momenta of the particles.

Lastly, there is no optimal choice of the time steps as it depends on the details of the simulations, such as redshift or force accuracy. However, some examples of time stepping schemes are: linearly or logarithmicaly spaced scale factor steps; schemes that decrease the time step with the evolution of the simulation; hierarchical time stepping schemes, where there are two kick operators, one for 'slow' particles and one for the 'fast' particles.

[29] We show the equation of motion using the conformal time τ, but Angulo & Hahn (2022) work with cosmic time t.

[30] One must also check for other quantities such as the total energy or total angular momentum.

Gravity solver

The most time consuming part of a N-body simulation is the computation of the gravitational interactions. Some important techniques are

1. Particle-Mesh (PM) based methods solve the Poisson equation in Fourier space, after the mass of each particle is assigned on a fixed grid. The technique is efficient when it can take advantage of the periodic boundary conditions, otherwise the grid must be zero padded, increasing thus the memory requirements. Finally, the resulting forces on the grid are reversely interpolated to the particles positions.

2. The direct Particle-to-Particle (P2P) summation technique calculates the interactions directly at the particle level. Therefore, the gravitational potential in which a particle i is found reads:

$$\Phi(\boldsymbol{x}_i) = -a^{-1} \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \left[\sum_{j=1, i\neq j}^{N} \frac{GM_j}{||\boldsymbol{X}_i - \boldsymbol{X}_j - \boldsymbol{n}L||} + \varphi_{\text{box},L}\left(\boldsymbol{X}_i - \boldsymbol{n}L\right) \right], \tag{1.136}$$

   summing the effects of all other particles. The uniform background density is taken into account into the box potential $\varphi_{\text{box},L}$. Thus, Φ is sourced by the density contrast $\rho - \overline{\rho}$. This summation scales with $\mathcal{O}\left(N^2\right)$, meaning that it can become easily very expensive from the computation point of view.

3. Hybrid methods split the potential into two terms: a long range potential that can be estimated using PM methods and a short range that can be calculated using P2P, or tree technique. The PM - P2P combination is called P^3M method.

4. The hierarchical tree methods organise the N particles in tree structures based on the distances between themselves. Therefore, one can compute the gravitational potential of groups (represented by a node in the tree) of particles and replace the P2P by particle-node interaction. Due to the fact that the depth of the tree is $\mathcal{O}\left(\log N\right)$, the complexity of the force calculation decreases from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}\left(N\log N\right)$.

FASTPM (Feng et al., 2016) is an N-body code that is used in Chapter 2. It uses a modified set of kick and drift operators that include – during the time step – an acceleration motivated by the ZA equation of motion. This allows for a description of the large scales that is in agreement with ZA, but with a significantly reduced number of steps compared to a full N-body simulation. Lastly, it uses the Particle-Mesh technique to compute the gravitational interactions.

Halo detection

Once the DM density field is evolved under the gravitational interaction using N-body simulations, one can detect bound structures such as haloes that can be further used to assign

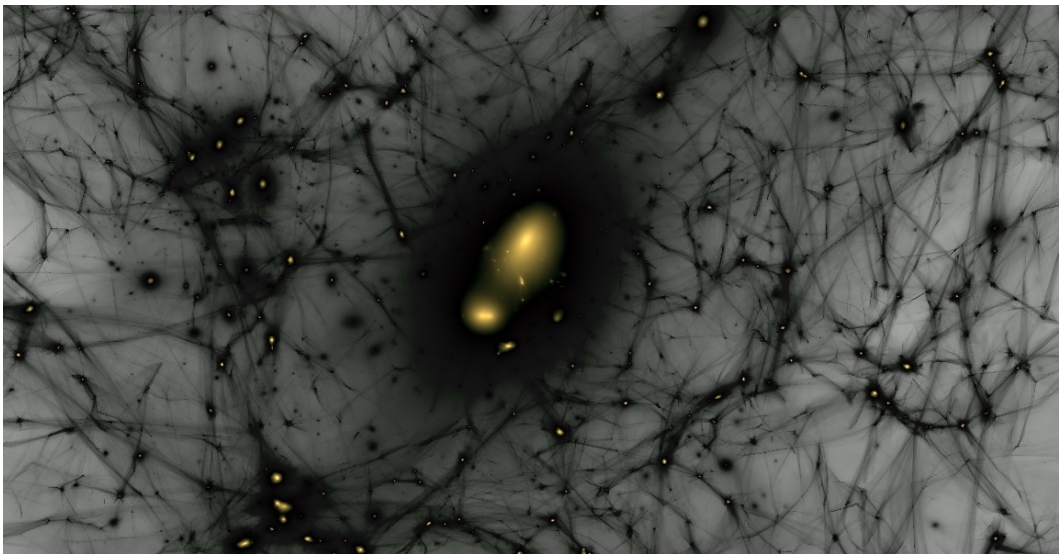


Figure 1.22: A CDM simulation with visible large-scale structures (haloes, filaments, etc). The DM haloes can host multiple luminous galaxies (yellow). Figure from https://news.fnal.gov/2020/04/the-milky-ways-satellites-help-reveal-link-between-dark-matter-halos-and-galaxy-formation/

galaxies as described in Chapter 2. There are many halo finders (see e.g. Knebe et al., 2011, for a comparison), but one can classify them in two main categories: particle collector algorithms (e.g. Friends-of-Friends (FOF) Davis et al., 1985; Behroozi et al., 2013) and density peak locator (e.g. spherical overdensity (SO) Warren et al., 1992; Hadzhiyska et al., 2022).

The FOF algorithms connect particles that are closer than a certain characteristic length and that are found in a region with a density above a threshold. The resulting collection of particles is considered the virialised[31] halo. FOF can be applied on the 3D configuration or 6D phase space.

The SO methods identity the density peaks and consider them as the centres of the haloes. Particles are added in a sphere whose size is increased until the enclosed density reaches a certain threshold (e.g. virialisation criterion).

The position of a halo can be considered to be the position of the maximum density peak or the average location of all the particles inside it. The halo velocity can be estimated as an average particle velocity. Lastly, the mass is simply the sum of all mass particles inside the halo. It is obvious that these properties depend strongly on the border (shape) of the halo. Finally, these properties play an important role in galaxy assignment on haloes (Wechsler & Tinker, 2018). Figure 1.22 illustrates a CDM simulation, in which the DM haloes host luminous galaxies (see Section 2.2 for methods to create galaxy simulations).

[31] A virialised object follows the virial theorem, i.e. a stable set of discrete particles at equilibrium bound by a conservative force must follow $2K + U = 0$, where U is the total potential energy of the system and K is the kinetic energy

1.4 Mapping the Universe

A rapidly evolving method to probe and better understand the Universe and its evolution is to map its structure through the three-dimensional positioning of matter tracers. On one hand, the 2D angular position on the sky is usually measured from wide field photometric surveys such as the Sloan Digital Sky Survey (SDSS; York et al., 2000) or the Legacy Surveys[32] (LS; Dey et al., 2019). On the other hand, the third dimension is represented by the redshift which can be related to distances, as discussed in Section 1.2.2

The redshift of a light-source can be estimated using photometric data, see e.g. Nishizawa et al. (2020); Zhou et al. (2021), however its precision is inferior to the redshift measured from a light-spectrum. Therefore, large scale spectroscopic surveys have been devised to measure the light-spectra of sources selected from photometric data.

The recently finished spectroscopic surveys part of SDSS – Baryon Oscillation Spectroscopic Survey (BOSS) and its extension extended-BOSS (eBOSS) – have mapped over 2 million[33] galaxies and quasars in more than ten years (Alam et al., 2021) with a 2.5 m telescope and 1000 optical fibres[34]. The on-going Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al., 2022) plans to map 40 million galaxies and quasars in five years with a 4 m telescope and 5000 optical fibres. DESI's proposed successor, MegaMapper(Schlegel et al., 2022), aims at measuring 100 million spectra in $2 < z < 5$ redshift range, using a 6.5 m telescope and 26000 optical fibres. A similar project, MUltiplexed Survey Telescope (MUST), for the northern sky is conceived by Zhang et al. (2023): a 6.5 m telescope with 20000 optical fibres, possibly located in Northwest China.

The Cosmology Redshift Survey (CRS; Richard et al., 2019), part of the 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al., 2012, 2019) consortium, is dedicated to mapping ≈ 7 million galaxies and quasars in the southern sky. Since DESI primarily targets the northern sky, CRS collects complementary data to the DESI. The quasar redshift range of the DESI is similar to the one of CRS, however CRS covers galaxies only up to $z \approx 1$ while DESI reaches $z \approx 1.6$ with the help of emission line galaxies.

The EUCLID (Laureijs et al., 2011; Euclid Collaboration et al., 2022) 1.2 m space telescope, which was recently launched, will explore the redshift range of $1 < z < 2$ and aims to measure 30 million spectroscopic redshifts in approximately six years. As a result, EUCLID effectively extends the CRS measurements to higher redshifts. In addition to the spectroscopic data, EUCLID will provide photometric information in four bands (one in the visible domain, $500 - 1000$ nm and three in the near infrared, $1000 - 2000$ nm) for approximately two billion galaxies.

[32] https://www.legacysurvey.org/

[33] Videos about the 3D map https://www.youtube.com/watch?v=VGA4NrqqYiU and https://www.youtube.com/watch?v=UTlYUxucEZA.

[34] The optical fibres guide the light from a target of interest to the spectrograph and the resulting spectrum is captured by a CCD. See Section 1.4.2 for more details

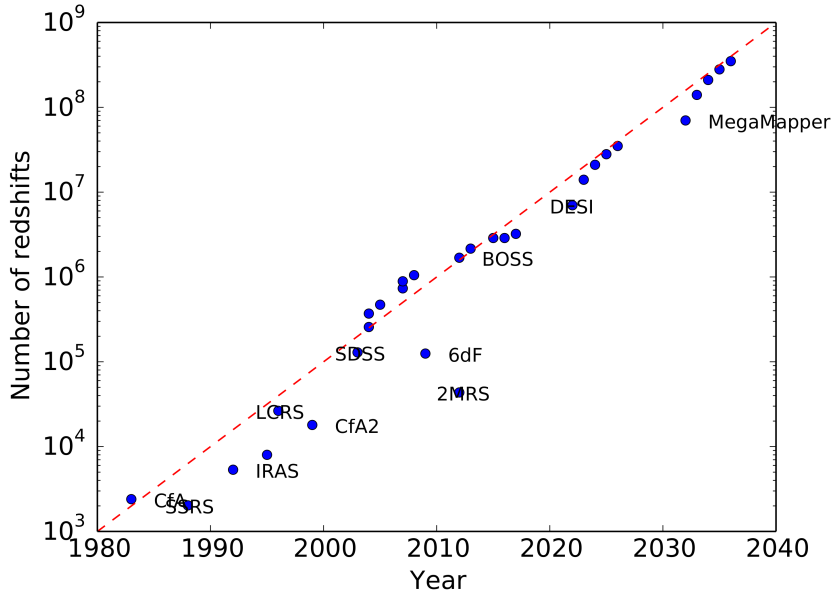


Figure 1.23: The number of galaxy redshifts for different spectroscopic surveys. The red line illustrates that every decade the sizes of different surveys increase by a factor of approximately ten. Figure 1 of Schlegel et al. (2022).

The Large Synoptic Survey Telescope (LSST; Ivezić et al., 2019) will have an 8.4 m primary mirror and will map 20 billion galaxies in the southern sky in six different optical bands, $320-1050\,\mathrm{nm}$. In this case, EUCLID can provide near infrared photometric data for the common targets. Finally, LSST and EUCLID can provide the photometric data to select the targets of interest for the future spectroscopic surveys such as MUST and MegaMapper. For cosmological forecasts of future surveys, one can consult e.g. Ivezić et al. (2019); Euclid Collaboration et al. (2020); d'Assignies D et al. (2023); Sailer et al. (2021).

A relatively new technique to probe the LSS is the 21-cm intensity mapping that traces the neutral hydrogen in the Universe. The 21-cm emission is caused by the hyperfine spin-flip transition of the electron in the neutral hydrogen. The probability of this transition is very low, however it is compensated by the large abundance of the neutral hydrogen in the Universe. Mapping the sky in different frequencies provides the 3D distribution of neutral hydrogen that can be used to compute the hydrogen clustering and detect the BAO (see e.g. Bull et al., 2015).

The proof-of-concept Canadian Hydrogen Intensity Mapping Experiment (CHIME; CHIME Collaboration et al., 2022) has measured for the first time the clustering amplitude of the neutral hydrogen from LSS (Amiri et al., 2023). This represents an important step for future radio-frequency experiments such as Packed Ultra-wideband Mapping Array (PUMA; Slosar et al., 2019), Hydrogen Intensity and Real-time Analysis eXperiment (HIRAX; Crichton et al., 2022) and Square Kilometre Array (SKA; Huynh & Lazio, 2013; Square Kilometre Array Cosmology Science Working Group et al., 2020) that aim to provide LSS measurements with a precision which is comparable to surveys like MegaMapper (Sailer et al., 2021).

The current thesis is focused on studying the LSS using 3D maps of matter tracers built by photometric and spectroscopic surveys together. Therefore, the first subsection is focused on the LS (Dey et al., 2019): Beijing-Arizona Sky Survey (BASS; Zou et al., 2017), Mayall z-band Legacy Survey (MzLS) and Dark Energy Camera Legacy Survey (DECaLS), that have been used as photometric precursors for the DESI. The second one presents BOSS/eBOSS and DESI – with a focus on the latter – and how they measure the redshifts. Lastly, starting from the 3D clustering of matter tracers, the BAO and the Redshift Space Distortions (RSD) are described as methods to measure cosmological parameters.

1.4.1 Photometric Surveys

The principle behind photometric surveys is simple: telescopes collect and reflect the photons arriving on the mirror onto the cameras (CCD array) found in the focal plane. Practically, one takes pictures of the sky and detects all bright enough light-sources found in the scanned footprint after a given exposure time.

However, one has to model the detected photometric signal in order to extract the useful information. The raw image $D(x, y)$ acquired by the telescope is determined by the flux as a function of the position (x, y) on the CCDs. This is modelled as follows:

$$D(x, y) = \big[I(x, y) + \text{Sky}(x, y)\big] F(x, y) + B(x, y), \tag{1.137}$$

where $I(x, y)$ is the scientifically interesting signal. $B(x, y)$ is called the "bias level" and it is a positive constant value set in the hardware in order diminish the readout noise and thus avoid negative values in the image. $F(x, y)$, the "flat-field" represents the response of the camera at the pixel level. Lastly, the Sky level comprises any source of light that is not of scientific interest (e.g. the Moon light, a distant town, repeated scattering of sunlight in the upper atmosphere, the airglow, the sunlight scattering off zodiacal dust grains in the solar system). While, F and B are measured, for the LS the sky-level is modelled together with the astronomical sources using TRACTOR[35].

Each light-source is modelled using an analytic profile (e.g point-source, exponential, Sérsic) creating thus a model image of the considered region, that is optimised using a χ^2 minimisation. The resulting catalogues of this routine include source positions, fluxes, shape parameters, and morphological quantities that can be used to discriminate extended sources from point-sources, together with errors on these quantities, see Dey et al. (2019).

A first important observation is that the sky is surveyed in different filters that set the wavelength (λ) intervals that are observed. This allows the probing of different regions of the light-spectrum, but in a relatively short time. The LS have mapped more than 20000 square degrees of the sky using four different filters (see Figure 1.24) in nine years. In practice, the flux f[J/cm^2/s] measured in each filter is converted into magnitudes m, such as the AB system for

[35] https://github.com/dstndstn/tractor

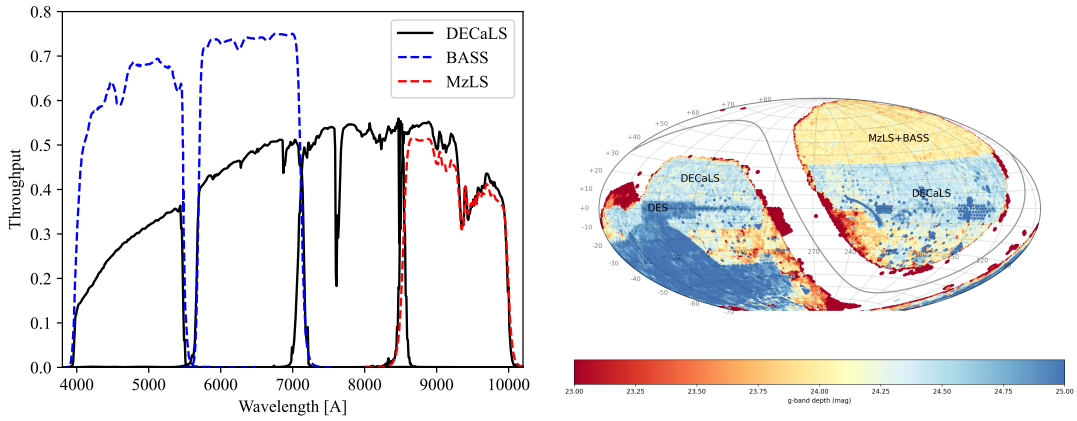


Figure 1.24: Coverage of Legacy Survey Data Release (DR) 9 and 10. The filters used by all Legacy Surveys. https://www.legacysurvey.org/status/

LS (Oke & Gunn, 1983):

$$m = 22.5 - 2.5 \log_{10} f, \tag{1.138}$$

where 22.5 is a calibration constant.

For LSS measurements, one requires to have a sample of matter tracers that are homogeneous as possible on the sky. However, there are systematic effects at the level of the photometric survey that can affect the homogeneity such as:

- the galactic extinction due to the presence of the dust (also called reddening);
- the finite exposure time that fixes the depth of the observation;
- the bright stars that can produce ghost images and thus spurious targets;
- the density of stars that can simply cover interesting targets;
- the presence of the atmosphere that degrade the image quality and disturbs the shapes of the sources (seeing).

Figure 1.25 depicts the reddening map computed by Schlegel et al. (1998) in the footprint of LS. One can observe as expected, that the extinction is more significant around the galactic plane, where the dust density is more significant. Nevertheless, one accounts for this effect in $m_{\text{corrected}}$:

$$m_{\text{corrected}} = m_{\text{measured}} - A_{\text{filter}} \times \text{EBV}, \tag{1.139}$$

where m_{measured} is the direct measured magnitude, A_{filter} is a correction factor dependent on the used filter (see Schlafly & Finkbeiner (2011)) and EBV is the value of the reddening at the

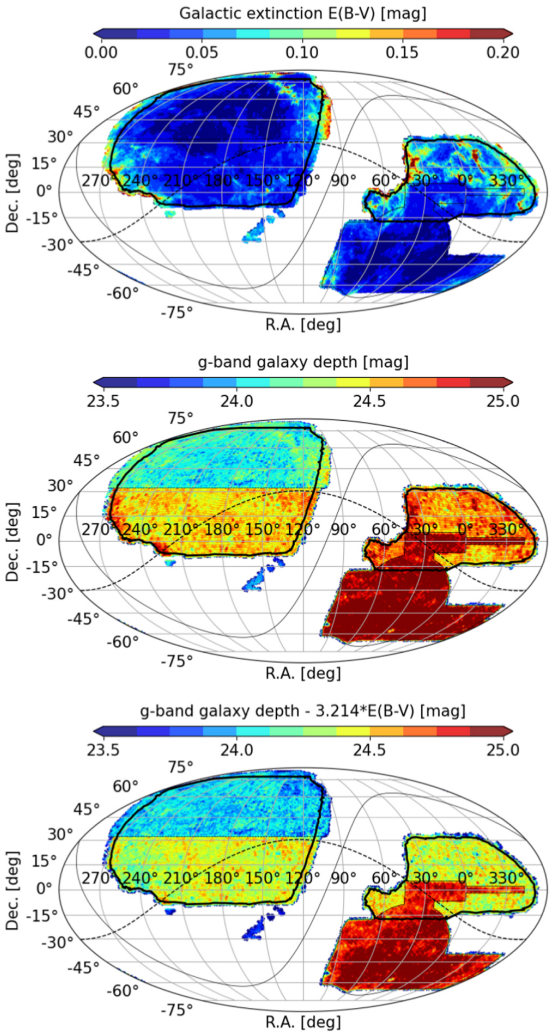


Figure 1.25: Upper panel: Dust reddening (EBV); middle panel: g-band depth of Legacy Survey; lower panel: the extinction corrected depth. Figure 1 of Raichoor et al. (2023)

position of the luminous target. In most cases, the corrected magnitudes are used for target selection, except when it is specifically mentioned.

The LS have adopted a dynamic observing strategy, meaning that they have adapted the exposure time to the sky conditions (e.g. transparency, sky brightness) and the positions on the sky (e.g. galactic dust reddening), in order to achieve a uniform depth. The second panel of Figure 1.25 shows the g-band galaxy depth, which shows that in the regions where the galactic extinction is more significant, the exposure time is larger. Therefore, the extinction-corrected depth map in the third panel is more homogeneous.

In addition, one can observe a systematic difference between the g-depth in the northern regions (BASS) and the southern ones (DECaLS). The g and r bands are shallower for BASS than for DECaLS, meaning that one has to account for this in the target selection, described in

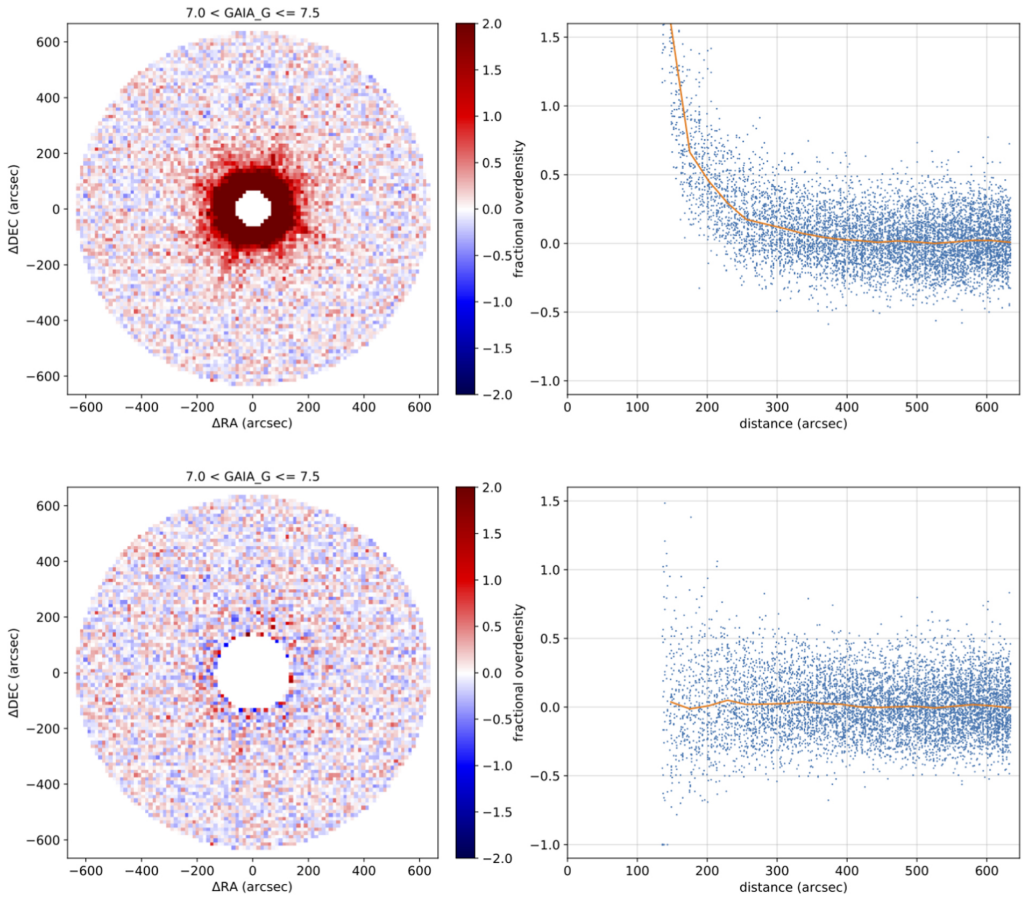


Figure 1.26: The effect of GAIA bright stars on the relative LRG densities. Left panels show fractional overdensity as function of RA and DEC. Right panels show the fractional overdensity as function from the distance to the bright star. Figure 22 of Zhou et al. (2023)

the next section.

Figure 1.26 illustrates the effect of a Gaia (Gaia Collaboration et al., 2016) bright star on the number density of targets. Around the bright star, the number of targets is much larger than it should naturally be (upper panels) due to ghost images, and scattering in the Earth's atmosphere and the telescope optics. Thus, putting a mask on stars below given magnitudes (lower panels) reduces the systematic overdensity to practically zero.

The systematic effects are studied in more details after the selection of interesting targets is performed. In practice, a certain weight corresponding to a systematic effect is attributed to each target such that the weighted sample is as close as possible to homogeneity, see e.g. Raichoor et al. (2023) for the Emission Line Galaxy sample.

Target Selection

The resulting number of unique sources in Data Release (DR) 9 of LS is approximately two billion[36]. In contrast, DESI plans to measure the spectra of ≈ 40 million extragalactic sources in five years, meaning that one has to select carefully the targets of interest from the photometric surveys. The studied targets are determined by the scientific goals and by the instrumental and physical constraints of DESI (discussed in more details in Section 1.4.2). Therefore, the selections are built to optimise the science goals given the constraints.

In practice, the selection is based on some quality and magnitude cuts and a colour selection, i.e. a difference between the magnitudes in two different filters, and morphological properties. Some examples of quality cuts include a high signal-to-noise ratio in all the bands needed for the colour selection and the removal of targets that are too close to bright stars or bright galaxies. The magnitude cuts impose a maximum exposure time required to measure the spectrum of a source with a high enough signal-to-noise ratio. Morphological properties can be used, for example, to discriminate between point sources and extended ones. Lastly, the colour selection can help to select the redshift interval of interest for the envisioned goal or to discriminate between types of galaxies.

DESI plans to study four types of extragalactic targets: Bright Galaxies (BG; Ruiz-Macias et al., 2020; Hahn et al., 2023), Luminous Red Galaxies (LRG; Zhou et al., 2020, 2023), Emission Line Galaxies (ELG; Raichoor et al., 2020, 2023) and Quasars (QSO; Yèche et al., 2020; Chaussidon et al., 2023). QSO have the highest fibre assignment priority, of the three "dark"[37] time tracers, followed by LRG and then by ELG. We provide a more detailed description of the ELG target selection as it represents the largest DESI galaxy sample and thus it is expected to yield the best constraints on cosmological parameters for measurements in $1.1 < z < 1.6$. Nevertheless, we briefly introduce the other targets.

Bright Galaxies. These targets have been selected to optimise the DESI survey during the "bright" time i.e. when the moon is bright enough, above the horizon. The plan is to create the most detailed map of the Universe for $z < 0.6$ using more than 10 million galaxies. There are three subsamples: one that is magnitude limited $r < 19.5$, a second magnitude limited one $19.5 < r < 20.175$, that is optimised using a colour-selection to achieve a high redshift efficiency and a low-redshift quasar sample. The BGS should provide the best BAO and RSD measurements for $z < 0.4$ to date.

Luminous Red Galaxies. The LRG sample is the lowest redshift "dark" time sample. The plan is to measure 8 million LRG redshifts in $0.4 < z < 1.0$, reaching a much higher density than former LRG surveys (e.g. BOSS, eBOSS), i.e. $5 \times 10^{-4}\,(h/\mathrm{Mpc})^3$. The colour selection is done using g, r, z bands from the LS and $W1$ infra-red band from WISE, but part of the

[36] https://www.legacysurvey.org/dr9/description/
[37] When the moon is not on the sky

LS catalogues. The resulting sample is robust against systematic effects and low in stellar contamination rate.

Quasars. In constrast to LRGs, QSOs cover the highest redshift intervals, i.e. $z > 0.9$. The sample of direct dark-matter tracers spans $0.9 < z < 2.1$, while the one used for Lyα forests [38] has $z > 2.1$. The QSO catalogue is selected using a random forest algorithm applied on the g, r, z, $W1$ and $W2$ magnitudes. The sample has a magnitude $< 16.5 < r < 23$ and a density of $310\,\mathrm{deg}^{-2}$ from which ≈ 70 per cent are true quasars, based on the visual inspection of the survey validation.

Emission Line Galaxies

One-third of all DESI tracers will be ELGs, that will probe the Universe in $0.6 < z < 1.6$, i.e. ≈ 80 per cent of cosmic history. The maximum $z = 1.6$ is imposed by the wavelenght coverage of the spectrographs and by the fact that the redshift measurement of ELGs is planned to be performed using the [OII] emission doublet[39], as it is an unambiguous signature in the galaxy spectrum.

The ELG target selection has been performed in two steps. Initially, a selection (see Figure 1.27) has been developed on a few hundreds of squared degrees using LS photometric data, HSC photometric redshifts(Nishizawa et al., 2020)[40] and using DEEP2 spectroscopic information for the [OII] flux, see Raichoor et al. (2020) for more information.

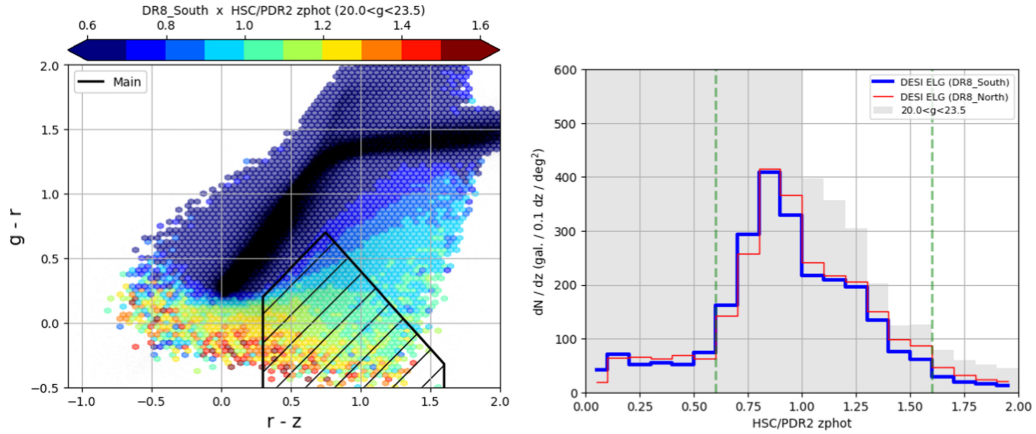
Finally, after the DESI Survey Validation (SV) program has provided spectroscopic data, the selections have been optimised, see more details in Raichoor et al. (2023). The ELG main target selection has been performed using the DR9 LS photometric data and it provides two ELG target subsamples: a) one with low fibre assignment priority (LOP) favouring $1.1 < z < 1.6$ and a target density of $1940\,\mathrm{deg}^{-2}$; b) one with very low priority (VLO) and a target density of $460\,\mathrm{deg}^{-2}$, that favours $0.6 < z < 1.1$. In practice, by randomly selecting 10 per cent of the ELG-LOP and ELG-VLO a third subsample has been defined, ELG-HIP, that has the same fibre assignment priority as the LRG.

The final selection is based on some quality criteria and a colour selection, see Figure 1.28. The latter is needed to impose the redshift range of interest and to choose targets that exhibit the [O II] doublet feature with a high enough signal-to-noise ratio – to secure a reliable measurement of the z_{spec}. The two subsamples share the quality criteria:

[38] Lyα forests are an imprint in the spectrum of a far QSO, caused by the neutral gas bubbles at different redshifts between the Earth the QSO. The neutral hydrogen strongly absorbs light at the Lyα wavelength $\lambda = 121.567\,\mathrm{nm}$, creating many absorption lines in the QSO spectrum.

[39] [OII] emission doublet are two spectral lines of oxygen at $\lambda \approx 372.6\,\mathrm{nm}$ and $\lambda \approx 372.9\,\mathrm{nm}$.

[40] Photometric redshifts are estimated using multiple magnitude bands and are less precise than the spectroscopic ones.



(a) (b)

Figure 1.27: The colour-colour diagram ($g-r$ as function of $r-z$) and the preliminary colour selection of ELGs for DESI. The colour bar on the top denotes the photometric redshift. The blue and red histograms contain the photometric redshifts of the selected targets for the south and north footprints, respectively. The grey histogram contains all objects within a given g magnitude interval. Figure 1 from Raichoor et al. (2020)

1. Each target must have at least one photometric observation in each of the three filters g, r, z;
2. The signal-to-noise ratio has to be positive in all three bands g, r, z;
3. The target must not be close to a bright star or a bright galaxy.

The ELG-LOP should, in addition, pass the following selections:

1. a magnitude cut so that it ensures the spectra can be measured in a given exposure time: $g > 20$ and $g_{\rm fibre} < 24.1$, where $g_{\rm fibre}$ is predicted from the g band flux of the object that can be observed using a 1.5" diameter optical fibre and g is the total g magnitude of the object[41];
2. $r-z > 0.15$ cut that rejects galaxies with a redshift $z > 1.6$ because the [OII] doublet is outside the DESI spectrograph;
3. $g-r < 0.5 \times (r-z) + 0.1$ selection to discriminate between stars or low redshift objects and the higher redshift targets;
4. $g-r < -1.2 \times (r-z) + 1.3$ selection to optimise the redshift range and to select targets that have a higher [OII] flux.

On the other hand, ELG-VLO should pass:

[41] Check for additional information https://www.legacysurvey.org/dr9/catalogs/

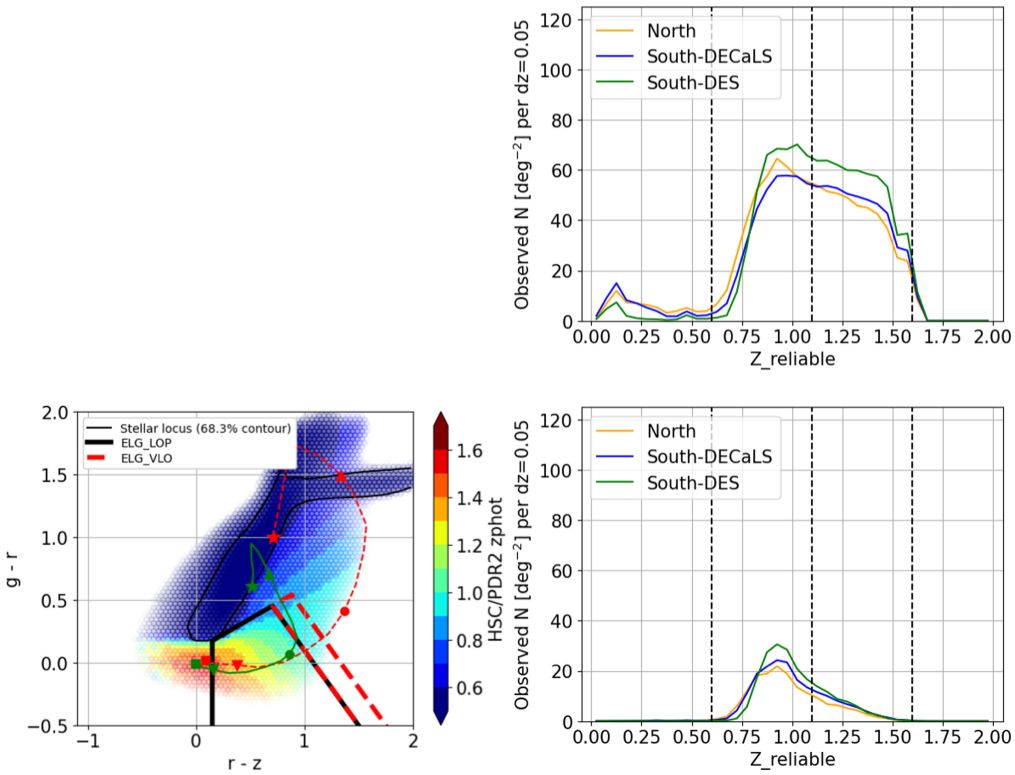


Figure 1.28: Left: The main target selection for low priority (LOP) and Very LOw priority (VLO) ELGs. The colours denote the photometric redshift. Right: The spectroscopic redshift distribution of the ELG-LOP (top) and ELG-VLO (bottom) samples after one year of DESI measurements. The vertical black lines denote the $z = 0.6$, $z = 1.0$, $z = 1.6$ redshfits. Figures 3 and Figure 21 of Raichoor et al. (2023)

1. $(g > 20)$ and $g_{\rm fibre} < 24.1$;
2. $r - z > 0.15$;
3. $g - r < 0.5 \times (r - z) + 0.1$;
4. a $(g - r > -1.2 \times (r - z) + 1.3)$ and $(g - r < -1.2 \times (r - z) + 1.6)$ selection to optimise the redshift range and to select targets the have a higher [OII] flux.

Figure 1.28 illustrates the spectroscopic redshift distribution of the ELGs selected using the previously described selections, after one year of DESI measurements. It proves the two sets of selection criteria favoured the two redshift ranges $1.1 < z < 1.6$ (LOP) and $0.6 < z < 1.1$ (VLO), respectively.

1.4.2 Spectroscopic Surveys

Photometric surveys can accurately provide the angular positions on the sky, right-ascension and declination (RA, DEC) of the targets of interest. Having measurements in different bands,

one can even estimate photometric redshifts (e.g. Nishizawa et al., 2020; Zhou et al., 2021), i.e. the third dimension[42]. Nevertheless, the current precision of these estimations is far from the requirements of LSS studies. Therefore, the 3D LSS studies rely on spectroscopic surveys to provide precise measurements of the redshift through the spectrum of the light sources (e.g. galaxies).

The latest completed large scale spectroscopic surveys are BOSS and eBOSS, both part of SDSS. The on-going DESI plans to map 40 million galaxies and quasars in five years, i.e. 20 times more targets than BOSS/eBOSS, during half the time needed by the previous generation. The measurement process is similar between the two generations: the light of a target is collected by an optical fibre and guided to a spectrograph to obtain the light-spectrum and detect it by CCDs. Nevertheless, the instrumentation has an important role in the survey speed through the exposure time and the number of spectra per observation, and the time between two successive exposures. Therefore, DESI has been designed to surpass BOSS/eBOSS.

BOSS/eBOSS Instrumentation

A brief description of the instrumentation of the two surveys is presented here. Nevertheless, for more details one can consult Gunn et al. (2006); Dawson et al. (2013, 2016).

BOSS and eBOSS have used the 2.5 meters Sloan telescope located at Apache Point Observatory and have taken advantage of 3° diameter field of view. The focal plane hosted a plate (see Figure 1.29) with 1000 holes each holding an optical fibre. The plate was designed for one exposure in a specific region of the sky, meaning that after each exposure, another plate with optical fibres would replace the former. In total there were ≈ 4000 unique plates for BOSS and eBOSS together. The 1000 optical fibres were pre-plugged during the day due to the ≈ 45 minutes long process, while switching between plates could be done with an overhead of 5 to 10 minutes during the night. Nevertheless, the instrumentation only allowed nine plates to be pre-plugged, strongly limiting the speed of the survey.

Lastly, two spectrographs – each having two arms – collected the light from the optical fibres and guided the spectrum to the CCDs, see Figure 1.29. They allowed the study of spectra in the $[350-1000]$ nm wavelength interval with a resolution $R \approx 2000$.

DESI Instrumentation

A more detailed description of DESI can be found in DESI Collaboration et al. (2016b, 2022); Silber et al. (2023) and the articles in preparation found at https://data.desi.lbl.gov/doc/papers/.

DESI is installed at the 4 m Mayall telescope at Kitt Peak National Observatory. The increased size of the telescope compared to Sloan allows for more light to be collected and thus deeper

[42] The redshift is directly connected to a distance, see equation (1.51)

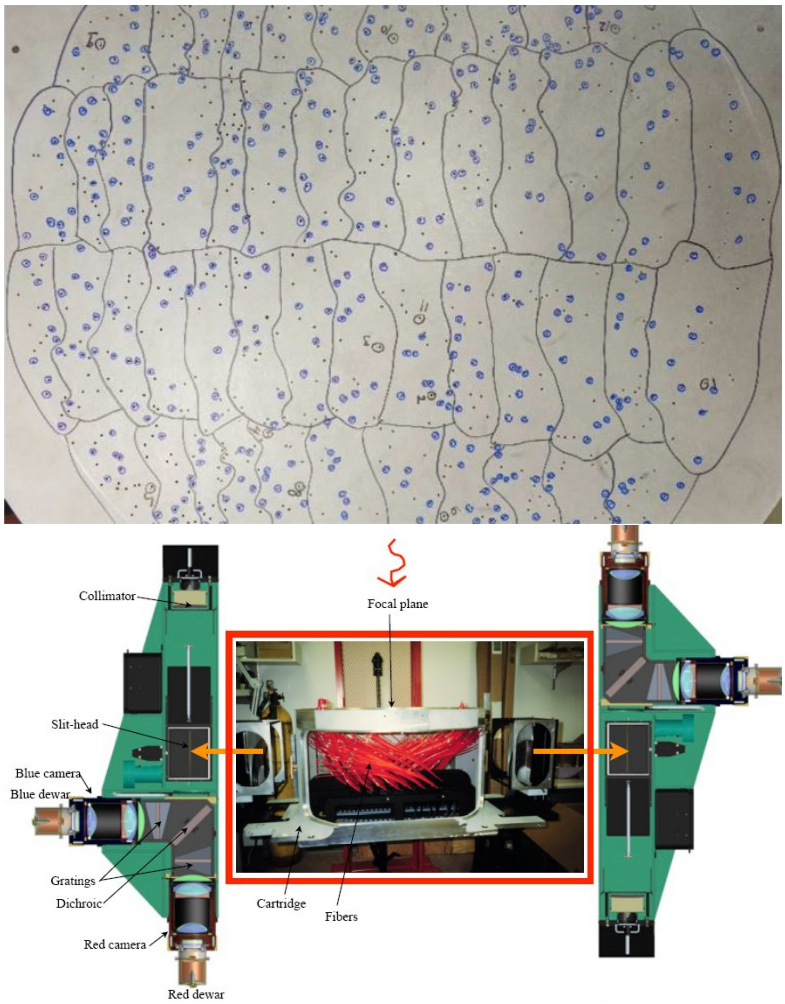


Figure 1.29: The plate (top) in the focal plane of the Sloan telescope. It has 1000 holes for the optical fibres that carry the light of luminous targets to the two spectrographs (bottom). Figure 2 from Dawson et al. (2013) and https://www.sdss4.org/instruments/boss_spectrograph/

observations. For bright enough sources DESI can also decrease the exposure time compared to the one Sloan would need. Nevertheless, the optical corrector allows for a 3.2° diameter field of view, similar as for BOSS/eBOSS.

The most important improvement compared to BOSS/eBOSS is the optical fibre system. The focal plane of the Mayall telescope hosts 5000 optical fibres, each being individually pointed by a pen-sized robot (see Figure 1.30). This allows for the measurement of approximately 5000 spectra at the same time, compared to the 1000 of BOSS/eBOSS. Due to the robotic system, DESI allows many more configurations per night compared to BOSS/eBOSS, even though the overhead time is only reduced from 5 to 2 minutes.

Due to the large number of optical fibres, ten identical spectrographs are needed to analyse the light-spectrum. Figure 1.31 shows a scheme of a DESI spectrograph. Each one has three

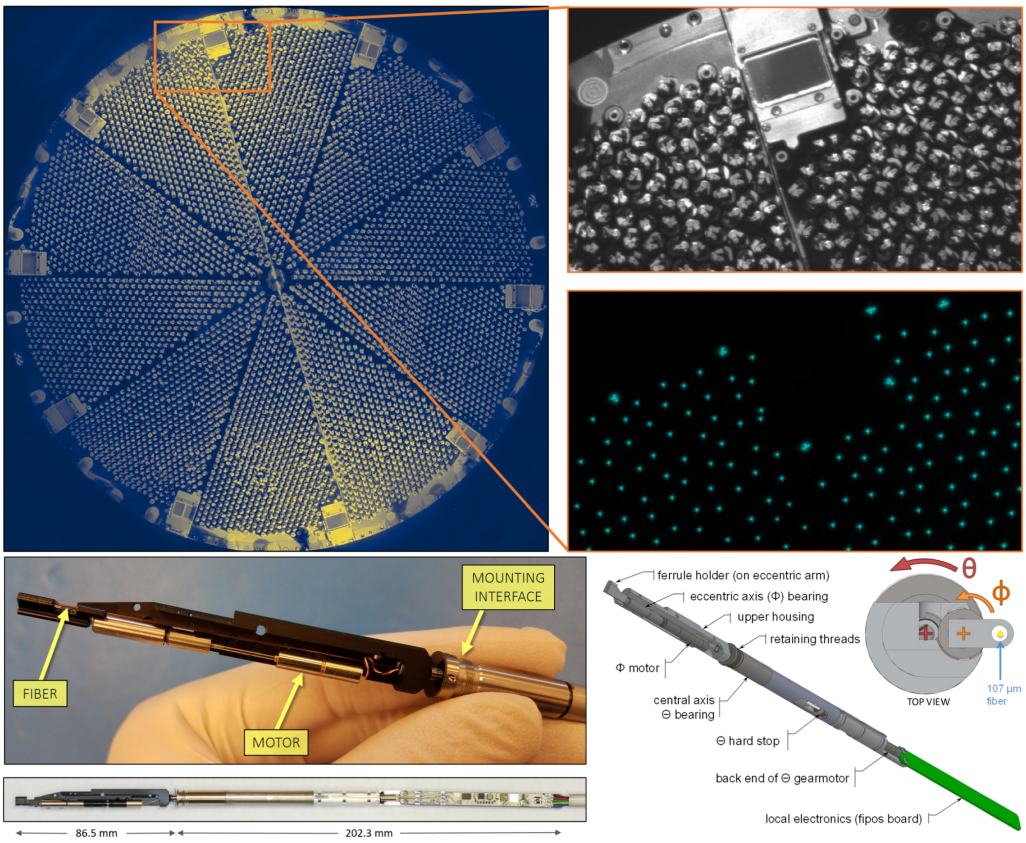


Figure 1.30: The 10 petals in the focal plane of the Mayall telescope used by DESI. Each petal contains 500 robots that position the optical fibres onto the targets of interest. Pictures and a diagram of a pen-sized robot and its two axes of rotation. Figure 3 and 17 from Silber et al. (2023)

arms recording the light wavelengths from 360 nm to 980 nm with a spectral resolution ranging from 2000 to nearly 5500, respectively.

Spectroscopic measurements

From December 2020 until June 2021, DESI has measured the spectra of 1.8 million targets and has published the Early Data Release (EDR; DESI Collaboration et al., 2023b). This shows the efficiency and the speed of the instrument to measure spectra. In comparison, BOSS has measured approximately the same number of spectra in 5 years. The entire DESI system (from the telescope to the spectrographs) has been optimised to detect and resolve the [OII] doublet of ELGs (galaxies, in general):

- within the $0.6 < z < 1.6$ redshift range;
- in 1000 seconds of effective exposure time, i.e. in reference conditions (zenith, dark sky, FWHM seeing of 1.1 arcsecond and no Galactic extinction);

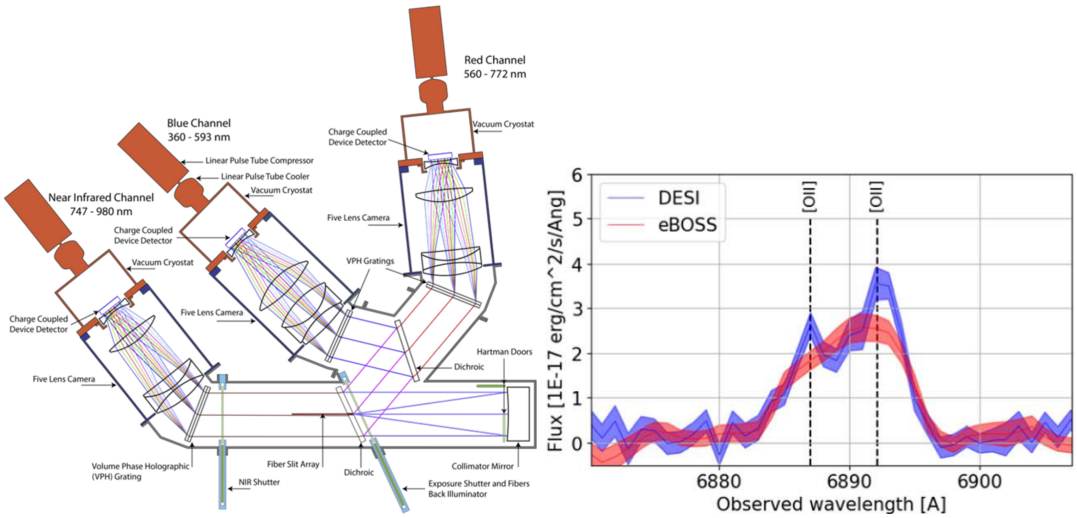


Figure 1.31: Left: the scheme of a DESI spectrograph. Right: the spectrum of an ELG measured by eBOSS (≈ 1 hour) and DESI (≈ 15 minutes), zoomed on the [OII] doublet. Figure 14 of DESI Collaboration et al. (2022) and Figure 11 of Raichoor et al. (2023)

- down to fluxes of 8×10^{-17} erg/s/cm^2.[43]

Figure 1.31 shows a comparison between the spectra of eBOSS and DESI for the same target, zoomed on the [OII] doublet. While the exposure time for that target was over one hour for eBOSS, DESI observed it for over 15 minutes, but it managed to resolve the doublet.

Figure 1.32 shows four measured DESI spectra, one for each extragalactic target type: BGS, LRG, ELG and QSO [44]. The shown spectra have been classified as high quality during the visual inspection due to the presence of absorption and emission spectral lines. The presence of such spectral lines allow for a robust redshift measurement. Notably, the ELG spectrum has not only the [OII] doublet, but also FeII and MgII absorption lines, see Lan et al. (2023). Nevertheless, not all spectra have the same quality and thus do not provide good quality redshift measurements.

Figure 1.33 shows the number of good quality redshifts per extragalactic tracer category and per redshift bin. According to DESI Collaboration et al. (2016a) the redshift measurements must have at least a precision of $\sigma_z/(1+z) \approx 0.0005$ per galaxy, in order to preserve the BAO feature along the line of sight[45]. DESI Collaboration et al. (2023a) have shown using the EDR data that BGS, LRG, ELG and QSO with $z < 2.1$ have a typical precision and accuracy much lower than the required ones for BAO studies, see Table 1.2. Nonetheless, Yu et al. (2023); Yuan et al. (2023) have observed that for LRGs and QSOs, these redshift uncertainties have an

[43] 1 erg = 10^{-7} J

[44] Stars are also observed by DESI, but their study is beyond the scope of the thesis.

[45] (Ishikawa et al., 2023) have shown that even with photometric redshifts that have three per cent precision, the BAO signature can be detected, but the expected uncertainties on cosmological parameters are much larger than the case with spectroscopic redshifts

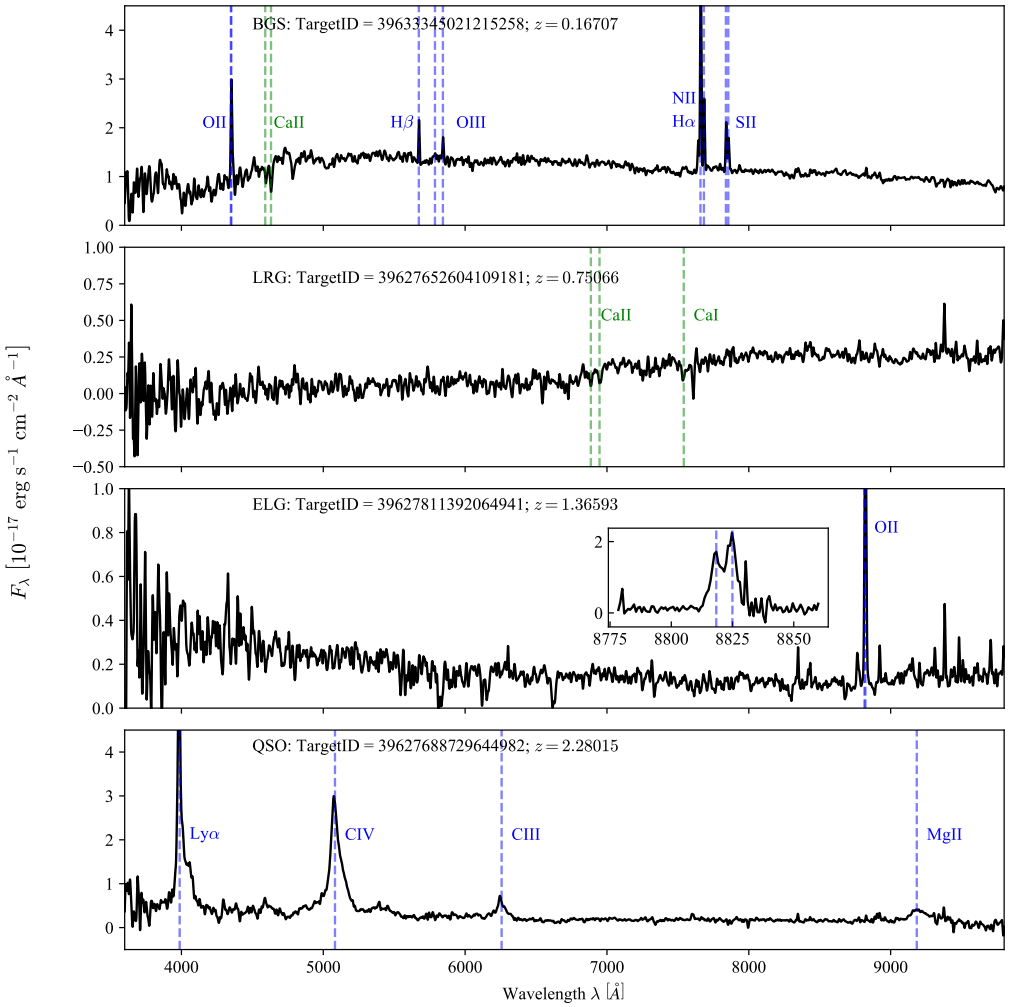


Figure 1.32: The measured spectra of the four DESI extragalactic target types and their best-fitting redshift. The blue lines illustrate the emission lines, while the absorption lines are shown in green. The wavelength range is limited by the spectrograph. Figure adapted from Lan et al. (2023); Alexander et al. (2023).

