Mining for cosmological information: Simulation-based methods for Redshift Space Distortions and Galaxy Clustering

CUESTA-LAZARO, CAROLINA

Mining for cosmological information: Simulation-based methods for Redshift Space Distortions and Galaxy Clustering

Carolina Cuesta-Lazaro

Abstract: The standard model of cosmology describes the complex large scale structure of the Universe through less than 10 free parameters. However, concordance with observations requires that about 95% of the energy content of the universe is invisible to us. Most of this energy is postulated to be in the form of a cosmological constant, Λ, which drives the observed accelerated expansion of the Universe. Its nature is, however, unknown. This mystery forces cosmologists to look for inconsistencies between theory and data, searching for clues. But finding statistically significant contradictions requires extremely accurate measurements of the composition of the Universe, which are at present limited by our inability to extract all the information contained in the data, rather than being limited by the data itself. In this Thesis, we study how we can overcome these limitations by i) modelling how galaxies cluster on small scales with simulation-based methods, where perturbation theory fails to provide accurate predictions, and ii) developing summary statistics of the density field that are capable of extracting more information than the commonly used two-point functions. In the first half, we show how the real to redshift space mapping can be modelled accurately by going beyond the Gaussian approximation for the pairwise velocity distribution. We then show that simulation-based models can accurately predict the full shape of galaxy clustering in real space, increasing the constraining power on some of the cosmological parameters by a factor of 2 compared to perturbation theory methods. In the second half, we measure the information content of density dependent clustering. We show that it can improve the constraints on all cosmological parameters by factors between 3 and 8 over the two-point function. In particular, exploiting the environment dependence can constrain the mass of neutrinos by a factor of 8 better than the two-point correlation function alone. We hope that the techniques described in this thesis will contribute to extracting all the cosmological information contained in ongoing and upcoming galaxy surveys, and provide insight into the nature of the accelerated expansion of the universe.

Mining for cosmological information: Simulation-based methods for Redshift Space Distortions and Galaxy Clustering

Carolina Cuesta-Lazaro

A thesis presented for the degree of
Doctor of Philosophy

Institute for Computational Cosmology
Department of Physics
Durham University
United Kingdom

August 2022

Dedicada a Don Luis Domínguez

El profesor de matemáticas que me metió
en este lío
“¡Hosti! No... estaría bonito”.

Contents

List of Figures

List of Tables

Declaration

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

The content presented in Chapter 3 has been published in

Carolina Cuesta-Lazaro, et al. *Towards a non-Gaussian model of redshift space distortions* Monthly Notices of the Royal Astronomical Society, 498(1), 1175

The content presented in Chapter 4 is undergoing peer review

Carolina Cuesta-Lazaro, et al. *Galaxy clustering from the bottom up: A Streaming Model emulator I*

The content presented in Chapter 5 is in preparation

Enrique Paillas, **Carolina Cuesta-Lazaro**, et al. *Constraining $\nu\Lambda$CDM through density-split clustering*

All aspects of the publication (idea development, code, analysis and writing) were co-authored by the author.

The author has had a major contribution to the content summarised in Chapter 6, with part of the work published in

- Aylett-Bullock, J., **Cuesta-Lazaro, C.**, and Quera-Bofarull, A. *XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets.* Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging vol 10953 (SPIE) pp 453–63 (2019)
- Aylett-Bullock, J., **Cuesta-Lazaro, C.**, Quera-Bofarull, A., et al. *June: open-source individual-based epidemiology simulation.* Royal Society Open Science 8 210506 (2021)

- Aylett-Bullock, J., **Cuesta-Lazaro, C.**, Quera-Bofarull, A., et al.*Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox's Bazar settlement.* PLOS Computational Biology 17 e1009360 (2021)
- Ian Vernon, Jonathan Owen, Joseph Aylett-Bullock, **Carolina Cuesta-Lazaro**, et al.*Bayesian Emulation and History Matching of JUNE* Philosophical Transactions A . (2022)

The author of the thesis was primarily responsible for all aspects of this publication.

Acknowledgements

First and foremost, I would like to thank my supervisors Baojiu and Carlton for all the support they have given me during the course of this Ph.D. Whenever I felt stuck, having you there always made me feel confident that we would figure it out together. Also, I was always treated as a colleague and I am very grateful for this. On a personal level, I thank you for the kindness and empathy you showed at times when I needed it.

To Alex and Pauline. Since I started working with you both, the PhD became much easier; thank you for your advice and friendship. You are amazing role models.

What I have definitely enjoyed most of this time has been working in collaboration with other people. Firstly, at Durham, where it has been a pleasure to work with Chen-Zong, Christian and Sownak. I am also very glad to have met Enrique Paillas through the DESI collaboration, whose passion and encouragement have made my day on multiple occasions, and to have been part of the Dark Quest collaboration. I thank Takahiro, for his invaluable insight and all the help he gave me during the job application period, and Masahiro, for the unforgivable experience of visiting Japan.

To my friends in the department, Aidan, Aoife, Ash, Christoph, Cesar, Chris, Ellen, Jack, Jake, James, Joaquin, Josh, Matteo, Miguel, Myles, Sergio, Louise, Vicky and many others. You have made working at the ICC a real blast. A special shout to the Durham data miners for all the fun we had as CDT students. Also to the people that keep the ICC running, in particular the ones I annoyed the most during these years, Shufei and Alastair. And thank you very much Aidan, for your kindness in helping me with any writing I need to do.

To the members of the JUNE collaboration for epidemiological modelling. I have never worked so intensely and been so stressed; it is definitely something I would never have done on my own. Also, thanks to my colleagues at IBEX and Amazon, who taught me so much and made me see that working in industry can be a lot of fun.

To those who made coming to Durham so hard. Ana, Arturo, Barreda, Elena, Sara, Joaquin, Jose and the Alfonso crowd, my friends back home that make me feel like time is frozen when I am back.

To those who make leaving Durham so hard. To my football lads: Alan, Bex, Dori, Gao, Holly, James, Kev and Ryan. You have made me experience Durham in ways I never thought possible. To my housemates over the last year, Miguel, Ellen and more recently Jack. It has been a pleasure to share dinners and time with you.

A special mention to Tom, Tilly and James. You made a lockdown be fun. James, thank you for welcoming me from my first day at Trevs and teaching me so much. You have made me feel at home in the UK.

A Arnau, trabajar contigo ha sido un placer pero nada comparado con vivir contigo. También te echaré de menos.

A mi familia. Edu, me has enseñado tanto que no se como sería si no te hubiese tenido como hermano. Eso sí, seguiría siendo del Atleti. Mamá y papá, sin vosotros nada de esto hubiese sido posible. Muchas gracias por las oportunidades y el ejemplo que me habéis dado.

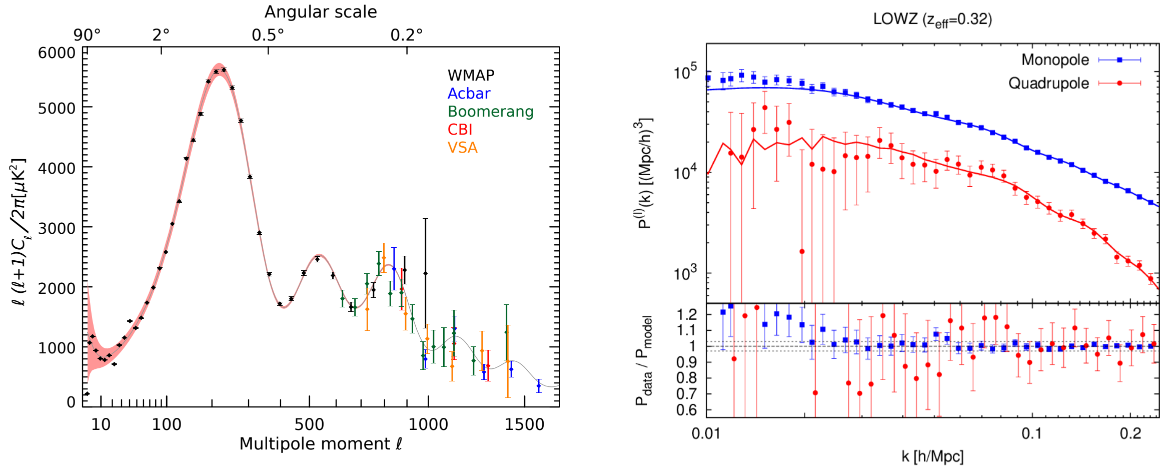
Chapter 1

The Times They Are A-Changin': An Introduction to Cosmology and its Conundrums

The standard model of cosmology is in trouble, and this is good news for science. The standard model ΛCDM (Lambda Cold Dark Matter) describes the evolution of cosmic structures through no more than ten free parameters. These parameters determine i) the statistical properties of the early universe, and ii) the universe's energy content, which shapes how matter clusters to form the structures that we observe today. With all its simplifying assumptions, ΛCDM provides a remarkably accurate description of a wide range of cosmological and astrophysical datasets, from the early Universe of small density perturbations to the highly structure cosmic web that we observe today. Figure 1.1 shows two example datasets from epochs of the Universe separated by almost 13 billion years, demonstrating the remarkable success of ΛCDM in describing both early and late observations.

But while it can reproduce observations, ΛCDM fails to provide an explanation for cosmology's most pressing questions, namely:

- *What is driving the accelerated expansion of the Universe?* By combining different experiments, such as supernova standard candles (Perlmutter et al., 1997; Riess et al., 1998), and cosmic microwave background (CMB) temperature anisotropies (Planck Collaboration et al., 2020a), astronomers have inferred that the expansion of the universe is accelerating. In the standard model of cosmology, the acceleration is driven by a cosmological constant, Λ, whose nature is not explained theoretically.



(a) **Early Universe**: The power spectrum of the cosmic microwave background radiation temperature anisotropy in terms of the angular scale (Commons, 2022). It measures the amplitude of fluctuations in temperature of the oldest electromagnetic radiation in the universe (see Section 1.1.1).

(b) **Late Universe**: The power spectrum of galaxies at $z = 0.43$ (Gil-Marín et al., 2016). The Monopole quantifies how clustered galaxies appear, whereas the Quadrupole measures the degree of anisotropy in the line-of-sight direction caused by peculiar motions (see Section 1.2.1).

Figure 1.1: Comparison of data (dots) and best-fit ΛCDM predictions (lines) for observations spanning almost 13 billion years in the evolution of the universe. It can be seen how ΛCDM can describe very accurately both early and late time observations.

- *What is dark matter made of?* Approximately 85% of the matter content of the universe is thought to not interact with electromagnetic radiation. The presence of dark matter has been inferred through its gravitational effects (Zwicky, 1933; Forman et al., 1979), but its nature also remains a mystery.

- *Did the Universe go through an inflationary period?* The leading paradigm to explain the origin of inhomogeneities in the early universe assumes that the universe underwent a phase of exponential expansion known as cosmic inflation (Guth, 1981). The detailed physical mechanism responsible for inflation is unknown.

These open questions push cosmologists to look for observations that are inconsistent with ΛCDM. In fact, there are now increasing statistically significant inconsistencies, also called tensions, between different datasets. By comparing observations of the early and late time universes, we can estimate the degree to which these observations are consistent with the expected evolution of a ΛCDM universe.

One of the most significant tensions is found in the inferred values of the expansion rate of the Universe, H_0. We can take the best-fitting model to the early Universe, and extrapolate the value of this parameter to the present time (Planck Collaboration et al., 2020a) assuming that ΛCDM is correct. If the model and the observations were fully consistent, this estimate would agree with that obtained from late-time observations, such as the Cepheid-Supernova distance ladder (Reid et al., 2019). However, Hubble parameter values that are more than 4σ away from one another are found by different combinations of early-late time probes (Di Valentino et al., 2021).

In addition, there are also tensions in the recovered values for the parameter that describes the strength with which matter is clustered in the Universe, σ_8. In particular, σ_8 is defined as the present day mass dispersion on a scale of 8 h^{-1}Mpc. Inconsistencies of more than $2-3\sigma$ have been found when comparing the matter clustering strength inferred from the early universe (through the cosmic microwave background) and the late universe (through weak gravitational lensing and galaxy clustering (Joudaki et al., 2016; Abbott et al., 2022; Philcox & Ivanov, 2022)). The latter prefer a lower degree of structure formation than those expected from CMB observations.

But are these tensions the result of systematic errors in the measurements, have we been statistically unlucky, or will they lead to the discovery of new physics? If the latter were true, the inconsistencies may give us clues as to how alternatives to ΛCDM should look, and these alternatives could, in turn, answer the biggest open questions that ΛCDM fails to explain.

Therefore, it is critical to constrain the standard model as precisely as possible. Our ability to do so is now more than ever limited by our theoretical and statistical techniques, rather than by the precision of the observations themselves. The research presented in this Thesis contributes to ongoing efforts to develop a theoretical framework that describes the large-scale structure of the Universe to the level of precision required to match that of ongoing and future galaxy surveys. We will show how a combination of N-body simulations and machine learning methods produces accurate theoretical models that can extract the cosmological information contained in the non-linear regime of structure formation, which, in the future, may help settle the debate around cosmological tensions and their origin.

1.1 ΛCDM: The Standard Model of Cosmology

Einstein, motivated by the equivalence principle, stated that gravity is a metric theory. Space-time is a four-dimensional manifold equipped with a metric $g_{\mu\nu}$. This metric is a dynamical field coupled to the matter and energy content of the Universe through the field equation

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.1.1}$$

which relates a measure of the curvature of spacetime, the Einstein tensor $G_{\mu\nu}$, to a measure of the energy content of the universe, the stress energy tensor $T_{\mu\nu}$. The definitions of these tensors can be found in Carroll (2019).