	BGS	LRG	ELG	Tracer QSO
$\frac{\sigma_z}{1+z}$	0.00003	0.00014	0.000026	0.00041
$\frac{\Delta z}{1+z}$	0.000022	0.00001	0.0000033	0.000087

Table 1.2: Random redshift error σ_z and typical systematic shift Δz of DESI tracers, see DESI Collaboration et al. (2023a) for more details.

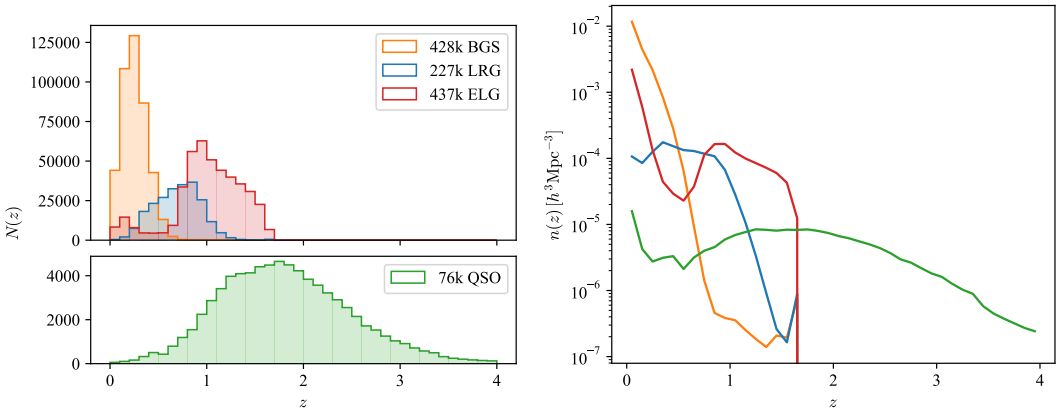


Figure 1.33: Number of DESI redshifts in the Early Data Release from a footprint of 1489 deg^2 and the number density of targets as function of redshift z, for a flat ΛCDM with $\Omega_{0\mathrm{m}} = 0.3166$, see Table 1.1. Figure adapted from DESI Collaboration et al. (2023b).

impact on the clustering measurements at small scales.

The same Figure 1.33 illustrates the radial number density of galaxies and QSOs. The $n(z)$ together with the footprint of the survey (see Figure 1.25 for the entire DESI footprint, and Figure 1.26 for an example of an added mask) are two examples of systematic effects that one must take into account when performing LSS studies. In practice, the FKP weights (Feldman et al., 1994) are computed based on the $n(z)$ to counterbalance the nonuniform distribution and thus to optimise the signal-to-noise in two-point clustering measurements (see Section 1.4.3). The footprint must be taken into account in the random catalogue and the window function as explained in Section 1.4.3.

As for the photometric surveys, spectroscopic surveys (such as DESI and eBOSS) suffer from specific systematic effects. The low signal-to-noise in some spectroscopic measurements can lead to failures in estimating some redshifts. These failures can introduce additional angular inhomogeneities (angular completeness) that must be taken into account through weights. Moreover, they also introduce artificial effects along the radial distribution affecting the $n(z)$, therefore specific weighs must be applied to account for these spectroscopic completeness as well, see e.g. Ross et al. (2020).

Another important systematic effect is introduced by the fibre assignment process which is illustrated in Figure 1.34. The fibre assignment implements a priority scheme, where QSO

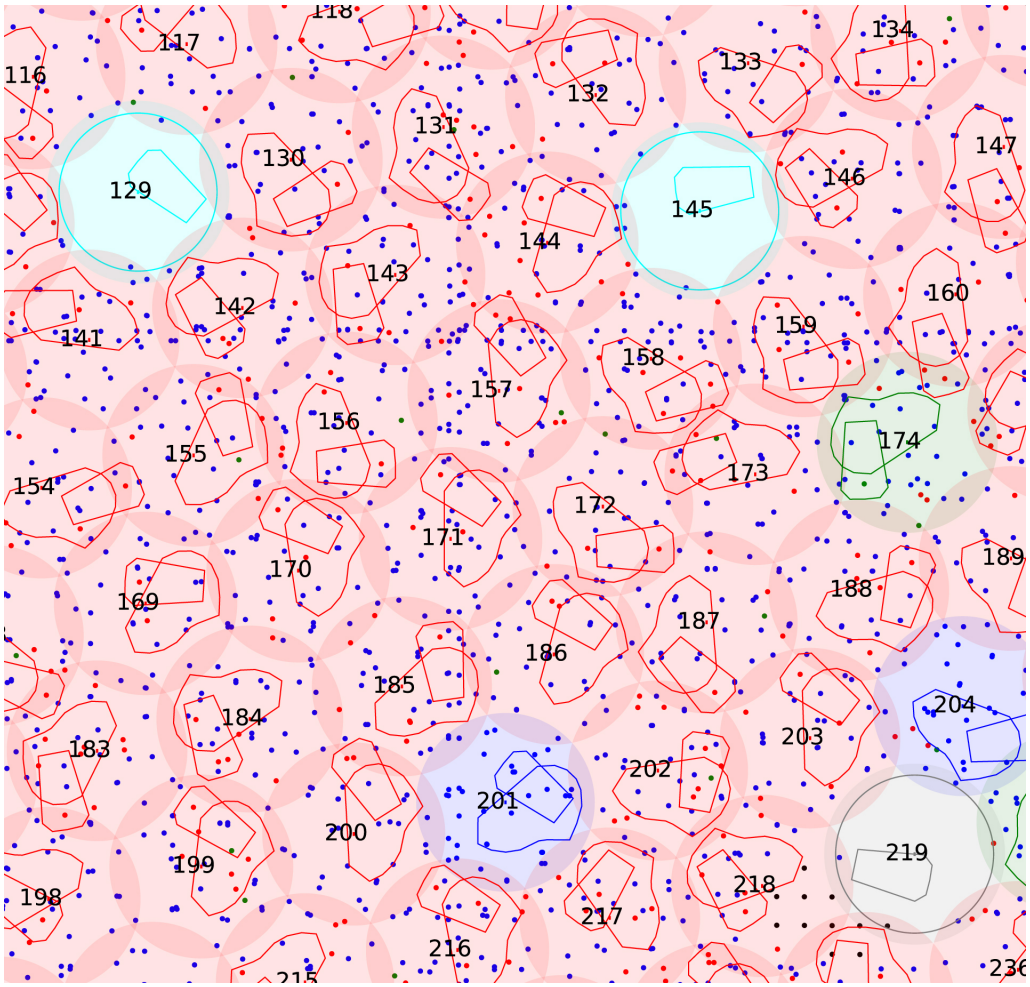


Figure 1.34: The overlapping circles denote the coverage regions of each fibre positioned by a robot: red – science fibres; green – standard stars; blue – sky fibres; cyan and grey – disabled robots. The dots represent the targets that can be reached by non-disabled robots. The colour scheme is the same as for the circles. Figure 21 of DESI Collaboration et al. (2022)

have the highest priority (due to low number density), then the LRGs are followed by the lower priority ELGs. In addition there are fibres that are pointed towards standard stars and blank sky regions in order to calibrate the system.

Due to the fact that neighbouring robots have overlapping coverage regions, there is a chance[46] that targets found in those overlapping regions cannot be observed during the survey observations. Nevertheless, one can account for this incompleteness effect using a weighting scheme, see e.g. Ross et al. (2020); DESI Collaboration et al. (2023b).

The final catalogue that includes:

[46] This chance depends strongly on how many repeated observations are performed on the same field-of-view. For example the SV3, part of EDR, has a very high fibre completeness, due to repeated observations, see DESI Collaboration et al. (2023b).

1. the 3D positions: RA, DEC (angular) and spectroscopic redshift z (radial);

2. weights: e.g. incompleteness, spectroscopic, FKP;

together with the information about the $n(z)$ and the footprint including masking (needed for the window function, see Section 1.4.3) can be used in clustering measurements.

1.4.3 Statistical measurements

In Section 1.3, we have defined the 2PCF and power spectra of a continuous matter density field. In contrast, surveys like BOSS/eBOSS or DESI provide 3D positions of galaxies or quasars. In this section, we briefly introduce the relation between galaxies and the matter field and explain how the 2PCF and power spectra are computed from a catalogue of discrete tracers. For a pedagogical overview of some of the next topics, one can consult (Percival, 2013).

Two-point correlation function

The 2PCF $\xi(s)$ measures the excess of probability that two galaxies or quasars are separated by a distance s, with respect to a random probability:

$$\mathrm{d}P = \overline{n}^2 \left[1 + \xi(s)\right] \mathrm{d}V_1 \mathrm{d}V_2, \tag{1.140}$$

where $\overline{n}$ denotes the average tracer number density and $\mathrm{d}V_1$ and $\mathrm{d}V_2$ represent the volume elements where the tracers are located. In practice, ξ can be estimated by counting the number of pairs – in separation bins $[s, s+\Delta s]$ – from a data (D) catalogue and compare them by the ones from a random sample. The random catalogue (R) must have the same footprint and redshift distribution (i.e. $n(z)$) of points as the data catalogue, but the number of objects (N_R) can and is recommended to be larger than the data (N_D), in order to decrease the Poisson noise.

Therefore, several examples of 2PCF estimators are: the Peebles–Hauser (ξ_PH; Peebles & Hauser, 1974), the Davis–Peebles estimator (ξ_DP; Davis & Peebles, 1983), (ξ_Ham; Hamilton, 1993) and Landy–Szalay (ξ_LS; Landy & Szalay, 1993):

$$\xi_\mathrm{PH}(s) = \frac{\mathrm{DD}}{\mathrm{RR}} - 1, \tag{1.141}$$

$$\xi_\mathrm{DP}(s) = \frac{\mathrm{DD}}{\mathrm{DR}} - 1, \tag{1.142}$$

$$\xi_\mathrm{Ham}(s) = \frac{\mathrm{DD} \times \mathrm{RR}}{(\mathrm{DR})^2} - 1, \tag{1.143}$$

$$\xi_\mathrm{LS}(s) = \frac{\mathrm{DD} - 2\mathrm{DR} + \mathrm{RR}}{\mathrm{RR}}, \tag{1.144}$$

$$\tag{1.145}$$

where, DD and RR represent the normalised number of pairs for the data and the random catalogues:

$$\mathrm{DD}(s) = \frac{N_{\mathrm{DD}}(s)}{N_{\mathrm{D}}(N_{\mathrm{D}}-1)} \qquad \mathrm{RR}(s) = \frac{N_{\mathrm{RR}}(s)}{N_{\mathrm{R}}(N_{\mathrm{R}}-1)} \qquad \mathrm{DR}(s) = \frac{N_{\mathrm{DR}}(s)}{N_{\mathrm{D}}N_{\mathrm{R}}}, \tag{1.146}$$

with $N_{\mathrm{DD}}(s)$, $N_{\mathrm{RR}}(s)$ and $N_{\mathrm{DR}}(s)$ being the number of data-data, random-random, data-random pairs separated by a distance s, respectively. Pons-Bordería et al. (1999); Kerscher et al. (2000); Vargas-Magaña et al. (2013) present comparisons between some of these estimators. It turns out that while the PH one is convenient for a cubic simulation, the LS is more adapted for a survey-like geometry (i.e. light-cone).

Often, the RR term must be estimated using a random catalogue[47] that has the same survey geometry as the data (footprint with masking and $n(z)$), however, for cubic simulations with periodic boundary conditions one can compute it analytically:

$$\mathrm{RR}(s) = \frac{4\pi}{3}\frac{s_{\mathrm{max}}^3 - s_{\mathrm{min}}^3}{2V}, \tag{1.147}$$

where V is the volume of the box, s_{max} and s_{min} determine the boundaries of a separation bin.

One should notice that this calculations assume that the clustering is isotropic, which is in agreement with the cosmological principle. Nevertheless, the RSD and the Alcock & Paczynski (AP Alcock & Paczynski, 1979) effect induce anisotropies with respect to the line-of-sight (LOS). Consequently, it is interesting to study how the clustering changes with respect to this direction.

Anisotropic two-point correlation function. Given two galaxies at $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, with respect to the Earth, the LOS can be approximated by $\boldsymbol{l} = (\boldsymbol{s}_1 + \boldsymbol{s}_2)/2$. Therefore, the 2D 2PCF is defined for $s_\perp$ and $s_\parallel$ separations, where $\perp$ and $\parallel$ denote the directions perpendicular to the LOS and along the LOS:

$$s_\parallel = \frac{\boldsymbol{s}\cdot\boldsymbol{l}}{|\boldsymbol{l}|} \qquad |s_\perp| = \sqrt{s^2 - s_\parallel^2}, \tag{1.148}$$

where $\boldsymbol{s} = \boldsymbol{s}_2 - \boldsymbol{s}_1$ is the separation between the two galaxies. $\xi(s_\perp, s_\parallel)$ can be computed with the same estimator as for the isotropic 2PCF, but the 1D separation bins are replaced by 2D separation grids. Lastly, $(s_\perp, s_\parallel)$ are usually replaced by (s, μ), where $\mu = \cos\theta = s_\parallel/|\boldsymbol{s}|$ and the $\xi(s,\mu)$ is projected using the Legendre polynomials $L_\ell(\mu)$ into multipoles:

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1} L_\ell(\mu)\xi(s,\mu)\mathrm{d}\mu, \tag{1.149}$$

[47] A random catalogue should contain 3D randomly and uniformly sampled positions, on which the data survey geometry can be applied.

where $\ell = 0$, $\ell = 2$ and $\ell = 4$ denote the monopole, quadrupole and hexadecapole, that are most often used in BAO and RSD analyses.

Two-point cross correlation function. When two different types of tracers share the same volume (e.g. LRG and ELG), besides studying only the individual auto-2PCF – as defined previously – one can also compute the cross-2PCF between the two tracers 1 and 2, e.g.:

$$\xi_{\mathrm{PH}} = \frac{\mathrm{D_1 D_2}}{\mathrm{R_1 R_2}} - 1, \tag{1.150}$$

$$\xi_{\mathrm{LS}} = \frac{\mathrm{D_1 D_2} - \mathrm{D_1 R_2} - \mathrm{D_2 R_1} + \mathrm{R_1 R_2}}{\mathrm{R_1 R_2}}, \tag{1.151}$$

$$\tag{1.152}$$

In this case, $\mathrm{D_1}$ and $\mathrm{D_2}$ denote the two data catalogues, with their corresponding $\mathrm{R_1}$ and $\mathrm{R_2}$ random catalogues. Moreover, both 1D (isotropic) and 2D (anisotropic) cross-2PCF can be computed, as well.

Power spectrum

In order to compute the power spectrum, one needs the density field $\delta(\boldsymbol{k})$, which is usually obtained using Fast Fourier Transform (FFT) of the configuration space density field $\delta(\boldsymbol{x})$. In practice, the $\delta(\boldsymbol{x})$ is estimated on a grid of size N_{G} using a grid sampling scheme (e.g. Nearest Grid Point or Cloud-In-Cell, see Sefusatti et al. (2016) for a comparison of several methods) starting from a catalogue of matter tracers.

Finally, the isotropic power spectrum is computed in k shells of a given width and volume V_{s}:

$$P(k) = \frac{V_{\mathrm{s}}}{N_k} \sum_{i=1}^{N_k} |\delta'(\boldsymbol{k}_i)|^2, \tag{1.153}$$

where N_k is the number of modes in a given k shell. In a similar way as the 2PCF, there is an anisotropic power spectrum $P(k,\mu)$, as well, that can be decomposed in multipoles $P_\ell(k)$ using Legendre polynomials.

In contrast to the 2PCF, one has to subtract the shot-noise from the isotropic $P(k)$ and from the monopole of the $P(k,\mu)$, due to the self-correlation of discrete objects. The Poisson shot-noise can be estimated as the inverse of the mean tracer number density.

In a similar way to the 2PCF, a random catalogue – having the same footprint and $n(z)$ must be used to compute a correctly normalised power spectrum of a data-like[48] catalogue. Despite

[48] light-cone

the correct normalisation using the random catalogue, the resulting $P(k)_{\rm obs}$:

$$P(k)_{\rm obs} = \int {\rm d}k' W(k,k') P(k')_{\rm true} \tag{1.154}$$

is a convolution of the true power spectrum with a window function $W(k,k')$. The window function includes the footprint, $n(z)$, the weights and the effect introduced by using a cosmological model to convert the redshift into distance (see Section 1.4.4), that may different than the "true" cosmology of the Universe. The true power spectrum is the actual physical power spectrum in the absence of any systematic effect. For details on how to compute the power spectrum and include the effect of the window function, one can consult Feldman et al. (1994); Cole et al. (2005); Percival et al. (2007); Beutler et al. (2017); Gil-Marín et al. (2020) and the references therein.

The bias function

The current understanding of the galaxy formation suggests that galaxies are formed into DM haloes that are found in the overdense (above a threshold) regions of the dark matter field. As a consequence, galaxies and haloes constitute biased samples of the matter density field. This means that mathematically, the matter tracer (tr, galaxy or halo) field at a given redshift $\delta_{\rm tr}(\boldsymbol{x},z)$ is connected to the underlying matter field $\delta_{\rm m}(\boldsymbol{x},z)$ through a bias function $\mathbb{B}$: $\delta_{\rm tr}(\boldsymbol{x},z) = \mathbb{B}\left(\delta_{\rm m}(\boldsymbol{x},z)\right)$. For an in-depth review of the galaxy bias, one can consult Desjacques et al. (2018).

In general, the bias of galaxies is different than the one of haloes due to the fact that there is no one-to-one match between galaxies and haloes. This is partly caused by the specific physical processes of the baryonic matter. Nevertheless, neglecting the baryonic physics (see more details in Chapter 2), one can study perturbatively the bias of haloes and galaxies in the same way. Therefore, the matter tracer density field up to the second order in the Eulerian PT is (Nicola et al., 2023):

$$1 + \delta_{\rm tr} = 1 + b_1 \delta_{\rm m} + \frac{b_2}{2}\left(\delta_{\rm m}^2 - \overline{\delta_{\rm m}^2}\right) + \frac{b_K}{2}\left(K^2 - \overline{K^2}\right) + \varepsilon, \tag{1.155}$$

where b_1 and b_2 are the linear and quadratic Eulerian biases and ε captures the stochasticity of the discrete tracers. Furthermore, $K^2 = K_{ij}K^{ij}$ and K is the tidal tensor with its bias b_K:

$$K_{ij} \equiv \frac{\partial^2 \Phi}{\partial x_i \partial x_j} - \delta^{\rm K}_{ij}\frac{\nabla^2\Phi}{3}. \tag{1.156}$$

The $\overline{\delta_{\rm m}^2}$ and $\overline{K^2}$ have been subtracted to ensure that the matter tracer field has mean zero. One can also define the Lagrangian bias with respect to the initial density field. However, the Eulerian and Lagrangian expansions are equivalent if all terms are considered up to a given order.

The resulting linear order galaxy (or halo) auto power spectrum $P_{\rm gg}$ and galaxy-matter cross power spectrum follow:

$$P_{\rm gg} = b_1^2 P_{\rm mm} + P_{\rm SN} \qquad P_{\rm gm} = b_1 P_{\rm mm}, \tag{1.157}$$

where $P_{\rm SN}$ at the lowest order can be approximated to the Poisson shot-noise introduced by the discrete nature of the galaxies. Numerically, $P_{\rm SN} = 1/\overline{n}_{\rm g}$, where $\overline{n}_{\rm g}$ is the average galaxy number density. Finally, the linear bias is a good approximation at scales $k \leq 0.03\,h/{\rm Mpc}$ at $z = 0$ within ≈ 10 per cent precision.

1.4.4 The Baryonic Acoustic Oscillations as Standard Ruler

As discussed in Section 1.2.3, BAO (see Weinberg et al. (2013) for a review) propagated until the decoupling of baryons from the photons, leaving an overdense spherical shell of size $r_{\rm s}$ around the initial fluctuation. Due to gravitational instability, matter accumulated in the overdense regions leading to galaxy formation. Consequently, one expects a higher probability to see galaxies separated by a distance $r_{\rm s}$ and thus an imprint in the 2PCF and power spectrum. In this subsection, we explain how the measured 2PCF and power spectrum are modelled to capture the BAO signature and thus constrain cosmological parameters. We focus on the isotropic model, but we discuss about some aspects of the anisotropic case, see e.g. Bautista et al. (2021); Gil-Marín et al. (2020) for more details.

Before computing the 2PCF $\xi^{\rm m}(s)$ or the power spectrum $P^{\rm m}(k)$ of the matter tracers as described in Section 1.4.3, the (RA, DEC and z) measurements provided by photometric and spectroscopic surveys must be converted to comoving Cartesian coordinates $\boldsymbol{x} = (x_1, x_2, x_3)$ using a fiducial cosmology[49]:

$$x_1 = \mathbb{X}(z)\cos({\rm DEC})\cos({\rm RA}) \tag{1.158}$$

$$x_2 = \mathbb{X}(z)\cos({\rm DEC})\sin({\rm RA}) \tag{1.159}$$

$$x_3 = \mathbb{X}(z)\sin({\rm DEC}), \tag{1.160}$$

where $\mathbb{X}$ is the transverse comoving distance at a redshift z, see equation 1.26. The top panels of Figure 1.35 show one of the earliest measurements of LRGs 2PCF side-by-side with the latest LRG clustering of the DESI EDR, both detecting the BAO signature. On one hand, the shape of the 2PCF has two peaks: the first peak represents the initial fluctuations, while the second one corresponds the BAO imprint caused by the propagating fluctuations, see Figure 1.11. On the other, the BAO signature in the power spectrum has a undulate shape. This is explained by the fact that the Fourier transform of a peak has a sinusoidal shape.

The lower panels of Figure 1.35 illustrate recent power spectrum and 2PCF measurements from the BOSS and eBOSS surveys, together with their best-fitting curves. The isotropic 2PCF

[49] For a study on the impact of the fiducial cosmology on BAO measurements, one can consult Carter et al. (2020).

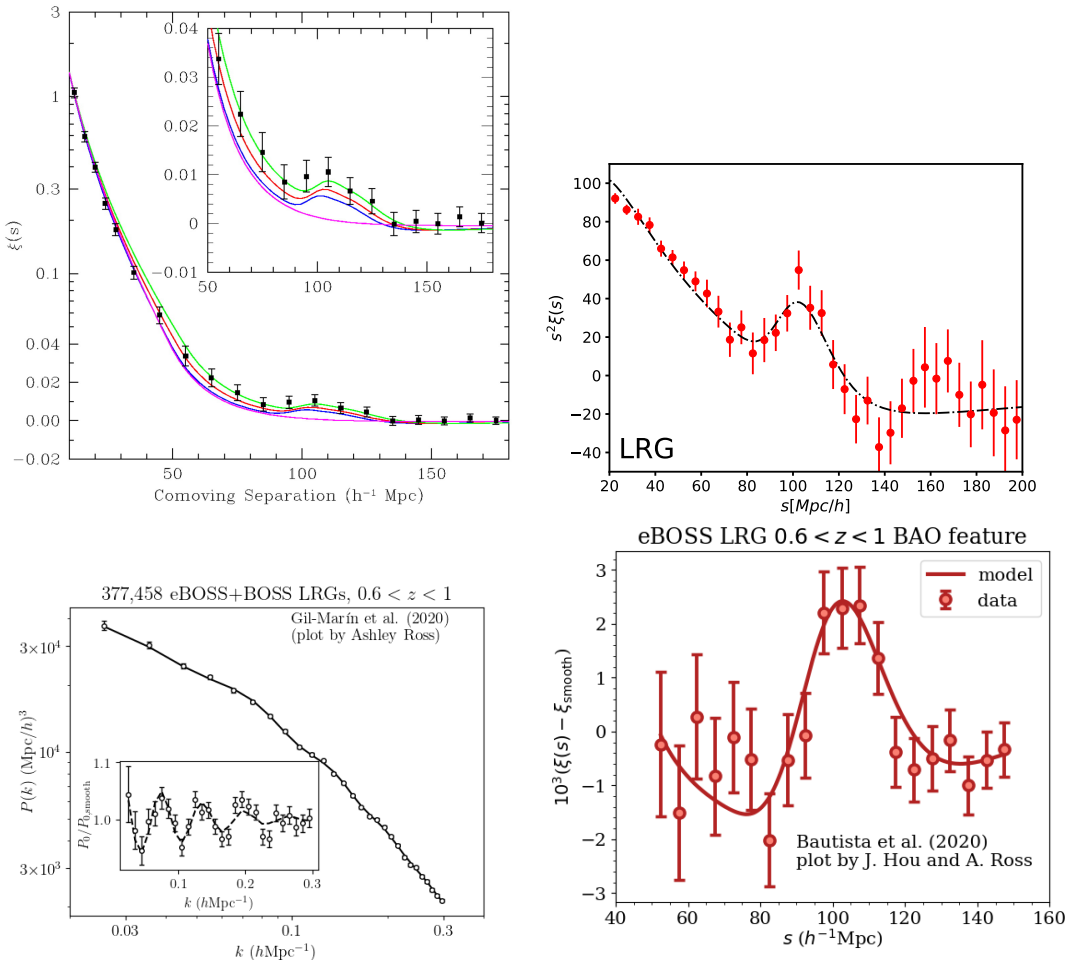


Figure 1.35: The Baryonic Acoustic Oscillations signature in the galaxy clustering measurements (and best-fitting models for the lower panels). Figure 2 of Eisenstein et al. (2005); Figure 5 from Moon et al. (2023); Figures from https://www.sdss4.org/science/final-bao-and-rsd-measurements/.

and power spectrum models are of the following kind (Xu et al., 2012):

$$\xi(s) = B^2(s)\xi^{\rm temp}(\alpha s) + A(s) \qquad P(k) = B^2(k)P^{\rm temp}(k/\alpha) + A(k), \tag{1.161}$$

where $B(s)$, $B(k)$, $A(s)$ and $A(k)$ are nuisance functions that describe the broadband shape and do not contain BAO information, hence one can marginalise them. Nevertheless, the choice of these functions may bias the BAO measurements. A typical choice for the configuration space, that should not bias the BAO measurements according to (Xu et al., 2012; Vargas-Magaña et al., 2014), is a free B parameter and a polynomial function for $A(s)$:

$$A(s) = a_0 s^{-2} + a_1 s^{-1} + a_2, \tag{1.162}$$

where a_0, a_1, a_2 are free parameters. More details about the Fourier space model can be found in e.g. Xu et al. (2012); Ross et al. (2015).

The cosmologically significant parameter is α, which is related to the Alcock–Paczynski parameter[50] (Alcock & Paczynski, 1979). Technically, α shifts the template 2PCF $\xi^{\rm temp}(s)$ or power spectrum $P^{\rm temp}(k)$, such that their BAO signature matches the one from the measured clustering. Furthermore, the template 2PCF can be obtained through the Hankel transform of a template power spectrum $P_{\rm t}(k)$ [51]:

$$\xi^{\rm temp}(s) = \frac{1}{2\pi^2}\int_0^\infty k^2 j_0(ks)P_{\rm t}(k){\rm d}k, \tag{1.163}$$

$$P_{\rm t}(k) = \left[\left(\frac{P_{\rm m}^{\rm L}(k)}{P_{\rm m}^{\rm L,nw}(k)} - 1\right){\rm e}^{-\frac{1}{2}k^2\Sigma_{\rm nl}^2} + 1\right] P_{\rm m}^{\rm L,nw}(k). \tag{1.164}$$

In the previous equation, both $P_{\rm m}^{\rm L,nw}(k)$ and $P_{\rm m}^{\rm L}(k)$ are linear power spectra computed with the fiducial cosmological parameters. However, while $P_{\rm m}^{\rm L}(k)$ is the result of codes such CAMB or CLASS, as described in Section 1.3.2, $P_{\rm m}^{\rm L,nw}(k)$ can be estimated using the formulas of Eisenstein & Hu (1998) and denotes the smooth (without BAO wiggles, nw) linear power spectrum. Lastly, $\Sigma_{\rm nl}$ takes into account the non-linear damping of the BAO feature (Eisenstein et al., 2007a) and $j_0(x)$ is the zeroth order spherical Bessel function.

In order to understand the cosmological information captured by α, one has to use the sound horizon $r_{\rm s}$ as a standard ruler, i.e. a fixed known scale within the entire Universe. Therefore, it is worth analysing separately the comoving sound horizon parallel $r_{\rm s}^{\parallel}$ and perpendicular $r_{\rm s}^{\perp}$ to the LOS through the equations (1.51) (1.60):

$$r_{\rm s}^{\perp} = (1+z)D_{\rm A}(z)\Delta\theta_{\rm s} \qquad r_{\rm s}^{\parallel} = \frac{c\Delta z_{\rm s}}{H(z)}, \tag{1.165}$$

where $\Delta\theta_{\rm s}$ and $\Delta z_{\rm s}$ are the angular and radial sizes of the sound horizon at redshift z, respectively. The cosmological principle assumes that the Universe is isotropic, therefore $r_{\rm s}^{\perp} = r_{\rm s}^{\parallel}$ for

[50] Sometimes α is called the Alcock–Paczynski parameter, however (Alcock & Paczynski, 1979) study $\frac{\Delta z}{z\Delta\theta}$.

[51] In practice, the isotropic model of the power spectrum is based on $P_{\rm t}(k)$ as well.

a given set of cosmological parameters. Nevertheless, we show that this may not be the case for the measured sound horizon $r_s^{\perp,m}$ and $r_s^{\parallel,m}$ from the measured clustering $\xi^m(s)$ or $P^m(k)$.

The values of $r_s^{\perp,m}$ and $r_s^{\parallel,m}$ depend on the real angular and radial sizes of the sound horizon ($\Delta\theta_s^{real}$, Δz_s^{real}) in the observable Universe, and the fiducial cosmology needed to compute D_A^{fid}, $H^{fid}(z)$:

$$r_s^{\perp,m} = (1+z) D_A^{fid}(z) \Delta\theta_s^{real} \qquad r_s^{\parallel,m} = \frac{c\Delta z_s^{real}}{H^{fid}(z)}. \tag{1.166}$$

Therefore, there are two important effects that lead to the $r_s^{\perp,m} \neq r_s^{\parallel,m}$ inequality:

1. due to the RSD effect, Δz_s^{real} is smaller than the actual radial size of the sound horizon, see Section 1.4.5;

2. the Alcock–Paczynski effect (Alcock & Paczynski, 1979) implies that $\frac{\Delta z_s}{z\Delta\theta_s}$ depends on the cosmological parameters, hence $\frac{\Delta z_s^{fid}}{z\Delta\theta_s^{fid}} \neq \frac{\Delta z_s^{real}}{z\Delta\theta_s^{real}}$[52], if the fiducial cosmology is different than the real one.

Furthermore, starting from equation (1.165) and considering the template clustering computed using the fiducial cosmological parameters:

$$r_s^{fid} = (1+z) D_A^{fid}(z) \Delta\theta_s^{fid} \qquad r_s^{fid} = \frac{c\Delta z_s^{fid}}{H^{fid}(z)}. \tag{1.167}$$

In this case, $r_s^{\perp,fid} = r_s^{\parallel,fid}$ because there is only one set of cosmological parameters involved in the computation and the RSD effect is not introduced into the template. Lastly, using equation (1.165) with the real cosmological parameters (they are unknown, but are the final product of the BAO measurements), one obtains:

$$r_s^{real} = (1+z) D_A^{real}(z) \Delta\theta_s^{real} \qquad r_s^{real} = \frac{c\Delta z_s^{real}}{H^{real}(z)}. \tag{1.168}$$

In contrast to the fiducial case, $r_s^{\perp,real} \neq r_s^{\parallel,real}$[53] due to the RSD effect. Nevertheless, the BAO reconstruction – presented in Section 1.4.6 – is used to remove the RSD effect, hence $r_s^{\perp,real} = r_s^{\parallel,real}$ for BAO studies.

[52] Using equation (1.167), $\frac{r_s^{\perp,m}}{r_s^{\parallel,m}} = \frac{\Delta\theta_s^{real}}{\Delta z_s^{real}} \frac{\Delta z_s^{fid}}{\Delta\theta_s^{fid}}$

[53] It is important to remark that the superscript "real" refers to the real cosmological parameters, but Δz^{real} is from real measurements. This means that there is a "true" sound horizon r_s^{true} that is not affected by RSD and is directly computed using the real cosmological parameters. This can also provide Δz^{true}. Ideally, the BAO reconstruction should bring the r_s^{real} close to r_s^{true}, thus we use interchangeably r_s^{true} and r_s^{real}. Nevertheless for RSD fitting, one must use r_s^{true} in equations (1.173) and the RSD model takes into account the anisotropy that makes $r_s^{\perp,real} \neq r_s^{\parallel,real}$.

In the isotropic case, it is more convenient to define an angle-averaged sound horizon[54] and spherically-averaged distance $D_V(z)$:

$$r_s = \left[\left(r_s^{\perp}\right)^2 \cdot r_s^{\parallel} \right]^{1/3}, \tag{1.169}$$

$$D_V(z) = \left[\frac{cz(1+z)^2 D_A^2(z)}{H(z)} \right]^{1/3}. \tag{1.170}$$

Considering now that when then template clustering fits the measured clustering, α should match the measured sound horizon with the fiducial one. Therefore, one obtains using the equations (1.166) (1.167):

$$\alpha = \frac{r_s^{\rm fid}}{r_s^{\rm m}} = \left(\frac{\Delta\theta_s^{\rm fid}}{\Delta\theta_s^{\rm real}} \right)^{2/3} \left(\frac{\Delta z_s^{\rm fid}}{\Delta z_s^{\rm real}} \right)^{1/3}. \tag{1.171}$$

Finally, replacing the equations (1.167) (1.168) (1.170) into the previous one:

$$\alpha = \left(\frac{D_A^{\rm real}(z)}{D_A^{\rm fid}(z)} \right)^{2/3} \left(\frac{H^{\rm fid}(z)}{H^{\rm real}(z)} \right)^{1/3} \frac{r_s^{\rm fid}}{r_s^{\rm real}} = \frac{D_V^{\rm real}(z)}{r_s^{\rm real}} \frac{r_s^{\rm fid}}{D_V^{\rm fid}(z)}. \tag{1.172}$$

Given the dependency of the Hubble parameter $H(z)$ and the angular diameter distance $D_A(z)$ on the cosmological parameters, one can notice that α quantifies how different the real cosmological parameters are compared to the fiducial ones. In practice, the value of α and its uncertainty are obtained by fitting a BAO model to the measured clustering, and $D_A^{\rm fid}(z)$ and $H^{\rm fid}(z)$ are computed directly using the fiducial cosmological parameters. Consequently, a constraint on α translates into constraints on the cosmological parameters through $D_A^{\rm real}(z)$ and $H^{\rm real}(z)$. While for low redshift [55] galaxy samples an isotropic BAO study performs similarly to an anisotropic one, at higher redshifts, it is better to fit separately (Anderson et al., 2014)[56]:

$$\alpha_{\parallel} = \frac{H^{\rm fid}(z) r_s^{\rm fid}}{H^{\rm real}(z) r_s^{\rm real}} \qquad \alpha_{\perp} = \frac{D_A^{\rm real}(z) r_s^{\rm fid}}{D_A^{\rm fid}(z) r_s^{\rm real}}, \tag{1.173}$$

where $\alpha = \left(\alpha_{\parallel}\alpha_{\perp}^2\right)^{1/3}$. One of the reasons is that the anisotropic BAO study can provide additional cosmological constraints at higher redshifts. Moreover, as previously discussed, the sound horizon is affected by the RSD along the line-of-sight.

If we replace the equations (1.68) (1.59) and the Hubble parameter $H(z)$ from Section 1.2.1 into the formulas of α, equation (1.172), one observes that there is a degeneracy between H_0 and the ρ_0 parameters of the cosmological components (or k for curvature)[57]. This means

[54] Due to anisotropies (RSD or AP) the BAO "sphere" of radius r_s is in fact an ellipsoid with two axes of size $r_s^{\perp}$ and one of size $r_s^{\parallel}$ and of volume $V = \frac{4\pi}{3} r_s^{\perp} \times r_s^{\perp} \times r_s^{\parallel}$. If the volume of the ellipsoid is transformed into a sphere of volume $V = \frac{4\pi}{3} r_s^3$, one obtains the angle-average sound horizon.

[55] At low redshifts, the different cosmological distances become similar.

[56] In addition, the signal was not strong enough for Anderson et al. (2014) to perform aniostropic measurements.

[57] For a given H_0, there is a set of ρ_0 and k parameters such that α remains unchanged.

that BAO measurements alone can provide constraints only on Ω_0 parameters. Nevertheless, one has to remove the dependency on the r_s, either by measuring the anisotropic BAO or by having measurements at different redshifts. Otherwise, one needs additional constraints such as CMB or BBN measurements.

Table 1.1 contains measurements from CMB anisotropies, BAO, BBN and different combinations of data-sets. One can observe that the latest BAO measurements have weaker constraints on the cosmological parameters than the CMB ones. Nonetheless, the combined BAO and CMB analysis improves significantly the precision on the Ω_0 parameters. Specifically in the case of oΛCDM, the combined BAO+CMB measurements improve the precision by almost an order of magnitude. In a similar way, for the wCDM, the precision on w is significanlty improved for the combined BAO+CMB analysis. These improvements are induced by the BAO due to the 3D[58] measurements at multiple redshifts. Lastly, the combined BAO+BBN[59] studies do not improve the measurements for the Ω_0 parameters, but they allow to constrain H_0 and $r_s \equiv r_{\rm drag}$ to values consistent with CMB measurements alone.

1.4.5 Redshift Space Distortions

Due to the expansion of the Universe, the redshift of a light-source is related to its distance with respect to the Earth, i.e. farther sources have higher redshifts, see Section 1.2.2. Nevertheless, light-sources have an additional peculiar motion due to the gravitational interaction (similarly to CDM particles in Section 1.3.3, see equation (1.86)). This peculiar velocity introduces Doppler redshift of the light, distorting the cosmological redshift (Redshift Space Distortions, Kaiser, 1987). Therefore, the measured redshift by spectroscopic surveys represents an overlap of the expansion and the peculiar motion.

Consequently,

- the distances estimated from the redshifts as in equation (1.51) are distorted:
  - if a light-source moves away from Earth, its estimated distance will be larger than the actual one;
  - if a light-source approaches Earth, its estimated distance will be smaller than the actual one.
- a 3D map of the light-sources will be anisotropic, as RSD affects the radial measurements (redshift) and not the angular ones.

Figures 1.36 and 1.37 illustrate the effect of peculiar motion on the distance estimation and thus on the galaxy 2PCF. Due to the RSD, the spherically symmetric BAO feature appears squeezed along the line-of-sight.

[58] Note that the CMB maps of temperature anisotropies are only angular (2D) maps at a single redshift.

[59] These studies include implicitly the CMB black body temperature measurement (Fixsen, 2009) $T_0 = 2.72548 \pm 0.00057$ K, that is needed to estimate the energy density of photons in the computation of r_s, equation (1.68).

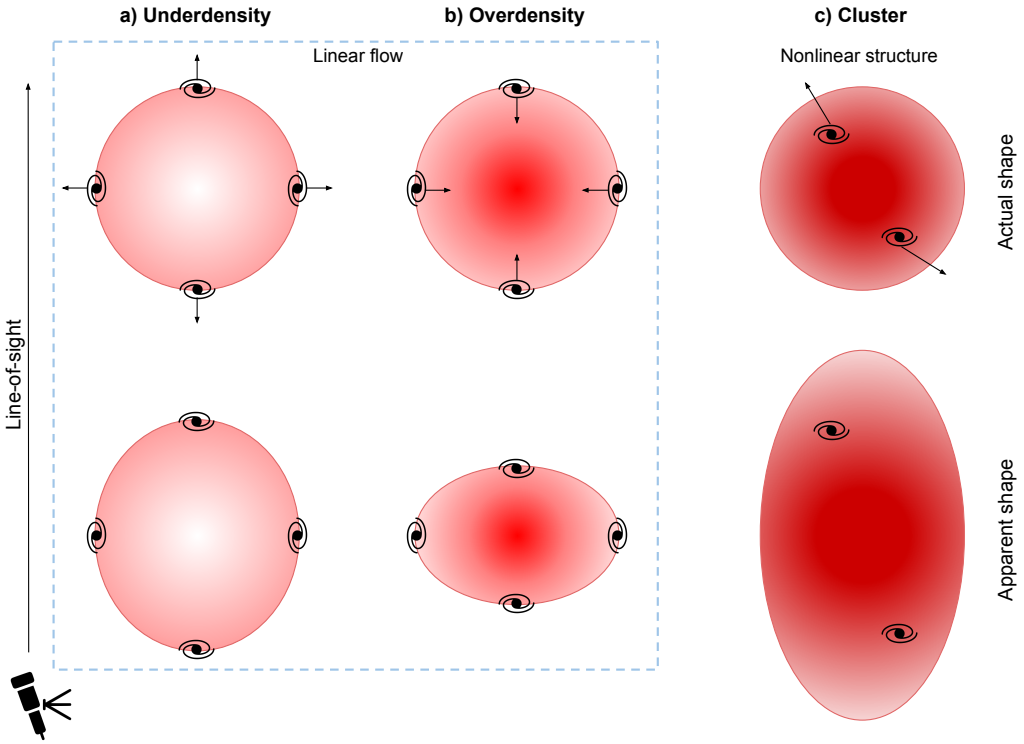


Figure 1.36: Redshift space distortions. Icons from https://www.onlinewebfonts.com/icon

At lower scales, i.e. $\approx 50\,\mathrm{Mpc}/h$, as galaxies are closer to the overdense region, their peculiar velocities cancel out the separation due to the Hubble flow. Therefore, galaxies that should be separated by a certain distance appear as collapsed along the radial direction. Lastly, at non-linear scales below $20\,\mathrm{Mpc}/h$, the virial motion of galaxies causes the Finger-of-God effect (Jackson, 1972).

The multipole decomposition of the 2D 2PCF is shown in the right-hand side of the Figure 1.37. The presence of the RSD effect makes the quadrupole different than zero. This suggests that instead of fitting the entire 2D 2PCF to extract cosmological parameters, one can fit the 2PCF (or power spectrum) multipoles, simplifying the process. Nevertheless, in practice, one creates a model of the 2D power spectrum $P(k,\mu)$ and then decomposes it into multipoles. The simplest 2D model power spectrum that accounts for the RSD effect is (Kaiser, 1987):

$$P(k,\mu) = (b(k) + f\mu^2)^2 P_{\mathrm{m}}(k). \tag{1.174}$$

In the former equation, $P_{\mathrm{m}}(k)$ is the matter power spectrum (that can be obtained from linear or non-linear PT), $b(k)$ is a scale-dependent bias and $f \equiv \mathrm{d}\ln D_1(a)/\mathrm{d}a$, equation (1.106). After the $P(k,\mu)$ model is decomposed in multipoles $P_\ell(k)$, one can directly fit the measured power spectrum multipoles. On the other hand, one can Hankel transform the model multipoles:

$$\xi_\ell(s) = \frac{i^\ell}{2\pi^2}\int_0^\infty k^2 j_\ell(ks) P_\ell(k)\mathrm{d}k \tag{1.175}$$

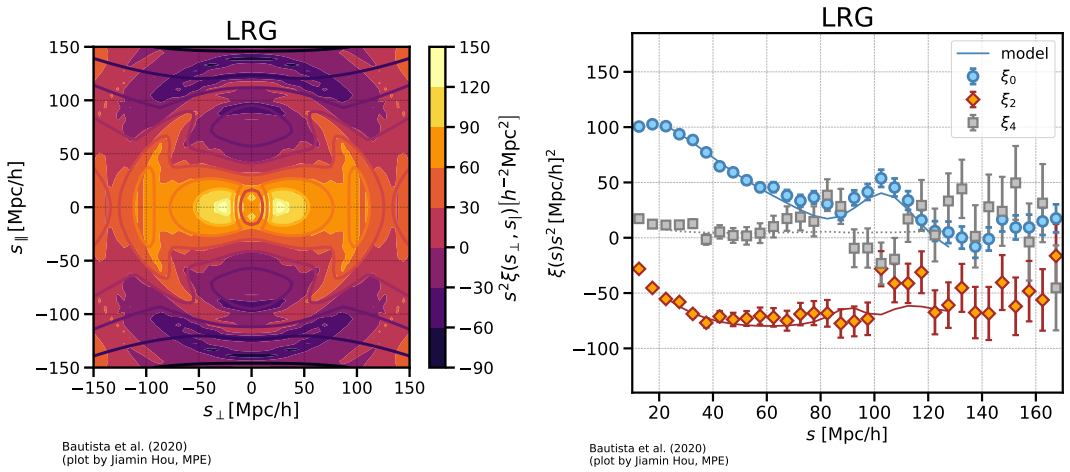


Figure 1.37: The effect of redshift space distortions on the 2D two-point correlation function and the BAO signature and the multipole decomposition of the 2D 2PCF. Figures from https://www.sdss4.org/science/final-bao-and-rsd-measurements/, (Bautista et al., 2021).

and then fit the measured 2PCF multipoles – with j_ℓ the spherical Bessel functions of order ℓ.

The combined study of BAO and RSD can be used to test the GR (e.g. the reviews Weinberg et al., 2013; Ishak, 2019). Nevertheless, more accurate models are needed (e.g. Gil-Marín et al., 2012). Recent BAO+RSD studies (e.g. Bautista et al., 2021; Gil-Marín et al., 2020) have shown that the measurements are in agreement with the GR and ΛCDM models.

1.4.6 Reconstruction

As discussed earlier in this section, after the decoupling of baryons and photons, the gravitational attraction leads to the formation of the LSS. In addition, this coherent flow of matter affects the BAO signature in two ways:

1. gives birth to the RSD effect previously discussed, squeezing the signature;
2. displaces the particles that form a "perfect" BAO spherical shell in the early Universe, such that the BAO signature is smeared out in time (top panels of Figure 1.38). From the point of view of the statistical description, part of the BAO signal in the two-point clustering statistics leaks into the higher-order clustering statistics (Schmittfull et al., 2015).

Eisenstein et al. (2007b) have introduced the BAO reconstruction technique to displace the galaxies back in time, in order to estimate the linear density field and thus increase the BAO signal – by restoring the information from higher-order statistics into the two-point clustering (Schmittfull et al., 2015), hence pushing the monopole and quadrupole close to zero –, see the bottom panels of Figure 1.38. Further studies (Padmanabhan et al., 2012; Burden et al., 2014,

2015; Seo et al., 2022, e.g.) have developed BAO reconstruction to remove the RSD effect as well.

Mathematically, one must solve the following equation to find the displacement field $\boldsymbol{\Psi}$ (of galaxies, for example) inspired from the ZA equation (1.121):

$$\nabla \cdot \boldsymbol{\Psi} + f \nabla \cdot (\boldsymbol{\Psi} \cdot \boldsymbol{s}) \boldsymbol{s} = -\frac{\delta_{\rm gal}}{b}. \tag{1.176}$$

The additional term $f\nabla \cdot (\boldsymbol{\Psi} \cdot \boldsymbol{s})\boldsymbol{s}$ to ZA is required to describe the RSD effect ($\boldsymbol{\Psi} \cdot \boldsymbol{s}$ is the displacement along the line-of-sight). The bottom left panel of Figure 1.38 illustrates the Lagrangian displacement field $\boldsymbol{\Psi}$ in blue, that is applied oppositely on the galaxies (particles). The result can be observed in the bottom right panel, where the BAO feature is much closer to the red ring, increasing thus the BAO signal.

The first application of the BAO reconstruction to galaxy surveys has been performed by Anderson et al. (2012); Padmanabhan et al. (2012). Padmanabhan et al. (2012) have observed that BAO reconstruction decreases the error on the BAO measurements by almost a factor of two. However, it is worth noting that the reconstruction has been performed on a low-redshift galaxy sample, where it is the most helpful. Therefore the same level of improvement is not expected at higher redshifts. Consequently, since then, this technique has been continuously used in BAO studies (e.g. Bautista et al., 2021; Gil-Marín et al., 2020; Alam et al., 2021; Zhao et al., 2022) to increase the precision and accuracy of the cosmological parameters.

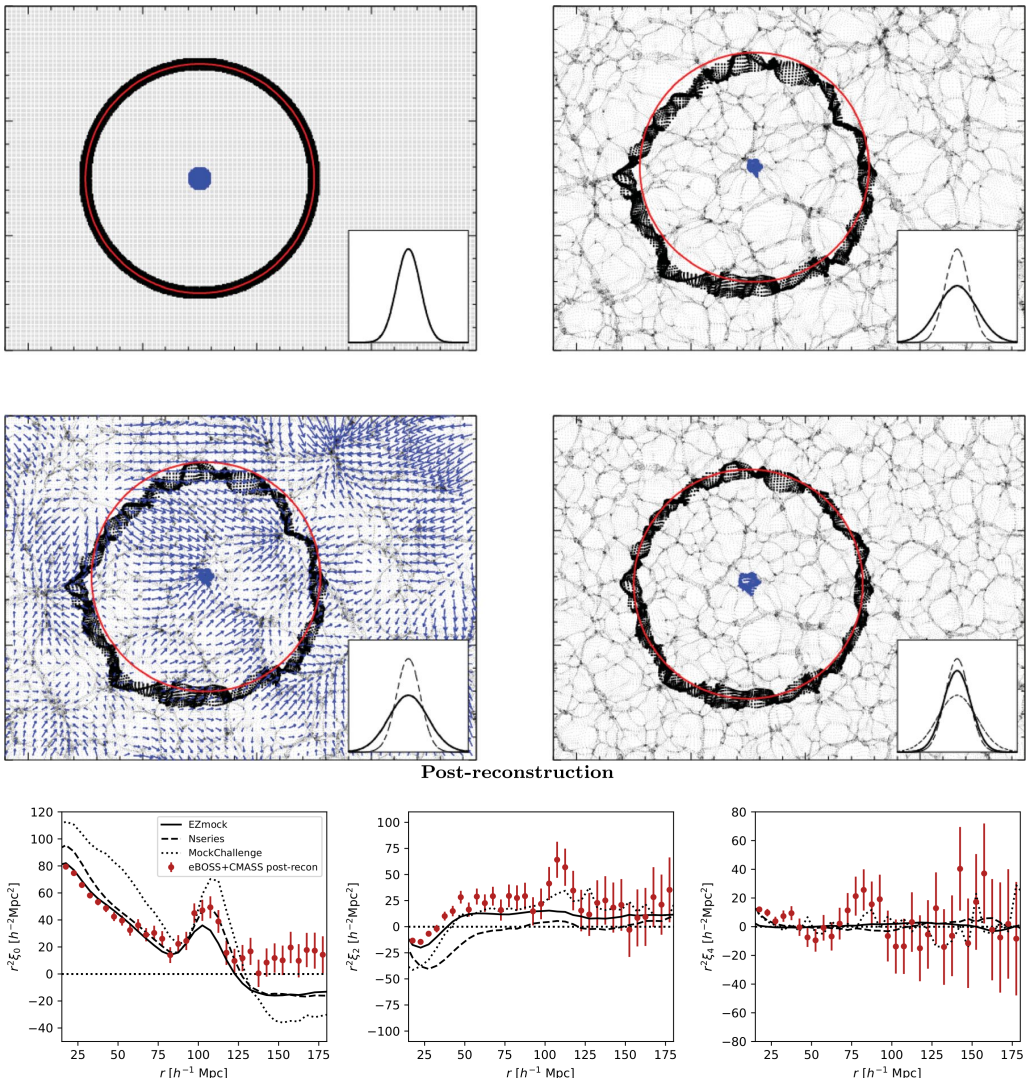


Figure 1.38: Top: A thin slice of a matter density field simulation: the blue points represent an overdense region; the black points around the red circle represent the BAO feature that is being spread out due to the gravitational evolution; the blue arrows illustrate the displacement field $\boldsymbol{\Psi}$. The inner panels show the radial profile of the BAO signature: the black continuous line show the radial profile of the black points around the red circle, whereas the long-dashed lines denote the initial radial profile. Bottom: The 2PCF multipole from a reconstructed galaxy catalogue. Figure 1 from Padmanabhan et al. (2012) and Figure 3 from Bautista et al. (2021).

2 Constructing galaxy catalogues for covariance matrix estimation

Due to the stochastic nature of the large scale structure, an implicit uncertainty – cosmic variance, whose estimation is a challenge in itself – is inherent in its measurement. The first section of this chapter introduces various techniques to estimate the cosmic variance for large scale structure clustering measurements, each having its own advantages and disadvantages. Therefore, a crucial task is to compare different methods for estimating covariance matrices – mathematical objects that describe the cosmic variance – for clustering measurements. Given that DESI will achieve an unprecedented level of statistical precision, making systematic effects potentially significant, the Cosmological Simulations Working Group (CosmoSimsWG) has initiated the DESI mock challenge (Chuang et al., 2023). This challenge aims to compare different methodologies for constructing covariance matrices and to assess the impact of certain systematic effects on the simulated data.

One method involves building numerous galaxy simulations to replicate multiple measurements sampled from the same intrinsic probability distribution. Given the fact that the full N-body simulations are computationally expensive, faster methods like FastPM have been developed to evolve the DM field and obtain haloes. As a consequence, the second section presents an overview of different methods to create galaxy catalogues starting from DM haloes.

The third section summarises my contribution to the generation of the First Generation Mocks for DESI. Practically, I have co-developed a code and applied it on cubic simulations to cut the survey geometry and make them more realistic.

Part of the DESI mock challenge, I have applied the Halo Occupation Distribution (HOD) technique to assign galaxies to FastPM haloes. Additionally, I have assessed the sensitivity of the estimated covariance matrices to the HOD fitting. The results are presented in the last section, which constitutes an article submitted to the Monthly Notices of the Royal Astronomical Society (MNRAS) (Variu et al., 2023a).

2.1 Covariance matrix estimation

As explained in Section 1.4.4, the cosmological parameters are constrained by fitting a dedicated model to the clustering statistics (e.g. power spectrum or 2PCF) for which a covariance matrix is needed. There are multiple methods to estimate the covariance matrix: the mock based technique in which the covariance matrix is computed from an ensemble of simulated datasets, internal estimators that resample the observed dataset (e.g. jackknife estimation), and analytical methods.

2.1.1 Sampled covariance matrix

In order to compute the sampled covariance matrix, one requires multiple clustering measurements. This can be achieved by building many simulations and measure their clustering statistics. Previous surveys have used faster approximate simulations such as PATCHY mocks (Kitaura et al., 2013) and EZMOCKS (Chuang et al., 2015; Zarrouk et al., 2021; Zhao et al., 2021) (for BOSS and eBOSS). DESI tests additional simulations such as BAM (Balaguera-Antolínez et al., 2020; Balaguera-Antolínez et al., 2019; Pellejero-Ibañez et al., 2020) and FASTPM. One issue is the fact that approximate techniques are less accurate at the non-linear scales which can affect the covariance matrix at the those scales. Additionally, the estimated matrix is sampled from a Wishart distribution which can affect the estimation of the parameter errors, see e.g. Hartlap et al. (2007); Percival et al. (2022).

Denoting by $Y(x)$ the clustering statistics as function of x (e.g. $P(k)$, $\xi(s)$), one can compute the sampled covariance matrix:

$$\mathbf{C}_{\rm s} = \frac{1}{N_{\rm mocks}-1}\mathbf{M}^{\rm T}\mathbf{M}, \tag{2.1}$$

where the components of the matrix $\mathbf{M}$ are defined

$$\mathbf{M}_{ij} = Y_i(x_j) - \bar{Y}(x_j), \qquad i = 1,2,...,N_{\rm mocks}, \qquad x_j \in [x_{\rm min}, x_{\rm max}]. \tag{2.2}$$

The Y_i denotes the vector corresponding to the $i-$th clustering realisation, $\bar{Y}$ represents the mean vector over all $N_{\rm mocks}$ realisations and $[x_{\rm min}, x_{\rm max}]$ defines the interval of points of interest.

2.1.2 Jackknife

In this Section, we only introduce the delete-one Jackknife (or just Jackknife) technique. Nevertheless, other resampling methods such as bootstrapping exist. For more details of internal estimators one can consult e.g. Norberg et al. (2009); Mohammad & Percival (2022).

The principle behind the delete-one Jackknife is to split the volume in $N_{\rm sub}$ sub-volumes and compute the clustering statistics for the total volume of $(N_{\rm sub}-1)$ sub-volumes. This means,

that there are $N_{\rm sub}$ clustering realisations each using a fractional $(N_{\rm sub}-1)/N_{\rm sub}$ volume of the total.

$$\mathbf{C}_{\rm JK} = \frac{N_{\rm sub}-1}{N_{\rm sub}}\mathbf{M}^{\rm T}\mathbf{M}, \tag{2.3}$$

where the components of the matrix $\mathbf{M}$ are similar to the ones of equation (2.2) except i, that runs from 1 to $N_{\rm sub}$ instead of $N_{\rm mocks}$.

The advantage of this technique is that all physical effects in the data are present in the covariance as well. In contrast, for mock based techniques one has to transform the cubic simulations into realistic one by adding systematic effects. The disadvantage is that the internal estimators can overestimate the true covariance matrix by 25 to 60 per cent, see e.g. Norberg et al. (2009). Nevertheless, Mohammad & Percival (2022) have shown that their weighting schemes can adjust the Jackknife to reliably estimate the covariance matrix for the 2PCF. Lastly, for these kind of methods, it is more difficult to estimate the window function (i.e. the shape of the volume) when computing the power spectrum from the Jackknife volumes.

2.1.3 Analytical covariance matrix

Considering that the primordial overdensities are sampled from a Gaussian distribution and all Fourier k modes of $\delta(k)$ grow independently, the covariance matrix between $\left(\xi_\ell(s), \xi_{\ell'}(s)\right)$ is (Xu et al., 2013):

$$C_{ij}^{\ell\ell'} = \frac{2(2\ell+1)(2\ell'+1)}{V}\int \frac{k^3 {\rm d}\log k}{2\pi^2} j_\ell(kr_i) j_{\ell'}(kr_j) P^2_{\ell\ell'}(k), \tag{2.4}$$

where V is the survey volume, $j_\ell(kr)$ is the spherical Bessel function of order ℓ and

$$P^2_{\ell\ell'}(k) = \frac{1}{2}\int_{-1}^{1}\left[P(k,\mu)+\frac{1}{\overline{n}}\right]^2 L_\ell(\mu)L_{\ell'}(\mu){\rm d}\mu. \tag{2.5}$$

$P(k,\mu)$ is the 2D power spectrum, $\overline{n}$ is the average galaxy number density and $L_\ell(\mu)$ are Legendre polynomials of order ℓ. This covariance matrix does not take into account the binning of the correlation function, however the binned version can be consulted in Xu et al. (2013).

As explained in Section 1.3, non-linear evolution introduces mode coupling. However, one can largely account for mode coupling by using non-linear models for the 2D power spectrum and shot-noise. Moreover, in the previous formula $\overline{n}$ is redshift independent, but Xu et al. (2013) provides a method to include $\overline{n}(z)$.

In terms of the power spectrum, one can find different models for its covariance matrix in e.g. Wadekar & Scoccimarro (2020); Wadekar et al. (2020); Blake et al. (2018). In this case, it can be complicated to account for the survey geometry (through the window function) and other effects. Moreover, higher-order correlations such as four-point (a.k.a trispectrum) may

be included to account for various non-Gaussian effects.

Finally, analytical covariance matrices are not affected by sampling noise and are cheaper from the computational cost point of view. Nevertheless, survey geometry (through the window function) is difficult to model analytically, making the mock-based covariance matrices a preferred approach from this point of view.

2.2 Galaxy-Halo connection

Section 1.3 presents the gravitational evolution of the CDM starting from the initial fluctuations in the CDM and primordial plasma of baryons and photons to the large scale structure of the Universe and collapsed objects such as DM haloes. However, that description does not include the specificity of the baryonic evolution after the baryon-photon decoupling. In this section, we briefly introduce the main modelling techniques to understand the evolution and creation of galaxies in the cosmological context, i.e. the galaxy-halo connection and how they can be used to create galaxy catalogues starting from DM haloes ones. For a more detailed presentation, we refer to Wechsler & Tinker (2018) and references therein.

After decoupling, the gas began to fell in the potential wells of the DM haloes. Furthermore, the gas cooled down enough to form stars and then protogalaxies, in the massive enough DM haloes. The further galaxy evolution is influenced by the evolution of DM haloes and energetic processes within galaxies such as feedback effects. The galaxy-halo connection incorporates both the physical and statistical links between halos and galaxies.

Figure 2.1 shows the large scale structure formed out of the CDM fluid and a biased galaxy distribution that is tuned to match clustering properties of an observed galaxy sample. The bias[1] implies that galaxies form only in some regions of the DM structure, usually in DM haloes about a certain mass threshold [2]. One of the reason is that due to different astrophysical effects, the gas in some regions cannot collapse to form stars.

Studies of the stellar-to-halo mass relation (i.e. the mass of a typical galaxy as function of its host halo mass) show that in fact, assuming all haloes contain $\Omega_{\rm b}/\Omega_{\rm m} \approx 0.17$ fraction of baryons, only 20 to 30 per cent – at its peak – of baryons have collapsed into stars. For more massive and less massive haloes, the percentage is even lower. The Active-Galactic-Nucleus (AGN) of galaxies can heat the halo gas hindering the star formation and thus decreasing the abundance of high mass galaxies in massive haloes. For lower mass haloes, feedback of massive stars such as stellar winds can eject gas or prevent it from falling into a galaxy, limiting the maximum galaxy mass.

In addition, Figure 2.1 contains a summary of main types of galaxy-halo connection models. These models can be used to produce simulated galaxy catalogues. It is important to mention

[1] See also Section 1.4.3.

[2] Note that not all haloes above this threshold contain galaxies.

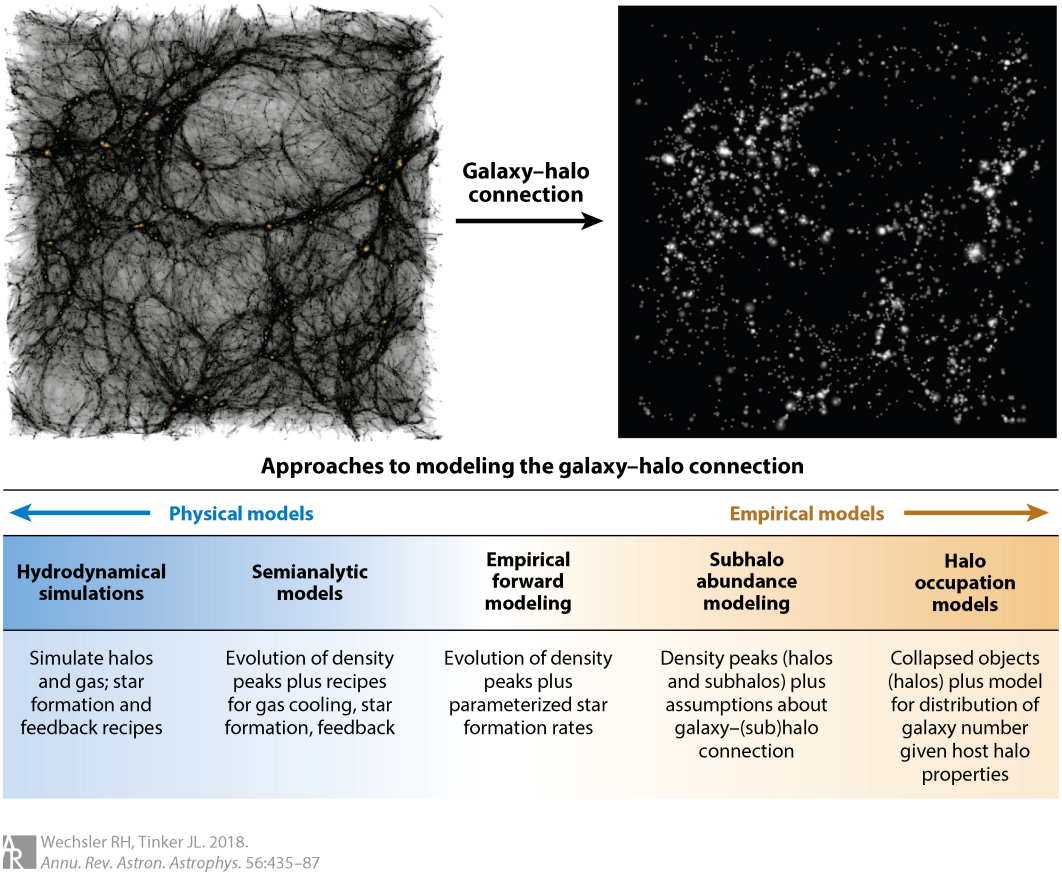


Figure 2.1: The simulated dark matter distribution and the corresponding galaxy distribution obtained using an abundance matching model. A summary of galaxy-halo connection models. Figure 1 of Wechsler & Tinker (2018)

that there is no approach that is agreed to as "strictly correct". Each method has its own limitations where it may not match the data in particular regimes. Therefore, we briefly introduce each of them.

2.2.1 Abundance Matching models

Abundance Matching (AM) models require a match between some properties (e.g. mass, size) of galaxies and some halo properties (e.g. mass, maximum circular velocity). The simplest and most intuitive match is the one of masses. One can assume that the most massive galaxies are hosted by the most massive haloes. Nevertheless, one has to consider a certain scatter between these two properties.

The CDM paradigm suggests that the DM haloes contain substructures, called subhaloes. As a consequence, a simple generalisation is that each halo and subhalo – above a certain threshold – host a galaxy whose mass (or other property) is matched by abundance to the property of its host. This is called subhalo abundance matching (SHAM).

AM can be regarded as a non-parametric technique that directly links the stellar mass function to the halo mass function, despite the necessity of including a scatter and finding the matching properties. Nevertheless, an important requirement of these models is high-resolution simulations capable to resolve the DM substructures and to accurately keep track of the history of the halo (i.e. merger tree).

2.2.2 The Halo Occupation Distribution

The Halo Occupation Distribution (HOD) method has been used in the article shown in Section 2.4. Briefly, the number of galaxies (central or satellite) is determined by a Probability Distribution Function, whose mean depends on the halo mass (or luminosity) through different functional forms.

A generalisation of this method is to include a conditional luminosity function (CLF) in order to describe the full distribution of galaxy luminosities for a given halo mass. In a similar way as for HOD, the distribution of central galaxy luminosities and the one of satellite galaxies are treated separately.

Compared to AM models, the HOD ones can be very complicated depending on the studied galaxy sample, i.e. whether it was selected by star formation rates or emission lines.

2.2.3 Empirical forward modelling

Empirical forward modelling is used to understand the galaxy evolution inside haloes through time. This is achieved by studying the galaxy-halo connection at each epoch. For example, one can perform AM at each epoch and follow the evolution in time of haloes through the

mass accretion histories (e.g. in a N-body simulation), in order to study the galaxy accretion history and star formation history.

Another technique is to parameterize the connection between galaxy star formation rate and the halo mass accretion rate and use simulated merger histories, in order to analyse and predict the galaxy evolution (such as star formation histories and statistical galaxy properties). A downside of this approach is that it requires high-resolution simulations to trace the evolution of DM haloes and subhaloes.

2.2.4 Hydrodynamical simulations

In contrast to the previous models, hydrodynamical simulations build galaxies by solving the equations of both gravity and hydrodynamics in an expanding Universe. This is done by including astrophysical processes such as stellar, black hole and supernovae feedbacks, gas cooling, and following the evolution of dark matter, gas and stars over time.

As it is impossible to simulate in a cosmological context, all physical phenomena down to the scales needed for galaxy formation, one needs to parameterize physical phenomena occurring below the resolution scale. These effects are included in the so-called "subgrid physics" domain. These parameterizations can be tuned using real measurements or using results of empirical models (e.g. AM, HOD) connected to the observations, as well.

The combined study of hydrodynamical simulations and empirical models such as HOD allows for robustness tests for both of them. On one hand, it is possible to test the assumptions in the empirical models by comparing them to the these simulations. On the other hand, having HOD models constrained from the data and measured from the hydrodynamical simulations, one can check how realistic hydrodynamical simulations are.

2.2.5 Semi-analytic models

The idea behind these models is to approximate some physical processes with analytic functions that can be used through the merging history of haloes, in order to decrease the computational cost for the study of galaxy formation and evolution. In practice, one can apply these analytic models through the merger trees of N-body simulations. The main disadvantage is that these models have a large number (10 to 30) of parameters. Therefore, the exploration of the parameter space becomes a challenge. In addition, due to the implicit simplifications of these models, one needs to continuously test them against full hydrodynamical simulations and data.

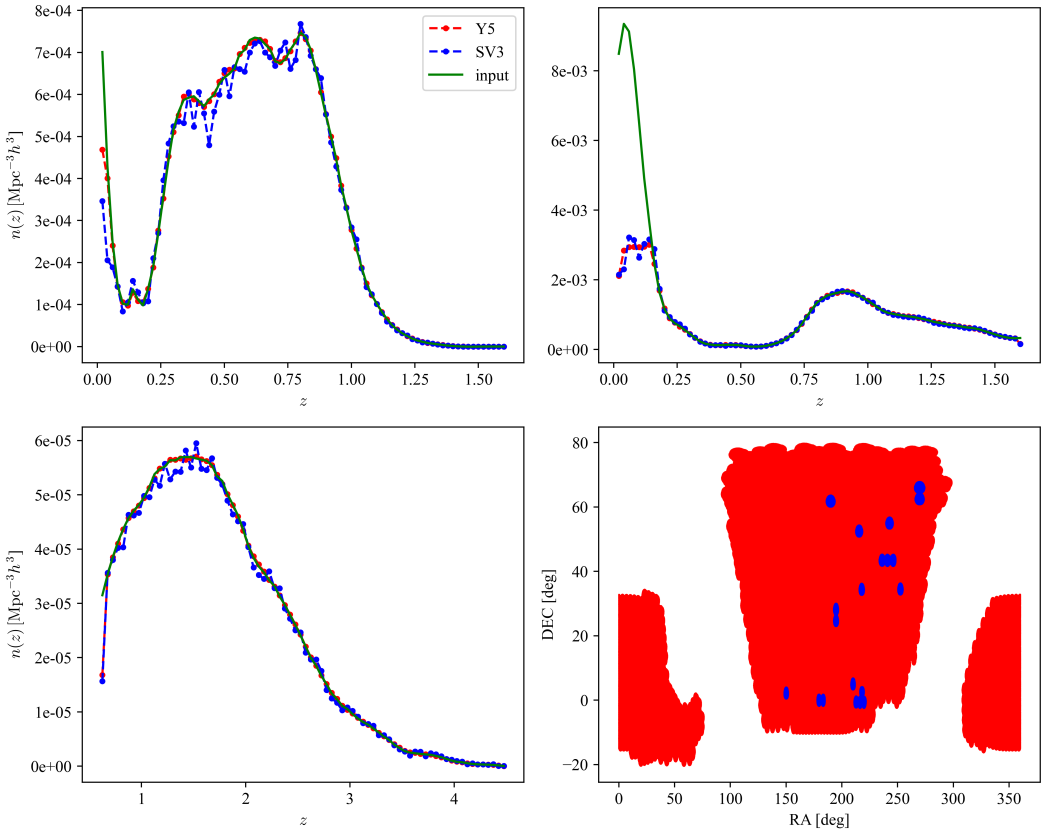


Figure 2.2: The SV3 radial number density $n(z)$ of the three tracers: LRG, ELG, QSO. The green curves denote the input $n(z)$ used to downsample the number of objects, the red and blue curves denote the measured $n(z)$ from a final mock: for the supposed DESI Year 5 (red) footprint and for the SV3 (blue) footprint .

2.3 First Generation Mocks for DESI

Numerical simulations are done in boxes and using (x, y, z) Cartesian coordinates. In contrast, the resulting galaxy catalogues of photometric and spectroscopic surveys contain positions in (RA, DEC, z). Moreover, the radial distribution of the measured targets is not uniform (see Figure 1.28) and the targets do not cover the entire sky, see Figure 1.25. These two observations describe what is called the survey geometry.

In order to correctly estimate the cosmic variance of measurements, the covariance matrix has to take into account the survey geometry. Therefore, at first order, the simulated catalogues must have the same survey geometry of the data. The mocks that have the same survey geometry as the data are called CutSky. Part of CosmoSimWG, I have been tasked to apply the survey geometry on the DESI First Generation Mocks (FirstGenMocks) for LRG, ELG and QSO. The FirstGenMocks are sets of BGS, LRG, ELG and QSO catalogues, whose clustering matches the one from the SV3 DESI data (DESI Collaboration et al., 2023b).

My task was to convert the (x, y, z) into (RA, DEC, z) and apply the survey geometry for 25 realisations of 2 Gpc/h ABACUSSUMMIT (Maksimova et al., 2021) simulation, 1000 cubic EZMOCKS (Chuang et al., 2015; Zarrouk et al., 2021; Zhao et al., 2021) of 2 Gpc/h and 2000 cubic EZMOCKS of 6 Gpc/h, for each of the three tracers: LRG, ELG, QSO.

To this end, I have co-developed an adaptable PYTHON code[3] (Generate-Survey-Mocks, GSM) that reads in parallel the sub-boxes of a simulation, converts in parallel the (x, y, z) to (RA, DEC, z) and applies the survey geometry, see Figure 2.2. The 2 Gpc/h boxes have a lower volume than the one surveyed by DESI for each of the tracer. Therefore, GSM applies the periodic boundary conditions and practically multiplies the box as much as needed to cover the requested volume and then it cuts the survey geometry. As a consequence, the cosmic variance of the CutSky computed from the 2 Gpc/h boxes is not correctly estimating the uncertainty in the real measurements, since the same regions of the box are used multiple times. Nevertheless, the 2 Gpc/h EZMOCKS have been conceived to replicate the cosmic variance of the ABACUSSUMMIT simulations.

In contrast, a 6 Gpc/h box is large enough – if rotated optimally – to cut the volume[4] of either the North Galactic Cap (NGC) or the South Galactic Cap (SGC) of the DESI Year 5 footprint. This is the reason why there are 2000 6 Gpc/h EZMOCKS: 1000 for the NGC and 1000 for the SGC. In the end, the covariance matrix for each of the three tracers is estimated using 1000 mocks. Similarly to the 2 Gpc/h, GSM uses the periodic boundary conditions on the optimally rotated 6 Gpc/h box to cut the survey volume. Nevertheless, due to the larger volume and the optimally chosen rotation, the final CutSky does not contain repeated volumes and thus the set of 1000 CutSky can estimate correctly the cosmic variance of the data.

It is important to notice that the rotation has the role of optimising the necessary cubic volume of the initial simulation. However, there are remapping techniques that transform a cubic simulation into an elongated box-like shape, optimising even more the necessary initial cubic volume (see e.g. Carlson & White, 2010).

The final mocks (survey geometry + additional masking and customisation) have been used by Moon et al. (2023) in the first detection of the BAO on the EDR DESI data (DESI Collaboration et al., 2023b).

2.4 Preprint version: "DESI Mock Challenge: Constructing DESI galaxy catalogues based on FASTPM simulations"

[3] https://github.com/Andrei-EPFL/generate_survey_mocks
[4] up to a redshift of 3

MNRAS 000, 1–18 (2021) Preprint 27 July 2023 Compiled using MNRAS LATEX style file v3.0

DESI Mock Challenge: Constructing DESI galaxy catalogues based on FastPM simulations

Andrei Variu,[1]★ Shadab Alam,[2,3]† Cheng Zhao,[1] Chia-Hsun Chuang,[4,5] Yu Yu,[6,7] Daniel Forero-Sánchez,[1] Zhejie Ding,[6,7] Jean-Paul Kneib,[1] Jessica Nicole Aguilar,[8] Steven Ahlen,[9] David Brooks,[10] Todd Claybaugh,[8] Shaun Cole,[11] Kyle Dawson,[4] Axel de la Macorra,[12] Peter Doel,[10] Jaime E. Forero-Romero,[13,14] Satya Gontcho A Gontcho,[8] Klaus Honscheid,[15,16,17] Martin Landriau,[8] Marc Manera,[18,19] Ramon Miquel,[20,19] Jundan Nie,[21] Will Percival,[22,23,24] Claire Poppett,[8,25,26] Mehdi Rezaie,[27] Graziano Rossi,[28] Eusebio Sanchez,[29] Michael Schubnell,[30,31] Hee-Jong Seo,[32] Gregory Tarlé,[31] Mariana Vargas Magana,[12] Zhimin Zhou,[21]
Affiliations are listed at the end of the paper

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Together with larger spectroscopic surveys such as the Dark Energy Spectroscopic Instrument (DESI), the precision of large scale structure studies and thus the constraints on the cosmological parameters are rapidly improving. Therefore, one must build realistic simulations and robust covariance matrices.

We build galaxy catalogues by applying a Halo Occupation Distribution (HOD) model upon the FastPM simulations, such that the resulting galaxy clustering reproduces high resolution N-body simulations. While the resolution and halo finder are different from the reference simulations, we reproduce the reference galaxy two-point clustering measurements – monopole and quadrupole – to a precision required by the DESI Year 1 Emission Line Galaxy sample down to non-linear scales, i.e. $k < 0.5\,h/\mathrm{Mpc}$ or $s > 10\,\mathrm{Mpc}/h$.

Furthermore, we compute covariance matrices based on the resulting FastPM galaxy clustering – monopole and quadrupole. We study for the first time the effect of fitting on Fourier conjugate [e.g. power spectrum] on the covariance matrix of the Fourier counterpart [e.g. correlation function]. We estimate the uncertainties of the two parameters of a simple clustering model and observe a maximum variation of 20 per cent for the different covariance matrices. Nevertheless, for most studied scales the scatter is between two to ten per cent

Consequently, using the current pipeline we can precisely reproduce the clustering of N-body simulations and the resulting covariance matrices provide robust uncertainty estimations against HOD fitting scenarios. We expect our methodology will be useful for the coming DESI data analyses and their extension for other studies.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

The study of Large Scale Structure of the Universe has significantly improved in the last two decades leading to Baryon Oscillation Spectroscopic Survey (BOSS; Alam et al. 2017) and extended-BOSS (eBOSS; Alam et al. 2021a) surveys. They have published the largest 3D map of over 2 millions galaxies and quasars (Alam et al. 2021a). This has allowed the measurement of cosmological parameters to a percent-level precision studying Baryonic Acoustic Oscillations (BAO) and Redshift Space Distortions (RSD).

Currently, the Dark Energy Spectroscopic Instrument (DESI; Levi et al. 2013; DESI Collaboration et al. 2022) is a five years long spectroscopic survey that will outperform previous surveys by a an order of magnitude(DESI Collaboration et al. 2016a), aiming to constrain the cosmological parameters with precision at a sub-percent level. With its 5000 robotically controlled optical fibres (Silber et al. 2023; Miller et al. 2023; DESI Collaboration et al. 2016b), DESI will scan a third of the sky to map 40 millions galaxies (Lan et al. 2023) and quasars (Alexander et al. 2023). Only after the five-month Survey Validation (DESI Collaboration et al. 2023a), DESI has measured the spectra of more than one million galaxies leading to the recent Early Data Release (EDR) (DESI Collaboration et al. 2023b).

Based on the DESI Legacy Imaging Surveys (Zou et al. 2017; Dey et al. 2019; Schlegel et al. 2023), there are five types of targets that are selected (Myers et al. 2023) on which optical fibres are assigned (Raichoor et al. 2023a) to measure and analyse their spectra (Guy et al. 2023; Bailey et al. 2023; Brodzeller et al. 2023): Milky Way Stars (MWS; Allende Prieto et al. 2020; Cooper et al. 2022), Bright Galaxies (BGS; Ruiz-Macias et al. 2020; Hahn et al. 2022),

★ E-mail: andrei.variu@epfl.ch
† E-mail: shadab.alam@tifr.res.in

Luminous Red Galaxies (LRG; Zhou et al. 2020, 2023), Emission Line Galaxies (ELG; Raichoor et al. 2020, 2023b), quasars (QSO; Yèche et al. 2020; Chaussidon et al. 2023). Such a complex system requires pipelines to optimise the observations (Schlafly et al. 2023; Kirkby et al. 2023).

The sub-percent precision measurements expected from ongoing and future surveys require careful analyses of the systematic effects. To this end, the DESI Mock Challenge was launched as a series of studies and projects to build and validate the methodology for the cosmological analysis. In particular, one must find the most robust way to estimate the uncertainty of the measurements (Chuang et al. 2023). To achieve this goal, one needs to create multiple realistic simulations of the large-scale structure, which is required to lower the noise on covariance matrix and to describe accurately the non-linear scales.

On one hand, the N-body simulations – e.g. (SLICS; Harnois-Déraps et al. 2018), (UNIT; Chuang et al. 2019) and (AbacusSummit; Maksimova et al. 2021) – are accurate, but they are computationally expensive. Therefore, they are mainly used in testing models and systematic effects, and it becomes impractical with the increase of mapped volume to have enough realisations to estimate and test covariance matrices. Consequently, faster but less accurate techniques have been developed – e.g. (EZmocks; Chuang et al. 2015; Zarrouk et al. 2021; Zhao et al. 2021), (PATCHY; Kitaura et al. 2013), (BAM; Balaguera-Antolínez et al. 2020; Balaguera-Antolínez et al. 2019; Pellejero-Ibañez et al. 2020) – to be run multiple times and estimate robustly the uncertainty.

In this study, we investigate the possibility to tune FastPM catalogues to reproduce the clustering of SLICS reference with the final goal of estimating the covariance matrix. In contrast to the other fast methods, FastPM uses accelerated particle-mesh solvers to evolve the dark-matter field, that should provide a higher accuracy of the large scale structure. The additional accuracy provided by FastPM can be important given the unprecedented statistical power of the DESI survey. Therefore, the FastPM covariance matrix is compared with different methods (BAM, EZmock, Jackknife (Zhang et al. 2023), analytical models (Xu et al. 2013; Wadekar & Scoccimarro 2020; Wadekar et al. 2020)) in a parallel DESI Mock Challenge paper (Chuang et al. 2023).

Fundamentally similar to standard N-body simulations, FastPM evolves the dark matter field into the cosmic web, the skeleton of the large scale structure in the Universe (e.g. Mo et al. 2010; Wechsler & Tinker 2018). After the dark matter haloes are selected, one must implement galaxy-halo connection models (Wechsler & Tinker 2018) to assign galaxies. There are more empirically inspired models such as the Halo Occupation Distribution (HOD; e.g. Benson et al. 2000; Seljak 2000; Peacock & Smith 2000; White et al. 2001; Berlind & Weinberg 2002; Cooray & Sheth 2002) and Sub-Halo Abundance Matching (SHAM; e.g. Kravtsov et al. 2004; Tasitsiomi et al. 2004; Vale & Ostriker 2004) and more physically inspired ones such as full hydro-dynamical simulations (e.g. Schaye et al. 2010, 2015; Dubois et al. 2014; McCarthy et al. 2017; Pillepich et al. 2018; Davé et al. 2019) or Semi Analytical Models (SAMs; e.g. Guo et al. 2011; Gonzalez-Perez et al. 2014). In this case, we adopt a HOD model as it is one the most efficient ways to create mock galaxy catalogues.

The purpose of the current paper is to show that the galaxy assignment process on FastPM halo catalogues with a HOD model can be adjusted to match the reference SLICS galaxy clustering. We thus compare the impact of different clustering statistics and examine the effects of various scales on the HOD fitting. Finally, we calculate covariance matrices for all the studied scenarios and perform a comparison to understand the influence of the HOD modelling on the parameter uncertainty.

In Section 2, we present the SLICS and FastPM simulations. The methodology that we follow is detailed in Section 3. We describe our results on the HOD fitting performance and the covariance matrix comparison in Section 4. In the end, Section 5 concludes the article.

2 SIMULATIONS

2.1 Scinet LIght-Cone Simulations

The Scinet LIght-Cone Simulations (SLICS, Harnois-Déraps & van Waerbeke 2015; Harnois-Déraps et al. 2018) consist of over 900 N-body mocks based on noise independent initial conditions. The large number of realisations is exploited to estimate the covariance matrices for weak lensing data (Joudaki et al. 2017; Hildebrandt et al. 2017; Martinet et al. 2018; Harnois-Déraps et al. 2022) and for combinations of weak lensing and foreground clustering data (Brouwer et al. 2018; van Uitert et al. 2018).

The cubic mocks – with $L_{\rm box} = 505\,{\rm Mpc}/h$ – simulate a flat ΛCDM cosmology, described by the cosmology of the WMAP9 + SN + BAO, i.e. ($\Omega_{\rm m}$, σ_8, $\Omega_{\rm b}$, w_0, h, $n_{\rm s}$) = (0.2905, 0.826, 0.0447, -1.0, 0.6898, 0.969). They are obtained by running the non-linear double-mesh Poisson solver cubep^{3}m (Harnois-Déraps et al. 2013) to gravitationally evolve 1536^3 particles – with a particle mass $m_{\rm p} = 2.88 \times 10^9\ M_\odot/h$ – on a 3072^3 grid from $z = 99.0$ up to $z = 0$.

The dark matter haloes have been selected by applying a spherical over-density halo-finder (Harnois-Déraps et al. 2013). Their mass function follows precisely the Sheth et al. (2001) fitting function, as shown in Figure 2 of Harnois-Déraps et al. (2018). The redshift of the halo catalogues included in this study is $z = 1.041$. Lastly, given that some halo catalogues have been corrupted at the run time, we are limited to only 139 independent mocks.

This study, together with the BAM (Balaguera-Antolínez et al. 2022), JackKnife and the DESI covariance matrix comparison papers (Chuang et al. 2023) focused on the DESI Emission Line Galaxies (ELGs) sample. Thus, one must assign galaxies on the SLICS halo catalogues. To this end, a HOD model adjusted for ELGs (Alam et al. 2020, 2021b) is implemented to create a set of 139 galaxy catalogues that are used as reference in all the studies mentioned before. More details about the SLICS galaxy catalogues production can be found in the DESI covariance matrix comparison paper (Chuang et al. 2023).

2.2 Fast Particle-Mesh

Accelerated Particle–Mesh (PM) solvers – such as the FastPM software (Feng et al. 2016) – are able to produce accurate halo populations with respect to the full N-body simulations. Thus, they are suitable to accurately simulate large volumes.

FastPM makes use of a pencil domain-decomposition Poisson solver and Fourier-space four-point differential kernel to compute the force. Additionally, the vanilla leap-frog scheme for the time integration is adjusted to account for the acceleration of velocity during a step, allowing for the accurate tracking of the linear growth of large-scale modes regardless of the number of time steps.

For the current analysis. we have run FastPM with two resolutions, resulting in one set of 778 Low Resolution boxes (LR; 1296^3 particles) and one set of 141 High Resolution (HR; 1536^3 particles) catalogues. Both sets output snapshots at the same redshift ($z = 1.041$), and have the same box side length ($L_{\rm box} = 505\,{\rm Mpc}/h$) and cosmology as the SLICS simulations. In contrast to SLICS, the

particle mass of the HR simulations is $2.86444 \times 10^9\ M_\odot/h$, while for LR it is $4.77 \times 10^9\ M_\odot/h$. The resolution of the force mesh is boosted by a factor of $B = 2$ compared to the number of particles per side, for both LR and HR. Lastly, 40 linear steps have been used to evolve the density field from $a = 0.05$ to $a = 0.96$.

Due to the small number of SLICS galaxy realisations, for 123 runs of the FastPM (LR and HR likewise), we use the SLICS initial conditions. This plays an important role to reduce the effect of the cosmic variance in the clustering statistics and thus in the HOD fitting. SLICS initial density field (initial conditions) has been estimated using the Zel'dovich approximation (Zel'dovich 1970): $\delta_{\rm IC}^{\rm HR}(\mathbf{q}) = -\nabla_{\mathbf{q}}\Psi_Z(\mathbf{q})$, where $\Psi_Z(\mathbf{q}) = q - q_{\rm G}$ is the difference between the Lagrangian particle coordinates q and the Lagrangian coordinates $q_{\rm G}$ of a 1536^3 regular grid. Lastly, the initial conditions had been downgraded to the LR by cutting in Fourier space the high frequency modes larger than the Nyquist frequency corresponding to the LR field.

The halos have been selected from the dark matter field with the Friends-of-Friends halo finder in nbodykit (Hand et al. 2018). During the galaxy assignment process – Section 3.3 – we only make use of halos with a minimum mass of $5.72 \times 10^{10}\ M_\odot/h$.

Finally, in Section 3, when we mention FastPM, we imply for simplicity both HR and LR. We only make the distinction in the results section, i.e. Section 4.

3 METHODOLOGY

3.1 Clustering computation

3.1.1 Two point correlation function

Mathematically, the two-point correlation function (2PCF) is a continuous function that can describe the clustering of galaxies. However, given the discrete nature of the galaxy distribution in the Universe, the 2PCF is measured using discrete estimators. In the case of cubic mocks, one can implement the natural estimator (Peebles & Hauser 1974):

$$\xi(s,\mu) = \frac{DD(s,\mu)}{RR(s,\mu)} - 1, \tag{1}$$

where $DD(s,\mu)$ and $RR(s,\mu)$ are the data and the random pair counts, respectively, as functions of the radial distance

$$s = \sqrt{s_\perp^2 + s_\parallel^2}, \tag{2}$$

and the cosine of the angle between $\mathbf{s}$ and the line-of-sight

$$\mu = \frac{s_\parallel}{s}. \tag{3}$$

In the previous equations, $s_\perp$ and $s_\parallel$ are the perpendicular ($\perp$) and parallel ($\parallel$) to the line-of-sight components of $\mathbf{s}$, respectively. While the DD term is evaluated directly on the data catalogue, RR is calculated theoretically.

In the present analysis, we run pyFCFC[1] the python wrapper of the Fast Correlation Function Calculator[2] (Zhao 2023, FCFC) to estimate the 2PCF. Lastly, we decompose the 2D 2PCF ($\xi(s,\mu)$) into 1D multipoles ($\xi_\ell(s)$) with the help of the Legendre polynomials $L_\ell(\mu)$ of order ℓ, as follows:

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1} \xi(s,\mu)L_\ell(\mu)d\mu. \tag{4}$$

[1] https://github.com/dforero0896/pyfcfc
[2] https://github.com/cheng-zhao/FCFC

3.1.2 Power spectrum

From the mathematical point of view, the power spectrum $P(\mathbf{k})$ is the Fourier Transform of the 2PCF. However, the limited volume of a survey or a simulation creates mode coupling and makes the two clustering measurements not completely equivalent. Consequently, $P(\mathbf{k})$ is computed starting from the density field in Fourier space $\delta(k)$, as follows:

$$\langle\delta(\mathbf{k})\delta(\mathbf{k}')\rangle = (2\pi)^3\delta_D(\mathbf{k}+\mathbf{k}')P(\mathbf{k}), \tag{5}$$

where δ_D is the Dirac delta function.

As for the 2PCF, we evaluate the multipoles ($P_\ell(k)$) of the power spectrum ($P(k,\mu)$):

$$P_\ell(k) = \frac{2\ell+1}{2}\int_{-1}^{1} P(k,\mu)L_\ell(\mu)d\mu, \tag{6}$$

where μ is the cosine angle between $\mathbf{k}$ and the line-of-sight, i.e.,

$$\mu = k_\parallel/k, \quad k = \sqrt{k_\perp^2 + k_\parallel^2}. \tag{7}$$

In practice, we harness the versatility of POWSPEC[3] described in Zhao et al. (2021) through its python wrapper[4] to calculate the power spectra and their multipoles starting from the galaxy catalogues. We estimate the density field on a grid of size 512^3, by applying the Cloud-In-Cell (CIC; Sefusatti et al. 2016) particle assignment scheme on the catalogues of galaxies. Lastly, we exploit the grid interlacing technique (Sefusatti et al. 2016) to reduce the alias effects at small scales.

In the current analysis, we show the monopole ($\ell = 0$), quadrupole ($\ell = 2$) and hexadecapole ($\ell = 4$) for both the 2PCF and the power spectrum.

3.1.3 Bi-spectrum

The power spectrum and the 2PCF are two-point clustering statistics, but higher order statistics are necessary to characterize more precisely the galaxy distributions. In this study, we also look at the three-point clustering statistics, namely the bi-spectrum $B(\mathbf{k_1},\mathbf{k_2},\mathbf{k_3})$, the Fourier pair of the three-point correlation function (e.g. Bernardeau et al. 2002):

$$\delta^D(\mathbf{k_1}+\mathbf{k_2}+\mathbf{k_3})B(\mathbf{k_1},\mathbf{k_2},\mathbf{k_3}) = \langle\delta(\mathbf{k_1})\delta(\mathbf{k_2})\delta(\mathbf{k_3})\rangle. \tag{8}$$

The three vectors $\mathbf{k_1},\mathbf{k_2},\mathbf{k_3}$ are chosen to form a triangle whose two of the three sides are fixed ($k_1 = 0.1 \pm 0.05$ and $k_2 = 0.2 \pm 0.05$), but the angle θ_{12} between $\mathbf{k_1}$ and $\mathbf{k_2}$ is varied from 0 to π. In practice, we run the bispec[5] code with a grid size of 512^3 to compute the monopole of the bispectra, .

3.2 FastPM HOD model

The galaxy population and its associated clustering covariance matrix can potentially be influenced by halo properties beyond just mass, as shown in Alam et al. (2023). Nonetheless, such effects are expected to be small for large volume surveys such as DESI and hence we plan to address them in future work. Additionally, the FastPM haloes are less accurate than the ones from a N-body simulation, thus we do not expect that the final HOD model and parameters maintain the same physical interpretation.

[3] https://github.com/cheng-zhao/powspec
[4] https://github.com/dforero0896/pypowspec
[5] https://github.com/cheng-zhao/bispec

As a consequence, we can adopt the simple five-parameter HOD model described in Zheng et al. (2005) to assign galaxies to the FastPM halo catalogues, as long as the resulting clustering and covariance matrix match the reference. Nevertheless, in future work one can study more complex and more adapted models for the studied ELG sample.

The current model assumes that each halo can host at most one central galaxy with a probability $\mathcal{B}(1) = \langle N_{\rm cen}\rangle(M_{\rm h})$ dependent on the halo mass $M_{\rm h}$, where $\mathcal{B}(x)$ denotes the Bernoulli distribution and:

$$\langle N_{\rm cen}\rangle(M_{\rm h}) = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M_{\rm h} - \log M_{\rm min}}{\sigma_{\log M}}\right)\right] \tag{9}$$

with erf the error function:

$${\rm erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2} du. \tag{10}$$

$\log M_{\rm min}$ is the halo mass at which the probability to host a central galaxy is one half and $\sigma_{\log M}$ controls the steepness of the transition from a probability of one to zero. Lastly, the positions and velocities of the central galaxies are precisely the values of their parent haloes.

In contrast, the number of satellite galaxies $n_{\rm sat}$ per halo is sampled from a Poisson distribution $\mathcal{P}(n_{\rm sat}|\langle N_{\rm sat}\rangle(M_{\rm h}))$ with the mean:

$$\langle N_{\rm sat}\rangle(M_{\rm h}) = \left(\frac{M_{\rm h} - M_0}{M_1}\right)^{\alpha}, \tag{11}$$

where M_0 is a minimum halo mass threshold below which haloes cannot host satellite galaxies and together with M_1 indicating the halo mass at which one halo hosts on average one satellite galaxy, and α is the power-law index. Furthermore, the positions and velocities of the satellite galaxies follow the Navarro-Frenk-White (Navarro et al. 1996, NFW) density profile.

In the interest of adjusting the smaller scales and the quadrupole, we introduce a velocity dispersion factor ($v_{\rm disp}$) for the velocity parallel ($\parallel$) to the line-of-sight (i.e oZ in the current case) of the satellite galaxies, in addition to the five HOD parameters:

$$v_{\parallel}^{\rm sat,new} = \left(v_{\parallel}^{\rm sat,old} - v_{\parallel}^{\rm halo}\right) \times v_{\rm disp} + v_{\parallel}^{\rm halo}, \tag{12}$$

where $v_{\parallel}^{\rm halo}$ is the velocity parallel to the line-of-sight of the satellites' parent halo. Finally, the six free parameters are fitted so that the resulting FastPM clustering matches the SLICS one.

3.3 HOD fitting

We would like to draw the attention of the reader to Table 1. It contains a summary of important symbols related to the HOD fitting.

With the aim of finding the best-fitting FastPM clustering, we run a HOD Optimization Routine (HODOR[6]). It uses the Halotools (Hearin et al. 2017) package to define and apply the HOD model and PyMultiNest (Buchner et al. 2014) the python wrapper of MultiNest (Feroz & Hobson 2008; Feroz et al. 2009, 2019) to sample the six HOD parameters.

MultiNest is a sampler based on Bayes' theorem that provides the maximum likelihood (best-fitting) parameters, as well as the posterior probability distribution of parameters alongside the Bayesian evidence. Bayes' theorem combines prior knowledge about the Θ parameters of a model M with information from the data D to calculate the posterior probability density of the Θ parameters:

$$p(\Theta|D, M) = \frac{p(D|\Theta, M)p(\Theta|M)}{p(D|M)}, \tag{13}$$

[6] https://github.com/Andrei-EPFL/HODOR

Notation	Meaning
$N_{\rm mocks}^{\rm cov} = 123$	The number of FastPM and SLICS pairs that share the same initial conditions. These catalogues have been used to compute $\mathbf{C}_s$, Eq.(19), part of $\Sigma_{\rm diff}$.
$N_{\rm mocks}^{\rm fit} = 20$	The number of FastPM and SLICS pairs for which we have computed the clustering during the HOD fitting described in Section 3.3.1 and Section 3.3.2.
$\Sigma_{\rm diag}$	Eq. (16): Diagonal matrix used during the first step of the HOD fitting, see Section 3.3.1.
$\sigma_{n_{\rm g}}$	Estimation of the galaxy number density noise used in $\Sigma_{\rm diag}$. Standard deviation of 139 SLICS mocks, divided by $\sqrt{139}$.
$\Sigma_{\rm diff}$	Eq. (20): Difference covariance matrix used during the second step of the HOD fitting, see Section 3.3.2.
$\sigma'_{n_{\rm g}}$	Estimation of the galaxy number density noise used in $\Sigma_{\rm diff}$. Standard deviation of 139 SLICS mocks, divided by $\sqrt{N_{\rm mocks}^{\rm fit}}$.
Σ_{χ}	Eq. (22): The covariance matrix used to compute the χ^2_v, Eq.(21). It is not used for fitting.

Table 1. A summary of some of the most important and possibly confusing notations and their meaning.

name	$\log \frac{M_{\rm min}}{M_\odot}$	$\sigma_{\log M}$	$\log \frac{M_1}{M_\odot}$	κ	α	$v_{\rm disp}$
min	11.6	0.01	9	0	0	0.7
max	13.6	4.01	14	20	1.3	1.5

Table 2. The limits of the uniform prior distributions included in the HOD fitting. Note that M_0 from Eq. (11) is $M_0 \equiv \kappa \times M_{\rm min}$. $M_\odot$ denotes the solar mass.

where $p(\Theta|M)$ is the prior distribution of Θ of the model M, $p(D|\Theta, M)$ is the likelihood, and $p(D|M)$ is a normalizing factor called Bayesian evidence.

The uniform prior distributions that we impose on all six parameters are shown in Table 2. Furthermore, we approximate the likelihood by a multivariate Gaussian:

$$p(D|\Theta, M) = \mathcal{L}(\Theta) \sim {\rm e}^{-\chi^2(\Theta)/2}, \tag{14}$$

with the chi-squared:

$$\chi^2(\Theta) = \mathbf{v}^{\rm T}\mathbf{C}^{-1}\mathbf{v}, \tag{15}$$

where $\mathbf{v}$ is the difference between the data and model vectors $\mathbf{v} = S_{\rm data} - S_{\rm model}(\Theta)$, and $\mathbf{C}$ is the covariance matrix.

The purpose of a covariance matrix $\mathbf{C}$ is to estimate the noise in the data, in the context of a noise-free model. Nevertheless, the peculiarity of this study is that both the model ($S_{\rm model}(\Theta)$, FastPM) and the data ($S_{\rm data}$, SLICS) are affected by noise. Due to the small volume of the SLICS and FastPM boxes, the cosmic variance component of the noise would be larger than the expected precision of ongoing surveys such as DESI. However, since the simulations have been run with matching initial conditions, the relevant noise factor is no longer the cosmic variance but rather the difference in the gravitational evolution. Hence, the mock covariance estimated by SLICS or FastPM substantially over-estimates the error for our fittings. The more suitable noise term is the accumulated noise due to gravitational evolution while starting with exactly the same initial conditions.

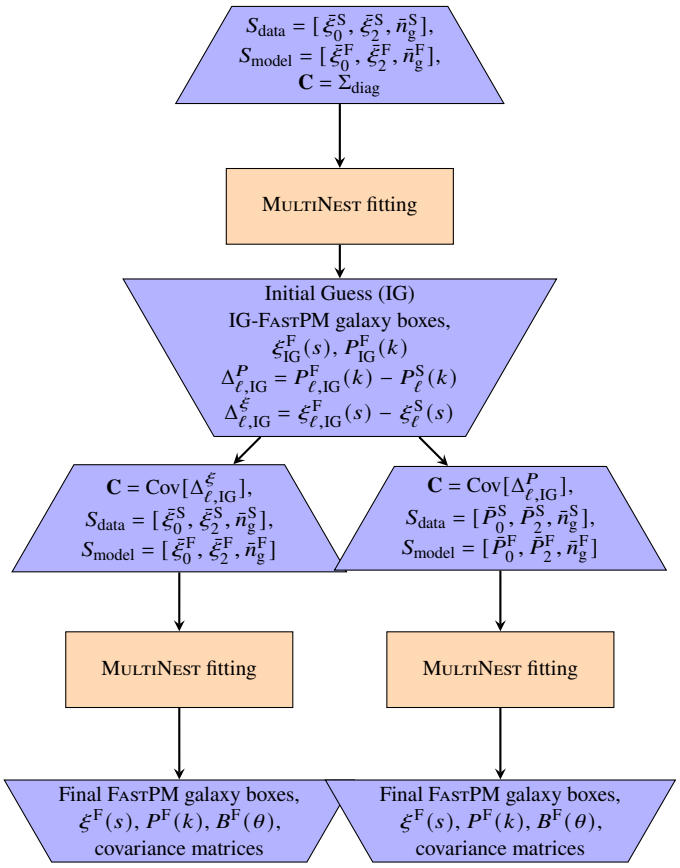


Figure 1. The two-step HOD fitting process that is detailed in Section 3.3.

Consequently, in order to more appropriately estimate the noise, we perform a two-step HOD fitting as schematically shown in Figure 1:

(i) we fit the monopole and quadrupole of the 2PCF $[\xi_0, \xi_2]$ and the galaxy number density $n_{\rm g}$ using a diagonal covariance matrix ($\Sigma_{\rm diag}$) and thus obtain an initial-guess (IG) best-fitting FastPM galaxy catalogues (IG-FastPM), see Section 3.3.1;

(ii) we compute the differences $[\Delta_0, \Delta_2]$ between the clustering (monopole, quadrupole) of the IG best-fitting FastPM and the SLICS galaxy catalogues; we use these differences to calculate a new covariance matrix ($\Sigma_{\rm diff}$) with which we perform again the fitting, see Section 3.3.2.

In both cases, we use 20 FastPM (F) and 20 SLICS (S) halo boxes ($N^{\rm fit}_{\rm mocks} = 20$) – sharing the same initial conditions – for the purpose of decreasing the noise. Nonetheless, the average $\bar{n}_{\rm g}^{\rm S}$ is computed using 139 realisations, while the average $\bar{n}_{\rm g}^{\rm F}$ is calculated using the 20 realisations included in the HOD fitting. There are three main reasons behind this discrepancy: first, it quickly becomes expensive to apply galaxies using HOD to more than 20 FastPM simulations; second, the number of SLICS reference simulations has to be the same as for FastPM, so that the cosmic variance is reduced in the clustering by the shared initial conditions; third, the noise in the galaxy number density is not reduced by the shared initial conditions, thus one needs more realisations to estimate a (practically) noiseless SLICS reference galaxy number density. The galaxy number density is an important constraint as it governs the shot-noise which has a significant role in the covariance matrix.

3.3.1 *The First Step*

Initially, we perform the HOD fitting on the monopole and the quadrupole of the 2PCF, together with the galaxy number density. Hence, the data vector $S_{\rm data}$ is formed by concatenating their respective averages for the SLICS (S) mocks: $S_{\rm data} = [\bar{\xi}_0^{\rm S}, \bar{\xi}_2^{\rm S}, \bar{n}_{\rm g}^{\rm S}]$. Similarly, the model vector $S_{\rm model}$ is determined from the FastPM (F) boxes: $S_{\rm model} = [\bar{\xi}_0^{\rm F}, \bar{\xi}_2^{\rm F}, \bar{n}_{\rm g}^{\rm F}]$.

Considering that the computing time of clustering measurements scales with the maximum separation, we need a large enough upper-limit to constrain relevant parameters, but small enough to keep a reasonable execution time for model evaluation during the HOD fitting. Additionally, since we are interested in capturing the non-linear effects, the lower-limit is set to 0. Consequently, the monopole and the quadrupole of the 2PCF are evaluated for $s \in [0, 50]$ Mpc$/h$, with a bin size of 5 Mpc$/h$. Thus, s is an array containing 10 elements $(s_1, \ldots, s_{10})$.

As previously argued, in the first step, there is no appropriate noise estimation. Therefore, we can use an approximate covariance matrix that enables us to proceed to the second step and calculate a more suitable one. In this regard, we create a diagonal covariance matrix:

$$\Sigma_{\rm diag} = \begin{pmatrix} \sigma_1^2 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & \sigma_{10}^2 & & & & & & & \\ & & & \sigma_1^2 & & & & & & \\ & & & & \ddots & & & & & \\ & & & & & \sigma_{10}^2 & & & & \\ & & & & & & \sigma_{n_{\rm g}}^2 \end{pmatrix}, \tag{16}$$

where the first 20 elements are defined as follows:

$$\sigma_i = \frac{3}{s_i^2}, \; i = 1, \ldots, 10. \tag{17}$$

This selection of the diagonal covariance matrix is based on an examination of the $s^2\sigma_{\rm SLICS}(s)$ values, where $\sigma_{\rm SLICS}(s)$ represents the standard deviation of the SLICS 2PCF. Notably, the highest value is approximately three; hence, we initially approximate all values as three for simplicity.

The last element $\sigma_{n_{\rm g}}$ is computed as the standard deviation of 139 SLICS galaxy number densities, divided by $\sqrt{139}$, so that it estimates the uncertainty corresponding to the average of 139 realisations. The strong constraint on the n_g improves the fitting time, as HODOR initially evaluates the goodness-of-fit based only on the $\bar{n}_{\rm g}^{\rm F}$ and $\bar{n}_{\rm g}^{\rm S}$, and does not compute the clustering if $\bar{n}_{\rm g}^{\rm F}$ is 10σ away from the reference. Additionally, the lack of covariance terms in the covariance matrix should, as well, decrease the convergence time.

Finally, we apply the best-fitting HOD model to all $N^{\rm cov}_{\rm mocks} = 123$ FastPM halo boxes that share the initial conditions with the SLICS mocks to obtain the IG-FastPM.

3.3.2 *The Second Step*

To examine the influence of smaller scales on the HOD fitting, we compute the following for both SLICS and FastPM:

(i) the power spectrum for $k \in [0.02, k_{\rm max}]\, h$/Mpc, with a bin size of 0.02 h/Mpc,

(ii) the 2PCF for $s \in [s_{\rm min}, 50]$ Mpc$/h$, with a bin size of 5 Mpc$/h$,

where the values of $k_{\rm max}$ and $s_{\rm min}$ are presented in Table 3. Consequently, we create the data and model vectors as follows:

name	Large	Medium	Small
$k_{\rm max}\,[h/{\rm Mpc}]$	0.5	0.4	0.3
$N^{\ell}_{\rm bins}$	24	19	14
$s_{\rm min}\,[{\rm Mpc}/h]$	0	5	10
$N^{\ell}_{\rm bins}$	10	9	8

Table 3. The fitting ranges for the HOD fitting process described in Section 3.3.2: $k \in [0.02,\ k_{\rm max}]\ h/{\rm Mpc}$ and $s \in [s_{\rm min},\ 50]\ {\rm Mpc}/h$. $N^{\ell}_{\rm bins}$ is the number of bins per multipole ℓ.

(i) $S_{\rm data} = [\bar{P}^{\rm S}_0, \bar{P}^{\rm S}_2, \bar{n}^{\rm S}_{\rm g}]$ and $S_{\rm model} = [\bar{P}^{\rm F}_0, \bar{P}^{\rm F}_2, \bar{n}^{\rm F}_{\rm g}]$;
(ii) $S_{\rm data} = [\bar{\xi}^{\rm S}_0, \bar{\xi}^{\rm S}_2, \bar{n}^{\rm S}_{\rm g}]$ and $S_{\rm model} = [\bar{\xi}^{\rm F}_0, \bar{\xi}^{\rm F}_2, \bar{n}^{\rm F}_{\rm g}]$.