Interestingly, Einstein's equations are a unique description of gravity under certain assumptions, as shown in (Lovelock, 1969): *The only second-order, local gravitational field equations derivable from an action containing solely the 4D metric tensor (plus related tensors) are the Einstein field equations with a cosmological constant.*

Given the complexity of Eq. (1.1.1), analytical solutions can only be found in systems with high degrees of symmetry. It turns out that the universe on large scales is one such system. When averaged on large scales, the complex cosmic web of galaxies and voids becomes isotropic. It is also statistically isotropic to any observer, regardless of where they are placed within the universe. This implies that the universe is statistically homogeneous as well. That is, although the distribution of matter is not homogeneous, it is when averaged over different realisations of the density field.

When combined with the laws of General Relativity (GR), the symmetries mentioned above single out a particular space-time geometry: that of an FLRW cosmology defined by the metric

$$ds^2 = dt^2 - a(t)^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right), \tag{1.1.2}$$

where $a(t)$ is the scale factor of the universe, which is determined by the matter energy content through the Einstein equations Eq. (1.1.1), k is a constant representing the curvature of space, and $d\Omega^2 = d\theta^2 + \sin\theta^2 d\phi^2$. Note that throughout this section we are using natural units, in which $c = 1$. The metric completely specifies the left-hand side of Eq. (1.1.1).

The right-hand side is determined by the density and flux of energy and momentum, $T_{\mu\nu}$. Given a statistically homogeneous and isotropic universe, the energy-momentum tensor of a perfect fluid as seen by a comoving observer is described by

$$T^{\mu}_{\nu} = g^{\mu\lambda} T_{\lambda\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & -P & 0 & 0 \\ 0 & 0 & -P & 0 \\ 0 & 0 & 0 & -P \end{pmatrix} \tag{1.1.3}$$

where ρ is the energy density and P is the pressure of the fluid. Manipulating the combination of Eq. (1.1.2) and Eq. (1.1.1) we find the so-called Friedmann equations

$$\left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G}{3} \rho - \frac{k}{a^2}, \tag{1.1.4}$$

Parameter	Best fit value	Uncertainty
$\Omega_b h^2$	0.02242	0.00014
$\Omega_c h^2$	0.11933	0.00091
H_0	67.66	0.42
Ω_Λ	0.6889	0.0056

Table 1.1: Cosmological parameters and their uncertainties estimated from the final full-mission Planck measurements of the CMB anisotropies (Planck Collaboration et al., 2020a).

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3P). \tag{1.1.5}$$

To simplify Eq. (1.1.4), we define the Hubble parameter as $H(a) = \dot{a}/a$ and H_0 as the current value of the Hubble parameter.

The first Friedmann equation Eq. (1.1.4) relates the Hubble parameter to the total energy density in the universe at a given time. Energy densities are usually expressed as ratios to the critical density, ρ_c, the total density of a spatially flat universe, as $\Omega = \rho/\rho_c$, where $\rho_c = \frac{3H^2}{8\pi G}$.

The universe is filled with different forms of matter. On the one hand, dark matter and baryons (ordinary matter) behave like nonrelativistic particles for which the pressure is much smaller than the energy density. Their energy density is parameterised in terms of Ω_c and Ω_b, respectively. On the other hand, photons and neutrinos behave like relativistic particles for which pressure is non-negligible. In the case of neutrinos, due to their small but non-negligible masses, they make the transition from a relativistic to a non-relativistic behaviour in the recent history of the universe. In both cases, their energy densities are negligible today compared to those of nonrelativistic particles. The element that dominates the energy content of the universe today is dark energy, which behaves like a negative pressure component $P = -\rho$. The dark energy density is parameterised as Ω_Λ and is introduced to drive the accelerated expansion of the universe within the ΛCDM model. In Table 1.1 we show our current best estimates of these parameters, as found in (Planck Collaboration et al., 2020a).

1.1.1 The early Universe

Although the universe is homogeneous and isotropic when averaged over large scales, the formation of structure, including the galaxy in which we live, implies that this is no longer the case on smaller scales. The current cosmological paradigm assumes that primordial density perturbations are the result of inflation (Guth, 1981; Linde, 1982; Linde & Mezhlumian, 1995), a period of rapid expansion that is thought to take place in the very early universe.

Microscopic quantum fluctuations that existed prior to inflation are believed to have been stretched during the period of inflationary expansion to serve as seeds for the structure that we observe today.

In fact, we can study the properties of these initial seeds in the cosmic microwave background radiation. The hot and dense early universe produced frequent particle interactions that would form a plasma. The photons were trapped inside this plasma by interactions with free electrons. But the expansion of the universe made it cool down to a temperature at which the first stable atoms could form, lowering the rate of scattering between photons and electrons. It is then that photons began to propagate freely through the Universe. Photons freed at the so-called time of recombination reach us today in the form of microwave radiation at a temperature of about 2.7 K.

By mapping the temperature of photons coming from different directions, we can study the homogeneity and isotropy of the early universe. We find small variations in temperature at the level of 1 in $100,000$. This means that we observe photons coming from different directions at slightly different temperatures. The ones that are slightly hotter were produced at denser regions than those that are slightly colder, and these denser regions would later on collapse to form the structure of galaxies and voids that we observe today.

Gaussian Random Fields

Another prediction of inflation is that initial density perturbations result from many independent quantum fluctuations. This implies that they are very nearly Gaussian distributed. The degree of deviation from Gaussianity is strongly constrained by observations of the CMB (Planck Collaboration et al., 2020b). In this section, we review the theory describing Gaussian random fields and the implications for cosmology.

Starting from the value of the matter density, $\rho(\mathbf{x})$, at a point in space, $\mathbf{x}$, we define the density contrast, also known as overdensity, δ, as

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}}, \tag{1.1.6}$$

where $\bar{\rho}$ is the average density field ρ taken over space. With this definition, the mean of the overdensity field vanishes. The probability density function of the overdensity for a Gaussian random field (GRF) is a multivariate Gaussian

$$P(\delta_1, ...\delta_d)\mathrm{d}\delta_1...\mathrm{d}\delta_\mathrm{d} = \frac{1}{\sqrt{(2\pi)^\mathrm{n} \det \mathrm{C}}} \exp\left\{ \left[-\frac{1}{2}\delta^\mathrm{T}\mathbf{C}^{-1}\delta \right] \right\}, \tag{1.1.7}$$

where δ_d is the value of the overdensity at the point $\mathbf{x}_d$. See (Wandelt, 2012; Leclercq et al., 2014) for a review of Gaussian random fields.

For a GRF, cumulants of higher than second order vanish. Therefore, the pdf of δ is fully characterised by its covariance, $\mathbf{C} = \langle \delta \delta^T \rangle$. Here the $\langle \rangle$ symbols denote the ensemble average, taken by drawing many realisations from the distribution

$$\langle X \rangle = \int X(\delta_1, \delta_2, ..., \delta_n) P(\delta_1, \delta_2, ..., \delta_n) \mathrm{d}\delta_1 \mathrm{d}\delta_2 ... \mathrm{d}\delta_n. \tag{1.1.8}$$

Given that we can access only one realisation of the universe through observations, we cannot measure ensemble averages. Instead, we can average over many distinct regions of space. This is defined as the volume average of one realisation of the distribution

$$\bar{X} = \frac{1}{V} \int_v X(\mathbf{x}) \mathrm{d}^3\mathbf{x}, \tag{1.1.9}$$

where V is some volume in the Universe. If the ensemble average coincides with the sample average, the system is said to be ergodic. In general, the validity of the ergodic hypothesis in cosmology depends on the ratio between the length scale over which we perform the spatial averaging and the scale at which spatial correlations become negligibly small. The assumption of ergodicity allows us to compute ensemble averages of fields in cosmology (Adler, 1981).

The covariance matrix, $\mathbf{C}$, describes the correlation between the amplitude of δ at two positions $\mathbf{x}_1$ and $\mathbf{x}_2$, and is also known as the correlation function, ξ. We can simplify the description of the correlation function by imposing homogeneity and isotropy as follows

$$\langle \delta(\mathbf{x}_1) \delta(\mathbf{x}_2) \rangle = \xi(\mathbf{x}_1, \mathbf{x}_2) \overset{\text{homogeneity}}{=} \xi(\mathbf{r}) \overset{\text{isotropy}}{=} \xi(r), \tag{1.1.10}$$

where $\mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1$. One can think of the correlation function as a compressed summary of the field. The information contained in the three-dimensional field $\delta(\mathbf{x})$ can be optimally summarised in its two-point correlation function. This compression step can reduce the dimensionality of the problem from about $\mathcal{O}(10^9)$ dimensions, if the overdensity is computed on a grid with 1024 cells, to only $\mathcal{O}(100)$ dimensions, the pair separation bins used to estimate the two-point correlation function.

The two-point correlation function can also be interpreted as the excess probability over random that two particles in volume elements dV_1 and dV_2 are separated by a distance r. If the particle number densities at two locations $\mathbf{x}_1$ and $\mathbf{x}_2$ separated by a distance r are n_1 and n_2, and given that the density contrast within an infinitesimal volume is $\delta(\mathbf{x_i}) = n_i/(n\mathrm{dV_i}) - 1$, where $\bar{n} = N/V$ is the average number density, then the two-point correlation function can be written as

$$\xi(r) = \langle \delta(\mathbf{x}_1)\delta(\mathbf{x}_2) \rangle = \frac{\langle n_1 n_2 \rangle}{\bar{n}^2 \mathrm{dV}_1 \mathrm{dV}_2} - 1, \tag{1.1.11}$$

where $\langle n_1 n_2 \rangle$ measures the average number of pairs within the volume elements dV_1 and dV_2. Therefore,

$$\langle n_1 n_2 \rangle = \bar{n}^2 (1 + \xi(r)) \mathrm{dV}_1 \mathrm{dV}_2. \tag{1.1.12}$$

If a distribution of particles is random, the average number of pairs is given by the product of the average number of particles in each volume, $\langle n_1 n_2 \rangle = \langle n_1 \rangle \langle n_2 \rangle = \bar{n}^2 \mathrm{dV}_1 \mathrm{dV}_2$, which corresponds to a vanishing correlation function.

It is useful to express δ as an integral of Fourier modes

$$\delta(\mathbf{x}) = \int \frac{\mathrm{d}^3\mathrm{k}}{(2\pi)^{3/2}} \delta(\mathbf{k}) e^{i\mathbf{kx}}, \tag{1.1.13}$$

and, equivalently,

$$\delta(\mathbf{k}) = \int \frac{\mathrm{d}^3\mathrm{x}}{(2\pi)^{3/2}} \hat{\delta}(\mathbf{x}) e^{-i\mathbf{kx}}. \tag{1.1.14}$$

Note that $\delta(\mathbf{k})$ is a complex random variable, but since $\delta(\mathbf{x})$ is real, the Fourier field must satisfy $\delta(-\mathbf{k}) = \delta^*(\mathbf{k})$.

The variance in Fourier space is denoted as the power spectrum, $P(k)$, and is defined as

$$\begin{aligned} < \delta(\mathbf{k}_1)\delta^*(\mathbf{k}_2) > &= \frac{1}{(2\pi)^{3/2}} \frac{1}{(2\pi)^{3/2}} \int \mathrm{d}^3\mathrm{x}_1 \int \mathrm{d}^3\mathrm{r} < \delta(\mathbf{x}_1)\delta(\mathbf{x}_1 + \mathbf{r}) > \mathrm{e}^{\mathrm{i}(\mathbf{k}_1 - \mathbf{k}_2)\mathbf{x}_1 - \mathrm{i}\mathbf{k_2 r}} \\ &\overset{\text{homogeneity}}{=} \frac{1}{(2\pi)^3} \delta_D(\mathbf{k}_1 - \mathbf{k}_2) \int \mathrm{d}^3\mathbf{r}\xi(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k_1 r}} \overset{\text{isotropy}}{=} \frac{1}{(2\pi)^{3/2}} \delta_\mathrm{D}(\mathbf{k}_1 - \mathbf{k}_2)\mathrm{P}(\mathrm{k}), \end{aligned} \tag{1.1.15}$$

where $k = |\mathbf{k_1} - \mathbf{k_2}|$.

1.1.2 Structure growth in a ΛCDM Universe

Gravity drives the overdensities and underdensities observed in the cosmic microwave background to collapse, forming the seeds for the galaxies, voids and clusters of galaxies that we observe today. Although it is possible to perform a fully relativistic treatment of how gravity results in structure formation (Bardeen, 1980; Mukhanov et al., 1992; Peebles, 1980a), here we will simplify the analysis by approximating dark-matter particles as a nonrelativistic fluid and ignoring general relativistic effects (i.e. working in the Newtonian limit in an expanding background).

Given that we focus on the behaviour of gravity on scales smaller than the Hubble radius, general relativistic effects such as curvature can be safely ignored. Moreover, although the nature of dark matter is still an open question in cosmology, its properties are strongly constrained by observations. In particular, in the standard model of cosmology dark matter is thought to be "cold", i.e. non-relativistic, since otherwise the seeds for the observed small-scale structure could not have formed in the early universe. Furthermore, we can ignore the discrete nature of the dark matter particles and treat dark matter as a collision-less fluid.

Under these approximations, we can derive the following equations from mass and momentum conservation

$$\frac{\partial}{\partial t}\rho + \nabla(\rho\mathbf{v}) = 0, \tag{1.1.16}$$

$$\frac{\partial}{\partial t}\mathbf{v} + \mathbf{v}\cdot\nabla\mathbf{v} = -\nabla\phi. \tag{1.1.17}$$

Eq. (1.1.16) relates the evolution of the density field to velocity fluxes, $\nabla\mathbf{v}$, while Eq. (1.1.17) shows that the velocity field evolves due to spatial variations in the gravitational potential, ϕ. These are known as the continuity and Euler equations, respectively. They both contain non-linear terms, which have been highlighted in purple. The non-linear terms limit our ability to describe growth of structure using analytical methods.