In order to estimate the noise in the context of shared initial conditions between SLICS and FastPM, we use the $N^{\rm cov}_{\rm mocks}$ galaxy boxes of both SLICS and IG-FastPM, along with their corresponding clustering measurements (power spectrum or 2PCF). Furthermore, we introduce $\Delta^{P}_{\ell,{\rm IG}} = P^{\rm F}_{\ell,{\rm IG}}(k) - P^{\rm S}_{\ell}(k)$ and $\Delta^{\xi}_{\ell,{\rm IG}} = \xi^{\rm F}_{\ell,{\rm IG}}(s) - \xi^{\rm S}_{\ell}(s)$, as well as the generic vector $\Delta^{\rm IG}(x) = [\Delta_{0,{\rm IG}}, \Delta_{2,{\rm IG}}]$ to express the difference between the SLICS and the IG-FastPM galaxy clustering that share the initial conditions. Here, the variable x represents either k or s.

Taking advantage of the previous definitions, we further define a matrix $\mathbf{M}$ with the following elements:

$$\mathbf{M}_{ij} = \Delta^{\rm IG}_i(x_j) - \bar{\Delta}^{\rm IG}(x_j),\ i = 1, 2, ..., N^{\rm cov}_{\rm mocks},\ x_j \in [x_{\rm min}, x_{\rm max}], \tag{18}$$

where $\Delta^{\rm IG}_i$ denotes the vector corresponding to the $i-$th (SLICS, IG-FastPM) pair, $\bar{\Delta}^{\rm IG}$ represents the mean vector over all (SLICS, IG-FastPM) pairs and $[x_{\rm min}, x_{\rm max}]$ defines the interval of points involved in the fitting, see Table 3. Starting from this matrix and its transpose, we calculate the sample covariance matrix $\mathbf{C}_s$ as follows:

$$\mathbf{C}_s = \frac{1}{N^{\rm cov}_{\rm mocks} - 1}\mathbf{M}^{\rm T}\mathbf{M}. \tag{19}$$

Lastly, we calculate the $\sigma'_{n_{\rm g}}$ as the standard deviation of 139 SLICS galaxy number densities, divided by $\sqrt{N^{\rm fit}_{\rm mocks}}$ – so that it estimates the uncertainty corresponding to the average of $N^{\rm fit}_{\rm mocks}$ realisations – and we attach it to the $\mathbf{C}_s$ to obtain the final covariance matrix used in the HOD fitting:

$$\Sigma_{\rm diff} \equiv \begin{pmatrix} \mathbf{C}_s & 0 \\ 0 & \sigma'^2_{n_{\rm g}} \end{pmatrix}. \tag{20}$$

Note that while the error estimate for the clustering is based on the difference in clustering due to matched initial condition, the error of the number density is directly computed from the SLICS realisations, as we aim to constrain the absolute number density, which has strong effect on the final clustering covariance.

3.3.3 Goodness-of-fit

In this section, we define a reduced $\chi^2 - \chi^2_\nu$ – that expresses the goodness-of-fit for the average of $N^{\rm fit}_{\rm mocks}$ FastPM galaxy clustering realisations with respect the SLICS reference, i.e. the $n_{\rm g}$ is not included:

$$\chi^2_\nu = N^{\rm fit}_{\rm mocks} \times \frac{\mathbf{\Delta}^{\rm T}\Sigma^{-1}_{\chi}\mathbf{\Delta}}{\nu}, \tag{21}$$

$\mathcal{K}\,[h/{\rm Mpc}]$	0.1	0.15	0.2	0.25
$\mathcal{S}\,[{\rm Mpc}/h]$	15	20	25	30

Table 4. The fitting ranges – $k \in [0.02, \mathcal{K}]\ h/{\rm Mpc}$ and $s \in [\mathcal{S},\ 200]\ {\rm Mpc}/h$ used in the clustering fitting described in Section 3.4

where $\mathbf{\Delta}$ denotes the difference between FastPM and SLICS clustering – monopole and quadrupole – and $\nu = N_{\rm bins} - N_{\rm params}$, with

(i) $N_{\rm params} = 6$ – the number of free parameters;
(ii) $N_{\rm bins} = 2{\times}N^{\ell}_{\rm bins}$ – the length of the $\Delta^{\rm IG}(x)$ vector, see Table 3.

The Σ^{-1}_{χ} is the unbiased estimate of the inverse covariance matrix (Hartlap et al. 2007):

$$\Sigma^{-1}_{\chi} = \mathbf{C}^{-1}_s \frac{N^{\rm cov}_{\rm mocks} - N_{\rm bins} - 2}{N^{\rm cov}_{\rm mocks} - 1}, \tag{22}$$

where $\mathbf{C}_s$ is defined in Eq. (19). Sellentin & Heavens (2016); Percival et al. (2022) have shown that this correction may not be the optimal choice for accurately determining the uncertainty of the parameters. However, since our main focus is on obtaining the best-fitting clustering and assessing its goodness-of-fit, it remains a reasonable correction.

Finally, as we fit the average of $N^{\rm fit}_{\rm mocks}$ realisations, we must scale the covariance matrix $\mathbf{C}_s$ by a factor of $1/N^{\rm fit}_{\rm mocks}$. As a consequence, the $N^{\rm fit}_{\rm mocks}$ factor appears in Eq.(21).

3.4 Covariance matrix comparison

Given that the main goal is to have a robust estimation of the uncertainty on the cosmological parameters, we want to compare the constraining power of the covariance matrices. To this end, we fit the 123 individual SLICS clustering (monopole and quadrupole) with the following models:

$$P^{\ell}_{\rm model}(k) = b_\ell \times \bar{P}^{\ell}_{123,{\rm SLICS}}(k) \tag{23}$$

and

$$\xi^{\ell}_{\rm model}(s) = b_\ell \times \bar{\xi}^{\ell}_{123,{\rm SLICS}}(s), \tag{24}$$

where $\bar{P}^{\ell}_{123,{\rm SLICS}}(k)$ and $\bar{\xi}^{\ell}_{123,{\rm SLICS}}(s)$ are averages of the 123 realisations and b_ℓ denotes the two free parameters.

Moreover, the covariance matrices are computed similarly to the Eq. (22), but using 778 LR FastPM realisations. The fitting is performed using PyMultiNest, for different fitting ranges ($k \in [0.02, \mathcal{K}]\ h/{\rm Mpc}$ and $s \in [\mathcal{S},\ 200]\ {\rm Mpc}/h$, see Table 4) for the purpose of comparing the effect of the covariance matrices at different scales. The largest fitting intervals are chosen so that they cover the nominal scales included in the BAO and RSD analyses, i.e. $\mathcal{K} \approx 0.2\,h/{\rm Mpc}$ and $\mathcal{S} \approx 20\,{\rm Mpc}/h$ (e.g. Tamone et al. 2020; de Mattia et al. 2021). Finally, the shown values are the average (b_ℓ) and standard deviation (σ_{b_ℓ}) of the marginalised posterior $p(b_\ell)$ and covariance ($\mathcal{R}[b_0, b_2]$) of the posterior distribution of b_0 and b_2, $p(b_0, b_2)$. By construction, the values of b_ℓ should be one.

The main reason why we perform such a simplified test is to avoid the systematic errors that can arise due to the modelling. Consequently, the comparison between the quoted σ_{b_ℓ} and $\mathcal{R}[b_0, b_2]$ should be directly related to the differences in FastPM covariance matrices. We, nevertheless, reckon that these comparisons do not show how the errors on the parameters of a realistic BAO/RSD model would behave.

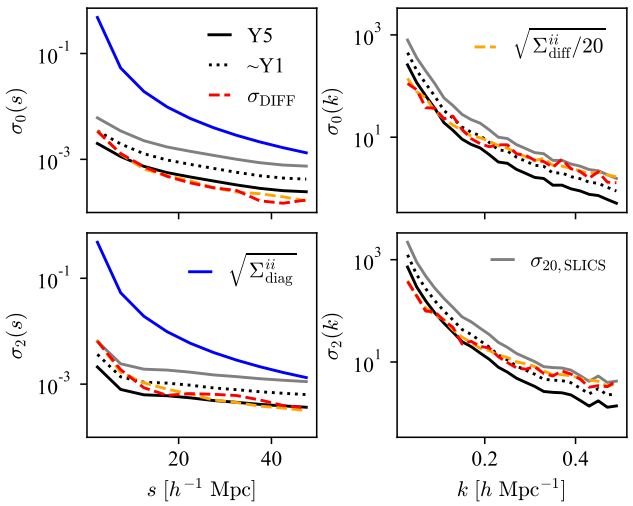


Figure 2. Monopole and quadrupole error bars: left panels – 2PCF; right panels – power spectrum. Black – estimated uncertainty for the entire DESI survey; Dotted black – estimated uncertainty for the Year 1 DESI survey; Blue – square root of the $\Sigma_{\rm diag}$'s terms; Dashed orange – square root of the $\Sigma_{\rm diff}$'s diagonal terms, divided by $\sqrt{20}$, $N^{\rm fit}_{\rm mocks} = 20$; Dashed red – standard deviation of the differences between the best-fitting FastPM clustering (from the second HOD fitting step, one HOD fitting scenario) and SLICS ($N^{\rm fit}_{\rm mocks}$ realisations), further divided by $\sqrt{N^{\rm fit}_{\rm mocks}}$; Grey – standard deviation of $N^{\rm fit}_{\rm mocks}$ SLICS clustering realisations, further divided by $\sqrt{N^{\rm fit}_{\rm mocks}}$.

4 RESULTS

One of the challenges of HOD fitting is addressing the high precision imposed by large volume surveys such as DESI because it requires prohibitively many large volume simulations. Figure 2 illustrates this issue as a comparison between $\sigma_{20,\rm SLICS}$[7] the noise corresponding to the average of $N^{\rm fit}_{\rm mocks} = 20$ SLICS clustering realisations and the expected DESI Y5[8] and Y1[9] errors of the ELG sample. It is obvious that $N^{\rm fit}_{\rm mocks}$ SLICS realisations do not reach the required precision[10].

In order to overcome this issue, we employ the novel matched initial conditions simulations (SLICS and FastPM). In this case, the effect of the cosmic variance on the clustering difference is mostly removed. Therefore, as discussed in Section 3.3, the relevant error estimate is given by the covariance matrix of the clustering difference between the two simulations. Given the fact that we use $N^{\rm fit}_{\rm mocks}$ pairs to perform the HOD fitting, the covariance matrix must be rescaled by $N^{\rm fit}_{\rm mocks}$. The square root of the diagonal of the resulting covariance matrix is illustrated with an dashed orange line in Figure 2. One can observe that the matched initial conditions significantly reduce the noise to values below $\sigma_{20,\rm SLICS}$.

Furthermore, we would like to highlight that the precision depicted by the dashed orange line is either better than or equal to DESI Y1 precision up to $k \approx 0.25\,h/\rm Mpc$. Consequently, the results presented in this paper are precise enough with respect to the requirements of further DESI Y1 analyses. Nonetheless, it might be necessary to readdress this study for the full DESI sample, to account for even lower noise levels. For this, one could use the 1800 AbacusSummit (Maksimova et al. 2021) N-body $0.5\,\rm Gpc/\mathit{h}$ cubic boxes.

In addition, Figure 2 illustrates the comparison between $\sigma_{\rm DIFF}$ and the square root of the diagonal elements of $\Sigma_{\rm diff}$. In this context, $\sigma_{\rm DIFF}$ represents the standard deviation of the differences between the best-fitting FastPM (obtained from the second HOD fitting step) and SLICS clustering, further divided by $\sqrt{N^{\rm fit}_{\rm mocks}}$. Ideally, an iterative HOD fitting process should be performed to ensure a robust $\Sigma_{\rm diff}$, but the close agreement between $\sigma_{\rm DIFF}$ and the diagonal elements of $\Sigma_{\rm diff}$ suggests that $\Sigma_{\rm diff}$ has approximately converged after a single iteration. A more detailed argument in support of the convergence of $\Sigma_{\rm diff}$ is presented in Section A.

As pointed out in Section 3.3, it is important that the FastPM galaxy catalogues reproduce the SLICS shot-noise. Examining the FastPM galaxy number densities of all HOD fitting cases, we observed that the largest deviation, $|\bar{n}^{\rm S}_{\rm g} - \bar{n}^{\rm F}_{\rm g}|/\sigma'_{n_{\rm g}}$, is approximately $0.5\,\sigma$, but most values are below $0.2\,\sigma$. This strongly supports that the galaxy number density is well constrained and that it is safe to define a χ^2_ν without including $n_{\rm g}$ – see Eq.(21).

Furthermore, the values of the χ^2_ν are subject to uncertainties due to the finite number of realisations used to estimate the covariance matrix and the limited number of HOD realisations per halo catalogue. The most significant uncertainty, $\approx$ 27 per cent, arises from the limited number of HOD realisations. The remaining values are below 20 per cent, see Section B for more details. The χ^2_ν is simply used as a metric to evaluate the goodness-of-fit. For this reason it is important to consider that it is affected by a large uncertainty when comparing its magnitude to the expected value of one.

The primary focus of this paper is to investigate the limits of the FastPM capabilities to model the non-linear scales captured by N-body simulations. Furthermore, we study the effect of fitting to successively more non-linear scales and either Fourier or configuration space statistics on the FastPM covariance matrix.

4.1 Power spectrum fitting

Figure 3 shows the results of the HOD fitting performed on the power spectrum for three different k intervals, defined in Table 3. The second, third and fifth rows display the difference in the clustering scaled by the difference error. We remind the reader that this error is smaller than the expected one for the given volume, due to the matched initial conditions between the two simulations, see Figure 2.

The best fitting monopoles and quadrupoles are within $\pm1\sigma$ for most scales. Moreover, the results for the HR FastPM – presented with dashed line – are only marginally better than the ones for LR FastPM. Given the modest difference between the performances of the two resolutions, we believe that the LR FastPM is precise enough to describe the two-point clustering to non-linear scales for the DESI Y1 ELG-like galaxies.

Considering that we only fit the first two even multipoles, there is no guarantee that the third one would match the reference. Nevertheless, the fifth row of Figure 3 illustrates that fitting the monopole and quadrupole to smaller scales improves the agreement of the hexadecapole. For instance, fitting on the Large interval pushes the $\ell = 4$ multipole within $\pm2\sigma$ for $k < 0.4\,h/\rm Mpc$, whereas for Medium and Small intervals, the hexadecapole is placed within $\pm2\sigma$ only for $k < 0.3\,h/\rm Mpc$ or $k < 0.2\,h/\rm Mpc$, respectively.

[7] This would be the noise level in a hypothetical case where SLICS and FastPM would not share the initial conditions.
[8] The DESI Year 5 error is estimated by rescaling $\sigma_{20,\rm SLICS}$ to match the Y5 ELG sample volume, which is assumed to be $24\,\rm Gpc^3\,\mathit{h}^{-3}$.
[9] The DESI Year 1 error is estimated by rescaling $\sigma_{20,\rm SLICS}$ to match the Y1 ELG sample volume, which is assumed to be one third of the Y5 volume.
[10] A simple calculation reveals that one would need 192 SLICS realisations to meet the DESI Y5 precision requirements.

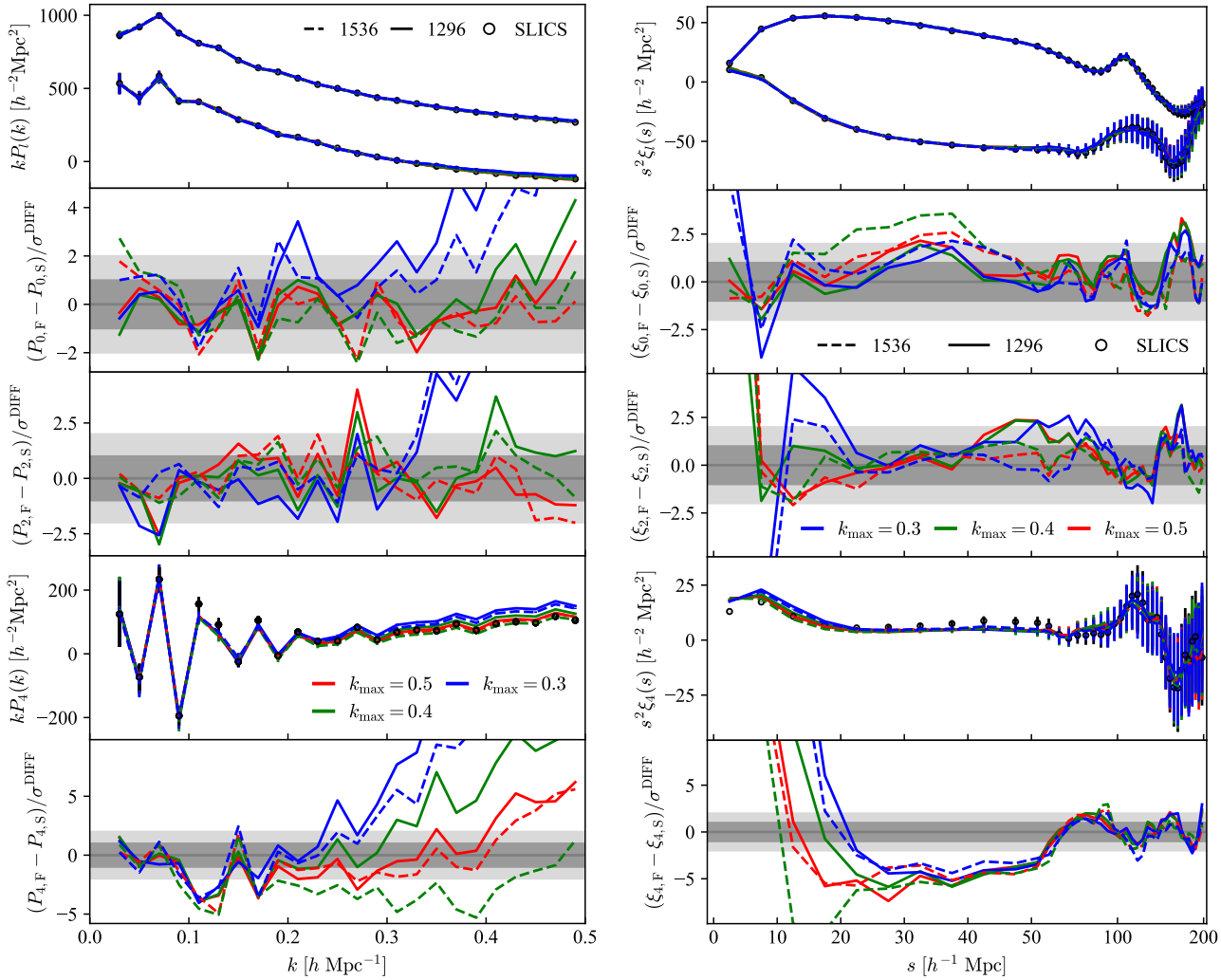


Figure 3. The average of 20 SLICS (reference-black) and 20 FastPM (model-colours) clustering realisations and the tension ($\sigma_{\rm DIFF}$ is shown in Figure 2) between them: left - power spectrum and right - 2PCF. FastPM mocks share the white-noise through the initial conditions with the SLICS ones. The fitting has been performed: 1) on the monopole and quadrupole of the power spectrum; 2) for three different fitting ranges, see Table 3; 3) using HR (dashed) and LR (continuous) FastPM realisations. The Ox axis of the 2PCF panels has a linear scale from 0 to 50 Mpc/h and a logarithmic scale above this limit.

Due to the fact that the power spectrum is affected by the window function, it is not obvious that a good matching in Fourier space translates as a good matching in Configuration space. Thus, we compute and display the corresponding 2PCF in the right-hand side of Figure 3. Most monopoles and quadrupoles agree within $\pm 2\sigma$ with SLICS for separations larger than 20 Mpc/h. This suggests that it is possible to obtain a reasonable 2PCF above a certain minimum separation, even when performing HOD fitting on the power spectrum. However, fitting on the Medium and Large intervals, the 2σ matching goes down to a separation of 10 Mpc/h.

In contrast, for separations smaller than 5 Mpc/h, the non-linear effects become dominant, making it difficult to replicate the velocity field. This is why increasing the fitting range up to $k_{\rm max} = 0.5$ can improve the monopole but not the quadrupole. Lastly, the 2PCF hexadecapole exhibits a bias of over 3σ for $s < 50$ Mpc/h in all six cases.

After a more qualitative description of the results, we present the χ^2_ν values in the upper panels of Figure 4. Generally, the HR FastPM produces lower χ^2_ν values than the LR, as expected from Figure 3. However, $\chi^2_\nu[P(0.02, k_{\rm max})] \simeq 1$, which reiterates that by fitting the monopole and quadrupole of the power spectrum up to the three $k_{\rm max}$ values, one can achieve a good match with the SLICS reference, within the DESI Y1 precision even with LR. In addition, $\chi^2_\nu[\xi(20, 50)] \simeq 2$ for the small fitting interval of the LR power spectrum, reinforcing the fact that one can get a reasonable 2PCF above a certain minimum separtion threshold when the fitting is performed on the power spectrum.

Additionally, we can observe the behaviour of χ^2_ν when it is estimated on different intervals than those used for the fitting. When the fitting is performed on the Large interval, the $\chi^2_\nu \simeq 1$ for all smaller intervals, regardless of the resolution. However, fitting on the Medium interval shows that the difference between HR and LR becomes more significant for $k > 0.4\,h$/Mpc (see also Figure 3): the $\chi^2_\nu \simeq 2$ for LR, while for HR, it is close to one. These findings imply that fitting up to $k \leq 0.4\,h$/Mpc is satisfactory for HR FastPM, whereas smaller scales play a more significant role in LR.

Furthermore, fitting on the Small interval shows that although $\chi^2_\nu[P(0.02, 0.3)] \simeq 1$, it is much larger for $k > 0.3\,h$/Mpc, indi-

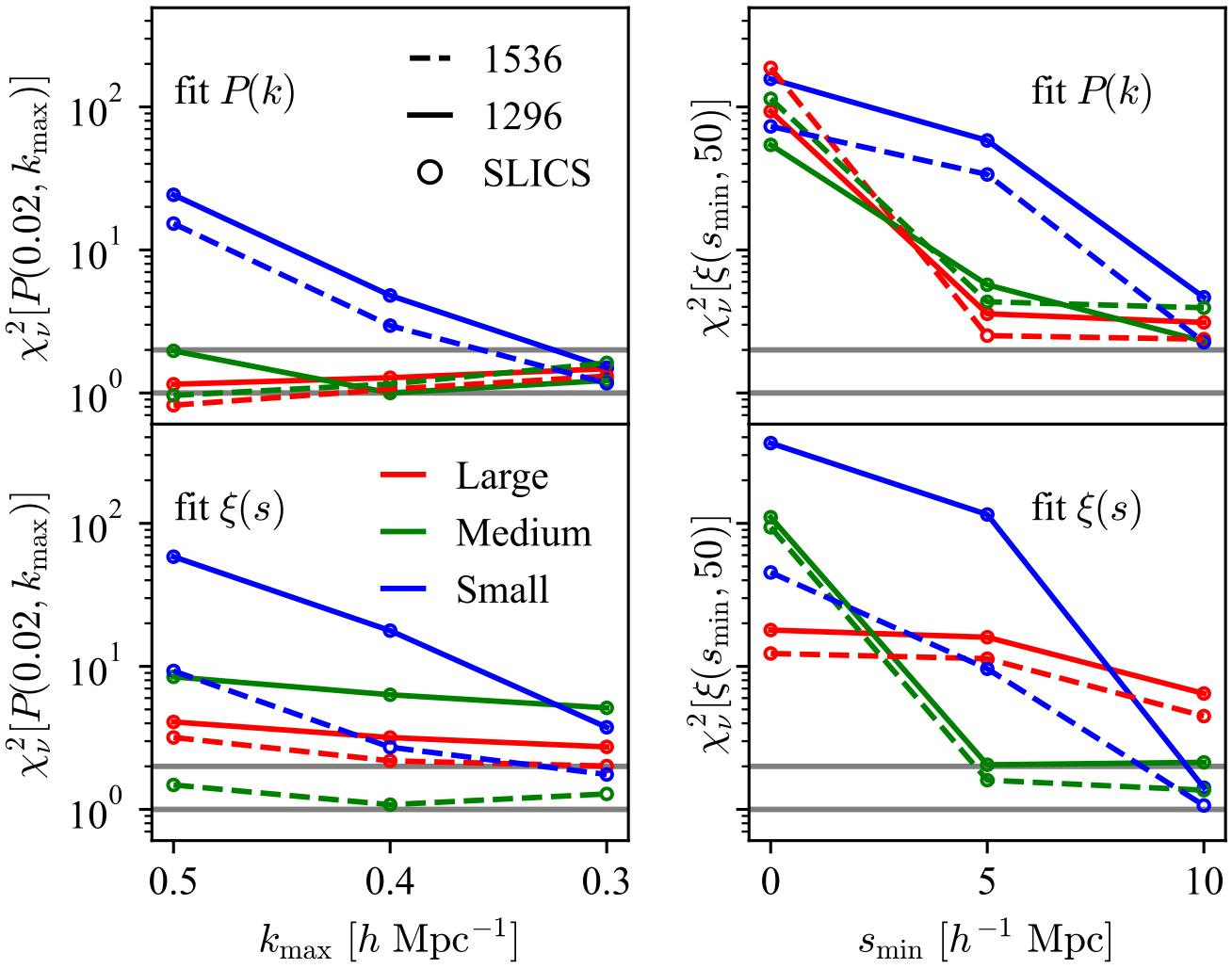


Figure 4. The χ^2_ν as defined in Section 3.3.3. We compute χ^2_ν: 1) for different intervals (see Oy and Ox axes) of the clustering statistics (left panels - power spectrum; right panels - 2PCF); 2) for different fitted clustering (upper panels - power spectrum, see Section 4.1; lower panels - 2PCF, see Section 4.2); 3) for different fitting ranges (see Table 3).

cating strong clustering divergence beyond that value (see Figure 3). Therefore, both LR and HR benefit from considering the clustering information contained in smaller scales $k > 0.3\,h$/Mpc.

4.2 2PCF fitting

When the HOD fitting is performed on the power spectrum, the minimum 2PCF χ^2_ν is $\chi^2_\nu[\xi(10, 50)] \approx 2$. While this translates to a 2σ agreement down to the separation of 10 Mpc$/h$ between FastPM and SLICS 2PCF, we test whether fitting directly the 2PCF can improve the results. Therefore, in this section, we analyse the outcomes of the HOD fitting performed on the 2PCF monopole and quadrupole, for $s \in [s_{\rm min}, 50]$ Mpc$/h$, see Table 3.

Figure 5 presents the monopole, quadrupole and hexadecapole of the 2PCF computed for $s \in [0, 200]$ Mpc$/h$ as well as the tensions between the FastPM and SLICS. The FastPM clustering typically falls within 2σ of the reference for scales larger than 50 Mpc$/h$ and is largely unaffected by the fitting scenario. However, the HR monopoles are consistently closer to the reference than LR monopoles by approximately 0.5σ at scales larger than ≈ 150 Mpc$/h$.

Including the smallest scales (Large interval) in the HOD fitting, we observe a 1 to 2σ agreement with the reference for $s < 10$ Mpc$/h$ in both the monopole and quadrupole. However, at intermediate scales $s \in [10, 50]$ Mpc$/h$, the monopole is significantly biased, exhibiting a deviation of 3σ. In contrast, for the Medium and Small scenarios, we notice that the tensions for the monopole and quadrupole at intermediate scales drop to 1σ, while the smallest scales can get biased by more than 3σ. Nevertheless, they match better the reference than the power spectrum HOD fitting case. Lastly, the hexadecapole does not depend on the resolution nor the fitting range and is strongly biased for $s < 60$ Mpc$/h$, showing no improvement compared to the power spectrum fitting.

As in the previous subsection, we test the clustering statistics of the best-fitting FastPM boxes that were not included in the HOD fitting, i.e. the power spectrum in the Figure 5. The first observation is that these FastPM power spectra do not fit as well the reference as the ones from Figure 3. On one hand, for the HR case and Medium and Small

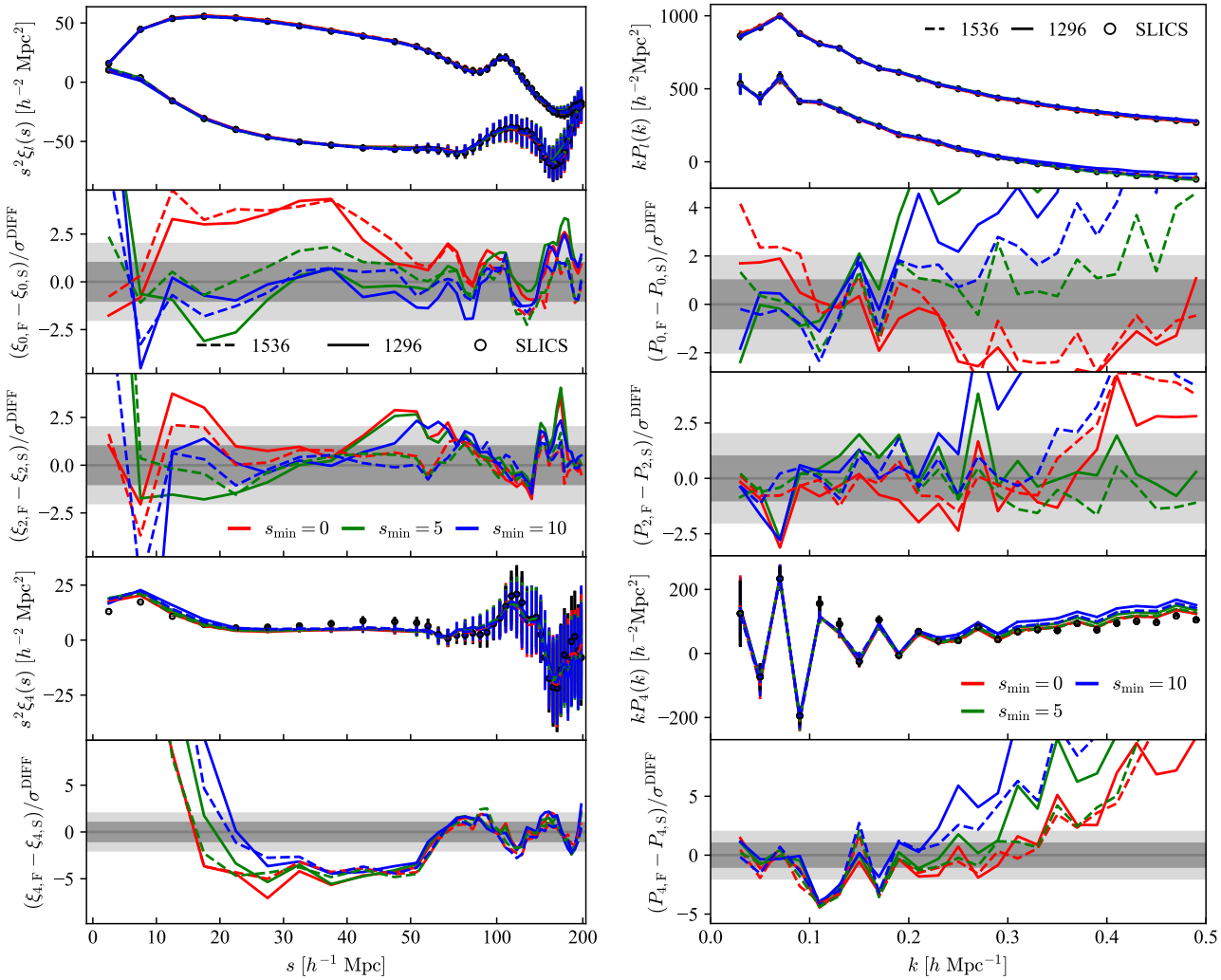


Figure 5. Same as Figure 3, but the fitting is done on the monopole and quadrupole of the 2PCF.

fitting intervals a $\pm1\sigma$ matching is possible up to $k = 0.4\,h/$Mpc and $k = 0.3\,h/$Mpc, respectively. On the other hand, the LR FastPM allows a good matching up to $k \approx 0.2\,h/$Mpc for the same fitting intervals. While the $s_{\rm min} = 0$ case has a good matching quadrupole up to $k \approx 0.4\,h/$Mpc, its monopole follows similar trend to the 2PCF monopole, i.e. the intermediate scales $k \in [0.25, 0.4]\,h/$Mpc are biased and the rest are mostly within 2σ deviation. Lastly, the hexadecapole is within $\pm2\sigma$ up to $k \approx 0.3\,h/$Mpc for the Large fitting interval and up to $k \approx 0.2\,h/$Mpc for the other cases.

A quantitative evidence that directly fitting the 2PCF yields superior matching of the 2PCF compared to fitting the power spectrum is displayed in Figure 4. The majority of the χ^2_ν values in the lower-right panel are lower compared to those in the upper-right panel. Furthermore, fitting on the Small interval ($s_{\rm min} = 10$), the $\chi^2_\nu \approx 1$, indicating that the 2PCF is in good agreement with the SLICS reference above a certain minimum separation. The almost constant χ^2_ν for the Large fitting interval in the lower-right panel of Figure 4 is explained by the discrepancy at the intermediate scales of the monopole for the Large fitting interval in Figure 5. Lastly, as in the previous fitting scenario, the HR FastPM generally provides a lower χ^2_ν than the LR. In contrast, only the HR simulations can provide a $\chi^2_\nu < 2$ to both the 2PCF and the power spectra, and only when fitting with the Medium and Small intervals to 2PCF. Although not shown in the aforementioned figure, it is important to note that $\chi^2_\nu[P(0.02, 0.2)] = 2.4$ for the 2PCF Small interval LR case.

4.3 Bi-spectrum comparison

Taking into account that the covariance matrix depends on the bi-spectrum (Baumgarten & Chuang 2018), we aim to understand its behaviour when incorporating various scales in the HOD fitting. Figure 6 compares the average bi-spectrum of the 20 best-fitting FastPM boxes with the one computed on the corresponding SLICS boxes. It is evident that by increasing the fitting range to include smaller scales, the FastPM bi-spectrum changes to the extent that for $k_{\rm max} = 0.5$, the tension ranges from 1 to 2σ. In contrast, when fitting the 2PCF the resulting bi-spectrum is more biased, i.e. the lowest deviation is $\approx 5\sigma$, for $s_{\rm min} = 0$ case. Finally, there is no significant improvement in terms of the goodness-of-fit between the HR and LR.

In the previous sections, we compare the HR and LR FastPM with SLICS using the two-point clustering of the 20 cubic mocks included in the HOD fitting. The HR simulations perform better than LR to model the extremely non-linear scales, such as $k \approx 0.5\,h/$Mpc, $s \approx$

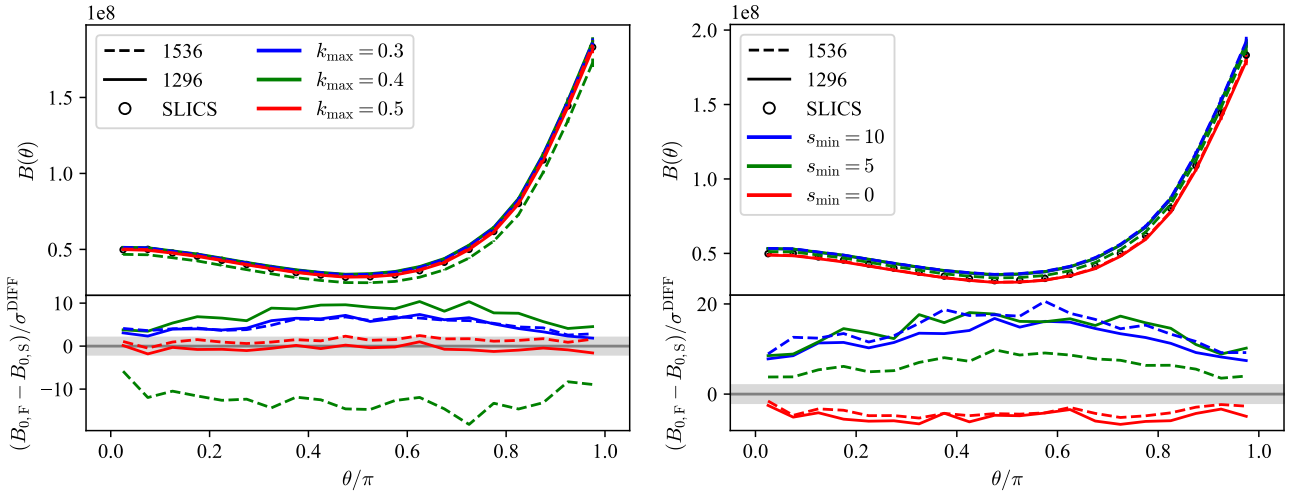


Figure 6. A comparison between the average SLICS bi-spectrum and the average FastPM bi-spectrum. The averages are computed from the 20 realisations used during the HOD fitting. The left panel shows the results from fitting the power spectrum, as in Figure 3. The right panel displays the results from fitting the 2PCF, as in Figure 5. The shaded area denotes $\pm 2\sigma$ deviation.

$0\,{\rm Mpc}/h$. In contrast, at mildly non-linear scales ($k \approx 0.3\,h/{\rm Mpc}$, $s \approx 10\,{\rm Mpc}/h$) that are more relevant to BAO and RSD analyses (e.g. Tamone et al. 2020; de Mattia et al. 2021), LR and HR show similar performance. Moreover, Figure 6 suggests that the bi-spectrum does not depend strongly on the resolution. Nevertheless, the computing cost of HR is significantly higher than for LR and given the small difference in precision, we argue it is optimal to use LR FastPM for further analyses.

Furthermore, in Figure 7, we compare the average bi-spectra – computed from 778 LR FastPM realisations – corresponding to the six HOD fitting cases, see Table 3. In this and the next figures, we choose the $k_{\rm max} = 0.5$ case as reference because:

(i) Figure 4 shows that the best-fitting power spectrum provides $\chi^2_\nu \approx 1$;
(ii) Figure 6 implies that the corresponding bi-spectrum is the closest to the SLICS reference.

One can notice that $s_{\rm min} = 0$ bi-spectrum is at most 5 per cent different than the reference, while the rest can reach 15 per cent discrepancies. The $k_{\rm max} = 0.3$ and $k_{\rm max} = 0.4$ cases are 1 to 2 per cent different from each other and similarly for $s_{\rm min} = 5$ and $s_{\rm min} = 10$.

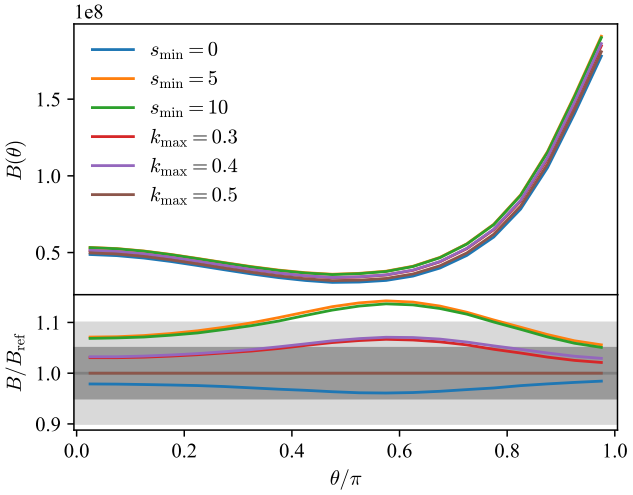


Figure 7. Average bi-spectra computed using 778 LR FastPM realisations for different HOD fitting cases, see Table 3. The reference for the bi-spectra ratios is the average bi-spectrum computed for the $k_{\rm max} = 0.5$ case.

4.4 Covariance comparison

Having studied the behaviour of the bi-spectra, we now want to understand their effect on the covariance matrices of the clustering (power spectrum and 2PCF).

4.4.1 Power spectrum covariance

Figure 8 presents the correlation matrices and the corresponding standard deviations σ_ℓ for the monopole and quadrupole of the power spectrum. The following pairs ($k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$), ($s_{\rm min} = 5$, $s_{\rm min} = 10$) and ($s_{\rm min} = 0$, $k_{\rm max} = 0.5$) have very similar correlation matrices, thus we only show three cases. However, we introduce all of them in Appendix C.

The similarity to the reference correlation matrix diminishes in the following order: $s_{\rm min} = 0$, $k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$, $s_{\rm min} = 5$ and $s_{\rm min} = 10$. However, for the largest scales of the quadrupole ($k < 0.15\,h/{\rm Mpc}$), the correlation coefficients are practically the same for all cases.

The standard deviations in the lowest panels show similar trends. The $s_{\rm min} = 0$ case is within two per cent of the reference case. The $k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$ cases overestimate the $\sigma_\ell(k)$ by approximately two per cent for $k < 0.27\,h/{\rm Mpc}$ and by ≈ 5 per cent for smaller scales. Nevertheless, these two cases are consistent with each other within one to two per cent. In contrast, $s_{\rm min} = 5$ and $s_{\rm min} = 10$ can overestimate the $\sigma_\ell(k)$ by ≈ 2 to 5 per cent for $k < 0.27\,h/{\rm Mpc}$ and by 10 to 20 per cent for smaller scales. These two cases are also consistent with each other for most scales, except for the quadrupole $k > 0.3\,h/{\rm Mpc}$. These findings are in agreement with the trends observed in the bi-spectrum comparison in Figure 7.

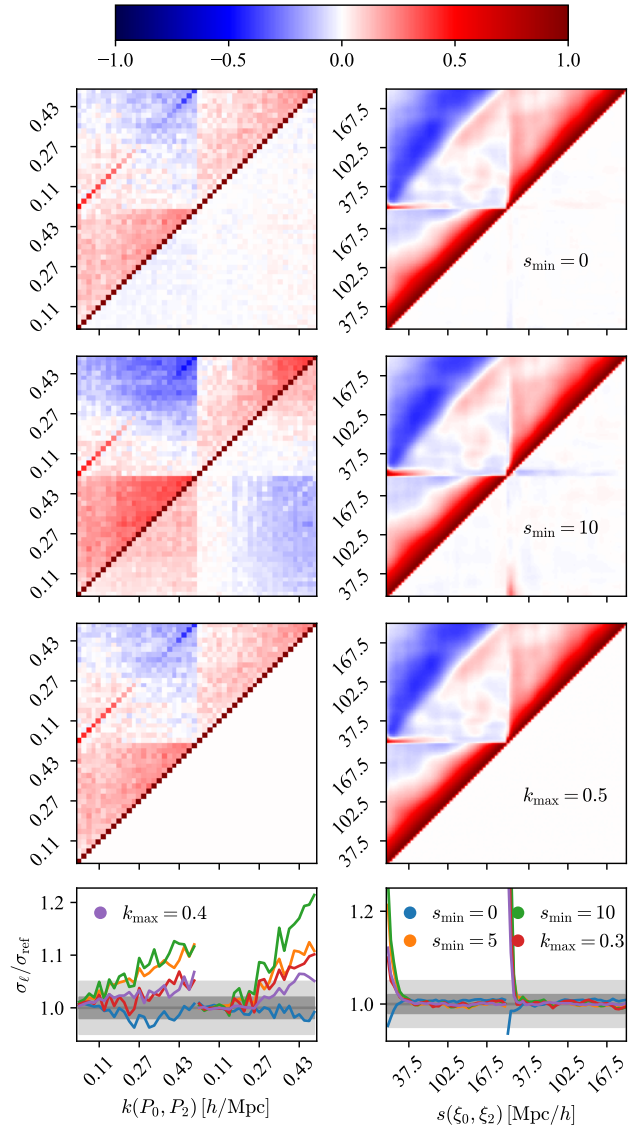


Figure 8. The correlation matrices and the standard deviations computed using the monopoles and quadrupoles of the power spectrum (left) and 2PCF (right). Given the similarity between some correlation matrices, only three out of six different HOD fitting cases – see Table 3 – are presented here. However, all cases can be found in Appendix C. The reference case corresponds to $k_{\rm max} = 0.5$. The upper triangular matrices display the correlation matrices, while the lower triangular ones show the differences between the correlation matrices and the reference one. The bottom panels illustrate the ratios between the standard deviations. The shaded regions denote two and five per cent limits.

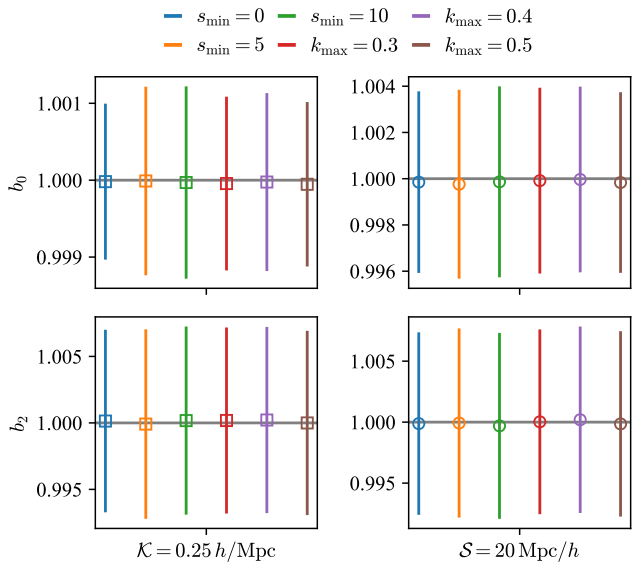


Figure 9. The average of the 123 fitting parameters (b_ℓ) obtained from 123 SLICS clustering realisations, as described in Section 3.4. The measurements performed on the power spectra with $k \in [0.02, \mathcal{K}]\,h$/Mpc are show on the left, while those based on the 2PCF with $s \in [\mathcal{S}, 200]\,$Mpc/$h$ are depicted on the right. The error bars are computed as the average of 123 σ_{b_ℓ}, divided by $\sqrt{123}$. Here, σ_{b_ℓ} represents the standard deviation of the b_ℓ posterior distribution. The colours correspond to the different FastPM covariance matrices illustrated in Figure 8. Due to the similar results, only one value for $\mathcal{K}$ and $\mathcal{S}$ are shown here. However, all tested values are presented in Appendix C

In order to quantify the differences between the covariance matrices we adopt the method described in Section 3.4 and thus obtain the results displayed in Figures 9 and 10. The first one reveals that none of the six covariance matrices bias the two fitting parameters, regardless of the fitting range. We only present here the results of one fitting range, however all cases can be found in Appendix C.

Examining the uncertainty on b_0 in Figure 10, we observe that including the smaller scales the discrepancy between the error estimates of the six covariance matrices increases, as we expect from Figure 8, reaching a maximum of $\approx$ 20 per cent larger error estimation for the $s_{\rm min} = 10$ covariance at $\mathcal{K} = 0.25\,h$/Mpc. Moreover, each of the following pairs ($k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$), ($s_{\rm min} = 5$, $s_{\rm min} = 10$) and ($s_{\rm min} = 0$, $k_{\rm max} = 0.5$) provide coherent estimations of the uncertainty, which is consistent with the observations on the correlation matrices and standard deviations. Lastly, a five per cent consensus between all six covariance matrices is achieved when we fit the power spectra on the $k \in [0.02, 0.1]\,h$/Mpc.

The agreement between covariance matrices on σ_2 is much better than σ_0. Given the error bars, the six methods estimate the uncertainty with a two per cent tolerance with each other for all $\mathcal{K}$ values.

Finally, all six covariance matrices provide values of $\mathcal{R}[b_0, b_2]$ that are consistent at the level of 5 per cent, given the error bars and up to $\mathcal{K} = 0.2\,h$/Mpc. For $\mathcal{K} = 0.25\,h$/Mpc, the largest discrepancy is shown by $s_{\rm min} = 10$ case which underestimates the value of $\mathcal{R}[b_0, b_2]$ by almost 50 per cent. The other cases underestimate $\mathcal{R}[b_0, b_2]$ by 10 to 20 per cent.

4.4.2 2PCF covariance

Comparing the correlation matrices obtained from 778 2PCF in Figure 8, one can observe that the largest differences occur at the smallest scales $s < 30\,$Mpc/h. Similarly to the power spectrum correlation matrices, the same pairs of cases show resembling behaviours at all scales. Equivalent qualitative comments can be made about the ratios of the standard deviations. Nonetheless, all cases are within $\approx$ 2 per cent from each other for $s > 30\,$Mpc/h, while at smaller scales, the differences can get larger than $\approx$ 20 per cent.

Following the method described in Section 3.4, we obtain the results shown in Figures 9 and 11. The first figure proves that all six

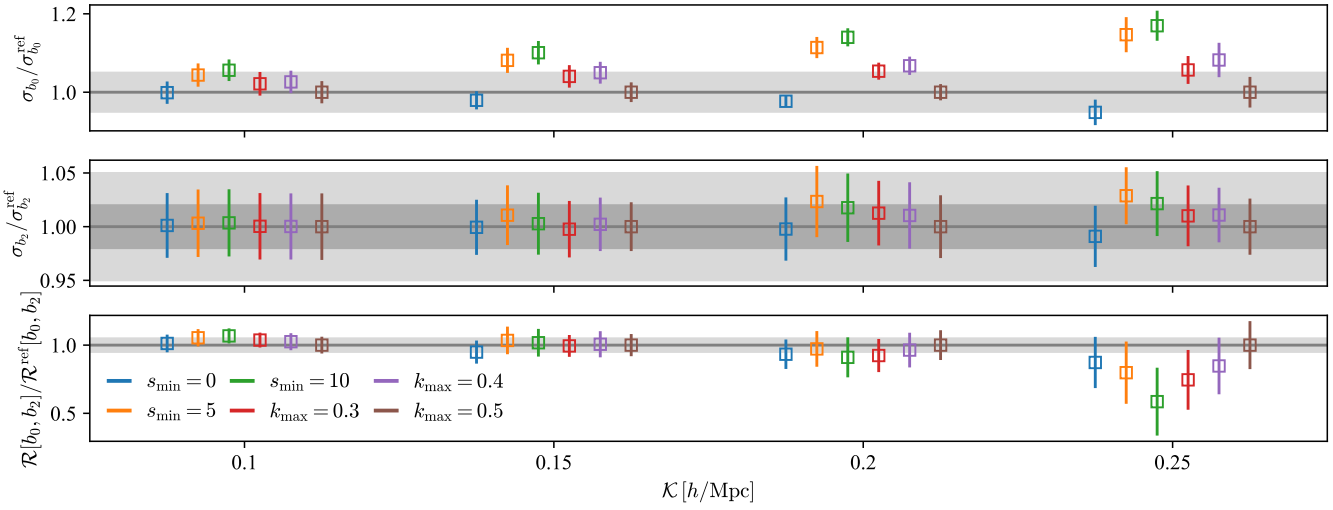


Figure 10. The averages of 123 σ_{b_ℓ} and 123 $\mathcal{R}[b_0, b_2]$ – the standard deviation of the b_ℓ posterior distribution and the covariance between b_0 and b_2, respectively, detailed in Section 3.4 – obtained from 123 SLICS power spectra fitted on $k \in [0.02, \mathcal{K}]\,h/{\rm Mpc}$. In order to estimate the error bars, we split the 778 FastPM realisations in six distinct sets of 123 realisations and compute for each set u a covariance matrix $\Sigma^u_{123,{\rm FastPM}}$ with which we fit the 123 SLICS clustering realisations. Having obtained 123 values of $\sigma^u_{b_\ell}$ per set, we compute their average $\bar{\sigma}^u$. Finally, the error bars are the standard deviation of the six $\bar{\sigma}^u$ divided by $\sqrt{6}$. The different colours stand for the different FastPM covariance matrices exhibited in Figure 8. The $k_{\rm max} = 0.5$ represents the reference, i.e. all values (σ_{b_ℓ} and its error bars) are scaled by the σ_{b_ℓ} corresponding to $k_{\rm max} = 0.5$ case. This is why all brown squares have the value of one. The shaded areas delineate the 2 and 5 per cent regions with respect to the reference.

covariance matrices provide unbiased measurements of b_ℓ parameters.

Resembling the power spectrum fitting case, the six estimations of σ_{b_0} in Figure 11 are in better agreement when the smallest scales are not included in the 2PCF fitting, however for $\mathcal{S} \geq 20\,{\rm Mpc}/h$ they are all within ≈ 5 per cent from each other. The largest discrepancy is around 10 per cent and occurs between $s_{\rm min} = 0$ and $s_{\rm min} = 10$ for $\mathcal{S} = 15\,{\rm Mpc}/h$. The values of σ_{b_2} are all consistent within ≈ 2 per cent, given the error bars. Interestingly, including the smaller scales, the $\mathcal{R}[b_0, b_2]$ values are more coherent, such that all discrepancies are within five per cent, given the error bars and for $\mathcal{S} < 30\,{\rm Mpc}/h$. In contrast, when $\mathcal{S} = 30\,{\rm Mpc}/h$ the $s_{\rm min} = 10$ and $s_{\rm min} = 5$ provide values $\mathcal{R}[b_0, b_2]$ that are approximately ten per cent larger than the reference, but nevertheless consistent within the error bars.

5 CONCLUSIONS

We have implemented an HOD model to assign galaxies on the FastPM halo cubic mocks, such that the resulting clustering – monopole and quadrupole – matches the SLICS reference one. In order to remove the cosmic variance, we have used 20 SLICS galaxy catalogues and 20 halo FastPM mocks (low resolution or high resolution) that share the initial conditions with the SLICS simulations. Given the shared white noise, the standard covariance matrix is obsolete, thus we have performed a two-steps HOD fitting:

(i) use a simple diagonal covariance matrix to get Initial-Guess best-fitting FastPM galaxy mocks;
(ii) compute the covariance matrix of the 123 realisations of the difference between the IG-FastPM and the SLICS clustering, and use it to perform the final HOD fitting.

The final HOD fitting has been performed on three different fitting ranges for both power spectrum ($k_{\rm max} = 0.5$, $k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$) and 2PCF ($s_{\rm min} = 0$, $s_{\rm min} = 5$, $s_{\rm min} = 10$).

On one hand, the HR FastPM generally performs better than the LR at modelling the SLICS clustering. On the other hand, LR is also able to provide a $\chi^2_\nu \approx 1$ for $k_{\rm max} = 0.5$, $k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$ and $s_{\rm min} = 10$. The $k_{\rm max} = 0.5$ case is one of the most valuable as it additionally offers 2σ matching:

(i) power spectrum hexadecapole for $k < 0.4\,h/{\rm Mpc}$;
(ii) 2PCF monopole and quadrupole for $s > 10\,{\rm Mpc}/h$;
(iii) bi-spectrum.

Nevertheless, fitting the 2PCF with $s_{\rm min} = 10$, produce a 1σ matching power spectrum monopole and quadrupole for $k \lesssim 0.2$, but a strongly biased bi-spectrum. In a similar way as the power spectrum, one must include the smallest scales to better reproduce the SLICS bi-spectrum, i.e. for $s_{\rm min} = 0$ the bi-spectrum tension drops from 20σ to 5σ. As a general remark, the power spectrum hexadecapoles can be slightly tuned by changing the values of $k_{\rm max}$ or $s_{\rm min}$, but the 2PCF hexadecapole is practically independent on the fitting range.

Finally, it could be interesting for future studies to perform a join fitting of both Fourier and Configuration clustering statistics to test for possible improvements in modelling non-linear scales.

In the second part of the study, we have focused on the 778 LR FastPM realisations corresponding to the six best-fitting cases, where $k_{\rm max} = 0.5$ is considered the reference. We have compared the resulting covariance matrices together with the differences in their constraining power using a simplified clustering model with two scaling parameters, i.e. b_0 and b_2 for the monopole and quadrupole. We focused on fitting intervals similar to the ones used in standard BAO and RSD analyses i.e. $\mathcal{K} \lesssim 0.20\,h/{\rm Mpc}$ and $\mathcal{S} \gtrsim 20\,{\rm Mpc}/h$(e.g. Tamone et al. 2020; de Mattia et al. 2021). In addition, we have analysed the bi-spectra from the point of view of the impact they have on the covariance matrices.

The $s_{\rm min} = 0$ bi-spectrum is at most five per cent different than

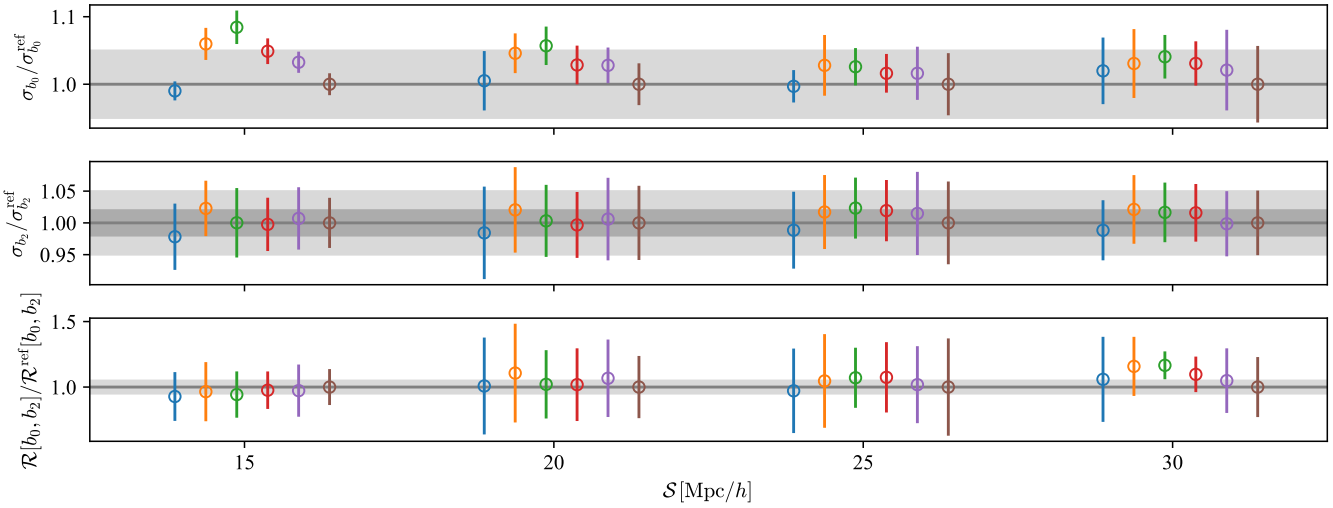


Figure 11. Same as Figure 10, but the fitting is performed on 123 SLICS 2PCF and $s \in [\mathcal{S}, 200]\,{\rm Mpc}/h$.

the $k_{\rm max} = 0.5$, while the other cases can reach a discrepancy of 15 per cent. However, each of these pairs ($k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$), ($s_{\rm min} = 5$, $s_{\rm min} = 10$) yield similar bi-spectra. These observations are in a good agreement with a qualitative description of the shown correlation matrices and the standard deviations.

Quantitatively, the power spectrum standard deviations of $s_{\rm min} = 0$ and ($k_{\rm max} = 0.4$, $k_{\rm max} = 0.3$) are within two percent from the reference for $k < 0.5\,h/{\rm Mpc}$ and $k < 0.27\,h/{\rm Mpc}$, respectively. Furthermore, the 2PCF standard deviations of all cases are within two percent from the reference for $s > 30\,{\rm Mpc}/h$.

Using the simplified clustering model, b_0 and b_2 are measured accurately for both power spectrum and 2PCF using all six covariance matrices. The six estimations of σ_{b_0} from the power spectrum fitting up to $\mathcal{K} = 0.20\,h/{\rm Mpc}$ are scattered within at most 20 per cent from the reference, whereas the values of σ_{b_2} are within two per cent agreement, given the error bars. Lastly, the covariances between b_0 and b_2 are scattered within 5 per cent from the reference.

In contrast, the estimations of σ_{b_0} from the 2PCF fitting down to $\mathcal{S} = 20\,{\rm Mpc}/h$ are found within five per cent from each other. Similarly to the power spectrum case, the σ_{b_2} values agree at the level of two per cent. Given the error bars, the covariances between b_0 and b_2 are consistent at the level of five per cent.

In conclusion, one can use an HOD model on the low resolution FastPM halo catalogues to tune the galaxy clustering such that it matches the SLICS reference down to certain minimum scales. Additionally, the HOD fitting intervals can have an impact on the final FastPM based covariances. This influence is observed as a scatter in the uncertainty estimation of up to 20 per cent for power spectrum and five per cent for 2PCF at the scales interesting for BAO and RSD analyses. Nevertheless, more accurate analyses could be performed in the future using actual BAO and RSD models and larger mocks, such as AbacusSummit.

ACKNOWLEDGEMENTS

AV, CZ, DFS acknowledge support from the Swiss National Science Foundation (SNF) "Cosmology with 3D Maps of the Universe" research grant, 200020_175751 and 200020_207379. SA acknowledge support of the Department of Atomic Energy, Government of india, under project no. 12-R&D-TFR-5.02-0200. SA is partially supported by the European Research Council through the COSFORM Research Grant (#670193) and STFC consolidated grant no. RA5496. The authors are thankful for the constructive comments of Joseph DeRose that have helped improved the clarity and quality of our article.

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico (CONACYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: https://www.desi.lbl.gov/collaborating-institutions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies.

The authors are honored to be permitted to conduct scientific research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

DATA AVAILABILITY

The SLICS and FastPM boxes used in this study can be provided upon reasonable request to authors. The clustering measurements can be found on Zenodo https://doi.org/10.5281/zenodo.8185822.

REFERENCES

Alam S., et al., 2017, MNRAS, 470, 2617
Alam S., Peacock J. A., Kraljic K., Ross A. J., Comparat J., 2020, MNRAS, 497, 581
Alam S., et al., 2021a, Phys. Rev. D, 103, 083533
Alam S., et al., 2021b, MNRAS, 504, 4667
Alam S., Paranjape A., Peacock J. A., 2023, arXiv e-prints, p. arXiv:2305.01266
Alexander D. M., et al., 2023, AJ, 165, 124
Allende Prieto C., et al., 2020, Research Notes of the American Astronomical Society, 4, 188
Bailey et al. 2023, in preparation
Balaguera-Antolínez A., Kitaura F.-S., Pellejero-Ibáñez M., Zhao C., Abel T., 2019, MNRAS, 483, L58
Balaguera-Antolínez A., et al., 2020, MNRAS, 491, 2565
Balaguera-Antolínez A., et al., 2022, arXiv e-prints, p. arXiv:2211.10640
Baumgarten F., Chuang C.-H., 2018, MNRAS, 480, 2535
Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, MNRAS, 311, 793
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Brodzeller A., et al., 2023, arXiv e-prints, p. arXiv:2305.10426
Brouwer M. M., et al., 2018, MNRAS, 481, 5189
Buchner J., et al., 2014, A&A, 564, A125
Chaussidon E., et al., 2023, ApJ, 944, 107
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621
Chuang C.-H., et al., 2019, Mon. Not. Roy. Astron. Soc., 487, 48
Chuang et al. 2023, in preparation
Cooper A. P., et al., 2022, arXiv e-prints, p. arXiv:2208.08514
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
DESI Collaboration et al., 2016a, arXiv e-prints, p. arXiv:1611.00036
DESI Collaboration et al., 2016b, arXiv e-prints, p. arXiv:1611.00037
DESI Collaboration et al., 2022, AJ, 164, 207
DESI Collaboration et al., 2023a, arXiv e-prints, p. arXiv:2306.06307
DESI Collaboration et al., 2023b, arXiv e-prints, p. arXiv:2306.06308
Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827
Dey A., et al., 2019, AJ, 157, 168
Dubois Y., et al., 2014, MNRAS, 444, 1453
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, The Open Journal of Astrophysics, 2, 10
Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264
Guo Q., et al., 2011, MNRAS, 413, 101
Guy J., et al., 2023, AJ, 165, 144
Hahn C., et al., 2022, arXiv e-prints, p. arXiv:2208.08512
Hand N., Feng Y., Beutler F., Li Y., Modi C., Seljak U., Slepian Z., 2018, AJ, 156, 160
Harnois-Déraps J., van Waerbeke L., 2015, MNRAS, 450, 2857
Harnois-Déraps J., Pen U.-L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, MNRAS, 436, 540
Harnois-Déraps J., et al., 2018, MNRAS, 481, 1337
Harnois-Déraps J., Hernandez-Aguayo C., Cuesta-Lazaro C., Arnold C., Li B., Davies C. T., Cai Y.-C., 2022, arXiv e-prints, p. arXiv:2211.05779
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hearin A. P., et al., 2017, The Astronomical Journal, 154, 190
Hildebrandt H., et al., 2017, MNRAS, 465, 1454
Joudaki S., et al., 2017, MNRAS, 465, 2033
Kirkby et al. 2023, in preparation
Kitaura F.-S., Yepes G., Prada F., 2013, MNRAS: Letters, 439, L21
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B. o., Primack J. R., 2004, ApJ, 609, 35
Lan T.-W., et al., 2023, ApJ, 943, 68
Levi M., et al., 2013, arXiv e-prints, p. arXiv:1308.0847
Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, Mon. Not. Roy. Astron. Soc., 508, 4017
Martinet N., et al., 2018, MNRAS, 474, 712
McCarthy I. G., Schaye J., Bird S., Le Brun A. M. C., 2017, MNRAS, 465, 2936
Miller T. N., et al., 2023, arXiv e-prints, p. arXiv:2306.06310
Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution. Cambridge University Press
Myers A. D., et al., 2023, AJ, 165, 50
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19
Pellejero-Ibañez M., et al., 2020, MNRAS, 493, 586
Percival W. J., Friedrich O., Sellentin E., Heavens A., 2022, MNRAS, 510, 3207
Pillepich A., et al., 2018, MNRAS, 473, 4077
Raichoor A., et al., 2020, Research Notes of the American Astronomical Society, 4, 180
Raichoor et al. 2023a, in preparation
Raichoor A., et al., 2023b, AJ, 165, 126
Ruiz-Macias O., et al., 2020, Research Notes of the American Astronomical Society, 4, 187
Schaye J., et al., 2010, MNRAS, 402, 1536
Schaye J., et al., 2015, MNRAS, 446, 521
Schlafly E. F., et al., 2023, arXiv e-prints, p. arXiv:2306.06309
Schlegel et al. 2023, in preparation
Sefusatti E., Crocce M., Scoccimarro R., Couchman H. M. P., 2016, MNRAS, 460, 3624
Seljak U., 2000, MNRAS, 318, 203
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Silber J. H., et al., 2023, AJ, 165, 9
Tamone A., et al., 2020, MNRAS, 499, 5527
Tasitsiomi A., Kravtsov A. V., Wechsler R. H., Primack J. R., 2004, ApJ, 614, 533
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
Wadekar D., Scoccimarro R., 2020, Phys. Rev. D, 102, 123517
Wadekar D., Ivanov M. M., Scoccimarro R., 2020, Phys. Rev. D, 102, 123521
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
White M., Hernquist L., Springel V., 2001, ApJ, 550, L129
Xu X., Cuesta A. J., Padmanabhan N., Eisenstein D. J., McBride C. K., 2013, MNRAS, 431, 2834
Yèche C., et al., 2020, Research Notes of the American Astronomical Society, 4, 179
Zarrouk P., et al., 2021, Mon. Not. Roy. Astron. Soc., 503, 2562
Zel'dovich Y. B., 1970, A&A, 5, 84
Zhang et al. 2023, in preparation
Zhao C., 2023, arXiv e-prints, p. arXiv:2301.12557
Zhao C., et al., 2021, MNRAS, 503, 1149
Zheng Z., et al., 2005, The Astrophysical Journal, 633, 791
Zhou R., et al., 2020, Research Notes of the American Astronomical Society, 4, 181
Zhou R., et al., 2023, AJ, 165, 58
Zou H., et al., 2017, PASP, 129, 064101
de Mattia A., et al., 2021, MNRAS, 501, 5616
van Uitert E., et al., 2018, MNRAS, 476, 4662

AFFILIATIONS

[1]Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland
[2]Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India
[3]Institute for Astronomy, University of Edinburgh, Royal Observatory,

Blackford Hill, Edinburgh EH9 3HJ, UK
[4]Department of Physics and Astronomy, The University of Utah, 115 South 1400 East, Salt Lake City, UT 84112, USA
[5]Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA
[6]Department of Astronomy, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China
[7]Key Laboratory for Particle Astrophysics and Cosmology(MOE)/Shanghai Key Laboratory for Particle Physics and Cosmology, China
[8]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[9]Physics Dept., Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA
[10]Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK
[11]Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK
[12]Instituto de Física, Universidad Nacional Autónoma de México, Cd. de México C.P. 04510, México
[13]Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, CP 111711, Bogotá, Colombia
[14]Observatorio Astronómico, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio H, CP 111711 Bogotá, Colombia
[15]Center for Cosmology and AstroParticle Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210, USA
[16]Department of Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210, USA
[17]The Ohio State University, Columbus, 43210 OH, USA
[18]Departament de Física, Serra Húnter, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
[19]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra Barcelona, Spain
[20]Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys, 23, 08010 Barcelona, Spain
[21]National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Rd., Chaoyang District, Beijing, 100012, P.R. China
[22]Department of Physics and Astronomy, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
[23]Perimeter Institute for Theoretical Physics, 31 Caroline St. North, Waterloo, ON N2L 2Y5, Canada
[24]Waterloo Centre for Astrophysics, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
[25]Space Sciences Laboratory, University of California, Berkeley, 7 Gauss Way, Berkeley, CA 94720, USA
[26]University of California, Berkeley, 110 Sproul Hall #5800 Berkeley, CA 94720, USA
[27]Department of Physics, Kansas State University, 116 Cardwell Hall, Manhattan, KS 66506, USA
[28]Department of Physics and Astronomy, Sejong University, Seoul, 143-747, Korea
[29]CIEMAT, Avenida Complutense 40, E-28040 Madrid, Spain
[30]Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
[31]University of Michigan, Ann Arbor, MI 48109, USA
[32]Department of Physics & Astronomy, Ohio University, Athens, OH 45701, USA

APPENDIX A: CONVERGENCE TESTS

Given the two-step approach to fit the HOD model, there is the important question of convergence that has to be answered. Thus, we study whether the $\Sigma_{\rm diff}$ estimates well the noise in our HOD fitting process.

Figure A1 illustrates that the magnitude of the errors estimated after the second HOD fitting step for all six hod fitting scenarios are consistent between themselves within at most 10 to 20 per cent. However, there seems to be a slight divergence between the error estimations when one approaches the lowest scales.

Finally, compared to the square root of the diagonal of $\Sigma_{\rm diff}$, the six standard deviations for both 2PCF and power spectrum are found within at most 20 per cent deviation. In order to quantify these discrepancies, we have built the difference covariance matrix using of each of the six best-fitting FastPM (of the second HOD fitting step, see Table 3). Furthermore, we have computed the χ^2_ν for the same best-fitting FastPM of the second HOD fitting step, but using these six new difference covariance matrices. The results are summarised in Figure A2.

Even though for most of the fitting cases, the six new difference covariance matrices seem to provide coherent biased χ^2_ν values compared to the $\Sigma_{\rm diff}$, the biases do not share the same sign between the fitting cases. In addition, most χ^2_ν values are within the error bars shown in Table B1 with respect to the reference. This suggests that given the error bars, the hypothetical best-fittings obtained using the six new covariance matrices, would be indistinguishable from the best-fitting FastPM of the second HOD step. Consequently, we argue that the $\Sigma_{\rm diff}$ is a good approximation of the noise in the difference of the (FastPM, SLICS) clustering, thus a hypothetical third step HOD would not drastically change the best-fitting FastPM compared to the ones after the second step.

APPENDIX B: UNCERTAINTY OF THE GOODNESS-OF-FIT

In this section, we are studying the uncertainty introduced by the covariance matrix and the finite number of HOD realisations per FastPM halo catalogue in the values of χ^2_ν, as defined in Eq. (21). The results are summarised in Table B1.

B1 Covariance matrix induced uncertainty

Due to the fact that we have only $N^{\rm cov}_{\rm mocks} = 123$ SLICS and FastPM realisations that share the same initial conditions, we are bound to use the JackKnife (JK) method to estimate the uncertainty introduced by the covariance matrix. Additionally, the HOD fitting is computationally expensive ($\approx$ 6000 CPU-hours), thus we are not able to perform hundreds of HOD fittings with different covariance matrices. Consequently, after obtaining one set of best-fitting HOD parameters, we computed the χ^2_ν with the same best-fitting FastPM clustering, but with $N^{\rm cov}_{\rm mocks}$ different covariance matrices.

The covariance matrices – Σ^i_χ, with i from 1 to $N^{\rm cov}_{\rm mocks}$ – are estimated using Eq. (22), but with only $N^{\rm cov}_{\rm mocks} - 1$ clustering realisations. Furthermore, we compute $\chi^{2,i}_{\nu,\rm JK}$ for each Σ^i_χ, as defined in Eq. (21) and we calculate the mean $\bar{\chi}^2_{\nu,\rm JK}$ and the variance $\sigma^2_{\chi,\rm JK}$:

$$\bar{\chi}^2_{\nu,\rm JK} = \frac{1}{N^{\rm cov}_{\rm mocks}} \sum_{i=1}^{N^{\rm cov}_{\rm mocks}} \chi^{2,i}_{\nu,\rm JK}, \tag{B1}$$

$$\sigma^2_{\chi,\rm JK} = \left[\frac{N^{\rm cov}_{\rm mocks} - 1}{N^{\rm cov}_{\rm mocks}} \sum_{i=1}^{N^{\rm cov}_{\rm mocks}} \left(\chi^{2,i}_{\nu,\rm JK} - \bar{\chi}^2_{\nu,\rm JK} \right)^2 \right]. \tag{B2}$$

B2 HOD induced uncertainty

During the HOD fitting, for each FastPM halo catalogue we create a single galaxy realisation, in order to reduce the optimisation time. As a consequence, we introduce additional noise in the HOD fitting process, that is not considered in the covariance matrix.

With the aim of estimating the effect of this noise on the χ^2_ν, we compute 100 galaxy realisations for a given set of best-fitting HOD parameters and per FastPM halo catalogue. Furthermore, using the 20 galaxy realisations corresponding to the halo catalogues used in the HOD fitting process and the same covariance matrix, we compute $\chi^{2,i}_{\nu,\rm HOD}$ as in Eq. (21), where i =

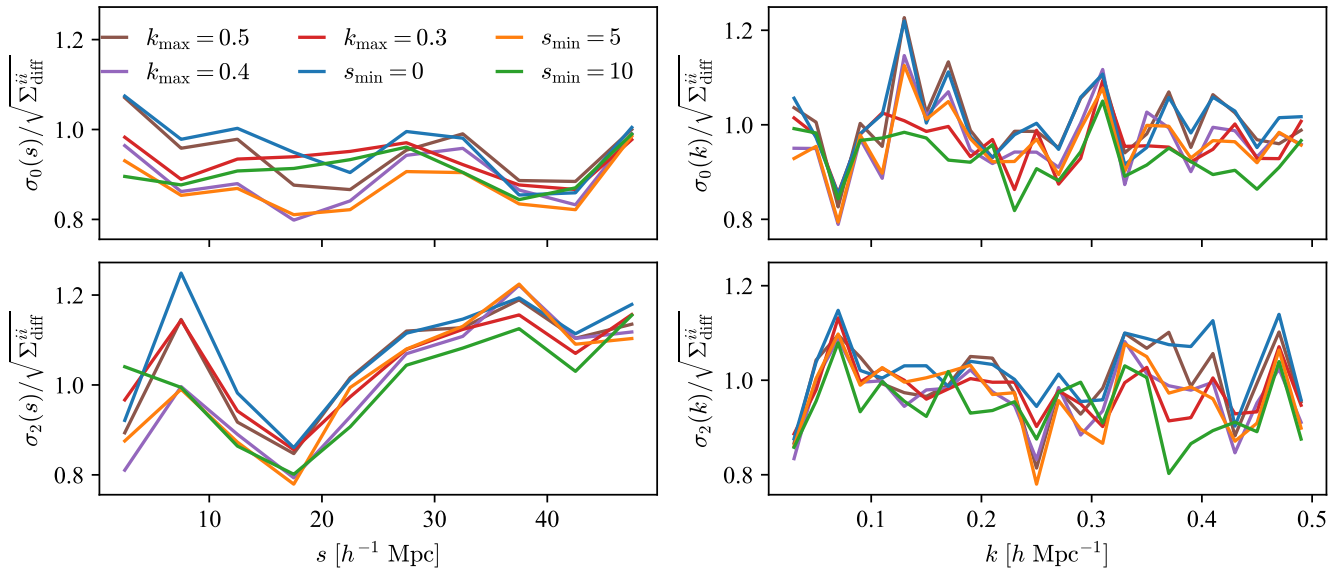


Figure A1. The ratios between the standard deviations computed on the differences of $N^{\rm cov}_{\rm mocks} = 123$ (LR FastPM, SLICS) clustering pairs and the square root of the diagonal of $\Sigma_{\rm diff}$, i.e. $\Sigma^{ii}_{\rm diff}$. The colours denote the HOD fitting scenarios in the second HOD fitting step, see Table 3. While the left panels include monopole and quadrupole of the 2PCF, the right ones display the power spectrum.

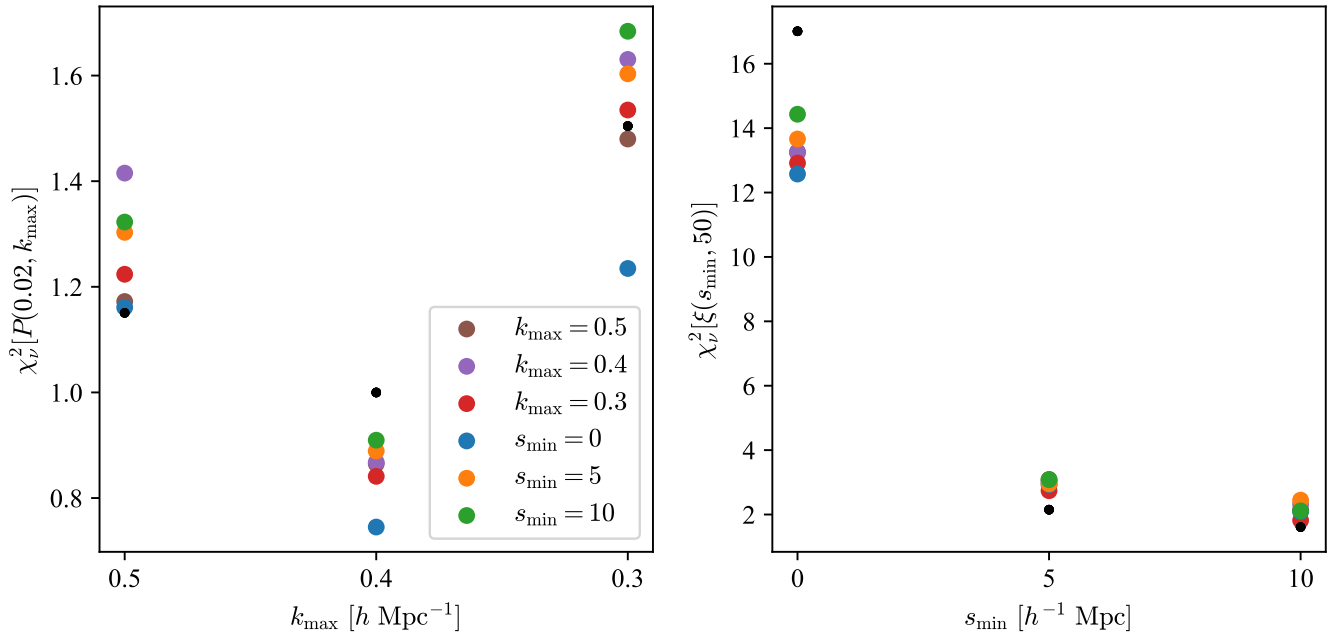


Figure A2. The χ^2_ν as defined in Section 3.3.3, but using different covariance matrices. We compute χ^2_ν: 1) for the six best-fitting FastPM cases, three cases for the power spectrum in the left panel (see Section 4.1), and three cases for the 2PCF in the right panel (see Section 4.2); 2) with the six difference covariance matrices obtained after the second HOD fitting step (the coloured dots). The black points show the best fitting χ^2_ν for the six cases that also appear in Figure 4.

1, ..., 100. Finally, we calculate the mean and the standard deviation of the 100 $\chi^{2,i}_{\nu,\rm HOD}$ values:

$$\bar{\chi}^2_{\nu,\rm HOD} = \frac{1}{100}\sum_{i=1}^{100}\chi^{2,i}_{\nu,\rm HOD}, \tag{B3}$$

$$\sigma^2_{\chi,\rm HOD} = \left[\frac{1}{100-1}\sum_{i=1}^{100}\left(\chi^{2,i}_{\nu,\rm HOD} - \bar{\chi}^2_{\nu,\rm HOD}\right)^2\right]. \tag{B4}$$

	$P(k)$	$\xi(s)$
Large		
χ^2_ν from Figure 4	1.15	16.94
$\bar{\chi}^2_{\nu,\mathrm{HOD}} \pm \sigma_{\chi,\mathrm{HOD}}$	1.38 ± 0.26	17.03 ± 1.65
$\bar{\chi}^2_{\nu,\mathrm{JK}} \pm \sigma_{\chi,\mathrm{JK}}$	1.16 ± 0.31	16.97 ± 2.68
Medium		
	1.00	2.16
	1.13 ± 0.23	2.19 ± 0.42
	1.00 ± 0.19	2.16 ± 0.32
Small		
	1.50	1.59
	1.50 ± 0.29	1.68 ± 0.41
	1.51 ± 0.25	1.59 ± 0.26

Table B1. The values of the χ^2_ν and their uncertainties introduced by the covariance matrix (Eqs. (B1) and (B2)) and the finite number of HOD realisations per FastPM halo catalogue (Eqs. (B3) and (B4)). The estimations have been performed on the LR (1296^3) FastPM galaxy catalogues and on both the 2PCF and the power spectrum for the three specific fitting ranges defined in Table 3.

APPENDIX C: COVARIANCE MATRIX COMPARISON

Analysing Figures C1 and C2 one can observe that the ($s_{\mathrm{min}} = 5$, $s_{\mathrm{min}} = 10$) and ($k_{\mathrm{max}} = 0.3$, $k_{\mathrm{max}} = 0.4$) pairs have very similar correlation matrices. Consequently, we only show $k_{\mathrm{max}} = 0.3$ and $s_{\mathrm{min}} = 10$ in the main text.

Figures C3 and C4 show the results of fitting the clustering with the simplified model detailed in Section 3.4. Since most results are consistent with the expected value of one, we only display the values for $\mathcal{K} = 0.25\,h/\mathrm{Mpc}$ and $\mathcal{S} = 20\,\mathrm{Mpc}/h$ in the main text.

This paper has been typeset from a TEX/LATEX file prepared by the author.

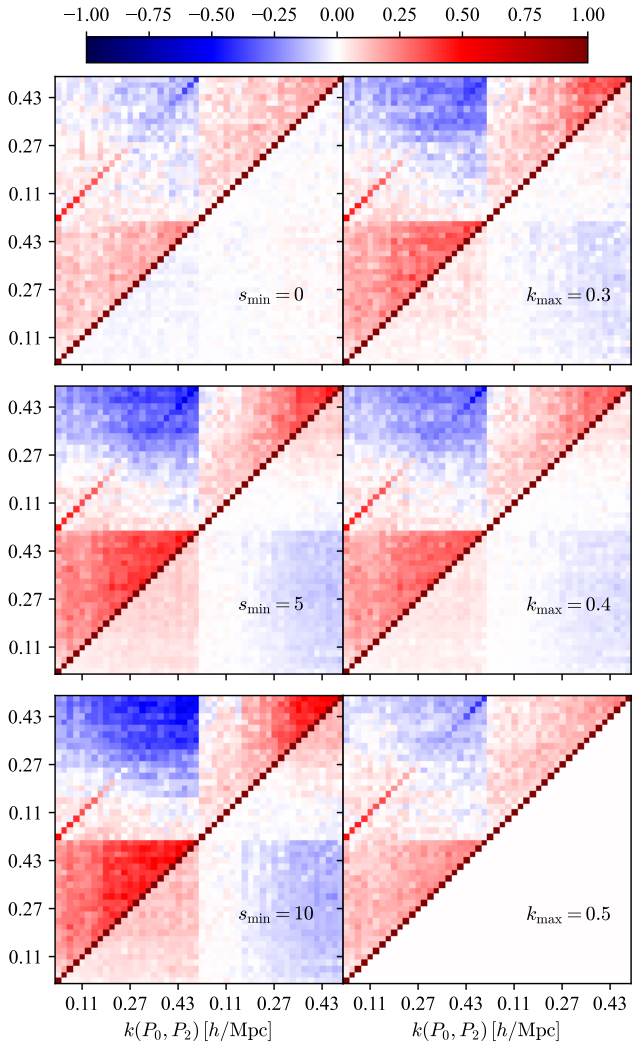


Figure C1. Upper triangular matrices: correlation matrices of the power spectrum monopole and quadrupole, for the fitting cases defined in Table 3. Lower triangular matrices: the difference between the shown correlation matrix and the reference one, i.e. $k_{\mathrm{max}} = 0.5$.

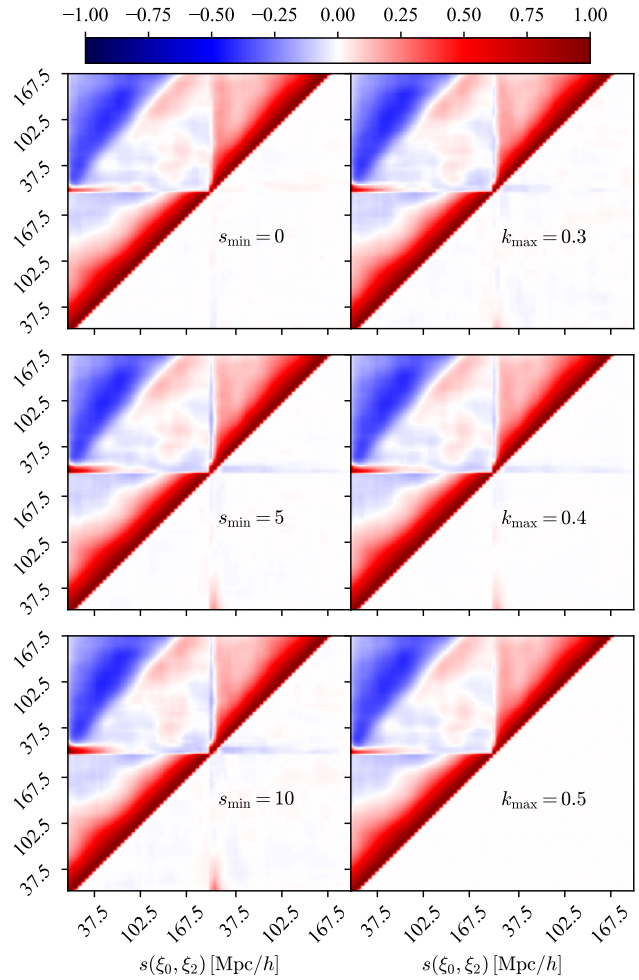


Figure C2. Same as Figure C1, but for 2PCF

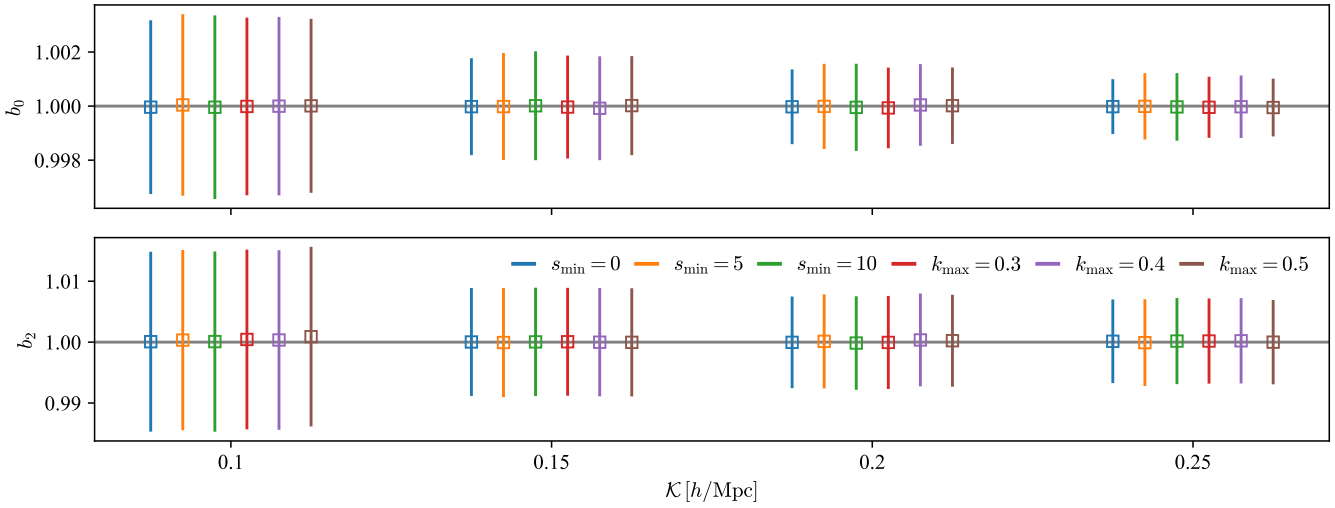


Figure C3. The average of the 123 fitting parameters (b_ℓ) detailed in Section 3.4, obtained from 123 SLICS power spectra fitted on $k \in [0.02, \mathcal{K}]\ h/\mathrm{Mpc}$. The error bars are computed as the average of 123 σ_{b_ℓ}, divided by $\sqrt{123}$, where σ_{b_ℓ} is the standard deviation of the b_ℓ posterior distribution. The different colours stand for the different FastPM covariance matrices exhibited in Figure C1.

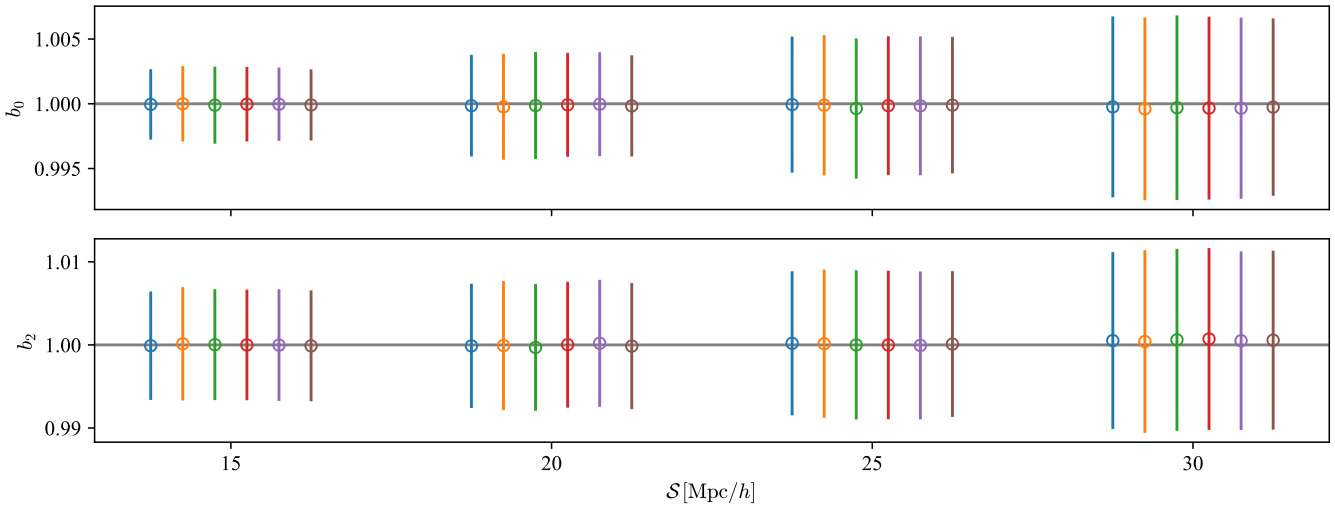


Figure C4. Same as Figure C3, but the fitting is performed on 123 SLICS 2PCF and $s \in [\mathcal{S}, 200]$ using the covariance matrices exhibited in Figure C2.

3 Void clustering models for cosmological measurements

An important limitation of the large scale structure analysis is cosmic variance which depends on the volume proved by the galaxy survey. Nevertheless, given a fixed survey volume, multi-tracer analyses have the potential to decrease the cosmic variance by including multiple biased tracers (for a brief review see Wang & Zhao, 2020). In this regard, Zhao et al. (2020) have shown that a multi-tracer BAO analysis of voids (that have a negative bias) and galaxies (that have a positive bias) improves the constraints compared to the galaxy alone. Therefore, showing the importance of voids in the present and future galaxy surveys that probe the large scale structure.

The first section introduces the concept of cosmic voids and two methods to detect them, with a focus on Delaunay Triangulation (DT). Moreover, it briefly presents the improvements brought by the latest multi-tracer BAO analysis (Zhao et al., 2022) with DT voids and galaxies. In this study, I have contributed to the construction of the numerical model required to describe the broadband shape of the DT void clustering affected by the exclusion effect.

The last section is dedicated to modelling the DT voids and it constitutes a published article (Variu et al., 2023b). In this work, I have co-developed a numerical model[1] of DT voids. In addition, I have performed an in-depth analysis and comparison between different methods for modelling DT voids to understand their impact on BAO measurements.

3.1 Cosmic voids

Qualitatively, cosmic voids are large volumes that do not contain luminous objects and that are found in underdense regions of the CDM field (Rood, 1988; Sheth & van de Weygaert, 2004; van de Weygaert & Schaap, 2009; van de Weygaert & Platen, 2011). Hoyle & Vogeley (2001, 2002) have measured an average void effective diameter of $\approx 30\,\mathrm{Mpc}/h$ and a matter density contrast of -0.92 to -0.96 for the 54 detected voids – defined as empty spheres detected from galaxies. Nevertheless, there are multiple practical definitions and hence algorithms to find

[1] https://github.com/Andrei-EPFL/SICKLE

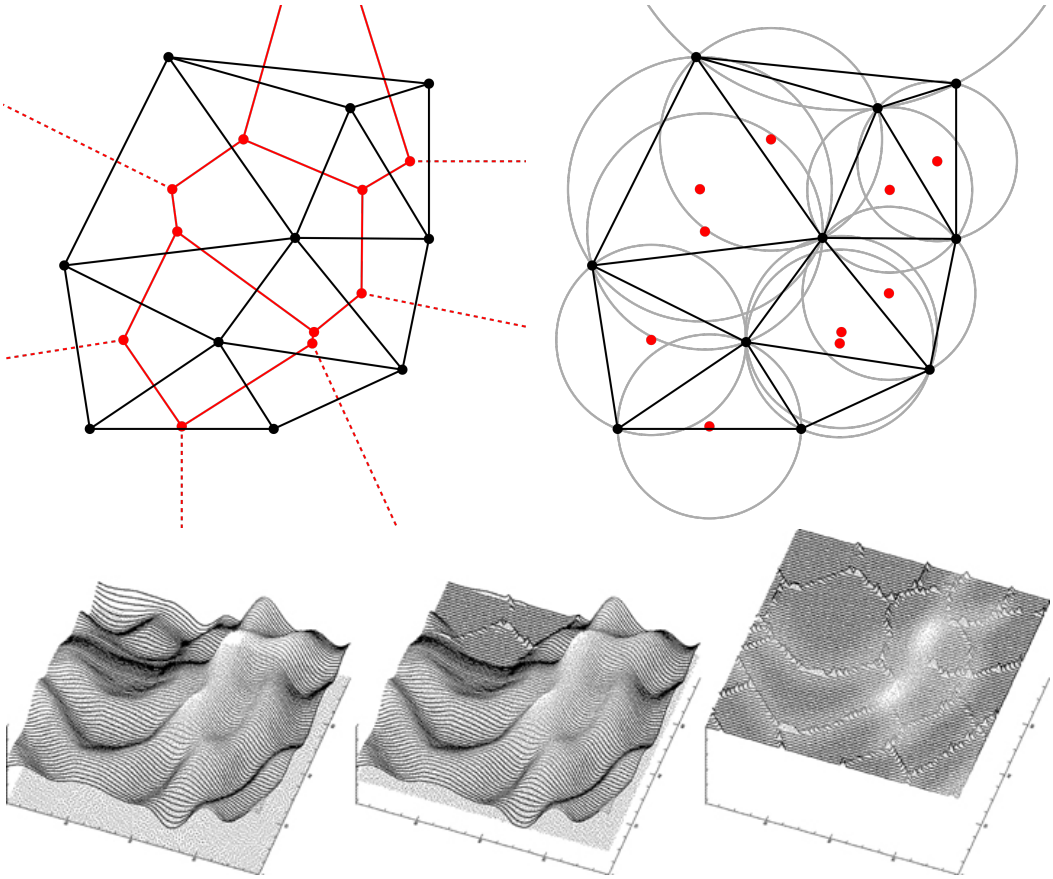


Figure 3.1: The upper panels present the Voronoi (left) and Delaunay (right) Tessellations. For a hypothetical 2D galaxy catalogue, the black points represent the galaxies, the red points are the centres of the circumscribing circles (grey lines) of the triangles (black lines) defined (detected by the Delaunay Tessellation) by three galaxies and the red lines denote the Voronoi cells. The lower panels illustrate the Watershed method. Figures from https://en.wikipedia.org/wiki/Delaunay_triangulation. Figure 1 of (Platen et al., 2007).

voids. Zhao et al. (2016) classify the cosmic voids in four groups:

1. regions with densities lower than the average, detected from the smooth DM, halo or galaxy density field (e.g. Colberg et al., 2005; Neyrinck, 2008);

2. regions that are expanding in time, in contrast to the gravitational collapse of matter (e.g. Hahn et al., 2007; Cautun et al., 2013);

3. regions found using the tessellations of the phase-space particle distribution, that do not contain shell crossings (e.g. Shandarin et al., 2012);

4. empty geometrical structures detected from the distribution of discrete tracers (e.g. El-Ad & Piran, 1997; Foster & Nelson, 2009; Zhao et al., 2016).

The diversity of void definitions and detection methods makes them versatile tools for various cosmological measurements. Their shape can be used to perform Alcock-Paczynski tests (e.g Sutter et al., 2012; Mao et al., 2017), the void-galaxy cross clustering can be utilised in RSD studies (e.g Hamaus et al., 2016; Hawken et al., 2020; Aubert et al., 2022), the void abundance, bias and profile offer tests of modified gravity (e.g Perico et al., 2019) and serve as probes for non-Gaussian primordial perturbations (e.g Kamionkowski et al., 2009). Additionally, their clustering is sensitive to massive neutrinos (e.g Kreisch et al., 2019), and their density profiles depend on the type of dark matter (e.g Yang et al., 2015). Lastly, voids have been used in BAO studies (e.g. Chan & Hamaus, 2021; Zhao et al., 2020, 2022).

A comparison between some of the algorithms to detect voids can be found in (Colberg et al., 2008). However, one commonly used algorithm for identifying voids – defined as underdense regions – is ZOnes Bordering On Voidness (ZOBOV[2]; Neyrinck, 2008). ZOBOV uses the Voronoi Tessellation Field Estimator (VTFE) to estimate the density field of matter tracers. The algorithm creates a cell with a volume $V(i)$ around each matter tracer i, defining it as "the region of space closer to matter tracer i than to any other tracer". The left panel of Figure 3.1 illustrates these Voronoi cells with red lines. Furthermore, the density around tracer i is estimated as $1/V(i)$.

After estimating the density field across the entire space, a watershed algorithm is used to detect the voids. The lower panel of Figure 3.1 metaphorically illustrates the watershed algorithm: water is poured into the valleys of the density field to fill them up. These valleys symbolise the voids, while the ridges correspond to the cosmic sheets and filaments. The Watershed Void Finder (WVF; Platen et al., 2007) employs a similar void detection method, although it estimates the density field with the Delaunay Tessellation Field Estimator (DTFE; Schaap & van de Weygaert, 2000).

DTFE is based on Delaunay Tessellation (also known as Delaunay Triangulation, DT; Delaunay, 1934), which is illustrated in 2D space in the right panel of Figure 3.1. In 3D, the DT detects the empty circumscribing spheres (DT spheres) of the tetrahedrons defined by four points. DT and Voronoi Tessellation are fundamentally related; thus, by connecting the centres of the DT spheres/circles, one can derive the Voronoi cells. In contrast to VTFE, DTFE uses uniformly-random sampled points in space to estimate the density field at those positions. The density field at a specific point is computed as the inverse of the volume of its surrounding DT tetrahedrons. This technique allows for continuous density field estimation, unlike VTFE.

Naturally, as depicted in the lower panel of Figure 3.1, the resulting ZOBOV/WVF voids are disjoint and can have arbitrary shapes associated with the cosmic web. In contrast, the DT identifies overlapping empty spheres. This represents the fundamental idea behind the concept of DT voids.

[2]Popular codes such as (VIDE; Sutter et al., 2015) or REVOLVER https://github.com/seshnadathur/Revolver are based on this algorithm as well

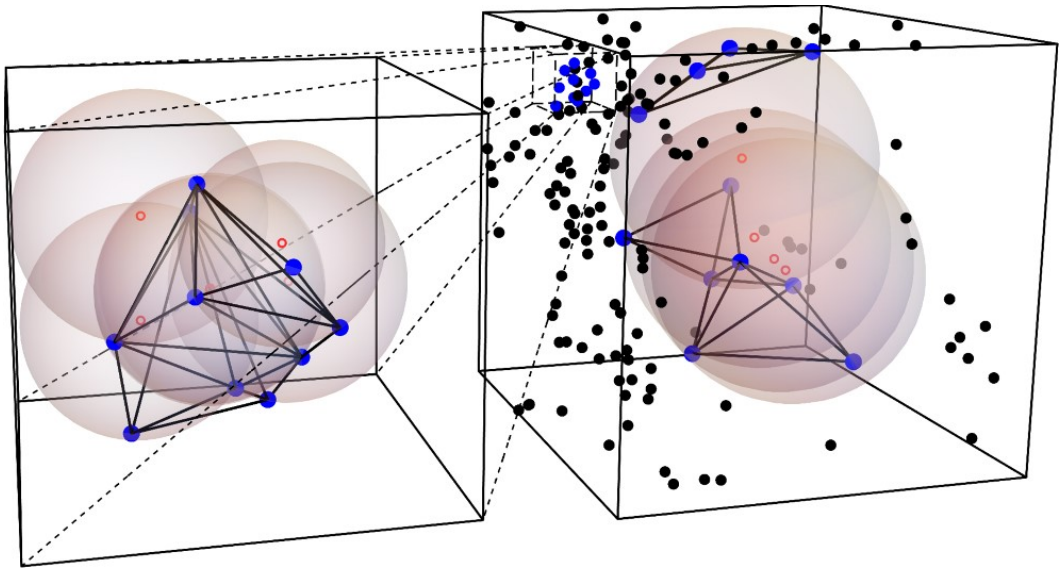


Figure 3.2: The black and blue points represent the haloes in a region of a cubic simulation. The blue points are – in addition – haloes that define tetrahedrons (black lines), whose circumscribing spheres (with red centres) are shown in pink. The left panel shows voids with a radius $R_V < 4\,\mathrm{Mpc}/h$, while the right panel illustrates voids with $R_V \in [26,27]\,\mathrm{Mpc}/h$. Figure 1 of (Zhao et al., 2016).

3.1.1 Delaunay Triangulation Voids

In the pursuit of identifying cosmic voids within a galaxy catalogue, the DT algorithm has been implemented into the Delaunay trIangulation Void findEr (DIVE; Zhao et al., 2016) code. Given a catalogue of 3D Cartesian positions as input, DIVE identifies a set of DT spheres, each characterised by a 3D position and a radius. Figure 3.2 presents a region of a simulated halo catalogue where the pink disks represent the overlapping DT spheres found by DIVE. Notably, these spheres exhibit a wide-ranging distribution of radii, and the radius of each sphere is strongly dependent on the local number density of the matter tracers (Zhao et al., 2016; Forero-Sánchez et al., 2022).

Zhao et al. (2016) have shown that the size of the DT spheres is correlated with the possibility of the spheres to trace underdense or overdense CDM regions. On one hand, Figure 3.3 shows that the large DT spheres are found in the underdense regions of the CDM field – as expected from Sheth & van de Weygaert (2004) – and are called "voids-in-voids". On the other hand, Zhao et al. (2016) explain that the small DT spheres trace the overdense regions (similarly to the haloes shown in Figure 3.3), therefore they are called "voids-in-clouds" (Sheth & van de Weygaert, 2004).

These observations are supported by the comparison between the small and large spheres in Figure 3.4. Firstly, the density contrast of large DT spheres is negative inside the spheres and positive outside and at the border, while for the small DT spheres, the density contrast is

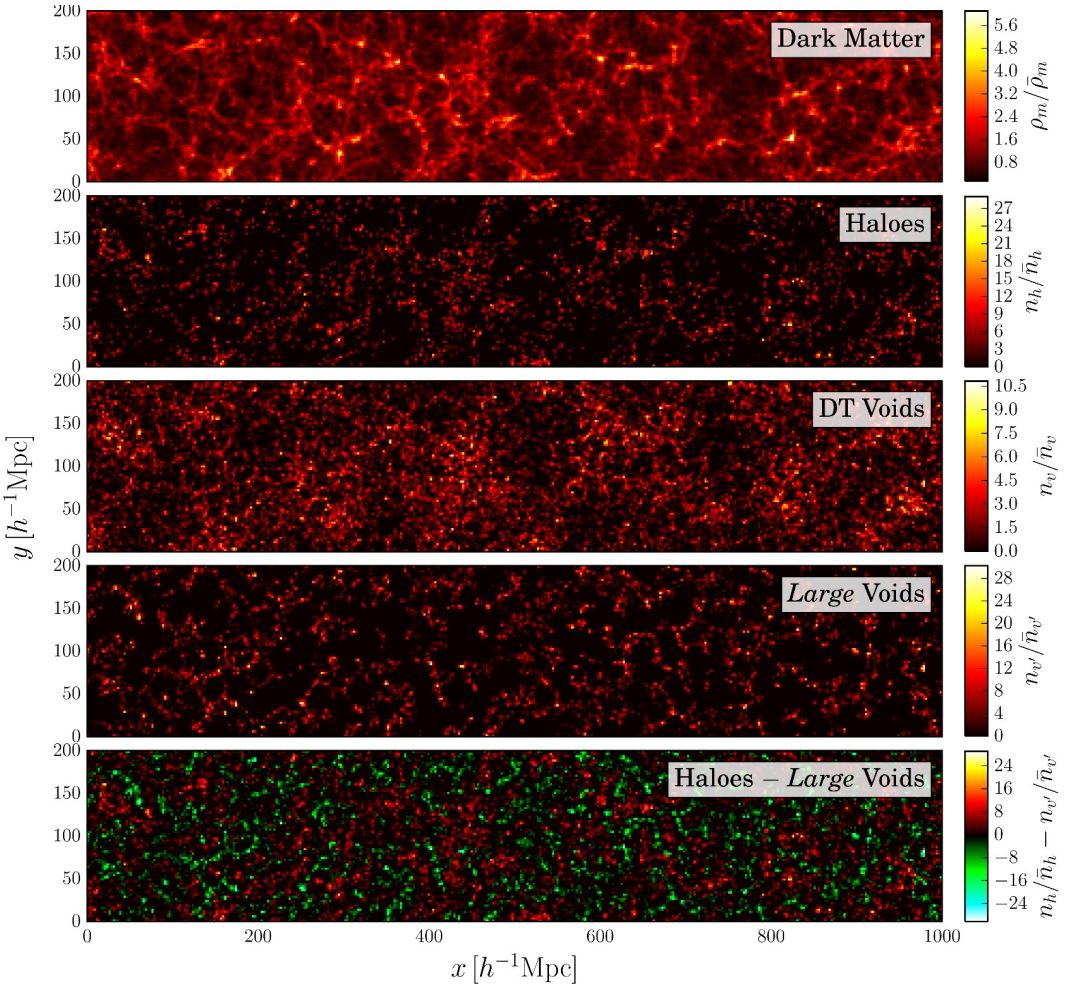


Figure 3.3: The spatial distribution of haloes and DT spheres in the dark matter density field. From top to bottom: the dark matter density field, the halo number density, the DT sphere number density, the DT void (with $R_V \leq 16\,\mathrm{Mpc}/h$) number density. The lowest panel contains both haloes (red) and DT voids (green). Figure 8 of (Zhao et al., 2016).