These two equations contain three unknowns ρ, $\mathbf{v}$, and ϕ. To solve for these quantities, we also need the Poisson equation to relate the gravitational potential to its source, the density field

$$\nabla^2\phi = 4\pi G\bar{\rho}\delta. \tag{1.1.18}$$

Due to non-linearities in the coupled system of equations defined by Eqs. (1.1.16), Eq. (1.1.17), and Eq. (1.1.18), we can only find analytical solutions for the evolution of the three fields in the restricted case of small density perturbations, $\delta \ll 1$, or in special cases, like spherical symmetry. Approximations to the exact solution can also be found when $\delta \approx 1$ using perturbation theory (Bernardeau et al., 2002). But to find solutions in the non-linear regime, where $\delta > 1$, we need to resort to N-body simulations, which will be introduced in the next section. Before that, we shall study the properties of linearised equations and their evolution, which are used to describe gravitational growth at early times.

Replacing ρ by the density contrast δ, working in comoving coordinates and linearising the equations, we find

$$\frac{\partial}{\partial t}\delta + \nabla\mathbf{u} = 0, \tag{1.1.19}$$

$$\frac{\partial}{\partial t}\mathbf{u} + 2H(a)\mathbf{u} = -\frac{\nabla\phi}{a^2}, \tag{1.1.20}$$

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \tag{1.1.21}$$

where $\mathbf{u} = \mathbf{v}/a$ is the comoving velocity.

Taking the time derivative of Eq. (1.1.19), and combining it with the spatial derivative of Eq. (1.1.20) and Eq. (1.1.21) we find

$$\boxed{\frac{\partial^2 \delta}{\partial a^2} + \frac{1}{a}\left(3 + \frac{\mathrm{d}\ln \mathrm{H(a)}}{\mathrm{d}\ln \mathrm{a}}\right)\frac{\partial \delta}{\partial a} = \frac{3\Omega_m(a)}{2a^2}\delta. \qquad (1.1.22)}$$

This equation describes the temporal evolution of the density contrast. The right-hand side can be thought of as a source term. Increasing Ω_m increases matter clustering, and therefore δ, whereas the second term of the left-hand side is a drag term; the faster the background universe expands, the harder it is to cluster, and the slower δ will grow. For structures to grow, they must collapse on a dynamical time scale smaller than $1/H(a)$. In simple terms, the growth of structure is the result of a competition between gravity, trying to push matter together, and the expanding background, trying to pull it apart.

An important consequence of Eq. (1.1.22) is that growth of structure is scale-independent in the linear regime, and we can separate the position and time dependence by writing $\delta(\mathbf{a}, t) = \delta(\mathbf{x}, t_i)\frac{D_+(t)}{D_+(t_i)}$, where $D_+(t)$ is known as the linear growth factor, where t_i is some reference time.

From the linearised continuity equation (1.1.19) we find

$$\mathbf{u} = -a\delta\frac{\mathrm{dD}_+}{\mathrm{dt}}\frac{1}{D_+} = -a\delta H f, \qquad (1.1.23)$$

where $f = \frac{\mathrm{d}\ln D_+}{\mathrm{d}\ln a}$, known as the linear growth rate.

While measurements of $H(z)$ test the global evolution of the universe and its energy density content at the background level, estimates of $f(z)$ allow us to test the evolution of inhomogeneities. It is particularly relevant to estimate $f(z)$ to test viable gravity theories other than general relativity. If these theories modify the Einstein-Hilbert action, they might introduce additional degrees of freedom that mediate an additional gravitational force, also known as the fifth force. Modified gravity theories might produce an expansion history very similar to that of general relativity, but a fifth force would affect the rate at which structures grow (Linder, 2005), and therefore precise measurements of $f(z)$ could rule out these alternative models.

1.1.3 The late Universe

So far we have described the density field as being Gaussian distributed, and we have shown how, at early times, its evolution can be described by linearising the continuity and Euler equations, resulting in a scale-independent perturbation growth. However, the growth of structure results in the distribution of δ becoming non-Gaussian, its evolution non-linear, and its growth inhomogeneous.

Given that δ is bounded by -1, since a region of the universe cannot have negative density, the distribution of δ values must develop skewness as the underdense and overdense regions grow. This also implies that growth must be position dependent. Moreover, solutions to the linearised equations cease to give accurate approximate solutions to the growth of structure when δ is not much smaller than one. Corrections can be included by writing a perturbative series as a function of the initial density field, but their accuracy is limited on small scales. See (Bernardeau et al., 2002) for a review of perturbation theory and (Carlson et al., 2009) for comparisons of the predictions of perturbation theory with N-body simulations.

An important consequence of the non-Gaussianity of the density field is that the two-point correlation function is no longer an optimal summary of the information contained in the three-dimensional field. There is now an active topic of research in cosmology dedicated to finding more informative, or complementary, summary statistics that improve our constraints on the cosmological parameters. Examples of these are the bispectrum (Sefusatti et al., 2006; Yankelevich & Porciani, 2019), the marked correlation function (Beisbart & Kerscher, 2000; White, 2016), and the void-galaxy cross-correlation (Nadathur et al., 2019). Alternatively, there have been substantial developments in the area of field-level inference (Leclercq & Heavens, 2021; Villaescusa-Navarro et al., 2021; Dai & Seljak, 2022), which leads to inference of the values of the cosmological parameters from the full density field rather than using a summary statistic.

N-body simulations and their evolving role in cosmology

To obtain fully non-linear predictions for the properties of the large scale structure we must resort to N-body simulations. These calculations simulate the gravitational evolution of a discrete set of particles that approximate the motion of a fluid in an expanding background, producing simulated universes that cover representative patches of the Universe such as those observed through galaxy surveys (of the order of up to a few h^{-1}Gpc per side). Given that

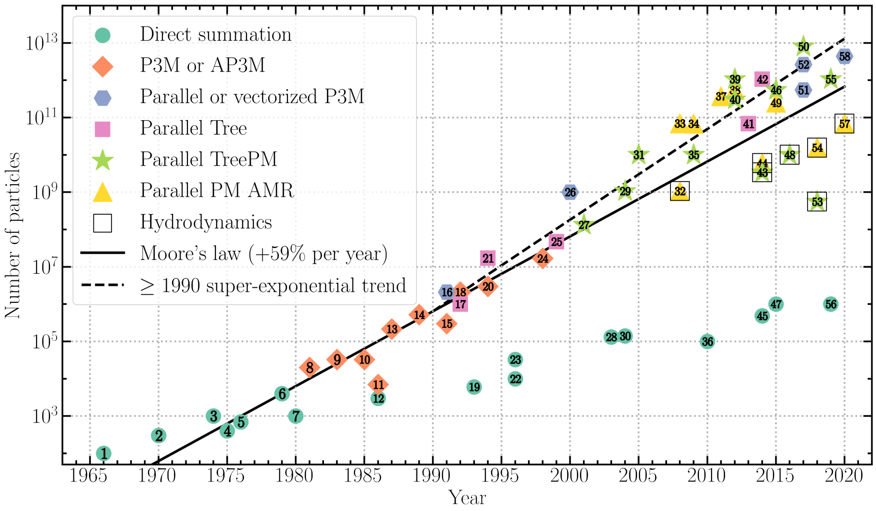


Figure 1.2: Number of particles used in cosmological N-body simulations and different algorithms. It can be seen that improvements with time are the result of both more efficient algorithms and more powerful computers. This image has been taken from http://florent-leclercq.eu/blog.php?page=2, see references therein for more information about the N-body codes

the size of these patches is still well below the horizon scale, relativistic effects can be safely ignored, and only Newtonian equations of motion need to be solved. However, there are codes capable of simulating the fully relativistic process of structure formation (Adamek et al., 2016; Barrera-Hinojosa & Li, 2020).

The computational cost of running an N-body simulation is largely determined by the number of particles used. In Figure 1.2, we show the historical evolution of the number of particles used to run the largest N-body simulation in a given year. The observed increase in number of particles is enabled by a combination of more powerful computers (as indicated by Moore's law) and more efficient algorithms, as can be seen from the points deviating from Moore's law. The particle mass is determined by the mean density of the universe in a given cosmological model. Therefore, for a fixed number of particles, there is a compromise between the volume of the box and the smallest mass object that can be resolved.

To run an N-body simulation, we need to determine a set of initial conditions describing the initial positions, velocities, and masses of the particles. If we run the simulations from a sufficiently high redshift, the density field can be approximated by a Gaussian Random Field. Therefore, the particle positions can be found by generating a GRF defined by the matter power spectrum at the desired redshift. The phases of the field are randomly generated, thus producing different realisations of the same random field. Running N-body simulations with different phases allows us to compute ensemble averages and reduce the noise introduced by cosmic variance. Moreover, at high redshift $\delta \ll 1$, which allows us to use the linearised

Euler equation Eq. (1.1.20) to convert overdensities into particle velocities. Current Initial Condition generation codes (Garrison et al., 2016; Jenkins, 2010) make use of either the Zel-dovich approximation (Zel'dovich, 1970) or second-order Lagrangian perturbation theory (Crocce et al., 2006) to obtain accurate initial conditions.

From the initial conditions, different algorithms and approximations are used to compute the gravitational forces on the particles. See (Kuhlen et al., 2012; Vogelsberger et al., 2020) for reviews on the topic, including hydrodynamic simulations. Apart from the phase-space dark matter particle distributions, an important output of N-body simulations are the so-called dark matter halos. These are groups of gravitationally bound particles that will be hosts of the galaxies that we observe in galaxy surveys. However, defining these objects is tricky, and different definitions of halo finders are still being used (Behroozi et al., 2015).

N-body simulations have been widely used as cosmic laboratories to test the precision and robustness of analytical methods for the large-scale structure, together with the effects of systematic errors in our measurements. However, over the past decade, computational advances have allowed us to run a large enough number of N-body simulations covering a significant fraction of the cosmological parameter space, which allows us to use the simulations themselves as predictive models that directly constrain the cosmological parameters. These are called simulation-based methods. Some examples of large suites of N-body simulations with different parameters have been presented in (Nishimichi et al., 2019a; DeRose et al., 2019a; Maksimova et al., 2021a). We will show such an application of simulation-based methods in Chapter 4.

Of particular relevance to simulation-based methods is the accuracy of the N-body codes themselves. To run simulations with a large number of particles, different assumptions and approximations are made to make the problem computationally tractable. A recent code comparison (Grove et al., 2021) showed how clustering predictions for dark matter halos obtained with three different N-body codes (ABACUS (Garrison et al., 2021), GADGET (Springel, 2005a) and SWIFT (Schaller et al., 2016)) are only within the statistical errors of future surveys like the Dark energy Spectroscopic Survey (DESI) for scales larger than $20h^{-1}$Mpc. Understanding these differences will be relevant for simulation-based inference.

1.2 The Universe as we see it

We can divide cosmological observables into those that test the background expansion history, $H(a)$, and those that aim to measure the growth rate of the Large Scale Structure (LSS) as a function of cosmic time, $f(a)$. It is important to measure the temporal evolution of these two functions to constrain the dynamics of the dark energy equation of state.

Probes that test the background expansion history and the growth rate of structure are complementary, and crucial in distinguishing between alternatives to ΛCDM models (Linder, 2005). Some viable modified gravity theories can be tuned to reproduce the observed evolution of the scale factor with time and therefore are indistinguishable on the background level from general relativity. However, by including information on the rate at which cosmic structures grow, we can detect modifications to gravity. The growth of cosmic structure is the outcome of a competition between the expansion of the Universe and the gravitational pull, generated by inhomogeneities. If there is an additional fifth force, but the expansion is compatible with that observed, the rate at which structures in the Universe grow will be modified.

In this Thesis, we focus on probes of the growth of structure. We refer the reader to (Freedman, 2021; Di Valentino et al., 2021), for reviews on tests of the background expansion history and their relevance in assessing the statistical significance of the H_0 tension.

1.2.1 Probing the growth of structure

To measure the large-scale growth of structure in the Universe, we look at the statistics of 3-D galaxy maps made using spectroscopic surveys. These maps contain the angular position of galaxies in the sky, together with their redshifts. The angular coordinates and redshift can be converted into comoving distances through the angular diameter distance.

Assuming that galaxies are at rest, as the photons they emit travel towards us through an expanding Universe, their wavelengths stretch accordingly. Therefore, we observe the redshifted light of distant galaxies. We can translate this redshift into a comoving distance by introducing the Hubble factor, $H(z)$,

$$r(z) = \int_0^z \frac{dz'}{H(z')}, \tag{1.2.1}$$

where $r(z)$ is the comoving distance to the galaxy, and we have used the natural unit where the speed of light $c = 1$.

Nevertheless, there are several effects related to the distorted way in which we observe the universe that complicate this simple picture. In fact, much of the information that we obtain from 3-D galaxy maps about the laws of gravity does not come directly from the comoving map of galaxy positions, but from distortion effects that alter this map. It is often not possible to isolate the contribution of the different effects, and we must model them jointly. In this section, we review the main effects that will play an important role in this Thesis.

Geometrical distortions: Alcock-Pacynski

Converting angular coordinates and redshifts into comoving coordinates requires a fiducial cosmological model. If this fiducial model is different from the true one, an isotropic distribution of galaxies would be distorted to become anisotropic because of the different conversion of angular coordinates and redshift into comoving coordinates. This is known as the Alcock-Paczynski (AP) effect (Alcock & Paczynski, 1979).