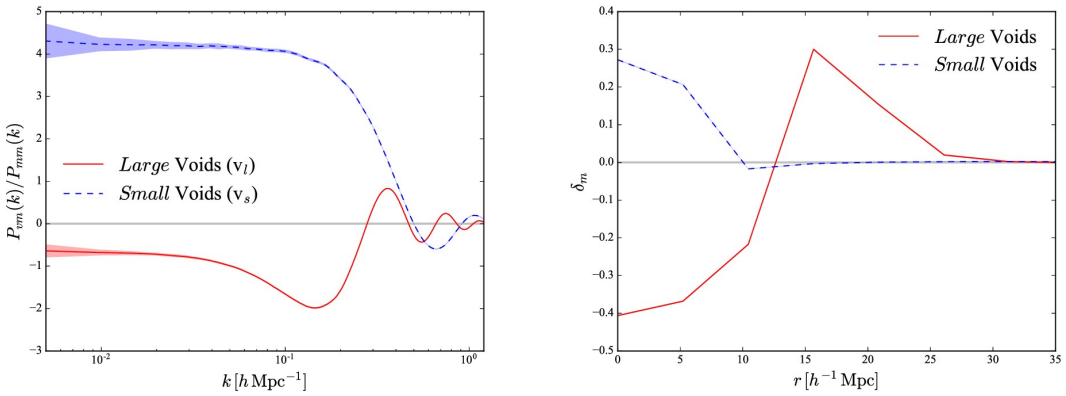


Figure 3.4: The left panel shows the bias of the small DT spheres (blue) and DT voids (red). The same colour scheme is used to illustrate the radial density profile in the right panel. Figure 13 of (Zhao et al., 2016).

positive inside. Secondly, the linear bias of the large spheres is negative, while the one of the small spheres is positive (Hamaus et al., 2014). As a consequence, the large DT spheres are named DT voids. It is important to distinguish the fact that DT voids are not true voids, but they are tracers of the underdense regions. Moreover, they are entirely geometrical structures as DT relies only on geometry and does not require any parameters as input.

Liang et al. (2016) have defined a process to optimally select DT voids based on their radius[3] such that the signal-to-noise ratio of the BAO signature in their clustering is maximised on mocks. Based on this methodology, Kitaura et al. (2016) have detected for the first time the BAO signal (see Figure 3.5) in the clustering of underdense regions, more specifically DT voids constructed from the BOSS galaxy catalogues. Furthermore, they have shown that the BAO feature is not detectable in the 2PCF of the disjoint DT voids[4], suggesting the importance of the overlapping feature of DT voids.

Due to the presence of the BAO feature in the DT void clustering, Zhao et al. (2020) have conducted a galaxy-void multi-tracer BAO study using BOSS DR12 data to assess the potential of DT voids to improve the constrains on cosmological parameters. Figure 3.6 presents a summary of their results: the combined galaxy-DT void sample increases the precision on the α parameter for most of the measurements on the simulated catalogues (black points). Nonetheless, due to cosmic variance, certain combined measurements perform worse than the galaxy measurements alone. This is notably the case for the BOSS galaxy sample at $0.5 < z < 0.75$ as well. In contrast, they have found a 18 per cent improvement in the precision of α for the $0.2 < z < 0.5$ sample by including DT voids.

The latest and most complete multi-tracer BAO study with DT voids and galaxies has been performed by Zhao et al. (2022), using the BOSS DR12 and eBOSS DR16 galaxy samples.

[3]This selection depends on the number density of matter tracers, see e.g. Forero-Sánchez et al. (2022).

[4]The disjoint DT voids form a subsample of all DT voids. They are selected one by one from the largest to the smallest, by excluding each DT void that overlaps with the previously selected disjoint ones.

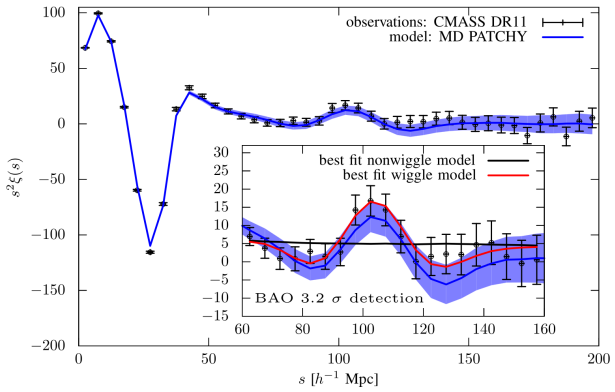


Figure 3.5: The first detection of the BAO feature in the clustering of DT voids. The DT voids have been identified in the galaxy sample of BOSS DR11. Figure 4 of Kitaura et al. (2016).

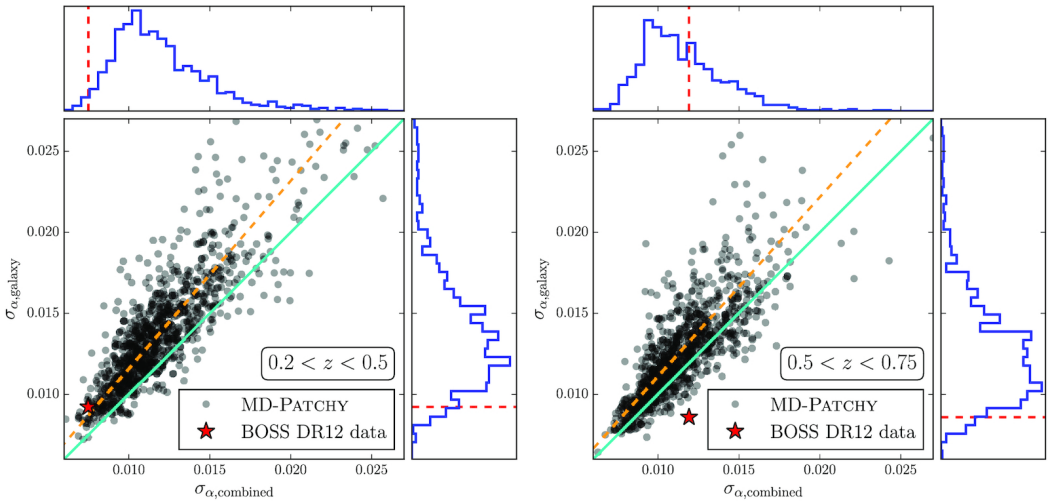


Figure 3.6: A comparison between the precision on the α parameter measured from the galaxy sample alone ($\sigma_{\alpha,\mathrm{galaxy}}$) and the precision from the combined galaxy-void sample $\sigma_{\alpha,\mathrm{combined}}$. The black points indicate the results from 1000 mocks. The red stars represent the measurements from BOSS DR12 data. The cyan lines denote $\sigma_{\alpha,\mathrm{galaxy}} = \sigma_{\alpha,\mathrm{combined}}$. The orange lines illustrate the results from fitting the average of 1000 mocks. Figure 19 of Zhao et al. (2020).

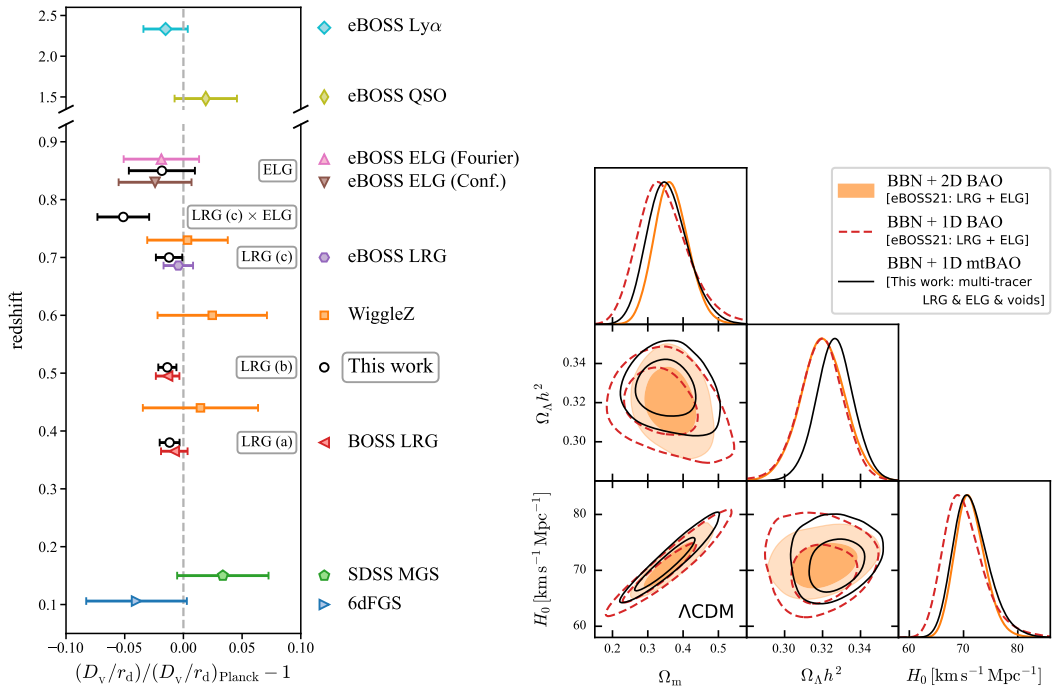


Figure 3.7: Constraints on the $D_V(z)$ by different galaxy samples from multiple spectroscopic surveys (colors). The DT void – galaxy multi-tracer (mtBAO) constraints are shown in black. Constraints on different cosmological parameters of a flat-ΛCDM model, using BBN and galaxy BAO or mtBAO. Figures 18 and 19 from Zhao et al. (2022).

The DT voids have been identified in four distinct galaxy samples: three LRG samples and one set of ELGs. The resulting distance measurements from the isotropic BAO fitting of the combined galaxy-DT void samples are shown in Figure 3.7 using black lines. The anisotropic BAO measurements from BOSS/eBOSS have been converted to the spherically-averaged D_V in order to facilitate a comparison with the multi-tracer results. One can observe that by including the DT voids, the distance measurements are improved in all cases. Numerically, the precision increases by five to fifteen per cent for each of the four samples.

The right panel of Figure 3.7 presents a comparison between the cosmological parameters obtained from:

- the anisotropic BAO measurements of Alam et al. (2021) – in orange;
- the isotropic results estimated from the combination of the anisotropic Alam et al. (2021) measurements – in dashed red lines;
- the isotropic multi-tracer (galaxy + DT voids) BAO measurements – in black lines.

One can observe that the multi-tracer isotropic measurements provide better constraints than the galaxy-only isotropic ones. However, the 2D analysis can have even tighter constraints. As

a summary (see Table 1.1 as well), including the rest of the SDSS samples (MGS, QSO, and Lyα, see Alam et al. (2021); Zhao et al. (2022) for more details), the uncertainties of the H_0, $\Omega_{0\mathrm{m}}$ and $\Omega_{0\Lambda}h^2$ are decreased by 6, 6 and 17 per cent, respectively, compared to the galaxy-only measurements in Alam et al. (2021).

Kitaura et al. (2016); Zhao et al. (2022) explain that DT voids – being defined by tetrahedrons formed out of four galaxies – include information from the higher order statistics, suggesting the reason why the combined study of galaxies and DT voids provides tighter constrains. Forero-Sánchez et al. (2022) have shown that the DT voids should improve the BAO constraints, provided that the galaxy density field does not become fully Gaussian after BAO reconstruction.

Recently, Tamone et al. (2022) have studied the possibility to conduct BAO measurements using the DT voids found in the QSO sample from eBOSS. They have found no improvements on the data, however the measurements on 70 per cent of the mocks have lead to tighter constraints.

Apart from the BAO signature, Figure 3.5 reveals the exclusion effect (Hamaus et al., 2014) on the 2PCF of DT voids at scales $s \in [R_\mathrm{V}, 2R_\mathrm{V}]$, where R_V is the radius cut to select the DT voids. This effect influences the broadband shape of the DT void power spectrum as well. Therefore, in order to perform the multi-tracer BAO fitting, Zhao et al. (2020) have adapted the template power spectrum – equation (1.164) – to account for void exclusion. In Section 3.2, I describe different methods to model the broadband shape of DT voids and test their effect on the BAO fitting.

3.2 Preprint version: "Cosmic void exclusion models and their impact on the distance scale measurements from large scale structure"

MNRAS 000, 1–19 (2022) Preprint 17 March 2023 Compiled using MNRAS LATEX style file v3.0

Cosmic void exclusion models and their impact on the distance scale measurements from large scale structure

Andrei Variu,[1]★ Cheng Zhao,[1]† Daniel Forero-Sánchez,[1] Chia-Hsun Chuang,[2] Francisco-Shu Kitaura,[3,4] Charling Tao,[5] Amélie Tamone,[1] Jean-Paul Kneib[1]
[1]Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland
[2]Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA
[3]Instituto de Astrofísica de Canarias, s/n, E-38205, La Laguna, Tenerife, Spain
[4]Departamento de Astrofísica, Universidad de La Laguna, E-38206, La Laguna, Tenerife, Spain
[5]CPPM, Aix-Marseille Université, CNRS/IN2P3, CPPM UMR 7346, F13288 Marseille, France

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Baryonic Acoustic Oscillations (BAOs) studies based on the clustering of voids and matter tracers provide important constraints on cosmological parameters related to the expansion of the Universe. However, modelling the void exclusion effect is an important challenge for fully exploiting the potential of this kind of analyses. We thus develop two numerical methods to describe the clustering of cosmic voids. Neither model requires additional cosmological information beyond that assumed within the galaxy de-wiggled model. The models consist in power spectra whose performance we assess in comparison to a parabolic model on Patchy cubic and light-cone mocks. Moreover, we test their robustness against systematic effects and the reconstruction technique. The void model power spectra and the parabolic model with a fixed parameter provide strongly correlated values for the Alcock-Paczynski (α) parameter, for boxes and light-cones likewise. The resulting α values – for all three models – are unbiased and their uncertainties are correctly estimated. However, the numerical models show less variation with the fitting range compared to the parabolic one. The Bayesian evidence suggests that the numerical techniques are often favoured compared to the parabolic model. Moreover, the void model power spectra computed on boxes can describe the void clustering from light-cones as well as from boxes. The same void model power spectra can be used for the study of pre- and post-reconstructed data-sets. Lastly, the two numerical techniques are resilient against the studied systematic effects. Consequently, using either of the two new void models, one can more robustly measure cosmological parameters.

Key words: software: simulations – methods: numerical – methods: data analysis – methods: statistical – cosmology: observations – large-scale structure of Universe

1 INTRODUCTION

In order to measure cosmological parameters and better understand the Universe and its expansion, multiple techniques have been developed and implemented; one of them is the study of the Baryonic Acoustic Oscillations (BAOs). They are oscillations in the primordial plasma that have altered the matter distribution in the early Universe, leaving an imprint that has been initially observed in the spectra of Cosmic Microwave Background (CMB) temperature anisotropies (e.g. Hinshaw et al. 2003; Planck Collaboration et al. 2020).

The large spectroscopic surveys provide complementary BAO constraints to CMB. Currently, the most precise BAO studies using the 3D clustering statistics of galaxies have been achieved by Baryon Oscillation Spectroscopic Survey (BOSS; Alam et al. 2017) and extended-BOSS (eBOSS; Alam et al. 2021). The ongoing Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016) plans to further improve the precision of the BAO measurements by increasing the number density of tracers and mapping larger volumes. Meanwhile, the future Cosmology Redshift Survey (CRS; Richard et al. 2019), part of 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019) survey, will provide complementary measurements to DESI by scanning different regions on the sky. In addition to the clustering of galaxies – e.g. luminous red galaxies (LRG; Ross et al. 2017; Beutler et al. 2017), emission line galaxies (ELG; Raichoor et al. 2020) – the BAO feature has been detected in the clustering of quasi-stellar objects (QSO; Ata et al. 2017), Lyman α forests (Lyα forests; Busca et al. 2013) and cosmic voids (Kitaura et al. 2016).

While the matter tracers – except Lyα forests – are directly observable, the cosmic voids are detected from the positions of the former. In general, cosmic voids are regions in space emptied of luminous objects that trace the under-dense zones of the density field (see review of van de Weygaert & Platen 2011). However, in practice, there are multiple definitions and thus different algorithms to detect them (e.g. Padilla et al. 2005; Platen et al. 2007; Neyrinck 2008; Sutter

★ E-mail: andrei.variu@epfl.ch
† E-mail: cheng.zhao@epfl.ch

et al. 2015; Zhao et al. 2016, and references therein). This allows for a greater diversity of cosmological measurements. For example, cosmic voids are part of BAO studies (e.g Zhao et al. 2020; Chan & Hamaus 2021; Zhao et al. 2022), their geometry is involved in performing Alcock-Paczynski tests (e.g. Sutter et al. 2012; Mao et al. 2017), their cross-clustering with galaxies has been used in Redshift-Space-Distortions (RSD) studies (e.g. Hamaus et al. 2016; Nadathur et al. 2019; Hamaus et al. 2020; Correa et al. 2022).

Multi-tracer analyses (Zhao et al. 2020; Zhao et al. 2022) of galaxies with voids determined using the Delaunay trIangulation Void findEr (DIVE; Zhao et al. 2016) – code that uses the Delaunay Triangulation (DT; Delaunay 1934) on the positions of the matter tracers – show improvements on the precision of Alcock–Paczynski parameter (α; Alcock & Paczynski 1979) of the order of 10 per cent compared to galaxy-only measurements. However, these studies imply the additional challenge of modelling the void clustering. Compared to the matter tracers, voids have large sizes, hence their exclusion has a stronger impact on the clustering (Hamaus et al. 2014a). In consequence, Zhao et al. (2020) have developed a more general model than the galaxy de-wiggled one (Xu et al. 2012) in order to correctly account for this difference.

The purpose of this paper is to introduce two numerical methods that can be used in the modified de-wiggled model to provide a description of the void exclusion effect. The principle behind the two methods is to first create a halo catalogue by assigning them directly on the density field corresponding to the initial conditions and then detect the voids. Finally, the computed void power spectrum represents the model for the void exclusion.

Section 2 presents the simulations involved in assessing the performance of the void model power spectra. The description of the two numerical techniques and the methodology employed in testing them are described in Section 3. Section 4 shows the results of the performance and robustness tests that have been effectuated on the numerical techniques. The last section concludes the current article.

2 DATA

2.1 PATCHY boxes

In this study, we use two sets of $2.5\,h^{-1}$Gpc cubic mock catalogues obtained using the PerturbAtion Theory Catalogue generator of Halo and galaxY distributions (PATCHY; Kitaura et al. 2013). This generator uses the Augmented Lagrangian Perturbation Theory (ALPT; Kitaura & Heß 2013) to model the structure formation and then it assigns biased tracers (e.g. haloes or galaxies) to the density field based on a bias model.

Both sets of PATCHY boxes are calibrated against the BigMultiDark (BIGMD) N-body simulation (Klypin et al. 2016). However, the set of 1000 boxes is tuned to match a BIGMD Sub-Halo Abundance Matching (SHAM) galaxy catalogue, whereas the set of 100 mocks is calibrated with a BIGMD halo catalogue.

The reference BIGMD dark-matter box has a side length of $2.5\,h^{-1}$Gpc and contains 3840^3 dark-matter particles with a mass of $2.359 \times 10^{10}\,h^{-1}\mathrm{M}_\odot$ each. The cosmology of the simulation is described by $h = 0.6777$, $\Omega_\Lambda = 0.692885$, $\Omega_\mathrm{m} = 0.307115$, $\Omega_\mathrm{b} = 0.048206$, $n = 0.96$, $\sigma_8 = 0.8228$ [1].

On one hand, the BIGMD SHAM mock is based on the dark-matter snapshot at redshift $z = 0.4656$ and has a galaxy density of $n = 3.976980 \times 10^{-4}\,h^3\mathrm{Mpc}^{-3}$. On the other hand, the BIGMD halo catalogue uses the snapshot at $z = 0.5618$ and has a number density of $n = 3.5 \times 10^{-4}\,h^3\mathrm{Mpc}^{-3}$.

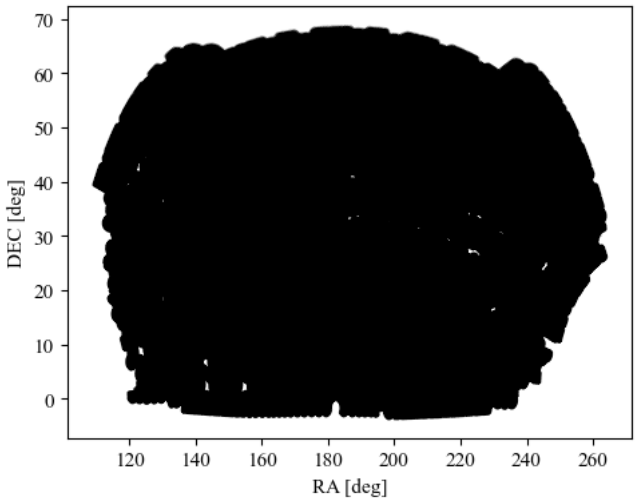


Figure 1. The NGC footprint of the BOSS DR12 (Alam et al. 2015) used to build the PATCHY light-cones.

2.2 PATCHY light-cones

In order to validate the suitability of the numerical models for survey-like data, we construct the Light-Cones (LC) of all the 1000 PATCHY galaxy boxes using the MAKE_SURVEY[2] (White et al. 2013) code. This implies:

- the conversion of the (X, Y, Z) euclidean coordinates to Right Ascension (RA), Declination (DEC) and redshift z;
- the cut of a survey geometry in (RA, DEC);
- the application of a radial selection function to sample tracers along the line-of-sight.

On one hand, the applied footprint (Figure 1) corresponds to the BOSS DR12[3] Northern-Galactic Cap (NGC) footprint (Alam et al. 2015). On the other hand, a Gaussian distribution (Figure 2) is used as a radial selection function, for $z \in [0.325, 0.775]$. This distribution is realistic enough for the current purpose and it allows for the flexibility of choosing the redshift range and the shape.

3 METHODOLOGY

3.1 BAO reconstruction

The BAO reconstruction technique (Eisenstein et al. 2007b) is used to increase the BAO signal (from the clustering of matter tracers) and thus improve constraints on the cosmological parameters (e.g. Anderson et al. 2014; Alam et al. 2017; Bautista et al. 2020; Raichoor et al. 2020; Alam et al. 2021).

The principle of this technique is to estimate the displacement of the biased matter tracers and then move them at positions corresponding to higher redshifts to linearise the density field. By implementation, this method affects the distribution and the clustering of

[1] https://www.cosmosim.org/cms/simulations/bigmdpl/

[2] https://github.com/mockFactory/make_survey

[3] https://data.sdss.org/sas/dr12/boss/lss/

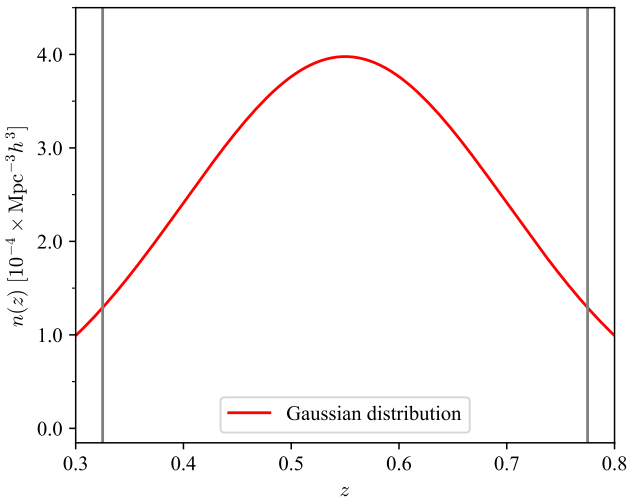


Figure 2. The theoretical radial selection function used to build the PATCHY light-cones. The used redshift range is $z \in [0.325, 0.775]$, between the two vertical grey lines.

the matter tracers, thus the distribution of the determined voids and their clustering also change. Given that the reconstruction has been used in multi-tracer analysis of voids and galaxies (Zhao et al. 2020; Zhao et al. 2022) and it changes the void clustering, it is imperative to test whether the numerical models can describe the voids obtained from reconstructed PATCHY mock catalogues.

In the current study, we adopt the iterative method proposed by Burden et al. (2015) to perform the reconstruction. In practice, we use the code REVOLVER[4] described in Nadathur et al. (2019). The required input parameters of the code are the number of iterations (three, in this study) the linear bias of the mock tracers $b = 2.2$, the growth rate $f = 0.743$ (corresponding to an effective redshift of the simulation boxes $z = 0.4656$), the smoothing scale $S = 15\,h^{-1}$Mpc and the grid size of 512^3 on which the density field is approximated using a Cloud-In-Cell (CIC; Sefusatti et al. 2016) mass assignment scheme.

3.2 Void detection

We apply the DIVE[5] code (Zhao et al. 2016) to the galaxy and halo catalogues to obtain the DT spheres. Similarly to other methods (e.g Sheth & van de Weygaert 2004; Hamaus et al. 2014b), Zhao et al. (2016) have shown that while the small DT spheres are mostly *voids-in-clouds* and have positive matter density contrast, the larger DT spheres (DT voids) are more probably *voids-in-voids* and exhibit a negative matter density contrast. Consequently, a radius based selection – which depends on the matter tracers' number density – can discriminate the true tracers of under-dense regions from the possible tracers of over-dense regions. Moreover, Liang et al. (2016) have proved that a radius based selection can be used to maximise the signal-to-noise ratio of the BAO signal from the clustering of DT voids.

In this study, we are interested in modelling only the DT voids as they have been used in multi-tracer analyses such as Zhao et al. (2020); Zhao et al. (2022) to improve the precision of BAO measurements. Thus, we select the DT spheres with a radius $\mathrm{R_v} \geq 16\,h^{-1}$Mpc to form the DT void sample. This radius cut is chosen by analogy to Zhao et al. (2020) and based on the studies of Liang et al. (2016); Forero-Sánchez et al. (2022). Forero-Sánchez et al. (2022) have shown that the void selection based on a constant radius cut yields unbiased BAO measurements when reconstruction is applied on the galaxy catalogue or when systematical effects – such as a small sample incompleteness – are present. Lastly, Zhao et al. (2016) have observed that by selecting the large DT spheres, the resulting DT void sample has a negative bias, consistently with the detailed results of Hamaus et al. (2014a).

3.3 Clustering computation

3.3.1 Two point correlation function

In order to compute the 2PCF we use the Fast Correlation Function Calculator[6] (FCFC) code (Zhao 2023), which accepts as input both boxes and light-cones and can employ any type of estimator. In the current study, several estimators have been necessary to correctly account for the specificity of the data sets.

- The natural estimator (Peebles & Hauser 1974) is used to compute the void auto-2PCF and void-galaxy cross-2PCF from pre-reconstructed boxes and the void auto-2PCF from post-reconstructed boxes:

$$\xi(s) = \frac{\mathrm{D_vD_v}(s)}{\mathrm{R_vR_v}(s)} - 1, \tag{1}$$

$$\xi(s) = \frac{\mathrm{D_gD_v}(s)}{\mathrm{R_gR_v}(s)} - 1. \tag{2}$$

- The Landy–Szalay estimator (Landy & Szalay 1993) is needed to compute the void auto-2PCF and void-galaxy cross-2PCF for the light-cones:

$$\xi(s) = \frac{\mathrm{D_vD_v}(s) - 2\mathrm{D_vR_v}(s) + \mathrm{R_vR_v}(s)}{\mathrm{R_vR_v}(s)}, \tag{3}$$

$$\xi(s) = \frac{\mathrm{D_gD_v}(s) - \mathrm{R_gD_v}(s) - \mathrm{D_gR_v}(s) + \mathrm{R_gR_v}(s)}{\mathrm{R_gR_v}(s)}. \tag{4}$$

- A modified version of the Landy–Szalay estimator (Padmanabhan et al. 2012) – inspired from Szapudi & Szalay (1997) – is required to compute the void-galaxy cross-2PCF from the post-reconstructed boxes:

$$\xi(s) = \frac{\mathrm{D_gD_v}(s) - \mathrm{S_gD_v}(s) - \mathrm{D_gR_v}(s) + \mathrm{S_gR_v}(s)}{\mathrm{R_gR_v}(s)}. \tag{5}$$

On one hand, the letter D denotes the data catalogue of voids ($\mathrm{D_v}$) or galaxies ($\mathrm{D_g}$) and thus DD represents the data-data normalised pair counts. On the other hand, the random catalogue is expressed through the letter R that can be related to both voids ($\mathrm{R_v}$) and galaxies ($\mathrm{R_g}$). Consequently, RR and DR serve as the symbols for the random-random and data-random normalised pair counts, respectively. Lastly, $\mathrm{S_g}$ is referring to a galaxy random catalogue that was shifted by the same displacement field as the reconstructed galaxy catalogue and thus $\mathrm{S_gR_v}$ represents the random-random pair counts.

The data-data pair counts can be directly computed given the

[4] https://github.com/seshnadathur/Revolver
[5] https://github.com/cheng-zhao/DIVE
[6] https://github.com/cheng-zhao/FCFC

measured data catalogue. However, in order to compute the data-random and random-random pair counts, one has to construct the random part. For boxes, which implicitly have periodic boundary conditions, the RR term (R_vR_v; R_gR_v) can be computed analytically:

$$RR(s) = \frac{4\pi(s_2^3 - s_1^3)}{3}\frac{1}{V}, \tag{6}$$

where s_2 and s_1 are the boundaries of a separation bin ($s_2 > s_1$) and $s = (s_2 + s_1)/2$ for linearly separated bins.

In contrast, for light-cones, the RR term has to be evaluated on random catalogues which must include the same observational effects as the data catalogues. For galaxies, we initially create a random box (RB) of the same size as the BigMD and Patchy boxes, but ten times denser, by randomly sampling Cartesian positions. Afterwards, we apply make_survey with the same configurations as for the Patchy boxes in order to obtain a random LC that is ten times denser than the Patchy LC.

In the case of voids, we adopt a modified version of the 'shuffling' technique (Liang et al. 2016). de Mattia & Ruhlmann-Kleider (2019); Zhao et al. (2021) have shown that it is necessary to avoid having identical angular and radial positions of objects in the data and the random catalogues, otherwise, the measured clustering is affected. Consequently, to diminish this effect, we stack 100 void Patchy LC mocks. Furthermore, we shuffle the RA-DEC pairs in bins of redshift and void radius. This shuffling maintains the angular coverage, but breaks the correlation between the redshift-radius pair and the RA-DEC pair. Finally, we uniformly and randomly down-sample the resulting shuffled catalogue down to 20 times the void density of the Patchy LC. Having the void and galaxy random catalogues, one can compute the R_vR_v, R_gR_v, D_vR_v, R_gD_v, D_gR_v pair counts for LC.

The shifted galaxy random cubic catalogues S_g are computed during the reconstruction of the Patchy boxes by applying the displacement field that is estimated from the Patchy boxes on the random box RB. This creates a dedicated random catalogue to each of the Patchy boxes. In comparison with galaxies, the void random box is simply constructed by randomly and uniformly sampling Cartesian positions inside a box of side-length of 2500 h^{-1}Mpc, so that the density is ten times larger than the DT void sample.

We finally compute the pair counts and the 2PCF using 40 separation bins between 0 and 200 h^{-1}Mpc (i.e. a bin width of 5 h^{-1}Mpc).

3.3.2 *Power spectrum*

In the current study, we exploit the POWSPEC[7] code – described in Zhao et al. (2021) – to calculate the required power spectra. The density field is estimated using the Cloud-In-Cell (CIC; Sefusatti et al. 2016) particle assignment scheme and power spectra are computed in k bins of size 0.0025 h Mpc^{-1}.

The smoothness of the 2PCFs obtained through the Hankel transform (see Section 3.4.1) of power spectra depends on the range spanned by the wavenumber k and on the number of power spectra realisations. The large value of k is required to ameliorate the effect of the undulatory shape of the 0-order spherical Bessel function used in the Hankel transform, while the large number of realisations is needed to decrease the noise coming from cosmic variance. In order to achieve a large enough k interval, we use a grid size of 2048^3 to measure the power spectra. This provides a $k_{max} \sim 2.57\,h$ Mpc^{-1} for boxes and a $k_{max} \sim 1.88\,h$ Mpc^{-1} for light-cones.

[7] https://github.com/cheng-zhao/powspec

Abbreviation	Description
DW	de-wiggled model, Eq. (10)
PAR	parabolic model, Eq. (13)
PAR_U	PAR with uniform prior, Eq. (29)
PAR_G	PAR with a prior defined by Eq. (30)
fix c	PAR with a fixed c parameter, determined from the fit of the average 2PCF from 500 or 1000 realisations
SK	SICKLE, details in Sec. 3.4.1.2 and Tab. 4
SK_B	calibrated SK model based on Boxes having the same halo number density as the reference
SK_{def}	defective SK model, see Tab. 4
SK_{LC}	the model obtained by applying the survey geometry (Light-Cone) of the reference on the halo boxes corresponding to SK_B
CG	CosmoGAME, details in Sec. 3.4.1.3 and Tab. 4
CG_B	same as SK_B but for CG
CG_{def}	same as SK_{def} but for CG
CG_{LC}	same as SK_{LC} but for CG
CG_{80}	calibrated CG model based on boxes having a 20% lower halo number density than the reference
CG_{120}	calibrated CG model based on boxes having a 20% higher halo number density than the reference
gv	void-halo (galaxy) cross-clustering
vv	void auto-clustering

Table 1. The abbreviations of the studied models.

Given the fact that we need a large number of realisations to reduce variances, it is computationally-expensive to always use a grid size of 2048^3. Thus, we also calculate power spectrum realisations using a grid size of 512^3 in order to have a smoother power spectrum for lower wavenumbers (see Section A for more details). In this case, we use the grid interlacing technique (Sefusatti et al. 2016) to reduce the alias effects introduced by the particle assignments scheme.

3.4 BAO fitting

3.4.1 *BAO models*

The theoretical model used to fit the 2PCF is defined as follows (Xu et al. 2012):

$$\xi_{model}(s) \equiv B^2\xi_t(\alpha s) + A(s), \tag{7}$$

where B tunes the amplitude of the model, α is the Alcock–Paczynski (Alcock & Paczynski 1979) parameter that is related to the position of the BAO peak and $A(s)$ is a function required to describe the broad-band shape of the correlation function, which consists of three nuisance parameters a_0, a_1, a_2:

$$A(s) = a_0 + a_1 s^{-1} + a_2 s^{-2}. \tag{8}$$

Xu et al. (2012) and Vargas-Magaña et al. (2014) have shown that this function does not bias the measurement of α. Lastly, ξ_t is the Hankel transform of the template power spectrum $P_t(k)$ as described in Xu et al. (2012):

$$\xi_t(s) = \int \frac{k^2 dk}{2\pi^2} P_t(k) j_0(ks) e^{-k^2a^2}, \tag{9}$$

where j_0 is the 0-order spherical Bessel function of the first kind (i.e. the sinc function) and $a = 2\,h^{-1}$ Mpc is a factor for the Gaussian damping of the Bessel function's wiggles at high-k. A more detailed discussion on how the value of a was chosen is presented in Section A.

In the case of galaxies, the template power spectrum can be expressed by the typical de-wiggled (DW) model (Anderson et al. 2014):

$$P_{\rm t,DW}(k) = [P_{\rm lin}(k) - P_{\rm lin,nw}(k)]{\rm e}^{-k^2\Sigma_{\rm nl}^2/2} + P_{\rm lin,nw}(k), \quad (10)$$

where $P_{\rm lin}(k)$ is the linear power spectrum that can be obtained using CAMB[8] software (Lewis et al. 2000), $P_{\rm lin,nw}(k)$ is the linear power spectrum without the BAO feature (no wiggles, nw) computed using the formula of Eisenstein & Hu (1998), and $\Sigma_{\rm nl}$ is the damping parameter for BAO (Eisenstein et al. 2007a). In this work, we use the input power spectrum employed in the generation of the Patchy mocks as $P_{\rm lin}(k)$ for BAO fittings. This provides a predictable α value in the absence of any systematic effects (see Section 3.4.4.2 for a discussion on this topic).

Zhao et al. (2020) have shown that the de-wiggled model is not suitable for voids due to the improper accounting of the broadband shape. More precisely, the exclusion effect of voids (Hamaus et al. 2014a) affects significantly the clustering and thus the shapes of the 2PCF and power spectrum.

Consequently, Zhao et al. (2020) have introduced a more general template power spectrum that accounts for the exclusion effect:

$$P_{\rm t}(k) = \varphi(k) P_{\rm t,DW}(k) \quad (11)$$

and

$$\varphi(k) = \frac{P_{\rm t,nw}(k)}{P_{\rm lin,nw}(k)}, \quad (12)$$

where $P_{\rm t,nw}(k)$ is the non-wiggled tracer power spectrum, that can practically include the void exclusion effect.

In this paper, we study different methods to model the additional factor introduced in the template power spectrum, whose names and abbreviations are summarised in Table 1. The first method is introduced by Zhao et al. (2020) and it consists in approximating the factor with a parabola (parabolic model). The other two methods provide numerical models for the $P_{\rm t,nw}(k)$ term in three steps:

(i) create a halo catalogue using gauSsIan moCK tempLate gEnerator (SICKLE[9]) or Cosmological GAussian Mock gEnerator (CosmoGAME[10]);
(ii) apply DIVE on the constructed halo catalogues to get the DT voids;
(iii) measure the power spectra of the resulting DT void catalogues.

SICKLE and CosmoGAME are two C codes that:

(i) generate Gaussian random fields based on $P_{\rm lin,nw}(k)$, using the fixed amplitude (Angulo & Pontzen 2016) presented in Chuang et al. (2019), in order to decrease the sample variance of halo–halo and halo–void clustering;
(ii) assign haloes directly on the Gaussian fields without gravitational evolution.

Nonetheless, the two techniques differ in their halo assignment schemes.

By construction, our methods have the advantage of being generalisable for multiple definitions of voids as one needs to simply apply the required necessary void finder on the resulting SICKLE or CosmoGAME halo catalogue. However, the disadvantage is that they are computationally expensive compared to analytical models. Consequently, we may consider in future studies analytical models based on the pioneering work to model the void exclusion (Hamaus et al. 2014a) by Chan et al. (2014).

3.4.1.1 Parabolic model Zhao et al. (2020) have shown that the additional factor – $\varphi(k)$, Eq. (12) – can be approximated by a parabola (PAR):

$$\varphi(k) \sim 1 + ck^2, \quad (13)$$

where c is a free parameter, determined through the fitting process. In practice, when we fit the 2PCF, we force c to take values only inside a prior interval with a given probability distribution. More details about the prior distribution are discussed in Section 3.4.3.

3.4.1.2 SICKLE The code generates a Gaussian random field in Fourier space on a grid whose size can be tuned ($N_{\rm grid}$). The field is then scaled by a factor γ to encode the information about the linear growth and the bias parameter. The resulting field is in an approximation of the matter overdensity field in Fourier space $\tilde{\delta}(\mathbf{k})$. Furthermore, $\tilde{\delta}(\mathbf{k})$ is transformed to real space into $\delta_{\rm m}(\mathbf{r})$ using the implementation of the Discrete Fourier Transform in the FFTW[11] package.

Starting from the matter overdensity field $\delta_{\rm m}(\mathbf{r})$, haloes are selected by an iterative algorithm inspired from the CIC mass assignment scheme until the desired number of haloes is reached:

(i) obtain the (x, y, z) position of the maximum overdensity value;
(ii) scatter the (x, y, z) position using displacements sampled from a Triangular distribution ($\mathcal{T}(x) = \max(1-|x|, 0)$; given by the weight of the CIC scheme) to get a new (x', y', z') position;
(iii) assign a halo at (x', y', z');
(iv) compute the contribution of the assigned halo to the matter density field using the CIC scheme;
(v) subtract the previously computed contribution from the density field in order to emulate the exclusion of massive haloes;
(vi) go to (i).

The exclusion of massive haloes has a strong impact on the halo clustering, thus it must be taken into account when the halo catalogues are constructed (Somerville et al. 2001; Casas-Miranda et al. 2002; Baldauf et al. 2013; Zhao et al. 2015). In our Universe, it is mainly caused by the facts that:

- two or more haloes that are close enough can gravitationally collapse into a single more massive one;
- there is not enough matter to form multiple massive haloes on small scales.

For this method, the scaling factor γ and the size of the grid $N_{\rm grid}$ are the two parameters that can be tuned to influence the halo and void clustering. Nevertheless, the effects of these parameters on the resulting void power spectrum are not straightforwardly interpretable.

3.4.1.3 CosmoGAME Similarly to SICKLE, CosmoGAME estimates the density field in real space $\delta_{\rm m}(\mathbf{r})$ on which it assigns haloes. While $\delta_{\rm m}(\mathbf{r})$ is identical to the one estimated by SICKLE (except the γ factor), the halo selection process and the tunable parameters are analogous to the galaxy assignment step for the Effective–Zel'dovich mocks (EZmocks; Chuang et al. 2014; Zhao et al. 2021). It is important to re-emphasize the fact that whilst EZmocks include the

[8] https://camb.info/
[9] https://github.com/Andrei-EPFL/SICKLE
[10] https://github.com/cheng-zhao/CosmoGAME
[11] http://fftw.org/

Zel'dovich approximation to estimate the gravitational evolution of the density field, CosmoGAME uses directly the Gaussian random field to assign haloes.

One of the CosmoGAME's parameters used to select haloes is the critical density (δ_c). This variable plays the role of a threshold below which one cannot assign haloes (Percival 2005) and thus has an impact on the three-point clustering of haloes (Kitaura et al. 2015).

After picking the density field values above δ_c, random numbers are added to them in order to take into account the stochasticity of the tracers (Chuang et al. 2014):

$$\delta_t(\mathbf{r}) = H(\delta_m - \delta_c)\delta_m(\mathbf{r}) \times (1 + S), \tag{14}$$

where:

$$S = \begin{cases} G(\lambda),\ G(\lambda) \geq 0; \\ \exp(G(\lambda)) - 1,\ G(\lambda) < 0 \end{cases} \tag{15}$$

and $H(x)$ is the Heaviside step function. In the previous equation, $G(\lambda)$ is a random number sampled from a Gaussian distribution with a zero mean and a standard deviation λ – as a free parameter.

Lastly, a power-law probability density function (PDF) is used to assign haloes to the resulting density values:

$$\mathcal{P}(n_t) = Ab^{n_t}, \tag{16}$$

where $\mathcal{P}(n_t)$ is the probability to assign n_t haloes to a density peak. The fact that one has to ask for a fixed number of tracers puts a constrain on one of two parameters (i.e. A or b). Thus, we fix A (with $A > 0$) and treat b as the only free parameter within $0 < b < 1$. In practice, using the previous PDF, one computes the number of density values to which one should assign n_t tracers:

$$n_c(n_t) = \lfloor N_{cell}\mathcal{P}(n_t) \rceil, \tag{17}$$

where $N_{cell} = N_{grid}^3$ ($N_{grid} = 512$, in this study) is the total number of cells in the density grid and the $\lfloor \cdot \rceil$ operator obtains the nearest integer. Moreover, we compute the maximum number of haloes that can be possibly assigned to one density value as:

$$n_{t,\,max} = \min_{n_t > 0} \{n_t | N_{cell}\mathcal{P}(n_t) < 0.5\}. \tag{18}$$

The tracer assignment is performed – after the density values $\delta_t(\mathbf{r})$ are sorted in descending order – as follows:

(i) one assigns $n_{t,\,max}$ haloes to the highest $n_c(n_{t,\,max})$ density values;

(ii) one continues to assign $(n_{t,\,max} - i)$ haloes to the next $n_c(n_{t,\,max} - i)$ density values,

where i takes values from 1 to $n_{t,\,max}$. The positions of the assigned haloes are sampled from a uniform distribution inside each of the grid cells.

Another parameter of CosmoGAME, similarly to SICKLE, is the grid size N_{grid}. Nonetheless, by adjusting the other parameters, one can emulate the effect of a different grid size. Thus, it is not used in the tuning process.

Lastly, CosmoGAME has been already run to create the void model power spectrum for the multi-tracer cosmological analysis with SDSS data by Zhao et al. (2022).

3.4.2 Parameter inference

In order to infer the fitting parameters, we have written pyBAOfit[12]. The code uses a combination of PyMultiNest[13] – the python implementation of MultiNest (Feroz & Hobson 2008; Feroz et al. 2009, 2019) – and a Least-Square (LS) method (Press et al. 2007; Zhao et al. 2022) in order to decrease the computational time. While PyMultiNest samples the $(\alpha, B, \Sigma_{nl}, c)$ parameters, the LS determines the best-fitting nuisance parameters (a_0, a_1, a_2). MultiNest is a Bayesian Monte Carlo (MC) sampler, which provides not only the best-fitting parameters, but also the Bayesian evidence and the posterior distributions of the parameters. A more detailed discussion about the different treatment of the two sets of parameters is done in Section B.

The Bayesian inference is based on Bayes' theorem that provides a way to merge the prior information about the Θ parameters of a model M with the measurements from the data D. Mathematically, the theorem provides the posterior probability density of the Θ parameters, given the data D and the model M:

$$p(\Theta|D,\ M) = \frac{p(D|\Theta,\ M)p(\Theta|M)}{p(D|M)}, \tag{19}$$

where $p(\Theta|M)$ is the prior distribution of the Θ parameters (see Section 3.4.3), $p(D|\Theta,\ M)$ is the likelihood – related to the measurements from data D – and $p(D|M)$ is the Bayesian evidence – $\mathcal{Z}$, a normalising factor and a valuable tool in model selection.

In the current study, we approximate the likelihood with a multivariate Gaussian:

$$p(D|\Theta,\ M) = \mathcal{L}(\Theta) \sim e^{-\chi^2(\Theta)/2}, \tag{20}$$

where χ^2 is the chi-squared defined as:

$$\chi^2(\Theta) = \mathbf{v}^T\mathbf{C}^{-1}\mathbf{v}. \tag{21}$$

In the above formula, $\mathbf{C}^{-1}$ is the inverse of the unbiased covariance matrix (Hartlap et al. 2007), and $\mathbf{v}$ is the difference between the model and the data vectors, i.e. $\mathbf{v} = \boldsymbol{\xi}_{data} - \boldsymbol{\xi}_{model}(\Theta)$.

The unbiased covariance matrix $\mathbf{C}$ is related to the sample covariance matrix of mocks $\mathbf{C}_s$ as follows:

$$\mathbf{C}^{-1} = \mathbf{C}_s^{-1}\frac{N_{mocks} - N_{bins} - 2}{N_{mocks} - 1}, \tag{22}$$

where N_{mocks} is the number of mocks used to compute the covariance matrix and N_{bins} is the length of the data vector $\boldsymbol{\xi}_{data}$ included in the fitting process. Furthermore, $\mathbf{C}_s$ can be decomposed into a multiplication between a matrix $\mathbf{M}$ and its transpose:

$$\mathbf{C}_s = \frac{1}{N_{mocks} - 1}\mathbf{M}^T\mathbf{M}. \tag{23}$$

Finally, the elements of the matrix $\mathbf{M}$ are computed as:

$$\mathbf{M}_{ij} = \xi_i(s_j) - \bar{\xi}(s_j),\ i = 1, 2, ..., N_{mocks},\ s_j \in [s_{min}, s_{max}], \tag{24}$$

where ξ_i denotes the 2PCF of the $i-$th mock realisation, $\bar{\xi}$ represents the mean 2PCF of all mocks and $[s_{min}, s_{max}]$ represents the interval of data points involved in the 2PCF fitting.

The quoted values of the parameters are the medians of the posterior distributions, and the 1σ uncertainties are half the differences between the 84th and 16th percentiles, unless otherwise specified.

3.4.3 Parameter priors

The Bayesian inference method requires prior knowledge about the measured parameters, generally implemented as a probability distribution function. In our case, we have mainly assumed uniform distributions $\mathcal{U}_{[a,\,b]}(\Theta)$:

$$\mathcal{U}_{[a,\,b]}(\Theta) = \begin{cases} 0, & \Theta < a \\ \frac{1}{b-a}, & \Theta \in [a, b] \\ 0, & \Theta > b. \end{cases} \tag{25}$$

[12] https://github.com/Andrei-EPFL/pyBAOfit
[13] https://github.com/JohannesBuchner/PyMultiNest

$\Sigma_{\rm nl}$ h^{-1}Mpc	Auto	Cross
fix-c	9.03	9.77
$\rm PAR_G$	9.03	9.77
$\rm SK_B$	6.88	5.28
$\rm CG_B$	7.03	3.88
$\rm SK_{LC}$	7.68	6.77
$\rm CG_{LC}$	7.64	5.80

Table 2. Prior values of $\Sigma_{\rm nl}$ when fitting the individual 2PCF of light-cones. These are the best-fitting values of the average of 1000 2PCF computed from light-cones. Cross – void-galaxy cross 2PCF; Auto – void auto 2PCF.

While for the priors of α and B we have generally imposed:

$$p(\alpha) = \mathcal{U}_{[0.8,\ 1.2]}(\alpha), \tag{26}$$

$$p(B) = \mathcal{U}_{[0,\ 25]}(B), \tag{27}$$

the prior of $\Sigma_{\rm nl}$ depends whether the 2PCF has been measured from boxes or from light-cones. In the first case – i.e. for boxes – we implement a uniform prior:

$$p\left(\Sigma_{\rm nl}\right) = \mathcal{U}_{[0,\ 30]\ \ h^{-1}{\rm Mpc}}(\Sigma_{\rm nl}). \tag{28}$$

In the second case – i.e. for light-cones – we fix the values of $\Sigma_{\rm nl}$ to the ones in Table 2. The chosen intervals are large enough to not bias the measurements, as shown by (Zhao et al. 2020) and also obvious in Figures B1-B4.

The reason behind fixing the $\Sigma_{\rm nl}$ is that the light-cones have a smaller volume than the boxes, thus the corresponding 2PCF are noisier. Given the noisier 2PCF, $\Sigma_{\rm nl}$ is not properly constrained and the uncertainty of α is overestimated – see also Figure A3. Zhao et al. (2022) have shown that fixing this parameter does not bias the measurements and thus it is appropriate to do it for the light-cones. In order to accurately measure $\Sigma_{\rm nl}$, we have fitted the average of all 1000 2PCF realisations measured from light-cones with a covariance matrix corresponding to the average 2PCF – i.e. computed from 1000 realisations and rescaled by 1000 (rescaled covariance matrix) – and the uniform prior shown in Eq. (28), as performed by Zhao et al. (2022). The best-fitting $\Sigma_{\rm nl}$ values (Table 2) are then used in fitting the individual 2PCF from light-cones.

In the case of the parabolic model, as seen in Eq. (13), there is an additional parameter c, for which we consider three cases:

- a uniform prior for c ($\rm PAR_U$)

$$p(c) = \mathcal{U}_{[-10^4,\ 10^4]\ \ h^{-2}{\rm Mpc}^2}(c); \tag{29}$$

- a uniform prior with two Gaussian tails ($\rm PAR_G$), similar to the one used in Zhao et al. (2020)

$$p(c) = \begin{cases} 0, & c < c_{\rm min} \\ A' \exp(-\frac{(c-c_{\rm fmin})^2}{2\sigma_c^2}), & c \in [c_{\rm min}, c_{\rm fmin}] \\ A', & c \in [c_{\rm fmin}, c_{\rm fmax}] \\ A' \exp(-\frac{(c-c_{\rm fmax})^2}{2\sigma_c^2}), & c \in [c_{\rm fmax}, c_{\rm max}] \\ 0, & c > c_{\rm max}, \end{cases} \tag{30}$$

where $c_{\rm fmin} = -100\,h^{-2}{\rm Mpc}^2$, $c_{\rm fmax} = 900\,h^{-2}{\rm Mpc}^2$, $c_{\rm min} = -400\,h^{-2}{\rm Mpc}^2$, $c_{\rm max} = 1200\,h^{-2}{\rm Mpc}^2$ and $\sigma_c = 100\,h^{-2}{\rm Mpc}^2$;

- a fixed value of c (fix c, see Table 3), as in (Zhao et al. 2022).

c h^{-2}Mpc^2	Auto	Cross
light-cone	2193	477
pre-recon box	1064	216
recon box	4030	319

Table 3. Prior values of c when fitting the individual 2PCF with a parabolic model. These are the best-fitting values of the average 2PCF (from 1000 light-cones or 500 boxes). Cross – void-galaxy cross 2PCF; Auto – void auto 2PCF.

The uniform prior on c (Eq. (29)) has been always used when we have fitted the average 2PCF (of 1000 realisations from LC and of 500 realisations from boxes). For the individual 2PCF, we have either fixed the values of c – as in Table 3 – or used the $\rm PAR_G$ prior, Eq. (30).

Similarly to $\Sigma_{\rm nl}$, we have determined the value of c by fitting the average 2PCF (from 500 boxes or from 1000 light-cones) with the rescaled covariance matrix – corresponding to the average 2PCF – to mitigate the potential biases due to the cosmic variance of the mocks. The best-fitting values of c – shown in Table 2 – are used in the fitting of individual 2PCF. In contrast, to test the 2PCF fitting range, we use the covariance matrix corresponding to one 2PCF realisation (unscaled covariance matrix) together with the average 2PCF.

It is important to note that all the above priors have been used for fitting both the void auto-2PCF and the void-galaxy cross-2PCF. However, when we fit the void-galaxy cross-2PCF, we have to account for the negative bias of the DT voids (Zhao et al. 2016). Generally, the B^2 term in Eq. (7) should be replaced by the product of the galaxy bias with the void one: $B_{\rm galaxy} \times B_{\rm void}$, with $B_{\rm void} < 0$. However, in this work, we do not write the explicit form because we do not fit simultaneously the void auto-2PCF, void-galaxy cross-2PCF and galaxy auto-2PCF. Consequently, we simply replace B^2 with $-B^2$ in Eq. (7) for the parabolic and the DW models. In contrast, the numerical models contain the information of the void negative bias in the shape of the resulting power spectrum, see the cross-clustering in Figure 3.

3.4.4 Model comparison

In the next paragraphs, we define the parameters that we use to compare the models.

3.4.4.1 Bayes factor Apart from inferring parameters, Bayes' theorem can also be utilised to compare the quality of different models given prior probabilities of each models and their evidences:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)}, \tag{31}$$

where

$$\mathcal{Z}_i \equiv p(D|M_i) = \int \mathcal{L}(\Theta)p(\Theta|M){\rm d}\Theta \tag{32}$$

is the Bayesian evidence, $p(M_1)/p(M_2)$ is the prior probability ratio between the two models and $p(M_1|D)/p(M_2|D)$ is the posterior probability ratio of the two models given the data set D.

MULTINEST provides the natural logarithm of the Bayesian evidence, thus one can easily compute $\ln(\mathcal{Z}_1/\mathcal{Z}_2)$, i.e. the natural logarithm of the Bayes factor between any two tested models. Given that we consider the prior probabilities of the models to be equal $p(M_1) = p(M_2)$, the Bayes factor is a direct indication of whether

a model has a higher probability to be correct than another given a data set.

3.4.4.2 Tension parameter The most important aspect of a studied model is the capability to provide unbiased measurements of the Alcock–Paczynski parameter and its uncertainty. In order to have a quantitative description of the possible biases, we define the tension parameter $\tau(x, y|\sigma_x, \sigma_y)$ between two values x and y, given their uncertainties σ_x and σ_y, respectively:

$$\tau(x, y|\sigma_x, \sigma_y) = \frac{x - y}{\sqrt{\sigma_x^2 + \sigma_y^2}}. \quad (33)$$

Naturally, this parameter can quantify the differences between different models, however it can also show the bias with respect to a reference.

Given the fact that the input power spectrum of the PATCHY mocks takes also the role of $P_{\rm lin}(k)$ in Eq. (10) to perform the BAO fitting, the expected measured value of α should be equal to one, in the absence of the non-linear evolution of the BAO peak and if all systematic effects are taken into account. Nonetheless, Prada et al. (2016) has shown that the BAO can have a shift towards higher α values of ~ 0.25 per cent for halo samples with linear bias from 1.2 to 2.8. Nevertheless, in this analysis, we approximate the reference to one and thus we also study the values of $\tau(\alpha, 1|\sigma_\alpha, 0)$.

3.4.4.3 Relative difference We also formally define the relative difference in order to compare two quantities:

$$\mathcal{R}(x, y) = 100 \times \left(\frac{x}{y} - 1\right). \quad (34)$$

This tells us the difference in percentage between the two values.

3.4.4.4 Pull function In order to verify whether the uncertainties are correctly estimated, we define the pull function:

$$g(x) = \frac{x - \bar{x}}{\sigma_x}, \quad (35)$$

where $\bar{x}$ is the mean of a set of values x and σ_x is its standard deviation. If the histogram of the $g(x)$ values follow a standard normal distribution, one can conclude that the uncertainty of x is correctly estimated.

4 TESTS AND RESULTS

4.1 Analysis and comparison of void clustering models

We start by comparing the ratio $\varphi(k)$ Eq. (12) of all models – DW, PAR, SICKLE, COSMOGAME – to the one of pre-reconstructed PATCHY boxes. In Figure 3, the colour dotted curves denote the numerical models, while the black curves represent the reference computed from 500 PATCHY mocks. The horizontal dashed lines represent the DW model ($\varphi(k) = 1$) that unequivocally under-fit the exclusion-effect-dominated reference. In contrast, one can observe that for small values of k a parabola is a good approximation of the ratio, however it evidently fails for $k > 0.05h$ Mpc^{-1}. Unlike the previous models, the numerical models follow the reference up to $k = 0.6\,h$Mpc^{-1}.

Furthermore, we check the robustness of all four models to the fitting range on the average correlation function – computed from 500 mocks – by evaluating the tension $\tau(\alpha, 1|\sigma_\alpha, 0)$. Figure 4 contains the values of the tensions for the void auto-2PCF (left) and void-galaxy cross-2PCF (right) for different fitting intervals. Generally, the tension depends on the fitting range. However, its values are also influenced by the model and the studied clustering.

Obviously, in the case of the de-wiggled model, the values of α are strongly biased for most fitting intervals, reaching values of $\sim 1\sigma$ and above. This observation is consistent with the fact that this model is not suitable to describe the clustering of voids, as shown in Zhao et al. (2020). The parabolic model shows significant improvements with respect to the de-wiggled model as most values are within $\pm 0.2\sigma$ from zero. There are the clear outliers at $s_{\rm min} = 40\,h^{-1}$Mpc) for the void auto-2PCF, that do not appear for the void-galaxy cross-2PCF. An explanation might be that the exclusion effect in configuration space is present at smaller separations for the cross-clustering than for the auto-clustering.

The numerical models are more robust to the fitting ranges – compared to the other methods – given the fact that the tension of α

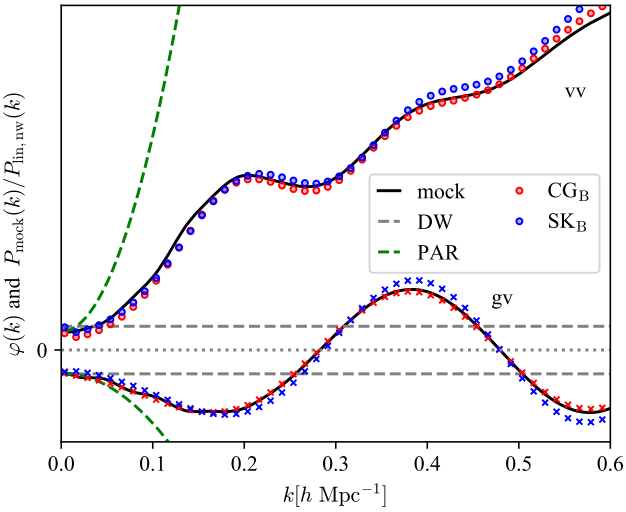


Figure 3. Comparison of $\varphi(k)$ – defined in Eq. (12) – with the ratio between the average mock power spectrum $P_{\rm mock}$ and $P_{\rm lin,nw}$ (black). $\varphi(k)$ is computed for different models: grey dashed - de-wiggled model; green - parabolic model; red and blue - numerical models. $P_{\rm mock}$ is obtained from 500 pre-reconstructed PATCHY cubic mocks. The numerical models were re-scaled to match $P_{\rm mock}$, so the y ticks are meaningless. See Table 4 for the tuning parameters of the numerical models and Table 1 for the abbreviations.

	CG$_{\rm B}$ / CG$_{\rm LC}$	CG$_{\rm def}$	CG$_{80}$	CG$_{120}$
δ_c	2.6 (1.8)	2.4 (1.6)	1.8 (1.2)	1.6 (1.8)
λ	1.0 (0.3)	2.0 (0.5)	0.4 (0.1)	1.5 (1.0)
b	0.44 (0.28)	0.28 (0.20)	0.32 (0.08)	0.52 (0.28)
	SK$_{\rm B}$ / SK$_{\rm LC}$	SK$_{\rm def}$		
$N_{\rm grid}$	1024	1024		
γ	0.075	0.3		

Table 4. Upper table: The values of the COSMOGAME's free parameters used to create the numerical models. A more detailed description of the parameters can be found in Section 3.4.1.3. The abbreviations are defined in Table 1. The values in brackets are for the void-halo cross-power-spectrum, while the rest are for the void auto-power-spectrum. Lower table: The values of the SICKLE's free parameters used to create the numerical models for both the void auto-power-spectrum and the void-halo cross-power-spectrum. More details can be found in Section 3.4.1.2

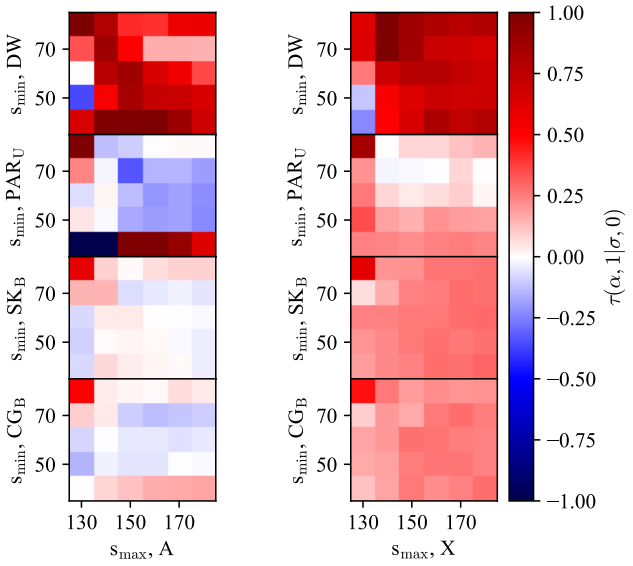


Figure 4. Comparison of different fitting ranges for four different models using $\tau(\alpha, 1|\sigma_\alpha, 0)$, Eq. (33). Both the average void auto-2PCF (left) and void-galaxy cross-2PCF (right) – computed from 500 individual PATCHY cubic mocks – are considered. The abbreviations are defined in Table 1.

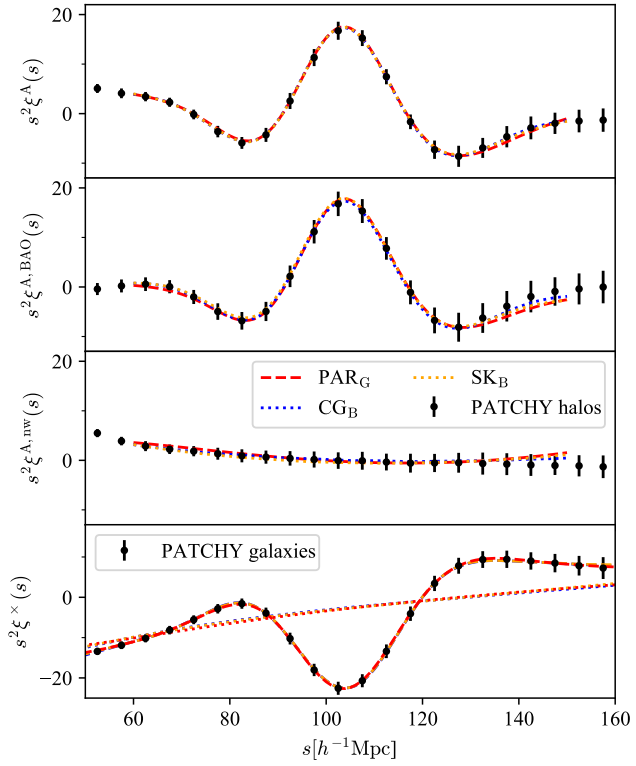


Figure 5. The best-fitting model curves for the average void auto-2PCF computed from 100 individual PATCHY halo boxes and for the average void-galaxy cross-2PCF computed from 500 individual PATCHY galaxy boxes. First panel: the complete auto-2PCF. Second panel: the BAO peak (i.e. $s^2[\xi(s) - \xi^{nw}(s)]$). Third panel: the 2PCF without the BAO peak. The fourth panel: the complete cross-2PCF with the best-fitting curves (with and without BAO peak). The abbreviations are defined in Table 1.

is more homogeneous across the fitting ranges. There is the obvious exception of the narrow $s \in [80, 130]\ h^{-1}$Mpc interval, which yields a strong bias given the lack of sufficient data points to describe well the peak. For most other fitting ranges, the results from the void auto-2PCF show little to no bias at all ($\pm 0.1\sigma$), whilst a more consistent, yet not significant bias is present for the void-galaxy cross-2PCF ($\sim 0.2\sigma$).

Due to the fact that around the $s \in [60, 150]\ h^{-1}$Mpc interval, the results are not sensitive to the fitting range, and this interval has been used in Zhao et al. (2020), we use it in the following tests.

Figure 5 presents the best-fitting curves of the average correlation function for three models: parabolic model, SICKLE and CosmoGAME. All three models are describing well both the BAO peak and the broadband shape. Looking at the BAO-free best-fitting curves (the third panel and the dotted lines in the fourth panel of Figure 5), one can ascertain that none of the models introduce any additional signal at the position of the BAO peak.

Figures 6 and 7 show a comparison of the four different models in terms of the measured α values from the 500 individual mocks. One can observe that the de-wiggled model induces a bias in the α values with respect to all other models for both void auto-2PCF and void-galaxy cross-2PCF.

The PAR_G model provides similar α values to the numerical models, but it is prone to fit poorly which leads to extreme values (the three points around the value of 0.8, in Figure 6). In contrast, the parabolic model with fixed c parameter is consistent with the numerical models for both the void auto-2PCF and the void-galaxy cross-2PCF. This suggests that a lack of a strong prior knowledge on c presents risks of extreme failure. Consequently, we consider only the fixed-c case in the further model comparison. Finally, the two numerical models are indistinguishable in terms of the resulting α values.

Analysing the average α of the 500 values from Figure 8, one can learn that the de-wiggled model introduces a bias of 0.4 to 0.7 per cent. In contrast, the bias shown by the numerical models and the parabolic model with the fixed c is around ± 0.1 per cent for the void auto-2PCF and around 0.15 per cent for void-galaxy cross-2PCF. Moreover, the α values for the void auto-2PCF tend to be lower than one, while the values for the void-galaxy cross-2PCF larger than one. This is consistent with the findings of McCullagh et al. (2013); Neyrinck et al. (2018): due to the gravitational evolution, the clustering of over-dense regions underestimates the length of the sound horizon, whereas with the under-dense regions, the sound horizon is overestimated. Additionally, one has to consider that the values of α are slightly over-estimated, given the noise in the individual 2PCF and the large prior interval for Σ_{nl}, as shown in Figure A3.

In order to more robustly check the tensions between the models, we compute $\tau(\alpha_x, \alpha_y|\sigma_x, \sigma_y)$ between all pairs of models and show the resulting histograms in the lower triangular plots of Figures 9 and 10. The mean tensions with respect to the de-wiggled model reach values of $\sim -0.7\sigma$ for void auto-2PCF, and $\sim -0.5\sigma$ for void-galaxy cross-2PCF, supporting previous claims. Moreover, despite the important differences between the numerical models and the parabolic model with the fixed c parameter observed in Figure 3, the actual tensions between the measured α values are not significant (at most $\sim 0.3\sigma$ and on average $\sim 0.1\sigma$). This is because the damping term a in the Hankel transform – defined in Eq. (9) – decreases the amplitude of the models sharply at high k, and thus the higher k discrepancies become less important.

While the tensions between the models can be informative on the

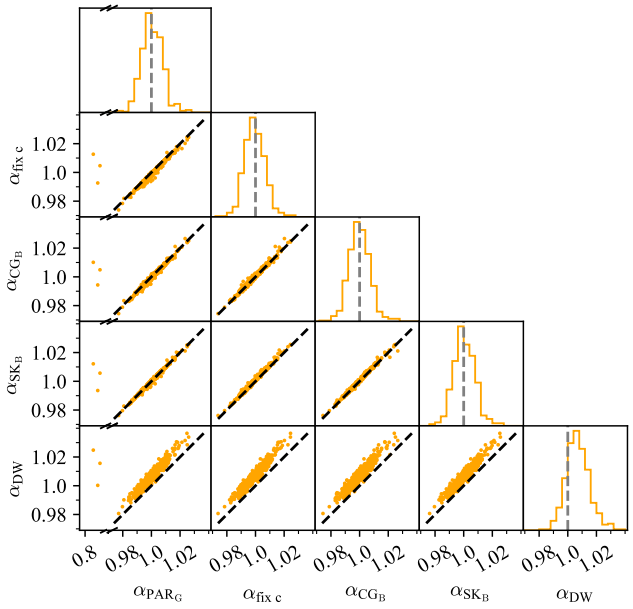


Figure 6. The α values obtained from the fitting of 500 individual void auto-2PCF computed from PATCHY cubic mocks. The abbreviations are defined in Table 1. Grey - the theoretical value of 1; Black - the diagonal $y = x$.

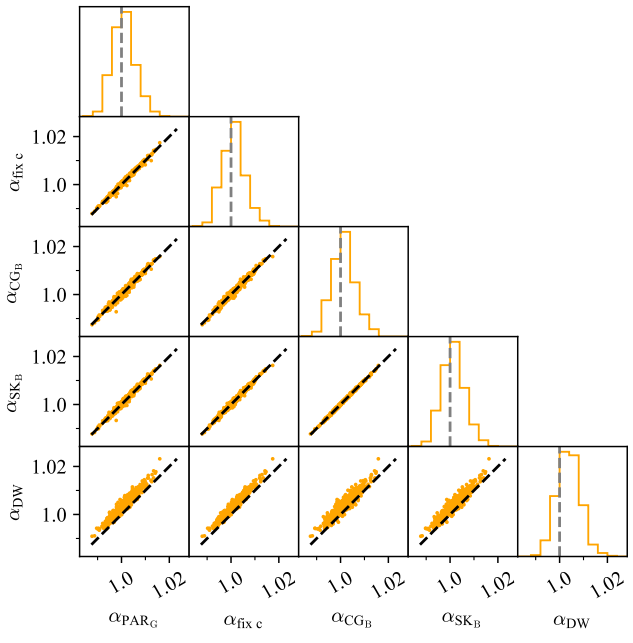


Figure 7. Same as Figure 6, but for the void-galaxy cross-2PCF.

possible introduced biases, the pull function $g(x)$ provides information about the uncertainty estimation. The resulting histograms can be observed along the diagonals of Figures 9 and 10. For both void auto-2PCF and void-galaxy cross-2PCF, one can estimate well the uncertainty σ_α with all models.

Finally, by studying the values of the Bayes factor for all pairs of models in the upper triangular panels of Figures 9 and 10, one can conclude that:

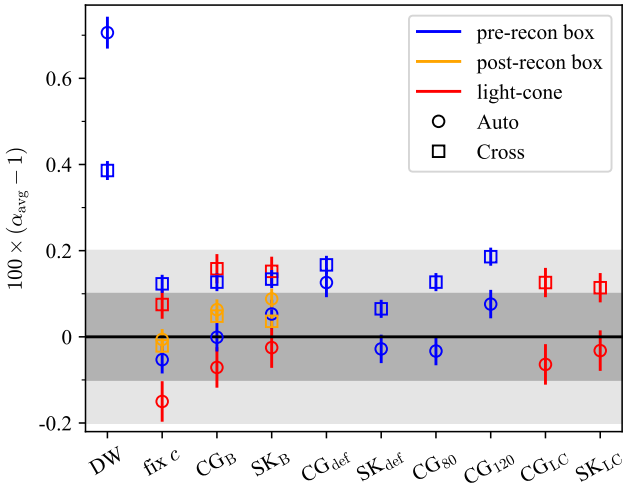


Figure 8. The average of 500 α values for PATCHY boxes and of 1000 α values for PATCHY light-cones measured from void auto-2PCF and void-galaxy cross-2PCF. The error bars are computed as the standard deviation of the 500 (1000) α values further divided by $\sqrt{500}$ ($\sqrt{1000}$). The black horizontal denotes the values of zero, while the grey shaded areas encompass the intervals of ±0.2% and ±0.1% from the reference. See Table 1 for abbreviations.

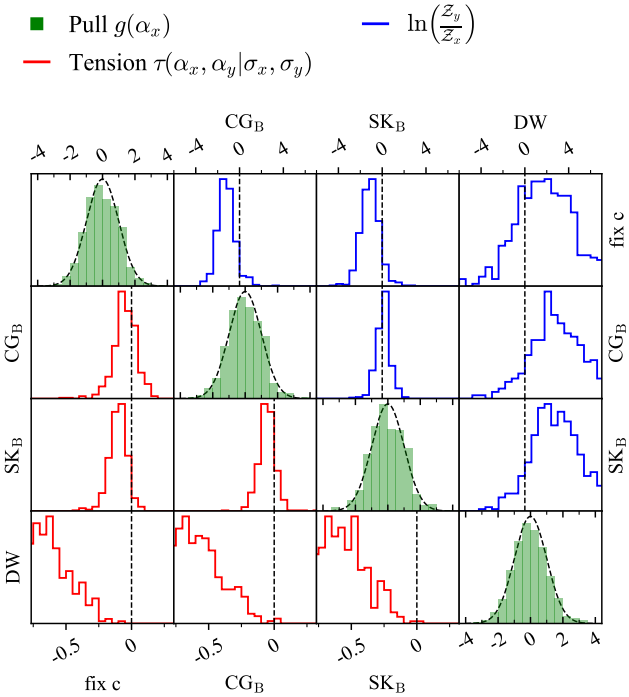


Figure 9. Diagonal panels: green - the histograms of the pull function $g(\alpha_x)$ values, Eq. (35); black - standard normal distributions. Lower triangular plots: the values of $\tau(\alpha_x, \alpha_y|\sigma_x, \sigma_y)$, Eq. (33), for all combinations of models. Upper triangular plot: the natural logarithm of the Bayes Factor $\ln(\mathcal{Z}_y/\mathcal{Z}_x)$ (see Section 3.4.4.1). The results correspond to the individual fittings of the 500 void auto-2PCF computed from the PATCHY cubic mocks. The abbreviations are defined in Table 1.

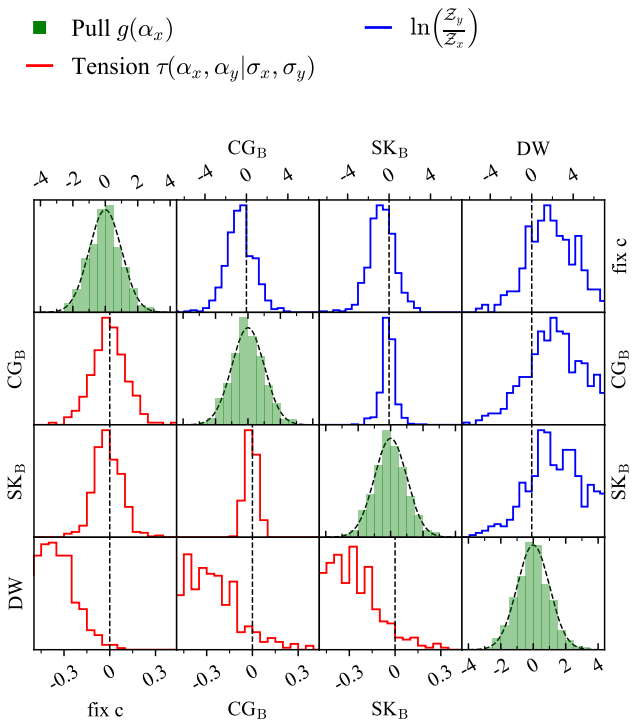


Figure 10. Same as Figure 9, but for the void-galaxy cross-2PCF.

(i) the DW model is the least likely to be true;
(ii) the parabolic model with a fixed c is slightly disfavoured with respect to the numerical models;
(iii) there is no preferential numerical model.

These observations can be naturally interpreted by analysing Figure 3:

(i) the DW model under-fits the exclusion wiggles;
(ii) the parabolic model is a better description of the wiggles than DW, but worse than the numerical models;
(iii) both numerical models follow similarly the exclusion feature up to $k = 0.6\,h\mathrm{Mpc}^{-1}$.

4.2 Robustness tests against systematic errors

In this section, we investigate the sensitivity of BAO measurements to possible systematic errors in the numerical models and the data. Initially, we examine the sensitivity of the measured α to the parameters of CosmoGAME and SICKLE by shifting them away from the fiducial values (see Table 4). As a result, the newly computed power spectra (defective models, SK$_{\rm def}$, CG$_{\rm def}$, see Figure 11) do not describe as well as the fiducial ones the reference clustering.

The second set of tests evaluates the robustness of the numerical models to potentially uncorrected systematic effects in the data. For example, the galaxy number density along the redshift is assumed to be isotropic, however, there are inhomogeneities across that sky, which means that the local number density of galaxies is not everywhere correctly estimated (see e.g. Appendix A of Zhao et al. 2021). This is important because a different matter density yields a different void size distribution (Zhao et al. 2016; Forero-Sánchez et al. 2022) that finally alters the exclusion pattern (Liang et al. 2016).

Another example of a systematic effect is the incompleteness in

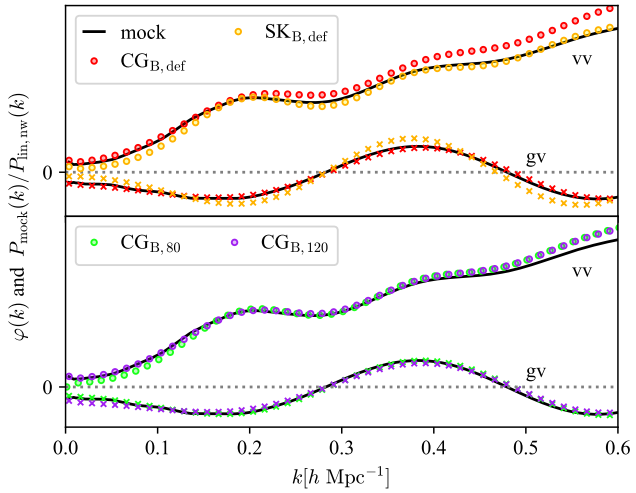


Figure 11. Same as Figure 3, but with different models. Upper panel: defectively calibrated numerical models. Lower panel: calibrated numerical models that are obtained from halo catalogues with a number density of 80% and 120% of the reference number density.

the data-set. For the SDSS data, on average, the incompleteness is lower than 5 per cent. In some sectors, the incompleteness can get as large as 50 per cent, but those regions cover small areas (Reid et al. 2016; Ross et al. 2020). Normally, these effects are included in the random and mock catalogues so that they compensate the ones in the data. However, the estimation of the galaxy number density might be imprecise, so the incompleteness effect might not be entirely removed. Consequently, we emulate these imprecise estimations by re-calibrating both codes' parameters (see Table 4), while asking for a halo number density that is different than the reference by −20 per cent (CG$_{80}$) and +20 per cent (CG$_{120}$). These considered differences are fairly conservative compared to the expected errors in galaxy density estimations.

Figure 12 shows how the numerical models shown in Figure 11 perform when the average void auto-2PCF (left) and the average void-galaxy cross-2PCF (right) from 500 mocks are fitted in different fitting ranges. On one hand, for the void auto-2PCF, the defective numerical models have generally a slightly larger bias compared to the fiducial ones (Figure 4), however most values remain within $\pm 0.1\sigma$ from zero. On the other hand, for the void-galaxy cross-2PCF, CG$_{\rm def}$ imposes a stronger bias on the measurement of α ($\sim 0.35\sigma$) than CG$_{\rm B}$, whereas SK$_{\rm def}$ decreases the bias from $\sim 0.2\sigma$ (SK$_{\rm B}$) to $\sim 0.15\sigma$. In the case of the void auto-2PCF, CG$_{80}$ and CG$_{120}$ remain within $\pm 0.1\sigma$ bias from zero. For the void-galaxy cross-2PCF, the bias induced by CG$_{80}$ is similar to the fiducial case, while CG$_{120}$ increases the bias to $\sim 0.3\sigma$.

Figure 13 contains a comparison between the results of the fiducial CG$_{\rm B}$ model and the CG$_{\rm def}$, CG$_{80}$ and CG$_{120}$ ones, for void auto-2PCF (in blue) and void-galaxy cross-2PCF (in red). In the case of the void auto-2PCF, the strongest tension occurs between CG$_{\rm B}$ and CG$_{\rm def}$, i.e. ~ 0.15 per cent or $\sim 0.15\sigma$ on average. In terms of the σ_α values, these three models are consistent with the fiducial CG$_{\rm B}$ within ± 1 per cent on average.

For SICKLE, we have only tested the sensitivity to the tuning parameters and we present the results in Figure 13. The bias introduced by SK$_{\rm def}$ with respect to the fiducial SK$_{\rm B}$ is on average −0.1 per cent

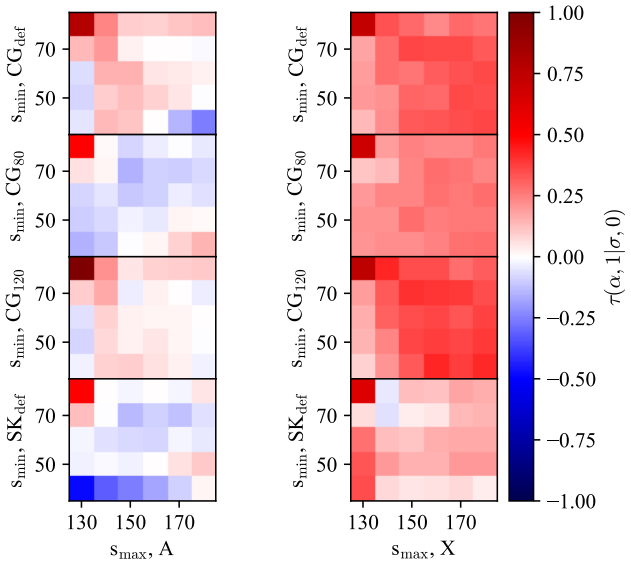


Figure 12. Comparison of different fitting ranges for four different cases using $\tau(\alpha, 1|\sigma, 0)$. Both the average void auto-2PCF (left) and void-galaxy cross-2PCF (right) – computed from 500 individual Patchy cubic mocks – are considered. The abbreviations are defined in Table 1.

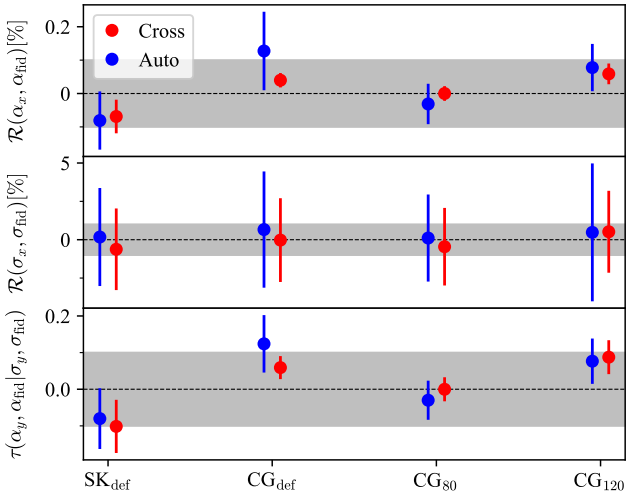


Figure 13. Comparison between the error-affected numerical models and the corresponding fiducial ones – CG_B or SK_B – using the clustering (blue for void auto-2PCF; red for void-galaxy cross-2PCF) computed from 500 individual pre-reconstructed Patchy cubic mocks. The abbreviations are defined in Table 1. First two panels: the relative difference (Eq. (34)) between the α values and σ_α values, respectively. Last panel: the tension from Eq. (33). The shown values and error bars are the averages and the standard deviations of 500 individual measurements. From top to bottom, the shaded areas delineate ±0.1%, ±1% and ±0.1σ, respectively.

or -0.1σ. In contrast, the uncertainties are consistent with fiducial case within ±1 per cent on average, as for CG.

Analysing the results of CG_{def}, CG_{120}, CG_{80} and SK_{def} in Figure 8, the average of the 500 α values is within $\sim$ ±0.1 per cent from the reference for four cases, while for the other four cases the bias is lower than $\sim$ 0.2 per cent. This suggests that even for larger survey such as DESI, the numerical models are robust enough to provide unbiased measurements of α.

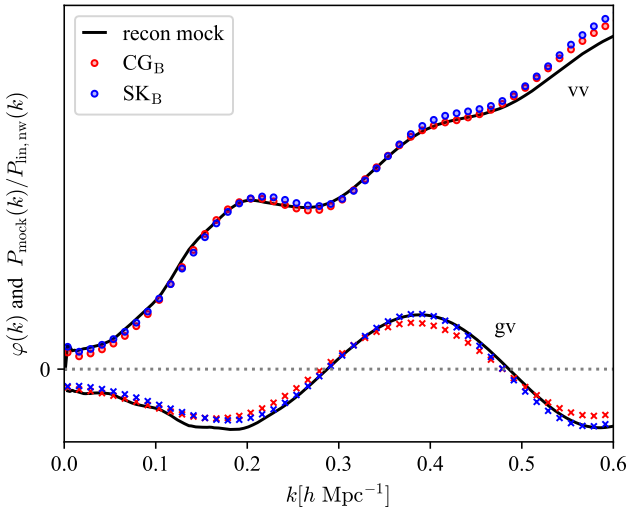


Figure 14. Same as Figure 3, but a different P_{mock} and only CG_B and SK_B. P_{mock} is computed from 500 Patchy reconstructed cubic mocks.

4.3 Robustness tests against BAO reconstruction

Figure 14 shows a comparison between the average power spectrum of 500 reconstructed Patchy catalogues and the numerical models presented in Figure 3, for both void auto-2PCF and void-galaxy cross-2PCF. It suggests that CG_B and SK_B can describe well the void clustering and be employed in BAO analysis.

After fitting the 500 individual void auto-2PCF (upper panel) and 500 void-galaxy cross-2PCF (lower panel), we compute the histogram of the pull $g(\alpha)$ values shown in Figure 15. In both cases, the distributions are consistent with a standard normal one (black dashed line), meaning fix c, CG_B and SK_B provide correct estimations of σ_α. Moreover, looking at Figure 8, the α_{avg} values corresponding to three previous models (orange points) are within ±0.1 per cent from the reference. One can also notice that for the void-galaxy cross-2PCF, the bias has systematically decreased by applying reconstruction on the galaxy catalogues, strengthening the observations of McCullagh et al. (2013); Neyrinck et al. (2018) that the gravitational evolution shifts the BAO peak of galaxies to lower separation.

Considering the fact that the reconstruction inverts the effect of the gravitational evolution and that the numerical models are based on Gaussian random fields – without any gravitational evolution – these models should describe better the reconstructed data. Thus, one should ideally calibrate the CosmoGAME and SICKLE for both post and pre-reconstructed data. Nonetheless, the current results show that the same set of void model power spectra (CG_B and SK_B) can be used in both scenarios.

4.4 Robustness tests against survey-geometry effects

In this subsection, we investigate the performance and robustness of the numerical models on light-cone data (described in Section 2.2). Given the smaller volume of the light-cone compared to the box, the correlation functions are noisier. Consequently, we have used 1000 Patchy realisations to reduce the noise. We have created two

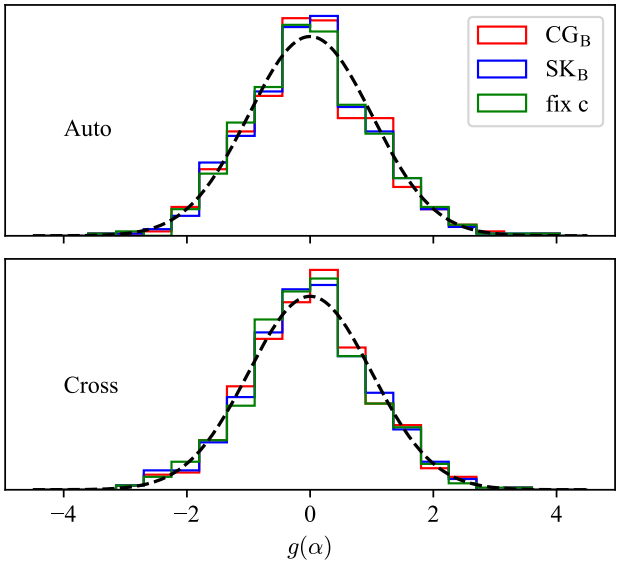


Figure 15. The histogram of the 500 pull $g(\alpha)$ values obtained from the individual fittings of the void auto-2PCF (upper panel) and void-galaxy cross-2PCF (lower panel) computed from reconstructed Patchy cubic mocks. The results are obtained using the two numerical models and the parabolic model with a fixed c parameter (coloured histograms, see Table 1 for abbreviations.). The black dashed line represents a standard normal distribution.

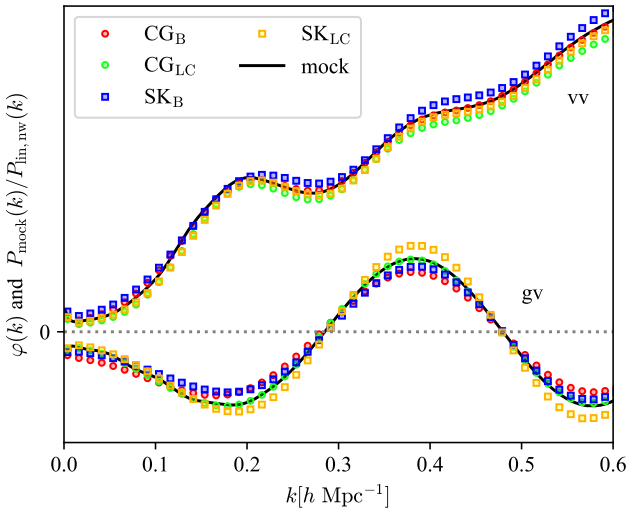


Figure 16. Same as Figure 3, but a different $P_{\rm mock}$ and additionally $\rm CG_{LC}$ and $\rm SK_{LC}$. $P_{\rm mock}$ is computed from 1000 Patchy light-cone mocks

additional numerical models ($\rm CG_{LC}$ and $\rm SK_{LC}$) by applying the survey-geometry on the cubic catalogues corresponding to $\rm CG_B$ and $\rm SK_B$. The resulting void model power spectra are shown in Figure 16.

At this stage, we only test $\rm CG_B$, $\rm SK_B$, $\rm CG_{LC}$, $\rm SK_{LC}$ and the parabolic model, given that the DW model is obviously insufficient to describe voids. Figure 17 shows similar results as Figure 4, most biases for the void auto-2PCF are within $[-0.1, 0.1]\sigma$ interval, while for the void-galaxy cross-2PCF, most values of the tension are lower than $+0.2\sigma$.

Studying the average of the 1000 α values in Figure 8, we observe

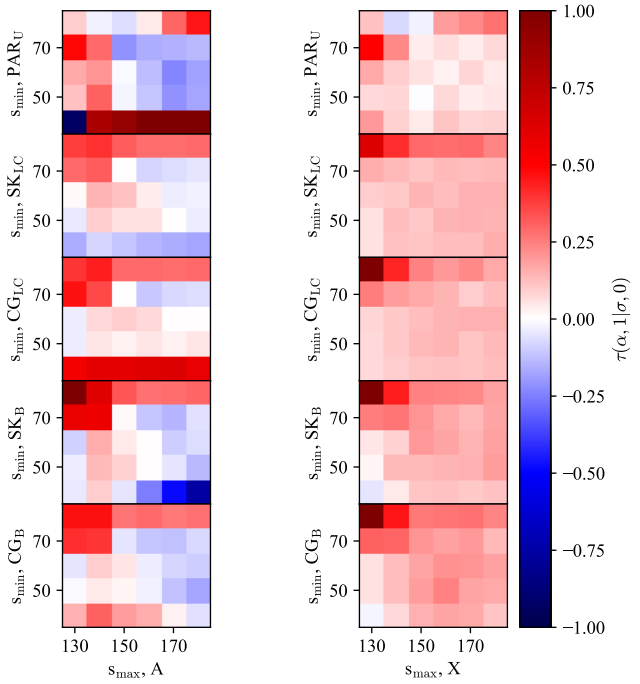


Figure 17. Comparison of different fitting ranges for five different cases using $\tau(\alpha, 1|\sigma, 0)$. Both the average void auto-2PCF (left) and void-galaxy cross-2PCF (right) – computed from 1000 individual Patchy light-cone mocks – are considered. The abbreviations are defined in Table 1.

that five points – $\rm SK_{LC}$, $\rm CG_{LC}$, $\rm CG_B$, $\rm SK_B$ for auto-2PCF and fix c for cross-2PCF)– are within ±0.1 per cent from the reference, while the remaining five are within ±0.2 per cent.

Analysing the tension parameter between the $\rm CG_{LC}$, $\rm CG_B$ and the fix-c models in Figures 18, 19, we observe that there is no significant tension: the mean values of the histograms are at most 0.1σ from 0, while the highest deviations are $\sim 0.3\sigma$. Moreover, the histograms of the 1000 pull $g(\alpha)$ values – diagonal panels of the same figures – additionally show that the uncertainties of α are correctly estimated by all models.

In terms of the most probable model for the void auto-2PCF, the logarithm of the Bayes Factor – upper diagonal panels of Figure 18 – suggests that the parabolic model with a fixed c parameter is slightly disfavoured against the numerical models. Furthermore, the light-cone numerical model is slightly preferred compared to the one constructed for boxes. In contrast, the results from void-galaxy cross-2PCF – Figure 19 – show that the parabolic model is slightly favoured with respect to the numerical models. Moreover, it shows that $\rm CG_{LC}$ is slightly disfavoured against the $\rm CG_B$.

We only show the results of CosmoGAME due to visibility reasons, however we have also analysed the results of SICKLE in Appendix C and shown that the same conclusions are available in this case. Moreover, there is no preference between the $\rm CG_{LC}$ and $\rm SK_{LC}$, nor between $\rm CG_B$ and $\rm SK_B$.

5 CONCLUSION

We have introduced two numerical techniques to model the DT void clustering: CosmoGAME and SICKLE. The main steps to construct the models are the following:

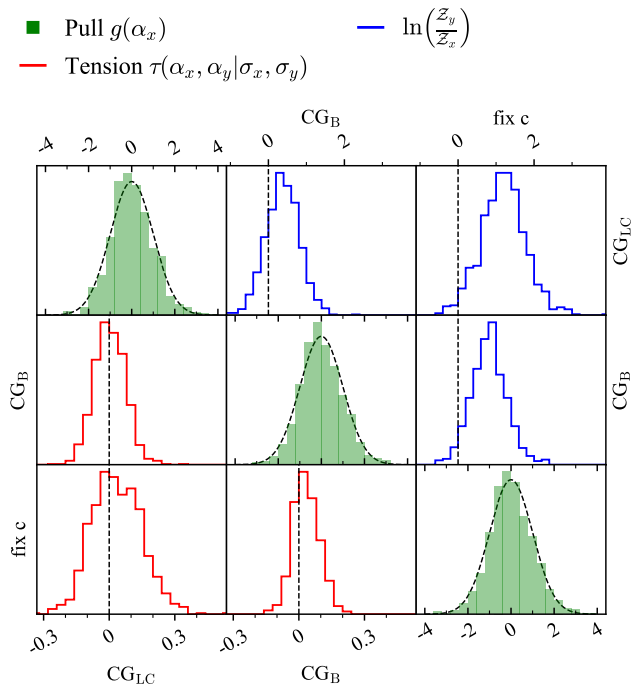


Figure 18. Same as Figure 9, but for 1000 void auto-2PCF computed from the Patchy light-cone mocks.

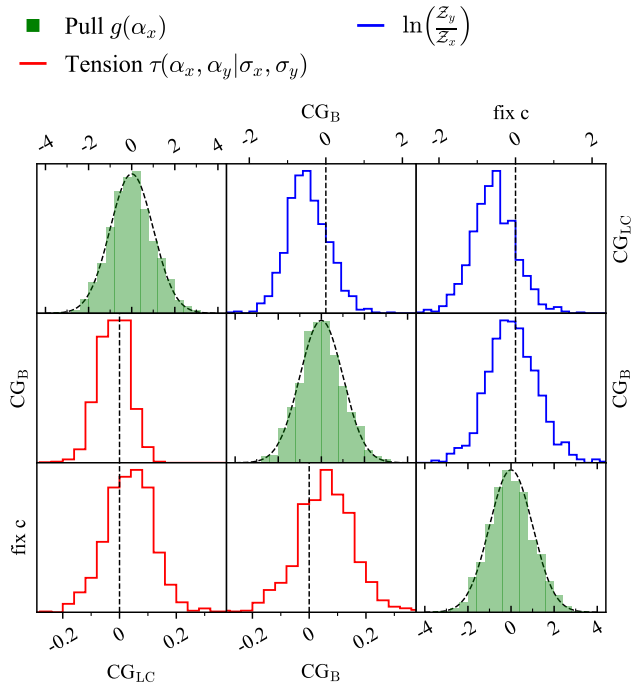


Figure 19. Same as Figure 9, but for 1000 void-galaxy cross-2PCF computed from the Patchy light-cone mocks.

- the initial conditions are built starting from a BAO free linear power spectrum;
- haloes are assigned directly on the density field corresponding to the initial conditions;
- voids are detected using DIVE;
- the void power spectrum is computed.

The difference between the two techniques lays into to the halo assignment process on the density field.

Furthermore, we have compared the performance of the two numerical models with the de-wiggled model of galaxies and a parabolic model introduced by Zhao et al. (2020) for the BAO analysis with DT voids. To this end, we have used 500 Patchy cubic mocks and 1000 Patchy light-cone mocks (similar to the BOSS DR12 LRG sample; Alam et al. 2015). On one hand, the de-wiggled model can bias the measurements of α by 0.4 to 0.7 per cent on average, when fitting the 2PCF from boxes. Thus, as also shown in Zhao et al. (2020), the de-wiggled model is not a viable model for voids. On the other hand, the parabolic model can provide unbiased results, however it tends to provide outlier values of α when the additional parameter c is not fixed. As a result, one has to fit the average of multiple mock 2PCF to precisely measure the value of c, so that it can be fixed when fitting individual 2PCF. Given that the cosmology of the mocks can be different from the one of the measured data, this might introduce a bias when fitting the clustering of data. In contrast, the numerical models can be directly calibrated on the void power spectrum computed from the measured data, as the exclusion pattern is much stronger than the noise.

By fitting the individual 2PCF from boxes, we have observed that the numerical models and the fixed c parabolic model are in agreement within $\sim 0.1\sigma$. Moreover, the histograms of the 500 values of $g(\alpha)$ are consistent with a standard normal distribution, meaning that all models estimate correctly the uncertainty of α. For the void auto-2PCF, the three models provide α values within ± 0.1 per cent from the reference, while for void-galaxy cross-2PCF the bias is below ~ 0.15 per cent. Studying the Bayes factor, the two numerical methods are favoured with respect to the parabolic model and there is no preferred numerical technique. Finally, the results provided by the two new models are less affected by the fitting range than the parabolic model.

We have analysed the robustness of the two numerical techniques to systematic errors such as incompleteness and defective calibration. The average of the 500 α values is within ~ 0.2 per cent from the reference value for all four cases affected by systematic effects. Thus, we can conclude that CosmoGAME and SICKLE are resilient to such systematic errors.

Given the fact that the BAO reconstruction is a standard procedure in BAO analysis, we study the behaviour of the two newly introduced techniques and the fixed c parabolic model on the reconstructed Patchy catalogues. We have observed that the values of α are consistent with one within ± 0.1 per cent and the uncertainty is well estimated, implying that CosmoGAME and SICKLE can be employed in modelling voids from both reconstructed and pre-reconstructed data-sets.

Lastly, we have tested CosmoGAME, SICKLE and the fixed c parabolic model on light-cones. In this case, the numerical models based on boxes have similar performances as the ones based on light-cones, i.e. uncertainties are well estimated and no tension between the models have been noticed. Slight discrepancies occur between the void auto-2PCF and void-galaxy cross-2PCF cases in terms of Bayes factors. For the void-auto 2PCF, the light-cone based numerical models have a higher evidence than the box based ones and all void model

power spectra are more likely to be correct than the parabolic model with a fixed c. In contrast, for void-galaxy cross-2PCF, the numerical models based on light-cones are slightly disfavoured against the ones based on boxes and the parabolic model. Analysing, the average of 1000 α values, we have noticed that most void model power spectra provide results within ±0.1 per cent from the reference and all of them are within ±0.2 per cent. This suggests that there is no bias introduced by the numerical models.

Even though, in the current case, the parabolic model with fixed c parameter has similar performances to the numerical models – in terms of estimating the α and its uncertainty – Tamone et al. (2022) have explained that for void quasars, that have a much stronger exclusion at even larger scales, the parabolic model cannot be used anymore. Therefore a better description of the void exclusion is necessary and the two numerical models can provide it. Moreover, the numerical models have the potential for even smaller biases due to the possibility of fine tuning the parameters to reach a better agreement at large values of k.

Finally, as explained by Zhao et al. (2020); Zhao et al. (2022), the combined 2PCF of voids and galaxies is preferred over multiple 2PCF due to a lower dimension of the data vector and thus a smaller required number of mocks. Consequently, for future studies, we will adapt the numerical models to the combined 2PCF for a multi-tracer cosmological analysis.

In conclusion, the usage of CosmoGAME or SICKLE in a BAO analysis with DT voids provides robust and unbiased measurements of the Alcock-Paczynski parameter. Moreover, the Bayes factor indicates a higher probability of these models to be true compared to the parabolic one. Nevertheless, we foresee the utility of these numerical methods in the study of different kind of voids or for different properties: e.g. void density contrast.

ACKNOWLEDGEMENTS

AV, CZ, DFS, AT acknowledge support from the Swiss National Science Foundation (SNF) "Cosmology with 3D Maps of the Universe" research grant, 200020_175751 and 200020_207379. FSK acknowledges the grants SEV-2015-0548, RYC2015-18693, and AYA2017-89891-P. CT is supported by Tsinghua University and sino french CNRS-CAS international laboratories LIA Origins and FCPPL.

DATA AVAILABILITY

The Patchy boxes used in this study can be provided upon request to CZ.

REFERENCES

Alam S., et al., 2015, ApJS, 219, 12
Alam S., et al., 2017, MNRAS, 470, 2617
Alam S., et al., 2021, Phys. Rev. D, 103, 083533
Alcock C., Paczynski B., 1979, Nature, 281, 358
Anderson L., et al., 2014, MNRAS, 441, 24
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Ata M., et al., 2017, MNRAS, 473, 4773
Baldauf T., Seljak U. c. v., Smith R. E., Hamaus N., Desjacques V., 2013, Phys. Rev. D, 88, 083507
Bautista J. E., et al., 2020, MNRAS, 500, 736
Beutler F., et al., 2017, MNRAS, 464, 3409
Burden A., Percival W. J., Howlett C., 2015, MNRAS, 453, 456
Busca N. G., et al., 2013, A&A, 552, A96
Casas-Miranda R., Mo H. J., Sheth R. K., Boerner G., 2002, MNRAS, 333, 730
Chan K. C., Hamaus N., 2021, Phys. Rev. D, 103, 043502
Chan K. C., Hamaus N., Desjacques V., 2014, Phys. Rev. D, 90, 103521
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2014, MNRAS, 446, 2621
Chuang C.-H., et al., 2019, MNRAS, 487, 48
Correa C. M., Paz D. J., Padilla N. D., Sánchez A. G., Ruiz A. N., Angulo R. E., 2022, MNRAS, 509, 1871
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036
Delaunay B., 1934, Bull. Acad. Sci. URSS, pp 793–800
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Eisenstein D. J., Seo H.-J., White M., 2007a, ApJ, 664, 660
Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007b, ApJ, 664, 675
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, The Open Journal of Astrophysics, 2, 10
Forero-Sánchez D., Zhao C., Tao C., Chuang C.-H., Kitaura F.-S., Variu A., Tamone A., Kneib J.-P., 2022, MNRAS, 513, 5407
Hamaus N., Wandelt B. D., Sutter P. M., Lavaux G., Warren M. S., 2014a, Phys. Rev. Lett., 112, 041304
Hamaus N., Sutter P. M., Wandelt B. D., 2014b, Phys. Rev. Lett., 112, 251302
Hamaus N., Pisani A., Sutter P. M., Lavaux G., Escoffier S., Wandelt B. D., Weller J., 2016, Phys. Rev. Lett., 117, 091302
Hamaus N., Pisani A., Choi J.-A., Lavaux G., Wandelt B. D., Weller J., 2020, J. Cosmology Astropart. Phys., 2020, 023
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hinshaw G., et al., 2003, ApJS, 148, 135
Kitaura F.-S., Heß S., 2013, MNRAS: Letters, 435, L78
Kitaura F.-S., Yepes G., Prada F., 2013, MNRAS: Letters, 439, L21
Kitaura F.-S., Gil-Marín H., Scóccola C. G., Chuang C.-H., Müller V., Yepes G., Prada F., 2015, MNRAS, 450, 1836
Kitaura F.-S., et al., 2016, Phys. Rev. Lett., 116, 171301
Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, MNRAS, 457, 4340
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Liang Y., Zhao C., Chuang C.-H., Kitaura F.-S., Tao C., 2016, MNRAS, 459, 4020
Mao Q., Berlind A. A., Scherrer R. J., Neyrinck M. C., Scoccimarro R., Tinker J. L., McBride C. K., Schneider D. P., 2017, ApJ, 835, 160
McCullagh N., Neyrinck M. C., Szapudi I., Szalay A. S., 2013, ApJ, 763, L14
Nadathur S., Carter P. M., Percival W. J., Winther H. A., Bautista J. E., 2019, Phys. Rev. D, 100, 023504
Neyrinck M. C., 2008, MNRAS, 386, 2101
Neyrinck M. C., Szapudi I., McCullagh N., Szalay A. S., Falck B., Wang J., 2018, MNRAS, 478, 2495
Padilla N. D., Ceccarelli L., Lambas D. G., 2005, MNRAS, 363, 977
Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132
Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19
Percival W. J., 2005, A&A, 443, 819
Planck Collaboration et al., 2020, A&A, 641, A6
Platen E., van de Weygaert R., Jones B. J. T., 2007, MNRAS, 380, 551
Prada F., Scóccola C. G., Chuang C.-H., Yepes G., Klypin A. A., Kitaura F.-S., Gottlöber S., Zhao C., 2016, MNRAS, 458, 613
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3 edn. Cambridge University Press, USA
Raichoor A., et al., 2020, MNRAS, 500, 3254
Reid B., et al., 2016, MNRAS, 455, 1553
Richard J., et al., 2019, The Messenger, 175, 50
Ross A. J., et al., 2017, MNRAS, 464, 1168
Ross A. J., et al., 2020, MNRAS, 498, 2354
Sefusatti E., Crocce M., Scoccimarro R., Couchman H. M. P., 2016, MNRAS, 460, 3624
Sheth R. K., van de Weygaert R., 2004, MNRAS, 350, 517

Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
Sutter P. M., Lavaux G., Wandelt B. D., Weinberg D. H., 2012, ApJ, 761, 187
Sutter P. M., et al., 2015, Astronomy and Computing, 9, 1
Szapudi I., Szalay A. S., 1997, arXiv e-prints, pp astro–ph/9704241
Tamone A., Zhao C., Forero-Sánchez D., Variu A., Chuang C. H., Kitaura F. S., Kneib J. P., Tao C., 2022, arXiv e-prints, p. arXiv:2208.06238
Vargas-Magaña M., et al., 2014, MNRAS, 445, 2
White M., Tinker J. L., McBride C. K., 2013, MNRAS, 437, 2594
Xu X., Padmanabhan N., Eisenstein D. J., Mehta K. T., Cuesta A. J., 2012, MNRAS, 427, 2146
Zhao C., 2023, arXiv e-prints, p. arXiv:2301.12557
Zhao C., Kitaura F.-S., Chuang C.-H., Prada F., Yepes G., Tao C., 2015, MNRAS, 451, 4266
Zhao C., Tao C., Liang Y., Kitaura F.-S., Chuang C.-H., 2016, MNRAS, 459, 2670
Zhao C., et al., 2020, MNRAS, 491, 4554
Zhao C., et al., 2021, MNRAS, 503, 1149
Zhao C., et al., 2022, MNRAS, 511, 5492
de Jong R. S., et al., 2019, The Messenger, 175, 3
de Mattia A., Ruhlmann-Kleider V., 2019, Journal of Cosmology and Astroparticle Physics, 2019, 036
van de Weygaert R., Platen E., 2011, in International Journal of Modern Physics Conference Series. pp 41–66 (arXiv:0912.2997), doi:10.1142/S2010194511000092

APPENDIX A: REDUCING THE NOISE OF THE NUMERICAL MODELS

Given the fact that each halo and void catalogues produced by CosmoGAME and SICKLE has an intrinsic noise, the measured power spectrum and its Hankel transform Eq. (9) are not smooth. In this section, we analyse how the number of realisations used to compute the void model power spectrum ($P_{\rm t,nw}(k)$) and the value of the damping factor a affect the Hankel transform of $P_{\rm t,nw}(k)$.