The distortions can be parametrised (Ballinger et al., 1996) by scaling the transverse and line-of-sight separation vectors

$$s_\perp = q_\perp s_\perp^{\text{fiducial}} \tag{1.2.2}$$

$$s_\parallel = q_\parallel s_\parallel^{\text{fiducial}}. \tag{1.2.3}$$

The q-scaling factors are related to the cosmological parameters through the comoving angular diameter distance, D, and the Hubble parameter, $H(a)$, by

$$q_\perp = \frac{D}{D^{\text{fiducial}}} \tag{1.2.4}$$

$$q_\parallel = \frac{H^{\text{fiducial}}}{H}. \tag{1.2.5}$$

Redshift Space Distortions

In Eq. (1.2.1), we assumed that the galaxies are at rest. However, galaxies also move because of the gravitational pull generated by the inhomogeneous distribution of matter around them. If a source that emits light moves, the wavelength of the emitted light becomes further shifted because of the Doppler effect. If we ignored this effect, then we would infer the wrong distance, $\mathbf{s}$, given by

$$\mathbf{s} = \mathbf{r} + \frac{\mathbf{v}(\mathbf{r}).\hat{\mathbf{z}}}{\mathcal{H}}\hat{z}, \tag{1.2.6}$$

instead of the real position of the galaxy, $\mathbf{r}$, where $\mathbf{v}(\mathbf{r})$ is the peculiar velocity of the galaxy, $\mathcal{H} = aH(a)$ is the comoving Hubble factor, and the inferred distance, $\mathbf{s}$, is called the redshift space distance. We have assumed that the observer is far away from the sources and therefore the line-of-sight direction can be fixed to a particular direction, which we, without loss of generality, set as the $\hat{z}$ axis. This approximation, known as the plane-parallel or distant observer approximation, has so far given results that lie within the statistical error bars of current surveys (Samushia et al., 2012).

The translation between redshift and distance is, in reality, more complex than that in Eq. (1.2.1), since we need to disentangle the combination of the position of the galaxy and its peculiar velocity along the line of sight. However, this complication turns out to be beneficial, since peculiar velocities are sourced by the gravitational pull of the inhomogeneous matter distribution. Peculiar velocities therefore allow us to detect the existence or constrain the strength of fifth forces by studying the growth of structure inferred from the statistics of the peculiar velocity field; see, e.g. (Gorski, 1988; Bose & Koyama, 2017).

To extract the growth rate, we measure the effect of peculiar velocities on the clustering properties of galaxies, known as redshift space distortions (RSD). Due to the peculiar motions of galaxies, we observe redshift space positions, **s**, instead of the real space positions, **r**, and thus we can only measure the redshift space two-point correlation function

$$\xi^S(s_\perp, s_\parallel) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{s})\rangle, \tag{1.2.7}$$

which depends both on the modulus of the pair separation vector, s, and on its inclination with respect to the line-of-sight direction. Throughout, we denote the separations perpendicular and parallel to the line of sight by $s_\perp$ and $s_\parallel$, respectively.

The redshift space correlation function is a combination of both real space clustering, $\xi^R(r)$, and the probability of finding a pair of galaxies with a given relative velocity along the line of sight, also denoted as the pairwise velocity distribution, as we will show in Chapter 3 using the Streaming Model of RSD, see e.g. (Fisher, 1995a; Scoccimarro, 2004). Since clustering in redshift space is affected by relative peculiar motions, it contains information about the growth of structure.

Biased tracers

In redshift surveys, we obtain 3-D maps of galaxies instead of the underlying dark matter field that we can describe theoretically. Modelling the mapping between dark matter and galaxies is challenging, since we lack a complete understanding of how galaxies form, how gas cools to form stars and galaxies, and what the dominant feedback processes are from baryons that alter the distribution of dark matter.

We can build physical models of galaxy formation by solving the equations of gravity and hydrodynamics simultaneously. This involves simulating gas cooling, feedback from Active

Galactic Nuclei (AGNs) and supernovae, and stellar driven winds by tracing the evolution of dark matter, gas, and star particles over time. However, we cannot simulate the full range of scales that one would need to resolve the relevant processes of galaxy formation in a cosmological context, nor do we have a full description of the different physical processes that affect the full range of scales. Hydrodynamical simulations resort to parameterisation of the relevant physics below a resolution scale that is tuned to match a wide set of observations. This is known as subgrid physics, and there is still uncertainty in the detailed implementations and tuning of subgrid models. See (Somerville & Davé, 2015) for a review of this topic.

It has been known for a long time that galaxies do not randomly sample the matter field, but form inside bound structures known as dark matter halos and are therefore biased tracers of the density field (e.g. White & Rees, 1978). See (Desjacques et al., 2018a) for a comprehensive review of galaxy bias.

Generally, galaxy bias can be expressed as a combination of halo bias, which describes the relation between the distribution of halos and that of dark matter, and the physics of galaxy formation, which is still poorly understood. Therefore, the relation between the galaxy field and the underlying matter field depends on all the variables relevant to galaxy formation. This relation can be complex (Sheth & Tormen, 1999; Tinker et al., 2010), and depends on the redshift and scale.

The Halo-Galaxy connection

We can bridge the gap between the observed galaxy distribution in galaxy surveys and the halos formed in dark matter only simulations through a statistical description of how galaxies populate dark matter halos. Such statistical modelling approaches are, by definition, empirical in the sense that we use data to constrain them. This means that they have limited predictive capabilities but are at the same time flexible enough to account for uncertainties in our theoretical descriptions of galaxy formation physics. We can therefore use them to obtain robust cosmological constraints, after marginalising over their free parameters.

In this Thesis, we focus on the halo occupation distribution (HOD) models. These models assume a functional form for the mapping of dark matter halo properties into galaxy occupation numbers. In particular, HOD models assume that the galaxy population can be split into central galaxies, living at the potential centre of the dark matter halo, and satellite galaxies, representing gravitationally bound structures orbiting within the dark matter halo.

The occupation of central galaxies is parameterized as a Bernoulli distribution, while that of satellites is Poisson distributed (Benson et al., 2000; Zheng et al., 2005).

Both distributions are described by their mean parameters. In the simplest HOD models, the mean number of galaxies is described as a function of dark matter halo mass only. Some commonly used parameterizations can be found here (Zheng et al., 2005; Reddick et al., 2013).

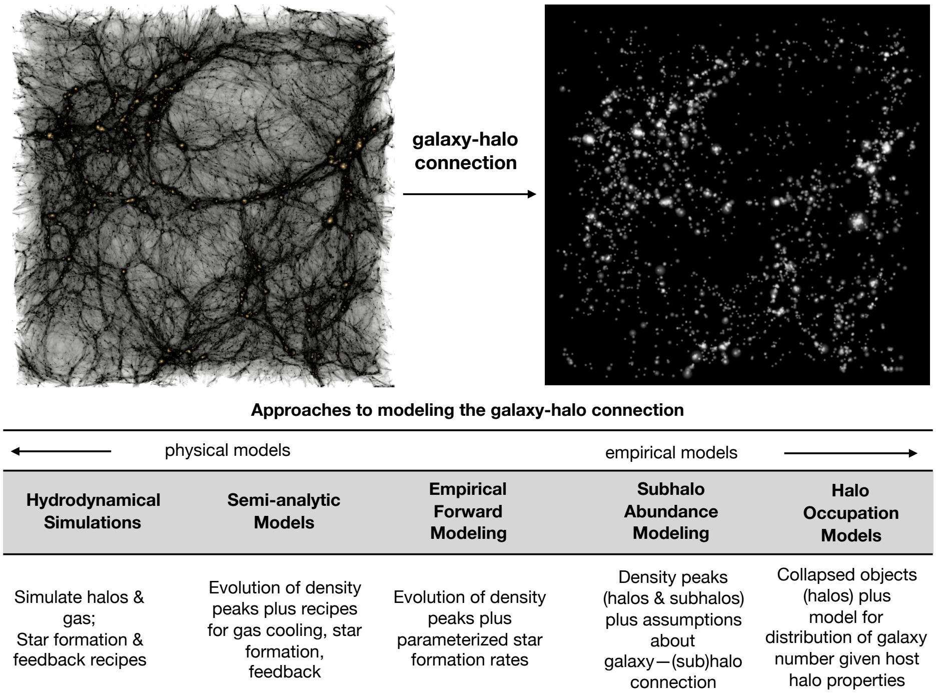
Predictions of galaxy clustering can be obtained through HOD models either by using the HOD to sample "galaxies" from dark matter halos in an N-body simulation, or by combining analytical models of dark matter halo abundance and clustering with HOD models to make predictions for summary statistics such as two-point clustering functions (van den Bosch et al., 2013; Tinker et al., 2005).

Although dark matter halo mass has a strong influence on galaxy clustering, we know that dark matter halos experience different assembly histories even at fixed halo mass. These different assembly histories influence the secondary properties of halos, and this might, in turn, affect the formation of galaxies. These effects are known as *halo* and *galaxy assembly bias* (Gao & White, 2007a; Croton et al., 2007). Although these two effects share the words assembly bias, they refer to different issues,

- *Halo assembly bias* refers to differences in the clustering of dark matter halos at a fixed halo mass. These differences depend on the secondary halo properties, which normally correlate with the formation history of the halo, albeit with some scatter.
- *Galaxy assembly bias* refers to differences in the number of galaxies within dark matter halos at a fixed halo mass, which in turn may depend on secondary halo properties.

The effect of halo assembly bias has been observed in simulations (Gao et al., 2005a; Gao & White, 2007a; Paranjape et al., 2018), and significant evidence has also recently been found through observations (Miyatake et al., 2016). Regarding galaxy assembly bias, studies of both semi-analytical models of galaxy formation and full hydrodynamical simulations have found evidence that the mean number of galaxies depends on secondary halo properties other than mass (Zhu et al., 2006; Zehavi et al., 2019; Yuan et al., 2022b).

In Figure 1.3, we show a comprehensive summary of the modelling techniques used to describe the halo-galaxy connection. In this section, we have only discussed the two ends of the spectrum of physical and empirical models, but a complete review can be found in (Wechsler & Tinker, 2018).



Approaches to modeling the galaxy-halo connection

Hydrodynamical Simulations	Semi-analytic Models	Empirical Forward Modeling	Subhalo Abundance Modeling	Halo Occupation Models
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Figure 1.3: This figure, taken from (Wechsler & Tinker, 2018), shows different approaches to modelling the galaxy-halo connection ranging from physical models based on hydrodynamical simulations to empirical models such as Halo Occupation.

Summing up, we do not have access to the comoving three-dimensional map of the dark matter field density but instead to the angular positions of biased tracers obtained through a particular cosmological model, and their redshift space line of sight distance to us. These distortions actually increase the information content of 3-D galaxy maps by also introducing information about the velocity field. The challenge is to model all of these accurately to obtain precise and unbiased constraints on the cosmological model.

1.3 Cosmological tensions: state-of-the-art constraints

We say that the parameters inferred from observations are in tension with each other if we find discrepant results even when accounting for the uncertainties of the measurement. The question that these discrepancies currently pose is: are tensions a result of systematic errors in the measurements, did we get statistically unlucky, or are they going to lead to the discovery of new physics?

Most often, the quoted σ deviations used to determine the significance of the tension measure the difference between the maximum marginalised posterior values of the cosmological parameters in units of the uncertainty, which is usually defined as the posterior errors of the experiments added in quadrature. This definition allows us to attach a probability estimate to the tension. Assuming a Gaussian distribution, a 2σ tension would translate into a 4.6% probability of exceeding the observed shift.

The two parameters responsible for most of the inconsistencies are H_0 and σ_8, when comparing their inferred values from observations of the early- and late-time universe. The σ_8 tension is often quoted in terms of the parameter S_8 since the two are related by $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$, and S_8 determines the amplitude of the weak lensing signal. It is also commonly referred to as the $f\sigma_8$ tension, as it is the combination of f and σ_8 that determines the amplitude of redshift space distortion measurements. Parameters f and S_8 are also closely related, due to the relationship between the growth rate and the density of matter, $f \approx [\Omega_m(z)]^{0.55}$, in General Relativity.

On the one hand, we find a 5σ discrepancy between the early-time CMB measurements led by the Planck Collaboration (Planck Collaboration et al., 2020a) (extrapolated to the present by assuming a ΛCDM model), and the larger late-time value of H_0 found by the SH0ES Collaboration (Riess et al., 2021). See Shah et al. (2021) for a review of the so-called H_0 tension.

On the other hand, measurements on the clustering of matter find discrepant values at a level of about 3σ for the parameters $f\sigma_8 - S_8$, where late-time Universe measurements, through weak gravitational lensing and galaxy clustering (e.g. Joudaki et al. (2017); Abbott et al. (2022); Philcox & Ivanov (2022); Heymans et al. (2021)), find that the strength of matter clustering is lower compared to the values inferred from the CMB.

In Figure 1.4 we show different measurements of the combination of parameters $f\sigma_8$ as a function of the redshift. The purple line shows the result of extrapolating Planck inferred values to later times by assuming that ΛCDM is correct. The time evolution of $f\sigma_8(z)$ is rather featureless, which means that its value is close to a constant over recent cosmic history, when we can get precise estimates from the data.

Figure 1.4 shows how several values inferred by clustering at late times are in $2 - 3\sigma$ tension with Planck. We have highlighted in broader lines the methods that make use of small-scale clustering ($s < 30\ h^{-1}$Mpc) to extract their constraints. These achieve the most stringent constraints on $f\sigma_8$, compared to analyses restricted to large-scale clustering.

To determine whether the observed tensions are statistically significant, it is crucial to reduce the estimated posterior errors in the late-time Universe measurements. This could be done through increasing the statistical power of the experiments themselves, upgrading

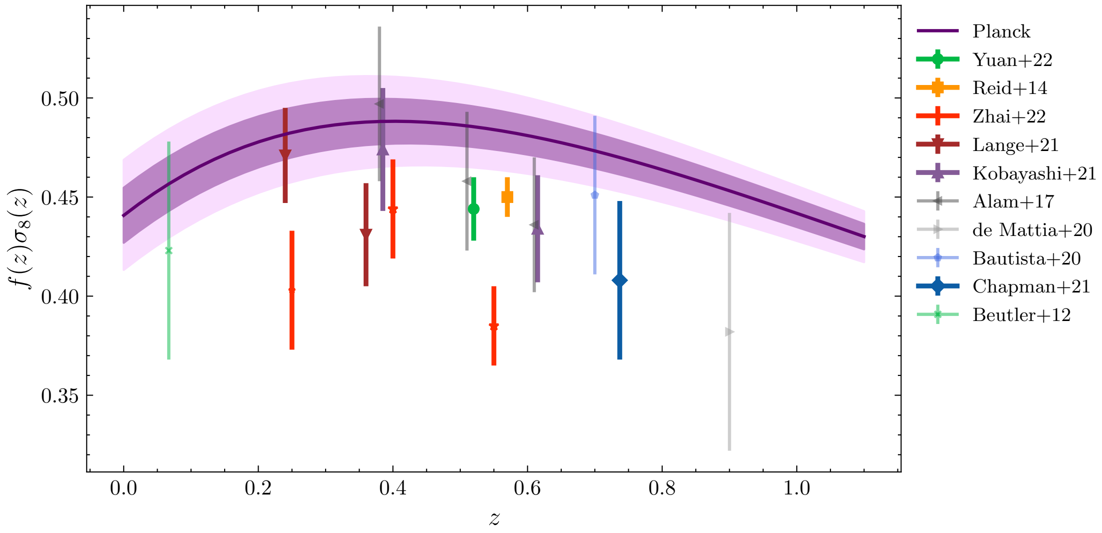


Figure 1.4: Marginalised constraints of the growth rate $f\sigma_8(z)$ obtained from the clustering of the Large Scale Structure, compared to those inferred by Planck. We have highlighted analyses that utilise small scale information with wider lines. This figure has been adapted from (Yuan et al., 2022a).

the instruments, and surveying larger patches of the sky, through the cross-correlations of different probes, or through the opening of new windows such as gravitational waves (Palmese et al., 2019) or the 21 cm line (Pritchard & Loeb, 2012). Given that future surveys will sample significant proportions of the observable universe, the limit on how much we can learn from observations is now mostly determined by our statistical methods and our inability to extract all the information contained in the data, rather than being limited by the data itself.

In this Thesis, we focus on investigating ways in which we can increase our precision in estimating the growth rate of structure by looking into: extracting cosmological information from the nonlinear regime inaccessible to perturbation theory techniques and studying alternative summary statistics to the two-point correlation function that aim at extracting information from the non-Gaussianity of the density field.

But before we present our results, we will introduce the basic concepts of Bayesian inference and machine learning in Chapter 2. In Chapters 3 and 4, we will show how we can use N-body simulations to model the effects that non-linear gravitational evolution have on the two-point correlation function and the inferred values of the cosmological parameters. In Chapter 5, we will determine the information content of environment dependent clustering and show how it could be used to constrain the standard model of cosmology. In Chapter 6, we will show the application of the same computational techniques used throughout the Thesis to other scientific domains, namely, medical imaging and epidemiology. Finally, in Chapter 7, we summarise the results of this Thesis and discuss future research avenues.

Chapter 2

Contrasting data with theory: statistics and Machine Learning

> If you need statistics, you did the wrong experiment
>
> *Attributed to Ernest Rutherford*

Ernest Rutherford lived between 1871 and 1937. At that time, the scientific method began with a research question, triggered by an observation that could not be explained in the prevailing scientific paradigm, and the formulation of a hypothesis that might be able to explain it. The scientist would then work on devising a controlled experiment that could falsify or confirm the hypothesis. Controlling the experiment meant being able to vary the theory's input variable X, and observing the theory's output variable, Y, with the lowest error possible.

Currently, we have access to vast amounts of data, which most of the time we cannot control. Although this is true in many areas of science, this issue is, in fact, at the very core of observational cosmology. We, as observers, are confined to a passive role, since it is impossible for us to modify the dynamics of the Universe in which we live. We can only test a hypothesis that can be falsified with experiments that effectively have already been carried out by the one Universe that we have at our disposal.

Uncontrolled experiments carry uncertainty not only from the instruments used to perform measurements, that introduces noise or unknown selection effects, but also from the inherent stochastic nature of the data generation process. The concept of uncertainty also expresses our capability to determine a given parameter from an observation, limited by the information about that parameter that the observable contains, or its uncertain relation to other parameters that are used to describe the data. In these cases, Bayesian statistics is used as a tool to confront theories with data in the presence of uncertainty, and to distinguish robust discoveries from statistical flukes.

Furthermore, in all areas of science, we have seen how the scientific method has changed from being purely driven by hypothesis to becoming increasingly driven by data (Hey et al., 2009; Succi & Coveney, 2019). This change has been triggered not only by the collection of huge amounts of data, but also by our ability to model complex systems through computational simulations and by advances in machine learning that allow us to efficiently extract patterns and insights from the data (Voit, 2019; Kitano, 2021; Lavin et al., 2021).

While the advent of the data-driven scientific method has generated controversy in various areas of science, cosmology can be argued to be data-driven since the beginning. Dark matter and the cosmological constant are phenomenological components of the overall model that lack an underpinning theory that explains them, and have only been introduced to explain observed datasets. This has been complemented by testing different hypotheses about their nature. It is this back and forth of theory and data that pushes cosmology to its most fascinating discoveries.

In this chapter, we will introduce the basic notions of Bayesian inference and machine learning that will be applied later in this Thesis to problems in cosmology (see Chapters 4 and 5). In the first section, we will introduce the basics of cosmological data analysis through Bayesian inference, whereas in the second part of this chapter we present an introduction to machine learning that will be used throughout this Thesis to accelerate Bayesian inference.

2.1 Solving inverse problems: Bayesian parameter inference

Probability, in a Bayesian context, describes a state of knowledge. It expresses a degree of belief in a proposition based on the available knowledge. In this section, we will show how we can infer the parameters of a given theory and their uncertainties from observed data.

An elementary rule of Bayesian probability defines the basis of parameter inference. The joint probability that two events A and B take place, $P(A, B)$, can be expressed as

$$P(A, B) = P(A|B)P(B), \tag{2.1.1}$$

where $P(A|B)$ is the conditional probability that the event A will take place given that B has taken place. Given that $P(B, A) = P(A, B)$, we can write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.1.2}$$

which is known as Bayes' theorem. This seemingly elementary result actually allows us to go from theoretical forward models $A(B)$ of a given observable, A, to inference of the theory parameters, B. This can be seen once we replace the event A by the parameters of a theory, $\boldsymbol{\theta}$, and the event B by the observed data, $\boldsymbol{d}$,

$$P(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{d})}. \tag{2.1.3}$$

The different components of this equation are known as the

- **Posterior probability**, $P(\boldsymbol{\theta}|\boldsymbol{d})$, that describes the degree of belief in the value $\boldsymbol{\theta}$ after seeing the data $\boldsymbol{d}$. It represents the inferred value of $\boldsymbol{\theta}$ and its estimated uncertainty.

- **Likelihood function**, $P(\boldsymbol{d}|\boldsymbol{\theta})$, which determines the probability of the observed data given a model M and its parameters $\boldsymbol{\theta}$. Note that it is called a function and not a probability, since it is normalised only on the data and not on the parameters.

- **Prior probability**, $P(\boldsymbol{\theta})$, defining the degree of belief in the value of $\boldsymbol{\theta}$ before seeing the data $\boldsymbol{d}$. It expresses our uncertainty in $\boldsymbol{\theta}$ prior to the analysis.

- **Evidence**, $P(\boldsymbol{d})$, the normalisation constant that ensures that the posterior is normalized to unity. It can be thought of as the probability of the data given a model M, integrated over all possible values of the model parameters $\boldsymbol{\theta}$.

The outcome of Bayesian inference is a full probability distribution over the parameters of the theory, the posterior, as opposed to a point estimate such as the value of $\boldsymbol{\theta}$ that maximises the likelihood. Since the number of parameters, $\boldsymbol{\theta}$, may be very large, representing the complete distribution is not always easy. Summaries of the posterior density (such as the mean or standard deviation) for each of the parameters after marginalising over all others are normally used to represent the constraints on the parameters.

Another useful summary is that of two-dimensional subsets of parameter combinations. In cosmology, untangling different effects can be very difficult, and degeneracies among different parameters are very common. For example, clustering of a universe with a high growth rate of structure, f, and a low amplitude of perturbations, σ_8, can resemble clustering of a different universe in which σ_8 is high, but the growth of structure is low. Untangling the growth rate of structure over time from the amplitude of perturbations is therefore very difficult from clustering measurements alone. The two-dimensional representation of the posterior distribution allows us to study such degeneracies between parameters.

2.1.1 Fisher Information

Alternatively, instead of estimating the full posterior distribution of the parameters given the data, we are also interested in estimating the amount of information a sample of the observable carries about the set of unknown parameters. In cosmology, we are often interested in estimating this information from a set of N-body simulations, to quantify the contribution of non-linear scales (Villaescusa-Navarro et al., 2020).

This has become especially relevant for comparing the information content of summary statistics beyond two-point correlation functions. Another use case is that of forecasts. Before the data has been collected, we are often interested in forecasting the constraints that a survey will achieve given its specifications. For both applications, cosmologists have used the Fisher matrix formalism (Fisher, 1935) extensively.

The Fisher information is defined as

$$\mathcal{F}_{ij}(\theta) = \left\langle \left(\frac{\partial}{\partial \theta_i} \log \mathcal{L}(\boldsymbol{s}|\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log \mathcal{L}(\boldsymbol{s}|\theta) \right) \right\rangle_{\boldsymbol{s}}, \tag{2.1.4}$$

where $\mathcal{L}(\boldsymbol{s}|\theta)$ is the likelihood of the data vector given the parameters $\boldsymbol{\theta}$. The expectation is taken over the data, as it is itself a random variable. The derivative of the likelihood with respect to the parameters is also known as the score function $s(\theta) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\mathbf{d}|\theta)$, which is zero at the maximum likelihood point. Eq. (2.1.4) can be interpreted as the variance of the score function, since the expected value of the score function is zero. A random variable that contains high Fisher information implies that the absolute value of the score is often high. The Fisher information is used to quantify the effect that small changes in θ have on the likelihood. If small changes in θ result in substantial variations in the likelihood, then we will be able to set tight constraints on the parameters and we say that the information content of $\mathbf{d}$ in θ is large.

When the likelihood can be differentiated twice, it can be shown (Lehmann & Casella, 1998) that the variance of the score is also related to the second derivative, and therefore to the curvature, of the likelihood function

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \mathcal{L}(\boldsymbol{s}|\theta) \right\rangle , \tag{2.1.5}$$

implying that a more peaked likelihood contains more information on the parameters than a flatter one.

The Cramér–Rao (Cramér, 1946; Rao, 1945) bound states that the diagonal elements of the inverse of the Fisher matrix are a lower bound on the variance of any unbiased estimator of $\boldsymbol{\theta}$

$$\sigma_{\theta_i} \geq \sqrt{(\mathcal{F}^{-1})_{i,i}} \, . \tag{2.1.6}$$

Therefore, we can use the Fisher matrix to estimate the expected error on the cosmological parameters, given the likelihood of the data. In Chapter 5, we will show an application of the Fisher formalism to estimate the information content of environment dependent clustering.

Although a Fisher analysis can be very useful in estimating the information content of different summary statistics, most of the time we are interested in the full posterior distribution over the parameters that allows us to obtain a precise estimate of the uncertainties on the parameters. In the following section, we will present the challenges one faces when trying to estimate posterior densities and explain how these can be overcome to perform parameter inference through Markov Chain Monte Carlo (MCMC) algorithms.

2.1.2 Estimating the posterior: The curse of dimensionality

Calculating the posterior through (2.1.3) is in practice intractable, since it involves a high-dimensional integral (with the dimension given by the parameter space) in the evidence

$$P(\boldsymbol{d}) = \int P(\boldsymbol{d}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{2.1.7}$$

When closed-form expressions for the evidence are lacking, we need to resort to numerical integration. However, in high dimensions, the number of grid points used to calculate the integral grows exponentially with the number of variables. This is known as the curse of dimensionality.

Even if $P(\boldsymbol{d})$ is known, turning it into something useful still requires computations in a high-dimensional space. For example, finding the expected values for $\boldsymbol{\theta}$ or their uncertainties would involve a numerical estimation of high-dimensional integrals.

The most common approach to avoid the curse of dimensionality consists of sampling the posterior directly, instead of computing the probability density function (PDF). An additional consequence of the curse of dimensionality is that most of the posterior mass becomes concentrated on a small proportion of the space in high dimensions. Therefore, we need methods that can exploit the structure of the posterior so that they focus on sampling the small proportion of the parameter space that has a significant posterior mass.

Here, we will present the most commonly used method used to sample the posterior distributions: Monte Carlo Markov Chain.

MCMC: Sampling the posterior

Markov chain Monte Carlo methods are a class of algorithms that are used to sample a probability distribution in a high-dimensional space. They allow us to characterise a distribution through random sampling.

Monte Carlo methods are general techniques that use random sampling to estimate a numerical result. In the context of Bayesian inference, we will use them to estimate a posterior distribution from a collection of N independent and identically distributed samples of the posterior, θ_n,

$$p(\theta|d) \approx \frac{1}{N} \sum_{n=1}^{N} \delta_{\theta_n}(\theta), \tag{2.1.8}$$

or estimate the expected value of a function of parameters, $f(\theta)$,

$$\mathbb{E}_{p(\theta|d)}\left[f(\theta)\right] = \int f(\theta) p(\theta|d) d\theta \approx \frac{1}{N} \sum_{n=1}^{N} f(\theta_n). \tag{2.1.9}$$

We will use Eq. (2.1.9) to compute summaries of the posterior, such as its mean or variance. Therefore, we are only short of a way to obtain a set of independent samples from the posterior distribution, now referred to as $\pi(\theta)$, through its unnormalised version $\pi_u(\theta) = Z\pi(\theta)$, where Z is the unknown normalisation constant in Eq. (2.1.7). In what follows, we explain how Markov chains and the Metropolis-Hastings algorithm (Hastings, 1970) can be used to obtain posterior samples, bypassing the need to estimate the evidence.

A Markov chain is defined as a chain of random samples "without memory", in which the probability of drawing the next sample, θ_{i+1}, depends only on the value of the previous sample, θ_i, and not on the chain of past samples. This is also known as the Markov property.