In Figure A1, one can see the 2PCF computed as the Hankel transform of the average void model power spectrum, for two different damping factors ($a = 1\ h^{-1}$Mpc and $a = 2\ h^{-1}$Mpc). The black curves in the upper panels represent best-fitting polynomials (BFP) of the $s^2\xi(s)$ curve – computed using Eq. (9) and the average of 2000 power spectrum realisations – for two different s intervals: $s \in (60, 150)\ h^{-1}$Mpc and $s \in (150, 200)\ h^{-1}$Mpc. The lower panels of Figure A1 contain the differences between $s^2\xi(s)$ curves and the BFP.

Apart from the visual inspection of the noise in the 2PCF, we also quantify it by computing:

$$\Phi = \frac{1}{n}\sum_{i=1}^{n}\left[s_i^2\xi(s_i) - \mathrm{BFP}(s_i)\right]^2, \tag{A1}$$

where n is the number of bins in the given interval and i is the index of the bin. One can observe from Figure A1 and Table A1 that the noise is drastically reduced when the number of realisations is increased from 100 to 2000.

As mentioned in Section 3.3.2, we need:

- a grid size of 2048^3 to measure the power spectrum for a large enough k interval,
- a large number of realisations to minimise the effect of the noise (cosmic-variance),

but achieving both conditions simultaneously is computationally-expensive. Thus, we create a stitched model by computing 2000 power spectra using a grid size of 512^3 (to decrease the noise at large scales) and 50 power spectra using a grid size of 2048^3 (to have a reasonably de-noised power spectrum up to a large value of k). Figure A1 and Table A1 suggest that the stitched model performs at least as well as the 500 case for the really large scales and reaches the precision of the 1000 case for the lower scales.

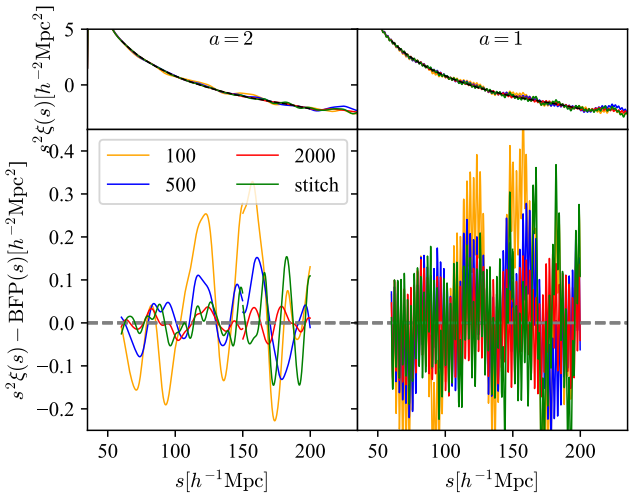


Figure A1. Upper panels: coloured curves - The result of the transformation expressed by Eq. (9) of the SICKLE power spectra computed as the average of 100, 500, 2000 realisations and by stitching the average of 2000 realisations with the one of 50 realisations (read text for details); black curve - the best-fitting polynomial of the red curve. Lower panels: the difference between the upper coloured curves and the black curve. The left and right panels correspond to a different damping parameter (Eq. (9)), i.e. $a = 2$ and $a = 1$, respectively.

One can also observe in Figure A1 and Table A1 that the damping factor a impacts the noise levels. By increasing it from $a = 1\ h^{-1}$Mpc to $a = 2\ h^{-1}$Mpc the amplitude of the noise is reduced by almost one order of magnitude. Consequently, we have tested whether the value of a can bias the measurement of α, by computing the tensions Eq. (33) between the α values corresponding to $a = 1\ h^{-1}$Mpc (α_1) and $a = 2\ h^{-1}$Mpc (α_2). Figure A2 shows that there is no tension between the two cases and for both cases, the histogram of the 500 $\tau(\alpha, 1|\sigma, 0)$ values are consistent with a standard-normal distribution, meaning there is no bias and the uncertainties are correctly estimated. Moreover, the relative difference $\rho_{\rm diff} = (\sigma_1 - \sigma_2)/[0.5 \times (\sigma_1 + \sigma_2)]$ shows that there is no bias in the uncertainty estimation between the two cases. Given the previous reasons, we fix $a = 2\ h^{-1}$Mpc in the current paper.

After fixing $a = 2\ h^{-1}$Mpc, we also test whether different number of realisations for the model power spectra and the stitching method affect the α measurements and the corresponding uncertainties. Figure A3 shows a comparison between the results of the model power spectra (SICKLE) computed from different number of realisations – 50, 100, 500, 1000, 2000 – and by stitching. We study three fitting scenarios:

(i) on the average of the 500 Patchy 2PCF with a rescaled covariance matrix (blue);
(ii) on the individual 2PCF, with the normal covariance matrix (red and green);
(iii) on the individual 2PCF, with the normal covariance matrix, but with a fixed $\Sigma_{\rm nl}$ (orange and cyan).

The shown α and σ_α corresponding to the three previous cases are, respectively: (i) the median of the posterior distribution and half the

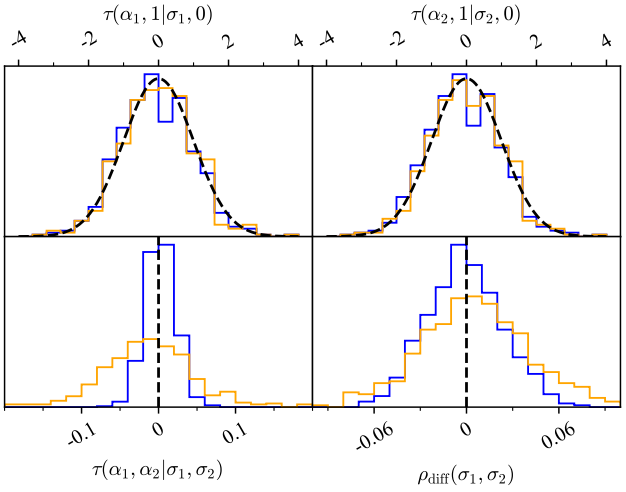


Figure A2. Upper panels: $(\alpha - 1)/\sigma$ for $a = 1\ h^{-1}$Mpc (α_1, σ_1, left) and $a = 2\ h^{-1}$Mpc (α_2, σ_2, right). Lower panels: the tension between the α values measured using $a = 1\ h^{-1}$Mpc and $a = 2\ h^{-1}$Mpc (left) and the relative difference between the uncertainties (σ) on α (right). Blue histograms: the results for the parabolic model with a $\mathrm{PAR_G}$ prior on c. Orange histograms: the results for the $\mathrm{CG_B}$ numerical model. a is the damping parameter from Eq. (9). The histograms contain the results of 500 individual 2PCF computed from Patchy cubic mocks.

Φ for $a = 1\ h^{-1}$Mpc	60–150 $\times 10^{-3}$	150–200 $\times 10^{-3}$
100	30.0	44.7
200	14.6	35.4
500	10.2	14.9
1000	7.60	9.60
2000	6.60	7.14
stitch	10.3	27.2
Φ for $a = 2\ h^{-1}$Mpc	60–150 $\times 10^{-4}$	150–200 $\times 10^{-4}$
100	189.0	306.0
200	50.0	235.0
500	29.5	73.1
1000	9.48	23.9
2000	3.56	6.16
stitch	8.66	74.5

Table A1. The Φ values defined in Eq. (A1) for two s intervals $s \in (60, 150)\ h^{-1}$Mpc and $s \in (150, 200)\ h^{-1}$Mpc and for two values of the damping factor $a = 1\ h^{-1}$Mpc and $a = 2\ h^{-1}$Mpc.

difference between the 84th and 16th percentiles; (ii) and (iii) the average and the standard deviation – divided by $\sqrt{500}$ – of the 500 α values (red and orange). Additionally, the cyan and the green points denote the mean of the 500 σ provided by the individual fittings of the 2PCF, divided by the $\sqrt{500}$. The uncertainties on the right panel from void auto-2PCF and void-galaxy cross-2PCF are divided by the corresponding blue σ_{2000}, which explains why the blue square and circle for the 2000 case are exactly positioned at one.

On one side, one can observe that starting from the '500' model, the α converges to the same value, for both void auto- and void-galaxy cross-2PCF and in all three fitting scenarios. On the other

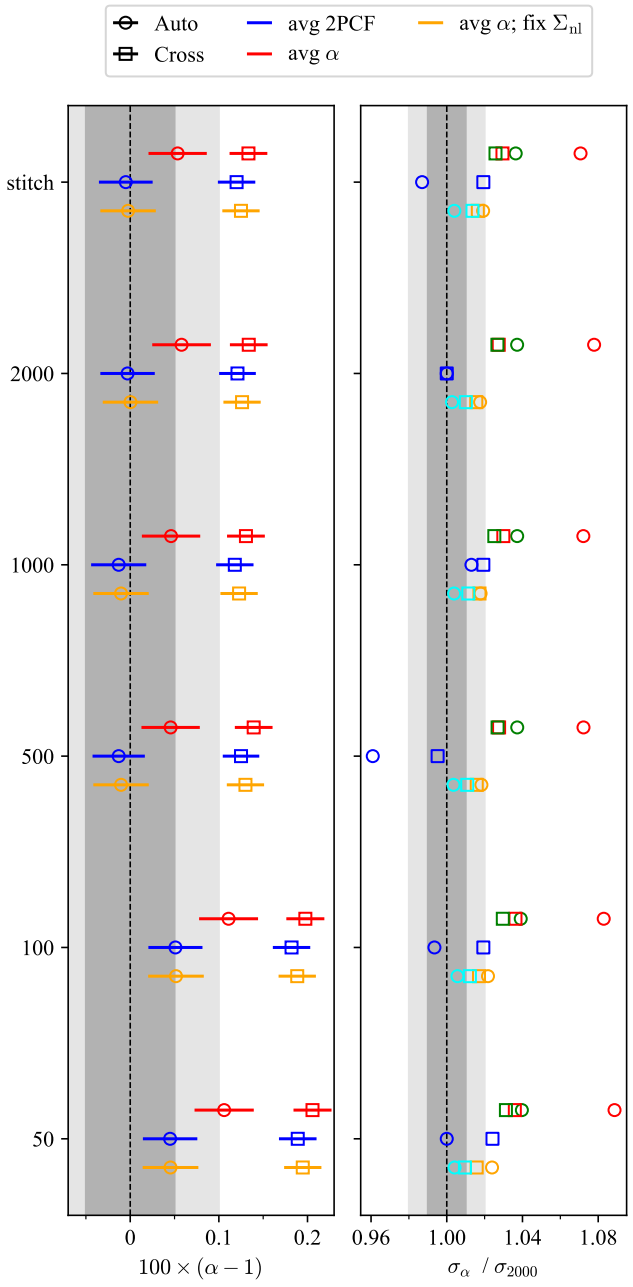


Figure A3. Comparison between the results of the model power spectra (SICKLE) computed from different number of realisations – 50, 100, 500, 1000, 2000 – and by stitching (see text), using the void auto- and void-galaxy cross-2PCF computed from 500 pre-reconstructed Patchy cubic mocks. First column shows the bias of α with respect one. The second column contains the ratios between different uncertainty estimations and the blue coloured σ_{2000}. The three colours denote the ways the fitting has been performed: blue - on the average of the 500 2PCF, with a rescaled covariance matrix (by 500), thus α is the median of the posterior distribution and σ_α is half the difference between the 84th and 16th percentiles; red and green - on the individual 2PCF, with the normal covariance matrix; orange and cyan - similarly to red and green, but with a fixed $\Sigma_{\rm nl}$. For red and orange, the shown α and σ_α are the average and the standard deviation – divided by $\sqrt{500}$ – of the 500 α values, respectively. For green and cyan, σ_α is the mean of the 500 σ provided by the individual fittings of the 2PCF, divided by the $\sqrt{500}$

	$a_0\,[10^{-4}]$	$a_1\,[10^{-2}\times h^{-1}\mathrm{Mpc}]$	$a_2\,[h^{-2}\mathrm{Mpc}^2]$
$\mathrm{CG_B}$	2.3 (4.9)	−8.7 (10)	7.6 (3.5)
$\mathrm{SK_B}$	5.4 (4.8)	−12 (−8.6)	7.4 (2.5)
$\mathrm{PAR_U}$	18 (0.69)	−47 (6.7)	31 (-7.9)

Table B1. The best-fitting nuisance parameters for three models. The fitting has been performed on the average void auto-2PCF and void-galaxy cross-2PCF (in brackets) computed from 500 pre-reconstructed Patchy boxes. The abbreviations are defined in Table 1.

side, all the ways to estimate the uncertainty provide σ_α values that are consistent within one to two per cent between all models and per method, except the '500' void auto-2PCF blue case, where the deviation is around four per cent. Consequently, the stitched method is chosen as the standard way to construct the void model power spectrum throughout this paper.

We also fit the individual 2PCF with a fixed $\Sigma_{\rm nl}$ – in Figure A3 because we have observed that the noise in the Patchy void 2PCF allows for larger values of $\Sigma_{\rm nl}$ to fit the data, which enlarges the posterior of α towards larger values. This slightly biases the measurement and overestimates the uncertainty. Given that throughout the paper we have not fixed $\Sigma_{\rm nl}$ for boxes, one has to consider this 0.05 per cent bias in the results of the main text.

APPENDIX B: THE STUDY OF THE NUISANCE PARAMETERS

Given the fact that the Least-Squares (LS) is much faster than PyMultiNest, in the main analysis, we use a two–fold approach in order to reduce the fitting time:

- PyMultiNest to fit α, B, $\Sigma_{\rm nl}$, c;
- LS to fit the nuisance parameters a_0, a_1, a_2.

In this section, we show that this approach does not bias the measurements of α, B, $\Sigma_{\rm nl}$, c and that there are no degeneracies between the nuisance parameters and α. To verify this, we fit the average void auto-2PCF and the average void-galaxy cross-2PCF computed from 500 pre-reconstructed Patchy cubic mocks, using a rescaled covariance matrix (i.e. divided by 500). Given that DW is not performing well, we only test the $\mathrm{CG_B}$, $\mathrm{SK_B}$ and $\mathrm{PAR_U}$ models.

Looking at the best-fitting nuisance parameters in Table B1, $\mathrm{SK_B}$ behaves similarly to $\mathrm{CG_B}$, thus we further focus on $\mathrm{CG_B}$ and $\mathrm{PAR_U}$. Figures B1, B2, B3 and B4 show the posterior distributions of the fitting parameters in two cases: red – all six or seven parameters are sampled by PyMultiNest; blue – the two–fold approach. In the first case, we used the following priors for the nuisance parameters: $p(a_0) = \mathcal{U}_{[-1,1]}(a_0)$, $p(a_1) = \mathcal{U}_{[-10,10]}(a_1)$ and $p(a_2) = \mathcal{U}_{[-100,100]}(a_2)$, that are wide enough to not influence the fitting results.

The same figures reveal that the measurements of α, B, $\Sigma_{\rm nl}$ and c are insensitive to the inclusion of the nuisance parameters in the PyMultiNest chain as the blue curves are consistent with the red ones. In the $\mathrm{PAR_U}$ case, there are slight degeneracies between α and a_1, a_2, however, they may be caused by the introduction of the c parameter and its strong degeneracy with a_1, a_2. In contrast, for $\mathrm{CG_B}$, α is not degenerate with the nuisance parameters. These results are consistent with the observations provided by Zhao et al. (2020); Zhao et al. (2022) and with the fact that the nuisance parameters should describe the broad-band shape.

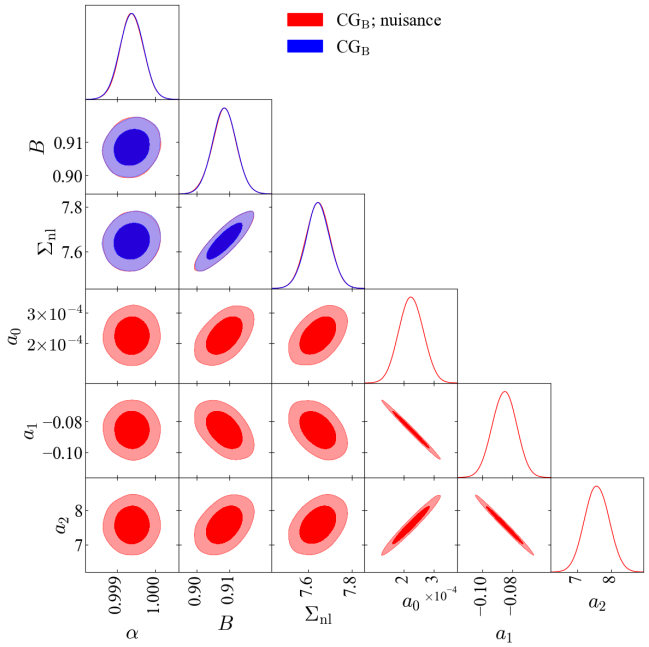


Figure B1. Triangle plot containing the posterior distributions of the fitting parameters described in Section 3.4.1. The fitting has been performed on the average void auto-2PCF computed from 500 pre-reconstructed Patchy cubic mocks using the $\mathrm{CG_B}$ numerical model. Red - all six parameters are given to PyMultiNest; Blue - only α, B and $\Sigma_{\rm nl}$ are given to PyMultiNest, while the nuisance parameters are fitted using a Least-Square method.

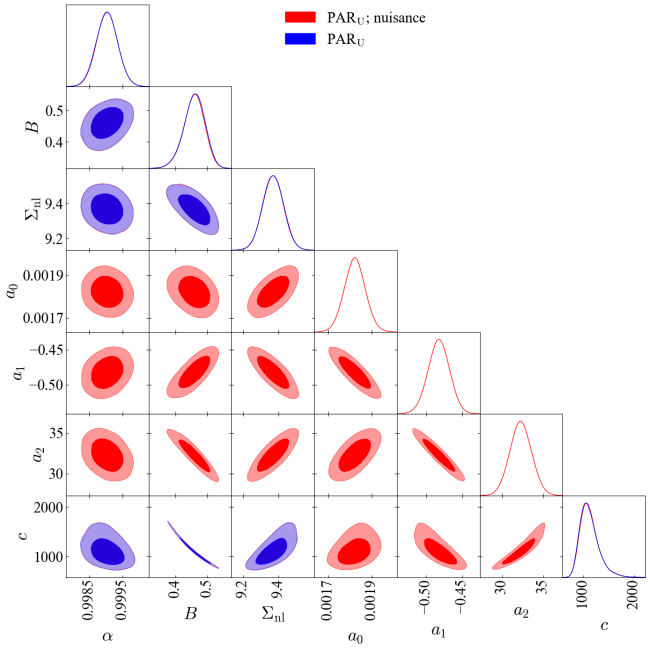


Figure B2. Same as Figure B1, but the model is $\mathrm{PAR_U}$.

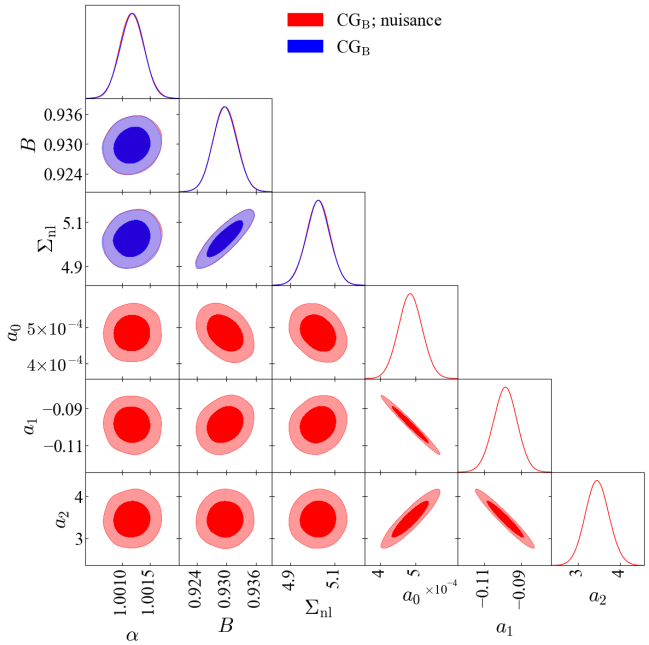


Figure B3. Same as Figure B1, but the reference is the average void-galaxy cross-2PCF.

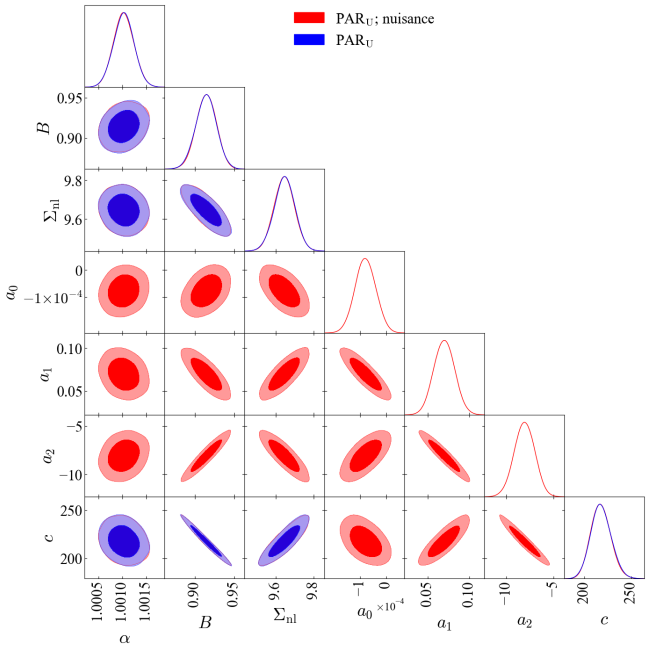


Figure B4. Same as Figure B1, but the model is $\mathrm{PAR_U}$ and the reference is the average void-galaxy cross-2PCF.

Consequently, we argue that one can safely use the combined PyMultiNest – LS approach in order to measure the fitting parameters.

APPENDIX C: LIGHT-CONE RESULTS

As mentioned in Section 4.4, we have only shown the results for CosmoGAME in the main text due to visibility reasons. Here, we show a comparison between all models $\mathrm{CG_B}$, $\mathrm{SK_B}$, $\mathrm{CG_{LC}}$, $\mathrm{SK_{LC}}$ and parabolic model with fixed c.

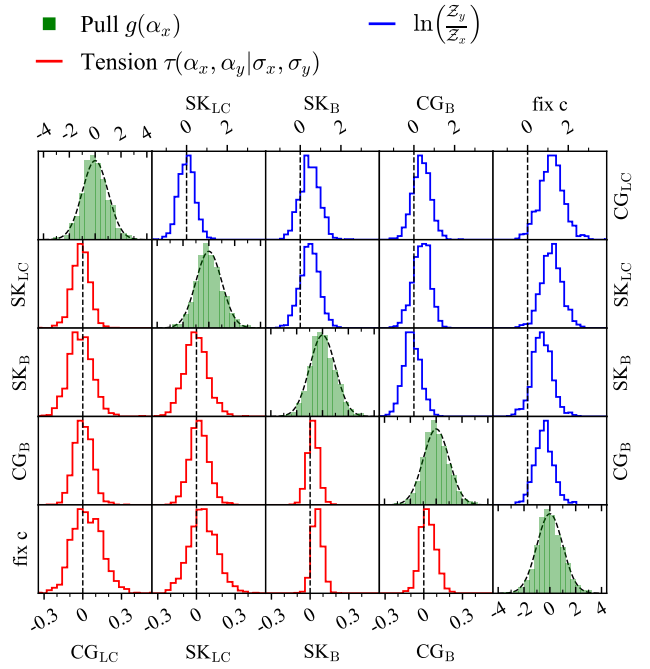


Figure C1. Diagonal panels: green - the histograms of the pull function $g(\alpha_x)$ values, Eq. (35); black - standard normal distributions. Lower triangular plots: the values of $\tau(\alpha_x, \alpha_y|\sigma_x, \sigma_y)$, Eq. (33), for all combinations of models. Upper triangular plot: the natural logarithm of the Bayes Factor $\ln\left(\mathcal{Z}_y/\mathcal{Z}_x\right)$ (see Section 3.4.4.1). The results correspond to the individual fittings of the 1000 void auto-2PCF computed from the Patchy light-cone mocks. The abbreviations are defined in Table 1.

Studying the tension in the lower diagonal plots of Figures C1 and C2, we observe that the box-based models and the light-cone based models provide highly consistent results. There is however a slight bias of the order of 0.1σ between the fixed c parabola and the numerical models. All models estimate accurately the uncertainty of α. The logarithm of the Bayes factor suggests that for the void auto-2PCF, the fixed c parabola is slightly disfavoured against the numerical models, while for the void-galaxy cross-2PCF, the reverse is true. Moreover, there are no siginificant differences between $\mathrm{CG_B}$ and $\mathrm{SK_B}$, nor between $\mathrm{CG_{LC}}$ and $\mathrm{SK_{LC}}$. Lastly, for the void auto-2PCF, the light-cone numerical models are slightly preferred compared to the ones constructed for boxes, while the opposite is valid for the void-galaxy cross-2PCF.

This paper has been typeset from a TEX/LATEX file prepared by the author.

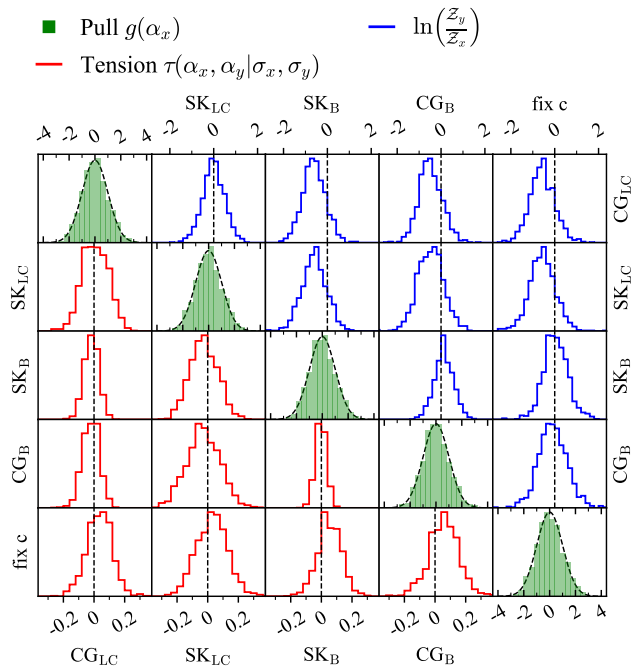


Figure C2. Same as Figure C1, but for void-galaxy cross-2PCF.

4 Conclusion

In 1929, Edwin Hubble published his work in which he measured the distances and radial velocities of 24 "Extra-Galactic Nebulae" (i.e. galaxies) providing compelling evidence for the expansion of the Universe. Seven decades later, two teams of astrophysicists extended this work by measuring the distances of tens of type-Ia supernovae up to a redshift of one ($\approx$ 10 billion light-years away Earth) and established that the expansion of the Universe was, in fact, accelerating. This finding was consistent with an expansion driven by a dark energy described by the famous cosmological constant Λ.

Today, thirty years after the discovery of the cosmic acceleration, the Λ Cold Dark Matter (ΛCDM) represents the standard cosmological paradigm. The high-precision measurements of the Cosmic Microwave Background (CMB) temperature anisotropies remain consistent with a flat ΛCDM Universe. Nevertheless, there are numerous open questions about the nature of dark energy that could potentially find answers through even more precise measurements.

The study of large-scale structure provides a third dimension to the CMB measurements, using the Baryon Acoustic Oscillations (BAO) as a standard ruler to measure distances at different redshifts and thus probing different epochs in the history of the Universe. Over the past two decades, the number of 3D-mapped galaxies and quasars has exponentially increased. This culminated in 2020 with the public release of a 3D map by the Sloan Digital Sky Survey (SDSS), containing over two million galaxies and quasars.

Recently, the on-going Dark Energy Spectroscopic Instrument (DESI) has published a map of over one million galaxies and quasars measured over a six-month period. Ultimately, DESI aims to create a 3D map of approximately 40 million extragalactic sources across 10 billion years of cosmic history, covering one-third of the sky during its five years of operation. Since the combined CMB+SDSS BAO measurements have already shown great improvements in precision, DESI has the potential to provide insights about the dark energy.

During my thesis, I have contributed to different stages of a large-scale structure study. Being part of DESI's Cosmological Simulations Working Group (CosmoSimsWG), I have participated

to the DESI mock challenge project that aims to compare different methodologies for constructing covariance matrices. This is an important task in the epoch of precision cosmology, as the systematic effects can become significant.

For this purpose, I have used a Halo Occupation Distribution (HOD) model to assign galaxies to the dark matter haloes identified in the FASTPM dark matter simulation. Furthermore, I have assessed the impact of the HOD fitting on the resulting covariance matrices of the galaxy clustering. The challenge has been to reproduce the two-point clustering statistics of a full N-body simulation, considering that FASTPM is an approximate method to gravitationally evolve the dark matter field.

Finally, I have built FASTPM galaxy catalogues whose two-point clustering is consistent with the full N-body reference, within the expected uncertainty of the DESI Year 1 ELG dataset. In addition, the resulting galaxy three-point clustering is in reasonable agreement with the reference one, without actually including it in the fitting process. Lastly, the estimated covariance matrices are robust against the details of the HOD fitting at scales of interest for BAO and RSD studies. This technique has the potential to provide galaxy catalogues that are accurate within the expected precision requirements of the entire DESI dataset. Thus it represents a tool for high-precision tests of systematic effects and for building high-precision covariance matrices, however further studies must be performed.

Part of CosmoSim, I have also contributed to the construction of the First Generation Mocks for DESI. I have been tasked to convert thousands of cubic simulations into light-cones and apply the survey geometry. This is a crucial step in creating realistic simulations that are needed for the final covariance matrix estimation.

A further step in a large-scale structure analysis involves testing and improving models. To this end, I have co-developed a numerical model for Delaunay Triangulation (DT) voids and conducted robustness tests on different methods to model the broadband shape of the DT void clustering statistics. I have shown the importance of properly accounting for the exclusion effect in the DT void modelling: the galaxy clustering model can bias the measurement of the α parameter by 0.7 per cent. This bias is highly significant given the expected sub-percent precision of surveys like DESI. In contrast, the two new numerical techniques recover α within 0.2 per cent for measurements on both cubic and light-cone simulations. Additionally, these two numerical methods are resilient against systematic effects, such as incompleteness and defective calibration.

Furthermore, the Bayesian analysis suggests that these two numerical models have similar probabilities of being true, but they are more likely to be correct than the previous models. Lastly, we foresee that these two new numerical techniques could be used to model different kinds of voids.

The final step in large-scale structure analysis is to measure the cosmological parameters by applying the set of tools that has been developed and tested on simulations to real data. To

this end, I have contributed to tuning the numerical model necessary for describing the DT void clustering in the latest multi-tracer BAO analysis of DT voids and galaxies based on the most recent SDSS data release. We have shown that the combined study of voids and galaxies improves the constraints on the cosmological parameters: H_0, $\Omega_{0\mathrm{m}}$ and $\Omega_{0\Lambda}h^2$ by approximately 6, 6 and 17 per cent, respectively. Therefore, in the domain of precision cosmology, the multi-tracer analysis of galaxies and DT voids – modelled by the robust numerical techniques – represents a valuable resource that will enhance the precision of DESI and future large-scale structure surveys measurements.

Bibliography

Agarwal S., Feldman H. A., 2011, MNRAS, 410, 1647

Alam S., et al., 2021, Phys. Rev. D, 103, 083533

Albrecht A., Steinhardt P. J., 1982, Phys. Rev. Lett., 48, 1220

Alcock C., Paczynski B., 1979, Nature, 281, 358

Alexander D. M., et al., 2023, AJ, 165, 124

Amiri M., et al., 2023, ApJ, 947, 16

Anderson L., et al., 2012, MNRAS, 427, 3435

Anderson L., et al., 2014, MNRAS, 439, 83

Angulo R. E., Hahn O., 2022, Living Reviews in Computational Astrophysics, 8, 1

Astropy Collaboration et al., 2022, ApJ, 935, 167

Aubert M., et al., 2022, MNRAS, 513, 186

Balaguera-Antolínez A., Kitaura F.-S., Pellejero-Ibáñez M., Zhao C., Abel T., 2019, MNRAS, 483, L58

Balaguera-Antolínez A., et al., 2020, MNRAS, 491, 2565

Bautista J. E., et al., 2021, MNRAS, 500, 736

Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109

Bennett C. L., et al., 2003, ApJ, 583, 1

Bernardeau F., 1994, ApJ, 427, 51

Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1

Beutler F., et al., 2017, MNRAS, 466, 2242

Blake C., Carter P., Koda J., 2018, MNRAS, 479, 5168

Blas D., Lesgourgues J., Tram T., 2011, J. Cosmology Astropart. Phys., 2011, 034

Brout D., et al., 2022, ApJ, 938, 110

Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21

Burden A., Percival W. J., Manera M., Cuesta A. J., Vargas Magana M., Ho S., 2014, MNRAS, 445, 3152

Burden A., Percival W. J., Howlett C., 2015, MNRAS, 453, 456

CHIME Collaboration et al., 2022, ApJS, 261, 29

Carlson J., White M., 2010, ApJS, 190, 311

Carroll S. M., 1997, arXiv e-prints, pp gr–qc/9712019

Carter P., Beutler F., Percival W. J., DeRose J., Wechsler R. H., Zhao C., 2020, MNRAS, 494, 2076

Cautun M., van de Weygaert R., Jones B. J. T., 2013, MNRAS, 429, 1286

Chan K. C., Hamaus N., 2021, Phys. Rev. D, 103, 043502

Chaussidon E., et al., 2023, ApJ, 944, 107

Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621

Chuang et al. 2023, in preparation

Colberg J. M., Sheth R. K., Diaferio A., Gao L., Yoshida N., 2005, MNRAS, 360, 216

Colberg J. M., et al., 2008, MNRAS, 387, 933

Cole S., et al., 2005, MNRAS, 362, 505

Cooke R. J., Pettini M., Steidel C. C., 2018, ApJ, 855, 102

Crichton D., et al., 2022, Journal of Astronomical Telescopes, Instruments, and Systems, 8, 011019

DESI Collaboration et al., 2016a, arXiv e-prints, p. arXiv:1611.00036

DESI Collaboration et al., 2016b, arXiv e-prints, p. arXiv:1611.00037

DESI Collaboration et al., 2022, AJ, 164, 207

DESI Collaboration et al., 2023a, arXiv e-prints, p. arXiv:2306.06307

DESI Collaboration et al., 2023b, arXiv e-prints, p. arXiv:2306.06308

Davis T. M., Lineweaver C. H., 2004, Publ. Astron. Soc. Australia, 21, 97

Davis M., Peebles P. J. E., 1983, ApJ, 267, 465

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

Dawson K. S., et al., 2013, AJ, 145, 10

Dawson K. S., et al., 2016, AJ, 151, 44

Delaunay B., 1934, Bull. Acad. Sci. URSS, pp 793–800

Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1

Dey A., et al., 2019, AJ, 157, 168

Dodelson S., 2003, Modern cosmology. Amsterdam (Netherlands): Academic Press. ISBN 0-12-219141-2, 2003, XIII + 440 p.

Dodelson S., Schmidt F., 2020, Modern Cosmology. Academic Press, doi:10.1016/C2017-0-01943-2

Eisenstein D. J., Hu W., 1998, ApJ, 496, 605

Eisenstein D. J., et al., 2005, ApJ, 633, 560

Eisenstein D. J., Seo H.-J., White M., 2007a, ApJ, 664, 660

Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007b, ApJ, 664, 675

El-Ad H., Piran T., 1997, ApJ, 491, 421

Euclid Collaboration et al., 2020, A&A, 642, A191

Euclid Collaboration et al., 2022, A&A, 662, A112

Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23

Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273

Feynman R. P., Leighton R., Sands M., 2006, The Feynman Lectures on Physics. Addison-Wesley

Fixsen D. J., 2009, ApJ, 707, 916

Forero-Sánchez D., Zhao C., Tao C., Chuang C.-H., Kitaura F.-S., Variu A., Tamone A., Kneib J.-P., 2022, MNRAS, 513, 5407

Foster C., Nelson L. A., 2009, ApJ, 699, 1252

Gaia Collaboration et al., 2016, A&A, 595, A1

Gamow G., 1946, Phys. Rev., 70, 572

Gamow G., 1948, Nature, 162, 680

Gil-Marín H., Wagner C., Verde L., Porciani C., Jimenez R., 2012, J. Cosmology Astropart. Phys., 2012, 029

Gil-Marín H., et al., 2020, MNRAS, 498, 2492

Gunn J. E., et al., 2006, AJ, 131, 2332

Guth A. H., 1981, Phys. Rev. D, 23, 347

Hadzhiyska B., Eisenstein D., Bose S., Garrison L. H., Maksimova N., 2022, MNRAS, 509, 501

Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, MNRAS, 375, 489

Hahn C., et al., 2023, AJ, 165, 253

Hamaus N., Wandelt B. D., Sutter P. M., Lavaux G., Warren M. S., 2014, Phys. Rev. Lett., 112, 041304

Hamaus N., Pisani A., Sutter P. M., Lavaux G., Escoffier S., Wandelt B. D., Weller J., 2016, Phys. Rev. Lett., 117, 091302

Hamilton A. J. S., 1993, ApJ, 417, 19

Hand N., Feng Y., Beutler F., Li Y., Modi C., Seljak U., Slepian Z., 2018, AJ, 156, 160

Harrison E. R., 1970, Phys. Rev. D, 1, 2726

Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399

Hawken A. J., Aubert M., Pisani A., Cousinou M.-C., Escoffier S., Nadathur S., Rossi G., Schneider D. P., 2020, J. Cosmology Astropart. Phys., 2020, 012

Hinshaw G., et al., 2013, ApJS, 208, 19

Hogg D. W., 1999, arXiv e-prints, pp astro–ph/9905116

Hoyle F., Vogeley M. S., 2001, arXiv e-prints, pp astro–ph/0110449

Hoyle F., Vogeley M. S., 2002, ApJ, 566, 641

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

Huynh M., Lazio J., 2013, arXiv e-prints, p. arXiv:1311.4288

Ishak M., 2019, Living Reviews in Relativity, 22, 1

Ishikawa K., Sunayama T., Nishizawa A. J., Miyatake H., Nishimichi T., 2023, arXiv e-prints, p. arXiv:2306.01696

Ivezić Ž., et al., 2019, ApJ, 873, 111

Jackson J. C., 1972, MNRAS, 156, 1P

Jackson J. D., 1999, Classical electrodynamics. New York : J. Wiley & Sons

Jeong D., Komatsu E., 2006, ApJ, 651, 619

Kaiser N., 1987, MNRAS, 227, 1

Kamionkowski M., Verde L., Jimenez R., 2009, J. Cosmology Astropart. Phys., 2009, 010

Karamanis M., Beutler F., 2021, arXiv e-prints, p. arXiv:2106.06331

Kerscher M., Szapudi I., Szalay A. S., 2000, ApJ, 535, L13

Kitaura F. S., Hess S., 2013, MNRAS, 435, L78

Kitaura F.-S., Yepes G., Prada F., 2013, MNRAS: Letters, 439, L21

Kitaura F.-S., et al., 2016, Phys. Rev. Lett., 116, 171301

Kittel C., 1973, Mechanics. McGraw-Hill

Knebe A., et al., 2011, MNRAS, 415, 2293

Kreisch C. D., Pisani A., Carbone C., Liu J., Hawken A. J., Massara E., Spergel D. N., Wandelt B. D., 2019, MNRAS, 488, 4413

Lan T.-W., et al., 2023, ApJ, 943, 68

Landy S. D., Szalay A. S., 1993, ApJ, 412, 64

Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193

Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473

Liang Y., Zhao C., Chuang C.-H., Kitaura F.-S., Tao C., 2016, MNRAS, 459, 4020

Linde A. D., 1982, Physics Letters B, 108, 389

Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, MNRAS, 508, 4017

Mao Q., Berlind A. A., Scherrer R. J., Neyrinck M. C., Scoccimarro R., Tinker J. L., McBride C. K., Schneider D. P., 2017, ApJ, 835, 160

Mohammad F. G., Percival W. J., 2022, MNRAS, 514, 1289

Mohayaee R., Mathis H., Colombi S., Silk J., 2006, MNRAS, 365, 939

Moon J., et al., 2023, arXiv e-prints, p. arXiv:2304.08427

Neyrinck M. C., 2008, MNRAS, 386, 2101

Nicola A., et al., 2023, arXiv e-prints, p. arXiv:2307.03226

Nishizawa A. J., Hsieh B.-C., Tanaka M., Takata T., 2020, arXiv e-prints, p. arXiv:2003.01511

Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009, MNRAS, 396, 19

Oke J. B., Gunn J. E., 1983, ApJ, 266, 713

Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132

Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton University Press

Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19

Peebles P. J. E., Yu J. T., 1970, ApJ, 162, 815

Pellejero-Ibañez M., et al., 2020, MNRAS, 493, 586

Penzias A. A., Wilson R. W., 1965a, ApJ, 142, 419

Penzias A. A., Wilson R. W., 1965b, ApJ, 142, 1149

Percival W. J., 2013, arXiv e-prints, p. arXiv:1312.5490

Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053

Percival W. J., Friedrich O., Sellentin E., Heavens A., 2022, MNRAS, 510, 3207

Perico E. L. D., Voivodic R., Lima M., Mota D. F., 2019, A&A, 632, A52

Perlmutter S., et al., 1999, ApJ, 517, 565

Planck Collaboration et al., 2016, A&A, 594, A13

Planck Collaboration et al., 2020a, A&A, 641, A1

Planck Collaboration et al., 2020b, A&A, 641, A6

Planck Collaboration et al., 2020c, A&A, 641, A9

Platen E., van de Weygaert R., Jones B. J. T., 2007, MNRAS, 380, 551

Pons-Bordería M.-J., Martínez V. J., Stoyan D., Stoyan H., Saar E., 1999, ApJ, 523, 480

Pospelov M., Pradler J., 2010, Annual Review of Nuclear and Particle Science, 60, 539

Raichoor A., et al., 2020, Research Notes of the American Astronomical Society, 4, 180

Raichoor A., et al., 2023, AJ, 165, 126

Rampf C., Hahn O., 2021, MNRAS, 501, L71

Refsdal S., 1964, MNRAS, 128, 307

Rhee G., 1991, Nature, 350, 211

Rich J., 2010, Fundamentals of Cosmology. Springer Berlin, Heidelberg, doi:10.1007/978-3-642-02800-7

Richard J., et al., 2019, The Messenger, 175, 50

Riess A. G., et al., 1998, AJ, 116, 1009

Rood H. J., 1988, ARA&A, 26, 245

Ross A. J., Samushia L., Howlett C., Percival W. J., Burden A., Manera M., 2015, MNRAS, 449, 835

Ross A. J., et al., 2020, MNRAS, 498, 2354

Rubin V. C., Ford W. Kent J., 1970, ApJ, 159, 379

Ruiz-Macias O., et al., 2020, Research Notes of the American Astronomical Society, 4, 187

Sachs R. K., Wolfe A. M., 1967, ApJ, 147, 73

Sailer N., Castorina E., Ferraro S., White M., 2021, J. Cosmology Astropart. Phys., 2021, 049

Schaap W. E., van de Weygaert R., 2000, A&A, 363, L29

Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Schlegel D. J., et al., 2022, arXiv e-prints, p. arXiv:2209.04322

Schmidt B. P., et al., 1998, ApJ, 507, 46

Schmittfull M., Feng Y., Beutler F., Sherwin B., Chu M. Y., 2015, Phys. Rev. D, 92, 123522

Schutz B., 2009, A First Course in General Relativity. Cambridge University Press

Sefusatti E., Crocce M., Scoccimarro R., Couchman H. M. P., 2016, MNRAS, 460, 3624

Seo H.-J., Ota A., Schmittfull M., Saito S., Beutler F., 2022, MNRAS, 511, 1557

Shajib A. J., et al., 2020, MNRAS, 494, 6072

Shandarin S., Habib S., Heitmann K., 2012, Phys. Rev. D, 85, 083005

Sheth R. K., van de Weygaert R., 2004, MNRAS, 350, 517

Silber J. H., et al., 2023, AJ, 165, 9

Silk J., 1968, ApJ, 151, 459

Slosar A., et al., 2019, in Bulletin of the American Astronomical Society. p. 53 (arXiv:1907.12559), doi:10.48550/arXiv.1907.12559

Smoot G. F., et al., 1992, ApJ, 396, L1

Somerville R. S., Lee K., Ferguson H. C., Gardner J. P., Moustakas L. A., Giavalisco M., 2004, ApJ, 600, L171

Square Kilometre Array Cosmology Science Working Group et al., 2020, Publ. Astron. Soc. Australia, 37, e007

Sutter P. M., Lavaux G., Wandelt B. D., Weinberg D. H., 2012, ApJ, 761, 187

Sutter P. M., et al., 2015, Astronomy and Computing, 9, 1

Tamone A., Zhao C., Forero-Sánchez D., Variu A., Chuang C. H., Kitaura F. S., Kneib J. P., Tao C., 2022, arXiv e-prints, p. arXiv:2208.06238

Tauber J. A., et al., 2010, A&A, 520, A1

Tosone F., Neyrinck M. C., Granett B. R., Guzzo L., Vittorio N., 2021, MNRAS, 505, 2999

Vargas-Magaña M., et al., 2013, A&A, 554, A131

Vargas-Magaña M., et al., 2014, MNRAS, 445, 2

Variu A., et al., 2023a, arXiv e-prints, p. arXiv:2307.14197

Variu A., Zhao C., Forero-Sánchez D., Chuang C.-H., Kitaura F.-S., Tao C., Tamone A., Kneib J.-P., 2023b, MNRAS, 521, 4731

Wadekar D., Scoccimarro R., 2020, Phys. Rev. D, 102, 123517

Wadekar D., Ivanov M. M., Scoccimarro R., 2020, Phys. Rev. D, 102, 123521

Walsh D., Carswell R. F., Weymann R. J., 1979, Nature, 279, 381

Wang Y., Zhao G.-B., 2020, Research in Astronomy and Astrophysics, 20, 158

Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405

Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435

Weinberg S., 1972, Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity. New York : J. Wiley & Sons

Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Phys. Rep., 530, 87

White M., 2014, MNRAS, 439, 3630

White M., Hu W., 1997, A&A, 321, 8

Workman R. L., et al., 2022, Progress of Theoretical and Experimental Physics, 2022, 083C01

Xu X., Padmanabhan N., Eisenstein D. J., Mehta K. T., Cuesta A. J., 2012, MNRAS, 427, 2146

Xu X., Cuesta A. J., Padmanabhan N., Eisenstein D. J., McBride C. K., 2013, MNRAS, 431, 2834

Yang L. F., Neyrinck M. C., Aragón-Calvo M. A., Falck B., Silk J., 2015, MNRAS, 451, 3606

Yèche C., et al., 2020, Research Notes of the American Astronomical Society, 4, 179

York D. G., et al., 2000, AJ, 120, 1579

Yu J., et al., 2023, arXiv e-prints, p. arXiv:2306.06313

Yuan S., et al., 2023, arXiv e-prints, p. arXiv:2306.06314

Zarrouk P., et al., 2021, Mon. Not. Roy. Astron. Soc., 503, 2562

Zel'dovich Y. B., 1970, A&A, 5, 84

Zeldovich Y. B., 1972, MNRAS, 160, 1P

Zhang Y., Jiang H., Shectman S., Yang D., 2023, PhotoniX

Zhao C., Tao C., Liang Y., Kitaura F.-S., Chuang C.-H., 2016, MNRAS, 459, 2670

Zhao C., et al., 2020, MNRAS, 491, 4554

Zhao C., et al., 2021, MNRAS, 503, 1149

Zhao C., et al., 2022, MNRAS, 511, 5492

Zhou R., et al., 2020, Research Notes of the American Astronomical Society, 4, 181

Zhou R., et al., 2021, MNRAS, 501, 3309

Zhou R., et al., 2023, AJ, 165, 58

Zou H., et al., 2017, PASP, 129, 064101

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

d'Assignies D W., Zhao C., Yu J., Kneib J.-P., 2023, MNRAS, 521, 3648

de Jong R. S., et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460T (arXiv:1206.6885), doi:10.1117/12.926239

de Jong R. S., et al., 2019, The Messenger, 175, 3

van de Weygaert R., Platen E., 2011, in International Journal of Modern Physics Conference Series. pp 41–66 (arXiv:0912.2997), doi:10.1142/S2010194511000092

van de Weygaert R., Schaap W., 2009, in Martínez V. J., Saar E., Martínez-González E., Pons-Bordería M. J., eds, , Vol. 665, Data Analysis in Cosmology. Springer, pp 291–413, doi:10.1007/978-3-540-44767-2_11

Andrei Variu

1008 Prilly, Switzerland

variuandrei@yahoo.com | https://github.com/Andrei-EPFL/ | https://www.linkedin.com/in/andrei-variu-9433a022b | https://orcid.org/0000-0001-8615-602X

Education

Swiss Federal Institute of Technology, Lausanne (EPFL) *Switzerland*

PhD in Astrophysics *Oct 2019 - Sep 2023*

- PhD Thesis: "Probing the nature of dark energy through the study of large scale structures using spectroscopic surveys and their simulations"
- **Courses:** Machine Learning, Deep Learning, Deep Learning for Optical Imaging, Parallel programming (MPI/C++)

Swiss Federal Institute of Technology, Lausanne (EPFL) *Switzerland*

Master's degree in Physics, score: 5.63/6 *Sep 2017 - July 2019*

- Excellence Scholarship
- Master's Thesis: "The study of baryonic acoustic oscillations using halos and voids"
- Specialisation project at CEA-Saclay, France: "Present and future large scale spectroscopic surveys"
- Social Science Group Project: "Prospects for AI startups in China"

University of Bucharest *Bucharest, Romania*

Bachelor's degree in Physics, score: 10/10 *2014 - 2017*

- Bachelor's Thesis: "Dynamics of the charge carriers in double phase noble gas detectors"

Experience

Swiss Federal Institute of Technology, Lausanne (EPFL) *Switzerland*

Machine learning group projects *Oct 2019 - Sep 2023*

- Labelling persons using images and audio samples. Implemented Residual Networks (ResNets) in TensorFlow for both the image and audio sets and combined the outputs using a multilayer perceptron. Pre-processed the audio samples by removing the very high frequencies to reduce the dimensionality.
- Building of a Deep Learning framework using only the PyTorch's tensors and without the inherent differentiation engine, i.e. autograd. The framework supports the implementation of fully connected layers and different activation functions and losses. Additionally, we have built a class that mimics some autograd features.
- Sorting pairs of MNIST digit images using a LeNet (Convolutional Neural Network – CNN) and a ResNet implemented in PyTorch. The purpose of the project was to test the effects of an auxiliary loss and of the weight-sharing technique.
- Detection of gravitation lenses using a ResNet and a CNN implemented in TensorFlow. Despite the limited GPU resources that imply a less deep ResNet and the short time-frame, we achieved a 70% detection accuracy on simulated data.
- AICrowd Higgs production event classification using six different regression algorithms. Implemented cross validation, hyper-parameter optimisation and data pre-processing such as feature expansion.

Swiss Federal Institute of Technology, Lausanne (EPFL) *Switzerland*

Teaching Assistant *Oct 2019 - Sep 2023*

- Observational Cosmology for the first year master students in physics. Created exercises using Python Jupyter notebooks that allow visualising some important notions in cosmology.
- Mechanics for the first year bachelor students in Life Sciences. Helped students to solve exercises and understand physics notions in French.

HGS-HIRe at GSI Darmstadt *Germany*

Summer Project: "Exploring precision limits of the HADES tracking system" *Jul 2017 - Sep 2017*

- Studied the effects on precision of new electronics using existing tracking system simulations.

University of Bucharest *Romania*

Teaching Module *2014-2017*

- Followed lectures about pedagogy and teaching practices for physics. Taught one physics class for high-school students.

Skills

Programming	Python (Pandas, PyTorch, TensorFlow, NumPy, Scikit-learn, Matplotlib), R, C/C++, Mathematica
Miscellaneous	Linux, Shell (Bash), Slurm, Git, LaTeX(Overleaf), Microsoft Office / LibreOffice, Google Slides
Soft Skills	Time Management, Teamwork, Problem-solving, Documentation, Engaging Presentation.

163

Languages

English	Fluent Spoken (C1) and written (C1/C2); TOEFL iBT (2016) 101/120
French	Intermediate level spoken and written (B1/B2), Centre de langues (EPFL, 2020); DELF B1 (2012)
German	Beginner level spoken and written (A1), Centre de langues (EPFL, 2022)
Romanian	Native language

Publications

Journal Articles

DESI Mock Challenge: Constructing DESI galaxy catalogues based on FastPM simulations
Andrei Variu, Shadab Alam, Cheng Zhao, Chia-Hsun Chuang, Yu Yu, Daniel Forero-Sánchez, Zhejie Ding, Jean-Paul Kneib, et al.
arXiv e-prints, arXiv:2307.14197 (July 2023) arXiv:2307.14197. 2023

Cosmic void exclusion models and their impact on the distance scale measurements from large-scale structure
Andrei Variu, Cheng Zhao, Daniel Forero-Sánchez, Chia-Hsun Chuang, Francisco-Shu Kitaura, Charling Tao, Amélie Tamone, Jean-Paul Kneib
Monthly Notices of the Royal Astronomical Society 521.3 (May 2023) pp. 4731–4749. 2023

Interests

Salsa	member of Cubaliente student association, replacing instructor
Guitar	ex-guitarists in a pop-rock and folk music band

Personal Information

Born on the 24th of May 1995, Romanian citizenship. Swiss driver's licence (Type B).