The Markov chain is then fully characterised by the probability of sampling its first element, $P(\theta_0)$, and the transition probability from one sample to another $T(\theta_{i+1}|\theta_i)$. If we were to take samples that were widespread throughout the chain, we would get nearly independent distributed samples. The goal of MCMC algorithms is to generate a Markov chain that produces samples from the target distribution $\pi(\theta)$.

This is where the Metropolis-Hastings algorithm comes in. It is designed to ensure that the samples of the Markov chain are drawn from the target distribution. It starts from a random sample, θ_0, and uses the so-called transition kernel or proposal distribution, $Q(y|\theta i)$, to draw a candidate sample from its previous one. This proposed sample y is accepted with probability $A(y|\theta i)$. If accepted, it enters the Markov chain $\theta_{i+1} = y$. If the sample is rejected, then $x_{i+1} = x_i$ and the previous step is repeated. The transition probability that characterises the Markov chain is given by $T(\theta_{i+1}|\theta_i) = Q(\theta_{i+1}|\theta_i)A(\theta_{i+1}|\theta_i)$.

The acceptance probability is defined as

$$A(x|y) = \min\left(1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}\right). \tag{2.1.10}$$

It is designed to ensure that given equal transition probabilities between two points in the parameter space, x and y, the sample y would be accepted only if it moves to a higher probability region where the ratio of $\pi(y)/\pi(x)$ is greater than unity. On the other hand, if the ratio $\pi(x)/\pi(y)$ is close to unity, the chain will not move to y if it is difficult to go back to x. In this way, we ensure that the chain will not move far away from high-probability states. Moving locally around regions of high posterior density is indeed the key to avoiding the curse of dimensionality.

Importantly, Eq. (2.1.10) can be evaluated through the unnormalised posterior density π_u, since π appears only as a ratio.

A Monte Carlo approach combined with the Metropolis-Hastings algorithm allows us to estimate the posterior density if we can evaluate the numerator of Eq. (2.1.3). It uses local movements around points of high posterior density to avoid the curse of dimensionality. However, this same feature also means that it is very difficult to sample multi-modal posteriors with Metropolis-Hastings. Moreover, the sequential nature of the algorithm implies that it is difficult to parallelise. These two reasons motivated the development of alternative sampling algorithms, such as nested sampling (Skilling, 2006).

However, regardless of the sampling algorithm, obtaining enough samples such that the posterior estimate converges might require a prohibitively large number of samples. This is especially true when the theoretical model needed to estimate the likelihood is slow to evaluate. In cosmology, obtaining fully non-linear predictions requires the use of N-body simulations. However, these are too slow to allow Bayesian inference of the cosmological parameters. For this reason, we need to develop faster surrogate models that allow us to estimate the cosmological parameters and their uncertainty from galaxy surveys. In the next section, we will introduce a set of machine learning algorithms that can be used to create the surrogate model and hence accelerate Bayesian inference.

2.2 A crash course in machine learning for avid cosmologists

A learning algorithm is described as one that can learn from data, which means that it can improve its performance on a given task through experience, which most of the time we equate to providing the algorithm with data. We can turn a learning problem into an optimisation one by defining a loss function that quantitatively estimates the algorithm's performance at solving the given task. In this context, learning is equivalent to minimising a loss function.

The challenge for machine learning algorithms is generalising to new data, i.e. performing well on new inputs other than those it learnt from. When training the model, we optimise the loss function in a training set. But our goal is to minimise the generalisation error (which will be denoted as the test error from now on). For this purpose, a fraction of the dataset, the test set, has to be left out of the optimisation process assuming that both the training and the test set are independent of each other and that they are identically distributed.

Therefore, learning consists of making the training error small and reducing the difference between the training error and the test error as much as possible. If we fail to lower the training error, we say that the model is underfitted. However, if the difference in training and test error is too large, the model is overfitting. Whereas the first issue can be overcome by modifying the model so that it is more flexible and can fit a wider variety of functions, this might also enable the model to memorise the training set and, therefore, increase the test error.

In the following section, we will introduce the most basic neural network algorithm, multilayer perceptrons, which serve as a building block for more complex models.

2.2.1 Multilayer perceptrons

Multilayer perceptrons (MLPs) or feedforward neural networks, approximate a function f such that

$$\mathbf{y} = f(\mathbf{x}|\boldsymbol{\theta}), \tag{2.2.1}$$

where $\mathbf{x}$ represents the dataset, y the desired outputs, and $\boldsymbol{\theta}$ the free parameters of the network, also known as trainable parameters. The optimization problem is solved by finding the set of $\boldsymbol{\theta}$ values that minimize the loss function.

Neural networks emerge when we chain several of these functions so that

$$\mathbf{y} = f^{(n)}(\ldots(f^{(2)}(f^{(1)}(\mathbf{x}|\boldsymbol{\theta}_1)|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_n). \tag{2.2.2}$$

Each level of the chain is called a layer in the neural network.

In the remainder of this section, we focus on i) How do we define the functions $f^{(n)}$ such that they are flexible enough to reproduce the dependencies observed in our dataset? ii) How do we find the set of optimal parameters $\boldsymbol{\theta}$ given a function f and a performance measure?

The perceptron

Let us begin with the first question. The simplest unit of a neural network, known as the perceptron, applies a non-linear function to a weighted combination of the inputs

$$y_i = \phi\left(\sum_j w_{ji}x_j + b_i\right), \tag{2.2.3}$$

where ϕ is a non-linear function applied element-wise, w_{ji} forms the weight matrix of parameters that we want to optimize, and b_i represents the bias terms which allow for shifts in the input data and are also trainable. The non-linear function is also called the activation and it is the element of the neural network that allows it to describe the non-linear relations between inputs and outputs.

Among the most commonly used activation functions is the rectified linear unit (Agarap, 2018), also known as ReLU

$$f(x) = \max(0, x), \tag{2.2.4}$$

where x is the input to a layer in the neural network.

Sometimes, however, applying (2.2.3) recursively is not the best solution for a problem. The only assumption we have made regarding the function's behaviour is that it will return a smooth interpolation over the inputs. In other cases, such as image data, we might also want to impose constraints on the space of learnable functions. For instance, one might want to learn a translation equivariant function for detecting an object in images such that if the object moves throughout the image, the network produces the same output at a different

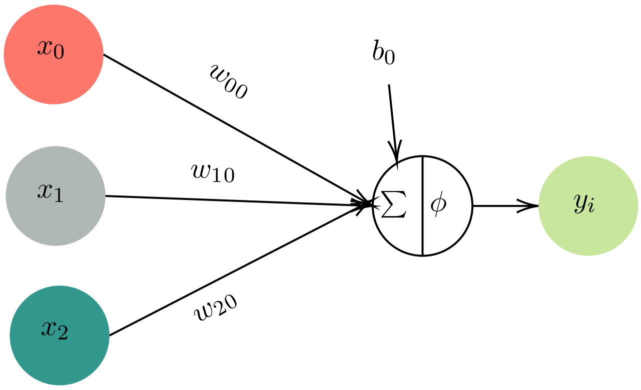


Figure 2.1: Depiction of a perceptron, the basic unit of a neural network that takes a set of inputs, x and maps them to a set of outputs y through a non-linear transformation shown in Eq. (2.2.3). The free parameters of the perceptron that are fitted to the data are the weights w_{ij} and the biases b_i.

position. An example of an architecture that is equivariant to translation is a convolutional neural network (LeCun et al., 2015). Recently, there have been efforts to design architectures that are equivariant (or invariant) to any given group action. For a review of the topic, see Bronstein et al. (2021). Given the importance of symmetries in physics, progress in this area will be fundamental in the development of machine learning methods that are capable of scientific discovery.

Learning as an optimisation problem

Finally, we focus on how to find the set of parameters, $\boldsymbol{\theta}$, that produce the best performance for the task at hand, measured by a scalar loss function, $\mathcal{L}(x, y|\boldsymbol{\theta})$. The nonlinearities introduced in f make the loss function non-convex, meaning that it will have multiple local minima as opposed to one local, and thus global, minimum. Non-convex optimization is still an unsolved problem, and currently iterative methods only aim at driving the loss function to a low value, rather than finding its global minimum.

Most iterative learning algorithms for non-convex loss functions are gradient based. Given that the gradient of the loss function gives the direction in which the loss increases the fastest, if the set of parameters $\boldsymbol{\theta}$ is modified in the direction opposite to the gradients of the loss function with respect to the parameters, it will lead to the largest decrease in the loss function at that point in the parameter space. This algorithm is known as gradient descent and proposes a new value of the neural network parameters

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \epsilon \nabla_{\boldsymbol{\theta}} \mathcal{L}(x, y|\boldsymbol{\theta}), \tag{2.2.5}$$

where ϵ determines the step size and is known as the learning rate. Optimization algorithms that only use the gradient are said to be of the first order.

2.2.2 Machine learning for cosmology

The era of big data in cosmology not only implies that we have to analyse an increasing number of datasets, and possibly their cross-correlations, but also that we are dealing with increasingly complex data that may require similarly complex theoretical methods to extract the relevant information needed to constrain the underlying physical models. Cosmology is now at a new phase in its development, where progress is not hindered by the size of available datasets but instead by our inability to make the most of the data that we have with current statistical techniques.

As stated in Section 1.2.1, N-body and hydrodynamical simulations will be necessary to exhaust the information content of current and ongoing surveys, and how to leverage them to learn about the composition of the universe is going to be an exciting area of development over the next few years.

For these reasons, there has recently been an increase in machine learning (ML) applications in the area of cosmology. We can generally categorize these into three different groups, based on their goals,

1. **Optimise data acquisition and processing**. These are ML models that either guide resource allocation of instruments such as telescopes, or process the observed data to improve its science return. See Cranmer et al. (2021) for an example of the first kind, where the authors trained a model to select which galaxies should be observed in a survey to optimise the constraints on cosmological parameters. Some examples of the second case can be found in the area of gravitational lensing. For example, Jeffrey et al. (2020) used ML to generate mass maps from weak lensing measurements, and Lin et al. (2021) did so to accurately detect strong gravitational lenses in large datasets.

2. **Accelerate predictions**. Given that N-body simulations are computationally very expensive, these cannot be included in a fully Bayesian data analysis pipeline. For example, constraining a 7 dimensional space of cosmology and galaxy occupation parameters with the two-point correlation function already requires $\mathcal{O}(10^4)$ likelihood evaluations at different input parameter values. Therefore, different techniques have been proposed

for emulating the outputs of N-body simulations, which range from emulating the density field (Rodriguez et al., 2018; He et al., 2019) to its summary statistics (Heitmann et al., 2013; Zhai et al., 2019a; Nishimichi et al., 2019a), or painting baryonic physics onto dark matter only simulations (Agarwal et al., 2018; Tröster et al., 2019).

3. **Maximise information gain**. Although we cannot quantify how much information is contained in a given observable (such as the set of tracer galaxy positions or weak lensing maps), different summary statistics have been shown to constrain the cosmological parameters to a better accuracy than two point functions. Recently, researchers have used ML models to learn a summary statistic for a particular dataset (Charnock et al., 2018; Fluri et al., 2021) that maximises its information content, or to directly constrain the cosmological parameters at the field level with neural networks (e.g. Dai & Seljak, 2022; Villaescusa-Navarro et al., 2021). Most applications of the latter type have been developed in the area of weak lensing (e.g. Gupta et al., 2018; Fluri et al., 2018), where deep learning models have been used to analyse real data and produce constraints on the cosmological parameters that improve those from the two-point function alone by about 30% (Fluri et al., 2019).

Although ML applications continue to grow in cosmology and continue to show the potential to revolutionise the field, only a handful of them have managed to transition from proof-of-concept applications using N-body simulations to generating insights from observations. This leap is particularly hard due to our inability to produce complete data models. For instance, in the area of galaxy clustering, we know that our knowledge of galaxy formation is incomplete, but we do not know what the impact of model specification could be when applying a model trained on a simulation to a real dataset. Developing robust models capable of producing calibrated uncertainties will be the next challenge of ML for cosmology.

Moreover, although improving our precision in estimating cosmological parameters is valuable to measure the consistency among different datasets and inform theoretical developments, we also want to understand *how* our models differ from the data. These differences may come from approximations, missing observational effects in our simulations, or, more interestingly, from unknown physics. So far, little work has been done in the direction of comparing simulations and data through interpretable models that could assist in the discovery of new physics in cosmology.

Note that interpretability is an ill-defined concept, and work will need to be done to describe its meaning within cosmology. For example, past research on interpretability and cosmology (Ntampaka & Vikhlinin, 2022) has focused on interrogating deep learning models to assess whether they would introduce biases when compared to data, and not so much on drawing comparisons between data and simulations. This definition of interpretability would, for instance, not enhance the model's capabilities for scientific discovery. Another line of research related to interpretability for the sciences is that of symbolic regression, developed by Udrescu et al. (2020) and Cranmer et al. (2020), which has, for instance, been applied to rediscover the analytical expression for Newton's gravity by learning the dynamics of solar system objects (Lemos et al., 2022).

In this Thesis, we will show an application of machine learning of the second kind listed above (i.e. accelerating predictions). In Chapter 4, we will use neural networks to learn the non-linear mapping between the cosmological parameters and summary statistics of dark matter halo properties, to emulate the outputs of N-body simulations at a speed that allows us to derive the posterior distributions of cosmological parameters with Bayesian inference.

Chapter 3

The real to redshift mapping on small scales

In Section 1.2.1, we show that much of the information that we obtain from 3-D galaxy maps comes from the distorted way in which we observe the universe. In this chapter, we will focus on one of these distortion effects, redshift space distortions, and in particular, on modelling the mapping from the real space two-point correlation function to its redshift space analogue.

The so-called redshift space correlation function, $\xi^S(s_\perp, s_\parallel)$, is a combination of both real space clustering, $\xi^R(r)$, and the probability of finding a pair of galaxies with a given relative velocity along the line of sight, also denoted as pairwise velocity distribution, as we show in Section 3.1 using the Streaming Model of RSD, see e.g. (Fisher, 1995a; Scoccimarro, 2004). Since clustering in redshift space is affected by relative peculiar motions, it contains information about the growth of structure.

State-of-the-art constraints on the growth factor are found measuring the two-point correlation function in redshift space (e.g., Satpathy et al. 2017 for galaxies from BOSS, Zarrouk et al. 2018 for eBOSS quasars), which have reported growth factors consistent with general relativity. The authors in Satpathy et al. (2017), used measurements of the two-point correlation function down to separations of $25\,h^{-1}$Mpc, beyond which theoretical predictions introduce larger systematic errors than the statistical errors of the measurement itself, thus biasing the estimate of the growth factor. For future surveys, the expected statistical errors will be significantly smaller (e.g., Huterer et al. 2015), and so we will need more accurate theoretical predictions down to small scales than those used in the analysis of current surveys, to improve constraints on the growth rate, and to avoid catastrophic interpretation errors (Jennings et al., 2011).

As such, increasing the accuracy of redshift space correlation function models would improve the accuracy of our growth rate estimates. The main hurdle that has to be overcome is the non-linear evolution of the density and velocity fields produced by non-linearities in the continuity and Euler equations that drive gravitational collapse. As we shall see, this is particularly relevant for describing the mapping of pairs from real to redshift space, which is necessary to model the two-point correlation function. The development of this mapping is the focus here.

The goal of this study is threefold. Firstly, we introduce an extension to the simplest Streaming Model, that assumes a Gaussian distribution of relative motions, that improves the accuracy of theoretical predictions for the clustering multipoles. Secondly, we show a comparison of state-of-the-art models for the streaming model ingredients with high resolution N-body simulations. Finally, we analyse the effect of the different velocity moments on the clustering multipoles and assess how accurate their theoretical predictions need to be for an RSD model that is at least as accurate as the measurements from future surveys.

3.1 The streaming model of redshift space distortions

The streaming model describes the mapping from the real-space two-point correlation function to the observed anisotropic two-point correlation function in redshift space. Since objects viewed in redshift space are the same as those in real space, but have been moved to different positions, we can relate their density contrasts by imposing mass conservation

$$\left(1+\delta^S(\mathbf{s})\right)\mathrm{d}^3\mathbf{s} = \left(1+\delta^R(\mathbf{r})\right)\mathrm{d}^3\mathbf{r}, \tag{3.1.1}$$

where superscript S denotes redshift space and R real space. This expression can be further manipulated (Scoccimarro, 2004) to obtain a relation between real and redshift space clustering,

$$1+\xi^S(s_\perp, s_\parallel) = \int_{-\infty}^{\infty}\mathrm{d}r_\parallel \,\left(1+\xi^R(r)\right)\mathcal{P}(v_\parallel = s_\parallel - r_\parallel|\mathbf{r}), \tag{3.1.2}$$

where $r^2 = r_\parallel^2 + r_\perp^2$, $s_\perp = r_\perp$, $\mathcal{P}(v_\parallel|\mathbf{r})$ is the pairwise velocity distribution, and $v_\parallel = v_{\parallel,1} - v_{\parallel,2}$, is the line-of-sight relative velocity of the pair of tracers. In our convention, $v_\parallel$ is defined as negative (positive) if the pairs are approaching (receding from) each other. Eq. 3.1.2 is known as the streaming model (Fisher, 1995a), which simply states that the probability of finding a pair of objects at a distance $\mathbf{s}$ in redshift space is given by the sum over all possible combinations of real space distances, $\mathbf{r}$, and velocities, $\mathbf{v}$, which would make us infer the redshift space position, $\mathbf{s}$. While the streaming model is one way to move forward, fully Eulerian perturbation theory treatments based on the same expression can also accurately describe redshift space clustering (Taruya et al., 2010) on linear or quasi-linear scales.

The plus one terms in Eq. (3.1.2) ensure that given a universe with randomly placed galaxies, if the pairwise velocity distribution is dependent on the pair separation, then we would still observe clustering in redshift space induced by the coherent velocity field. However, if the pairwise velocity distribution does not depend on pair separation, the plus one terms on both sides in Eq. (3.1.2) cancel out.

Note that Eq. (3.1.2) is exact, and the only approximation we have made so far is the plane-parallel approximation to select a particular line of sight. Nonetheless, the apparent simplicity of the streaming model may be deceptive, as the complexity of the gravitational dynamics is hidden in the shape of $\mathcal{P}(v_{\parallel}|\mathbf{r})$ and its dependence on pair separation. Broadly speaking, on small scales within dark matter haloes, virial motions produce a large velocity dispersion that reduces the amount of clustering along the line of sight; the size of this effect increases with halo mass. On larger scales, galaxies in-falling into larger structures shift the mean velocity to negative values, producing a change in the opposite sense to those on small scales, which increases the inferred clustering along the line of sight (Kaiser, 1987).

It has been known for a long time (Scoccimarro, 2004) that this scenario is further complicated by the non-Gaussian nature of the pairwise velocity distribution, which is evident from its nonzero skewness and kurtosis. There is no Gaussian limit for pairwise velocities on large scales, since velocity differences cancel out long-range contributions and leave only the local, nonlinear component of the velocity at the two different locations. Here, we focus on extending the streaming model to include these non-Gaussian features, as predicted by N-body simulations.

Throughout, we will use the relation between the full three-dimensional pairwise velocity and its line-of-sight projection. The line-of-sight pairwise velocity distribution can be obtained by integrating the full distribution $\mathcal{P}(v_r, v_t|r)$, where the radial velocity, v_r, and the transverse velocity, v_t, are defined as the velocity components parallel and transverse to the pair separation vector, respectively. Due to statistical isotropy, we only need to select one component of the two-dimensional transverse velocity. For ease of computation, we chose the one that will contribute to the line-of-sight projection, i.e., the one in the plane spanned by the galaxy pair and the observer; see Fig. 5.1. Thus,

$$v_{\parallel} = v_r \cos\theta + v_t \sin\theta, \tag{3.1.3}$$

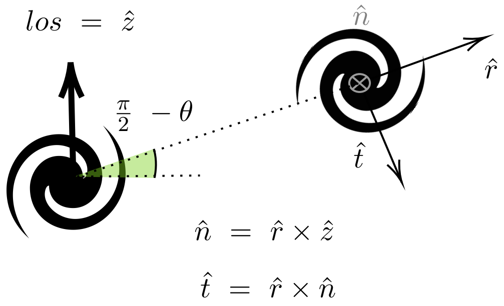

$los = \hat{z}$
$\hat{n}$
$\hat{r}$
$\frac{\pi}{2} - \theta$
$\hat{t}$
$\hat{n} = \hat{r} \times \hat{z}$
$\hat{t} = \hat{r} \times \hat{n}$

Figure 3.1: Decomposition of the three dimensional distance vector into a radial component along the pair distance, $\hat{\mathbf{r}}$, a normal component, $\hat{\mathbf{n}}$, which is perpendicular to both the line of sight direction and the pair separation vector, and a transverse component, $\hat{\mathbf{t}}$, which completes the basis formed by the radial and normal vectors. After projecting the distance vector onto the line-of-sight, only the radial and transverse component will give a non-zero contribution.

where θ is the angle between the pair separation vector and the line of sight, $\theta = \tan^{-1}\left(r_\perp / r_\parallel\right)$. Therefore,

$$\mathcal{P}(v_\parallel | r_\perp, r_\parallel) = \int \frac{\mathrm{d}v_r}{\sin\theta}\, \mathcal{P}\left(v_r, v_t = \frac{v_\parallel - v_r \cos\theta}{\sin\theta} \middle| r\right). \tag{3.1.4}$$

The relations between the moments of the two distributions are given by

$$c_n(r_\perp, r_\parallel) = \sum_{k=0}^{n} \binom{n}{k} \mu^k (1-\mu^2)^{\frac{n-k}{2}} c_{k,n-k}(r), \tag{3.1.5}$$

where c_n denotes the n-th central moment of the line of sight projected distribution, $\mathcal{P}(v_\parallel | r_\perp, r_\parallel)$, and $c_{k,n-k}$ the k-th radial moment, $(n-k)$-th transverse moment of $\mathcal{P}(v_r, v_t | r)$, and $\mu = \cos\theta$. The n-th moment about the origin is denoted as m_n.

3.1.1 The Gaussian Streaming Model

The commonly used model for the redshift space correlation function is known as the Gaussian streaming model (GSM; Fisher, 1995a; Reid & White, 2011). The radial and transverse components of the pairwise velocity are assumed to be independently Gaussian distributed. Therefore, the line-of-sight projection can be written as

$$\mathcal{P}_{\mathrm{G}}(v_\parallel | \mathbf{r}) = \frac{1}{\sqrt{2\pi\sigma_{12}^2(\mathbf{r})}} \exp\left[-\frac{\left(v_\parallel - v_{12}(\mathbf{r})\right)^2}{2\sigma_{12}^2(\mathbf{r})}\right], \tag{3.1.6}$$

where $v_{12}(\mathbf{r})$, denoted as $m_1(\mathbf{r})$ in our notation, and $\sigma_{12}(\mathbf{r})$, equivalent to $\sqrt{c_2(\mathbf{r})}$, are projections of the radial and transverse moments onto the line of sight, and are both dependent on the pair separation vector.

As explained in the previous section, a Gaussian distribution does not accurately describe the pairwise velocity distribution for an evolved matter distribution, even for large pair separations. However, this simplified assumption gives an accurate description of the clustering of dark matter haloes on scales larger than $30\,h^{-1}$Mpc (Reid & White, 2011; Wang et al., 2014). Later on, we shall illustrate how the accuracy of this model stems from the integral in Eq. 3.1.2 over the pairwise velocity distribution, which on large scales only receives contributions from the lowest-order pairwise velocity moments.

Nevertheless, an accurate model on smaller scales requires non-vanishing higher-order moments, mainly the skewness and kurtosis. Different approaches have been taken towards such a model in the literature. On the one hand Uhlemann et al. (2015) performed an Edgeworth expansion around a Gaussian distribution to add skewness and found improvements with respect to the Gaussian streaming model on scales smaller than $30\,h^{-1}$Mpc. We provide a more in-depth discussion of this model in the following section. On the other hand, a number of authors (e.g., Sheth, 1996; Tinker, 2007; Bianchi et al., 2015, 2016; Kuruvilla & Porciani, 2018) have all used mixtures of normal or quasi-normal distributions to model a skewed and heavy-tailed distribution. The first approach by Sheth (1996) modelled the one halo pairwise velocity distribution using a Maxwellian distribution that is then weighted by the Press-Schechter mass function. Tinker (2007) developed a similar approach using the halo model (Cooray & Sheth, 2002), but assuming that, at fixed environmental density around the halo pair, the pairwise velocity distribution of halos is Gaussian. The skewness is then developed by weighting these Gaussian distributions with the probability of finding a given density. The parameters of the model are calibrated using N-body simulations.

Further developments were introduced by Bianchi et al. (2015), who replaced the mixing distribution described above by another Gaussian, which assumes that the mean and standard deviations of the "local" Gaussian distributions are themselves jointly distributed according to a bivariate Gaussian. This model, however, cannot generate distributions that are sufficiently skewed to explain the halo pairwise velocity distribution. This limitation was later overcome by performing an Edgeworth expansion on the local distributions, which added skewness to the Gaussian distribution (Bianchi et al., 2016).

A more recent study by Kuruvilla & Porciani (2018) used a generalised hyperbolic distribution (GHD) to model the pairwise velocity distribution of N-body simulations. In this case the relation between the parameters of the distribution and velocity moments as a function of pair separation is not given, and the model requires five free parameters with a two dimensional dependency on the pair separation vector.

3.1.2 The Edgeworth Streaming Model

The Edgeworth Streaming Model introduced by Uhlemann et al. (2015), is one of the simplest extensions to the Gaussian Streaming Model. The authors used an Edgeworth expansion of the velocity PDF to extend the validity of the Gaussian Streaming Model towards smaller scales. The Edgeworth expansion is an asymptotic series expansion of a probability density function, which implies that there is no guarantee of convergence when more terms are added to the expansion. See Sellentin et al. (2017) for an interesting discussion on the Edgeworth expansion and its applications to cosmology.

Expanding the line of sight velocity PDF around a Gaussian distribution one obtains, to first order,

$$\begin{aligned}\mathcal{P}_{\mathrm{E}}(v_\parallel|\mathbf{r}) &= \frac{1}{\sqrt{2\pi\sigma_{12}^2(\mathbf{r})}}\exp\left[-\frac{\Big(v_\parallel - v_{12}(\mathbf{r})\Big)^2}{2\sigma_{12}^2(\mathbf{r})}\right] \\ &\times\left(1+\frac{\Lambda_{12}}{6\sigma_{12}^3}H_3\left(\frac{v_\parallel - v_{12}}{\sigma_{12}}\right)\right),\end{aligned} \tag{3.1.7}$$

where Λ_{12} is the third order cumulant of the velocity PDF projected onto the line of sight direction, and H_3 the third order probabilists' Hermite polynomial, $H_3(x) = x^3 - 3x$.

In the next section, we present a flexible model that we believe is simpler than the ones mentioned above and achieves similar or better levels of accuracy.

3.1.3 The Skewed Student-t (ST) Streaming Model

A study of the cluster-galaxy cross correlation by Zu & Weinberg (2013) found that the skewed Student-t distribution (ST; Azzalini & Capitanio, 2009) gives an accurate description of the cluster-galaxy pairwise velocity statistics predicted by simulations. The main advantage of using this distribution to model RSD is that its parameters can be written as functions of the four lowest-order moments. Here we use the ST distribution to model the redshift-space clustering of galaxy or halo pairs on all scales.

In recent years, there has been increasing interest in such flexible probability density functions that can accommodate different degrees of skewness and kurtosis. More specifically, a successful approach proposed by Azzalini & Capitanio (2009), found that a skewed, multivariate, distribution can be generated by combining a symmetric density function with a cumulative distribution function as follows

$$f(x) = 2\, f_0(x)\, G\left(w(x)\right),\ x \in \mathbb{R}^d, \tag{3.1.8}$$

where $f_0(x)$ is a symmetric PDF defined in $\mathbb{R}^d$, G is a one-dimensional cumulative distribution function, whose derivative satisfies $G^{'}(x) = G^{'}(-x)$, and w is a real-valued odd function in $\mathbb{R}$.

Since we are interested in a distribution that displays both skewness and extended tails, the symmetric function f_0 can be chosen to be a Student's t-distribution, hereafter referred to as the t-distribution, which in one dimension is given by

$$f_0(x) = t_1(x - x_c|w, \nu) := \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}w\Gamma(\frac{\nu}{2})}\left(1 + \frac{1}{\nu}\left(\frac{x - x_c}{w}\right)^2\right)^{-\frac{\nu+1}{2}}. \tag{3.1.9}$$

The t-distribution is characterised by three parameters: the location x_c, the shape parameter, w, and the number of degrees of freedom, ν. The latter controls the decay of probability in the tails, and therefore allows us to describe distributions with varying degrees of kurtosis.

The skewed multi-variate distribution which originates from the t-distribution by using Eq. 3.1.8 is known as the skew-t distribution, hereafter ST. Its density function for a one dimensional random variable, x, is,

$$\begin{aligned} &f_{\mathrm{ST}}(x|x_c, w, \alpha, \nu) := \\ &\quad \frac{2}{w} t_1(x - x_c|1, \nu) T_1\left[\alpha\frac{(x - x_c)}{w}\left(\frac{\nu + 1}{\nu + \left(\frac{x - x_c}{w}\right)^2}\right)^{1/2}; \nu + 1\right], \end{aligned} \tag{3.1.10}$$

where t_1 is the one dimensional t-distribution defined by Eq. 3.1.9, and T_1 is the one dimensional cumulative t-distribution with $\nu + 1$ degrees of freedom. The ST distribution has an extra skewness parameter, α, compared to the t-distribution.

The dependence of the distribution parameters on the pair separation vector, $\mathbf{r}$, has been omitted for clarity. The relation between these parameters and the four lowest order moments of the ST distribution can be found in Appendix A.1.

3.2 Comparison with N-body simulations

In this section, we assess the performance of the different RSD models by comparing them to a set of dark matter only ΛCDM simulations.

3.2.1 Simulations

We use the Dark Quest (Nishimichi et al., 2019b) set of simulations, which consists of fifteen independent realisations of the density fluctuations in a cosmological volume, adopting the best-fitting cosmological parameters given by the Planck CMB data (Planck Collaboration et al., 2016a)

$$\begin{aligned}\{\omega_b, \omega_c, \Omega_{\rm DE}, \ln\left(10^{10}A_s\right), n_s, w_{\rm DE}\} = \\ \{0.02225, 0.1198, 0.6844, 3.094, 0.9645, -1\},\end{aligned} \tag{3.2.1}$$

where $\omega_b \equiv \Omega_b h^2$, and $\omega_c \equiv \Omega_c h^2$ are the physical density parameters of baryons and cold dark matter, respectively, $\Omega_{\rm DE} = 1 - (\omega_b + \omega_c + \omega_\nu)/h^2$ is the dark energy density parameter (assuming a flat Universe and the neutrino density parameter, ω_ν corresponding to the total mass of 0.06eV for the three neutrino species), A_s and n_s are the amplitude and tilt of the primordial curvature power spectrum normalised at 0.05 Mpc^{-1}, and $\omega_{\rm DE}$ is the equation of state parameter of dark the energy.

The simulations follow the evolution of 2048^3 particles in a comoving box of size $L = 2h^{-1}$Gpc, which translates into a particle mass of $m_p = 8.158 \times 10^{10} h^{-1} M_\odot$, using the Tree-Particle Mesh code GADGET2 (Springel, 2005b). Halo catalogues were constructed using the publicly available ROCKSTAR halo finder (Behroozi et al., 2013a). Here, we focus on accurate predictions for massive central halos, with masses above $10^{13} h^{-1} M_\odot$, and leave the predictions for galaxies to future work. In all figures below, we show the mean simulation measurements over the fifteen independent realisations of the cosmological volume, with errorbars representing one standard deviation of the mean measurements. All results are shown for the $z = 0$ snapshots.

3.2.2 The ingredients of the streaming model

The streaming model, Eq. 3.1.2, takes as input both the pairwise velocity distribution and the real space two-point correlation function. In this subsection we will show the measurements of both ingredients from the simulations, together with their theoretical predictions for the given cosmological parameters.

Using the halo catalogues from the simulations, we measure the pairwise velocity distribution in bins of $0.5\, h^{-1}$Mpc size (note that the velocities are rescaled by $\mathcal{H}$ so that they have the unit of length). As mentioned above, in our convention the pairwise velocity is defined as negative (positive) when the members of the pair are approaching (receding from) each other. We show the measured pairwise distribution from the simulations, for a few selected cases of $(r_\perp, r_\parallel)$, in Fig. 3.2. The figure shows increasing $r_\parallel$ values from left to right, and increasing $r_\perp$ from top to bottom.

The black dots in Fig. 3.2 show the measured pairwise velocity of dark matter haloes, while the lines give the ST (red) and Gaussian (blue) distributions obtained by applying two different methods to find the best-fit parameters, which will be described in Section 3.2.3. It is evident that in all cases the Gaussian distributions are a poorer fit to the simulation measurements than the ST distributions. In particular, by comparing the symbols with the blue curves, we note that for all pair separations there is a significant kurtosis in the simulation data which a Gaussian distribution fails to capture.

In the cases of $r_\parallel = 5.25\,h^{-1}$Mpc and $r_\perp = 0.75$ or $5.25\,h^{-1}$Mpc, the pairwise velocity distributions are also very strongly skewed towards negative $v_\parallel$, which is because such close halo pairs are more likely to be found in high-density regions where haloes approach each other ($v_\parallel < 0$), than in void regions where haloes tend to move apart ($v_\parallel > 0$) . The skewness, however, decreases for much larger $r_\perp$ (e.g. $49.75\,h^{-1}$Mpc, the bottom panel of the central column of Fig. 3.2) or $r_\parallel$ (the right column of Fig. 3.2), because the probabilities of infalling and receding halo pairs tend to be even out for large separations. On the other hand, the left column of Fig. 3.2 shows that for very small $r_\parallel$ (e.g., $0.75\,h^{-1}$Mpc), the skewness is small again, which is because in this case the pair separation vector is nearly perpendicular to the line of sight, and $v_\parallel \approx v_t$, meaning that $v_\parallel$ has equal probability to be in any direction within the plane perpendicular to the pair separation vector due to statistical isotropy, that is, equal probability of $v_\parallel > 0$ and $v_\parallel < 0$.

In Fig. 3.3, we show the radial and transverse pairwise velocity distribution for the halos at different pair separations. The two components are not independent. This figure shows the same physical picture as Fig. 3.2. At small pair separations, but larger than halo size, the radial component has a non-zero (negative) mean, produced by tracers infalling towards larger objects. Given that the infall velocity is different in different environments (Tinker, 2007) and more pairs are likely to be found in high-density environments where members of a pair tend to approach each other, the radial distribution is skewed towards negative values (Juszkiewicz et al., 1998). Note that this would not be true for virial motions in high density regions around halos, but we only consider the motion of halo centres here. At larger pair separations, the radial skewness becomes smaller, but it still has heavy tails. Due to

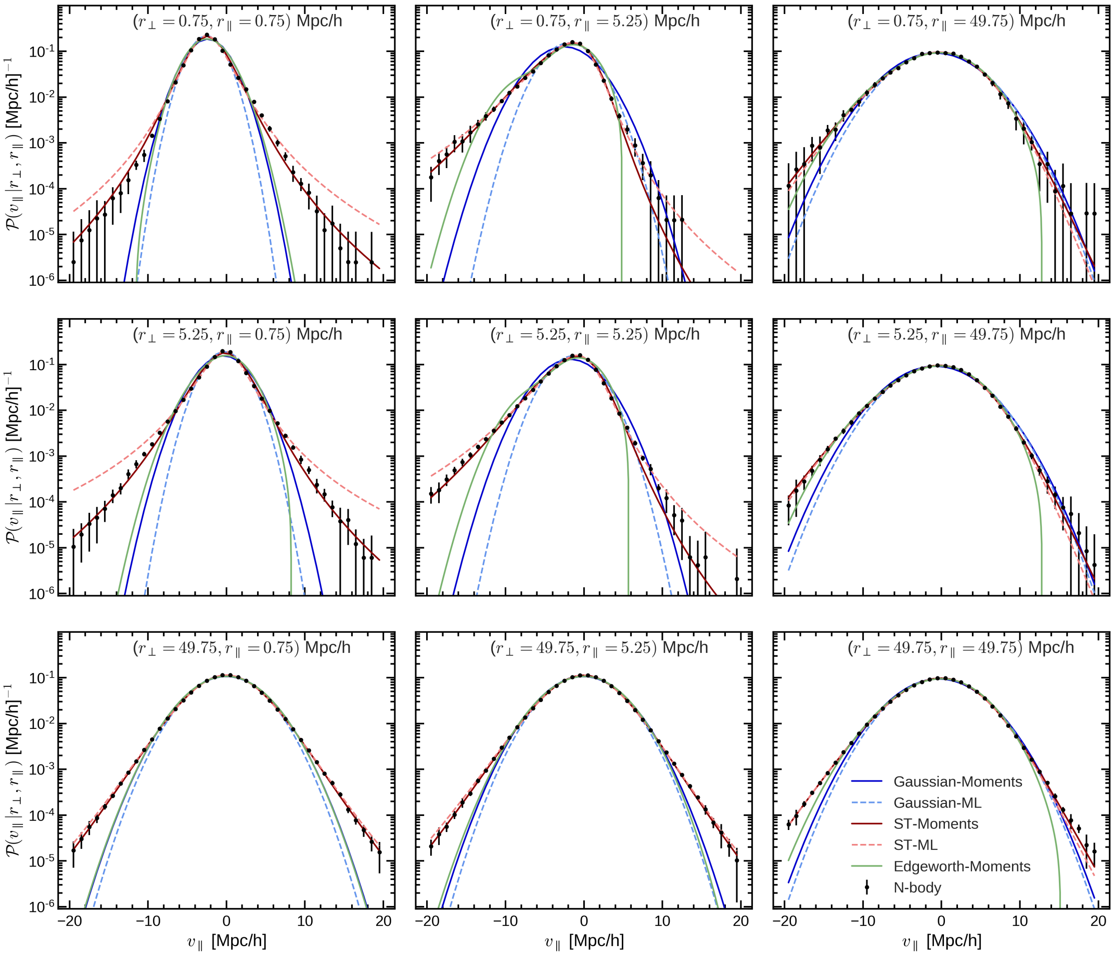


Figure 3.2: The pairwise line of sight velocity distribution for massive dark matter halos in the simulation at $z = 0$, evaluated at different pair-separations. Columns show increasing $r_\parallel$ separation, whilst rows show increasing $r_\perp$. The black dots show the mean measurements from the N-body simulation and their standard deviation, whilst the solid (dashed) curves show the different models found using the method of moments (maximum likelihood) estimate. The Gaussian model is shown in blue, the Edgeworth expansion model is shown in green, and the ST model, that includes skewness and kurtosis, is shown in red.

statistical isotropy, the transverse component is symmetric and has zero mean, although it also shows broader tails than a Gaussian distribution.

The moments of the different distributions are shown in Fig. 3.4, where the definition of the moments is mass-weighted, since the velocity field is only measured where there are tracers, by the number density of tracers at a given separation $r = |\mathbf{r}| = |\mathbf{x}_2 - \mathbf{x}_1|$,

$$m_{i\,j} = \frac{\langle (1+\delta(\mathbf{x}_1))(1+\delta(\mathbf{x}_2)) v_r^i v_t^j \rangle}{\langle (1+\delta(\mathbf{x}_1))(1+\delta(\mathbf{x}_2)) \rangle}, \tag{3.2.2}$$

where i and j denote the order of the moments in the radial and transverse components respectively. For instance, the radial mean is denoted as m_{10}, the second order transverse moment as m_{02}, and the third order cross-correlation between the radial and the squared of the transverse component as m_{12}. The central moments are analogously defined by

$$c_{i\,j} = \frac{\langle (1+\delta(\mathbf{x}_1))(1+\delta(\mathbf{x}_2)) (v_r - m_{10})^i (v_t - m_{01})^j \rangle}{\langle (1+\delta(\mathbf{x}_1))(1+\delta(\mathbf{x}_2)) \rangle}. \tag{3.2.3}$$

Statistical isotropy in the transverse plane implies that only moments with even powers of the transverse component are non-zero. That is c_{12} for the third order moment, and c_{22} for the fourth.

Although it is not the objective of this chapter to develop the relations between the cosmological parameters and the ingredients of the streaming model (the real-space two-point correlation and the pairwise velocity moments), for completeness we show the predictions from different methods as a summary of the recent progresses in perturbation theory. This exercise will show what stage we have reached in our efforts to predict these quantities and what still needs to be done. So far, only predictions for the first two moments of the velocity field have been successfully obtained from perturbation theory:

- *Linear perturbation theory* – Fisher (1995a) shows that the mean pairwise velocity in linear theory is determined by the correlation between the density and velocity fields, $\langle \delta v \rangle$, due to the mass-weighting factors in Eq. 3.2.2. The variance, however, is determined by the velocity-velocity coupling (Gorski, 1988). In the simplest flavour of Eulerian perturbation theory, there are two free parameters: the linear bias and the growth factor. Higher order corrections to the mean and the variance were computed in Reid & White (2011) by expanding the continuity and Euler equations in powers of the linear density field up to fourth order. They also used a local Lagrangian prescription for the bias (Matsubara, 2008), which turned out to be very important to reproduce the real space correlation function, since a local bias in Lagrangian space introduces a

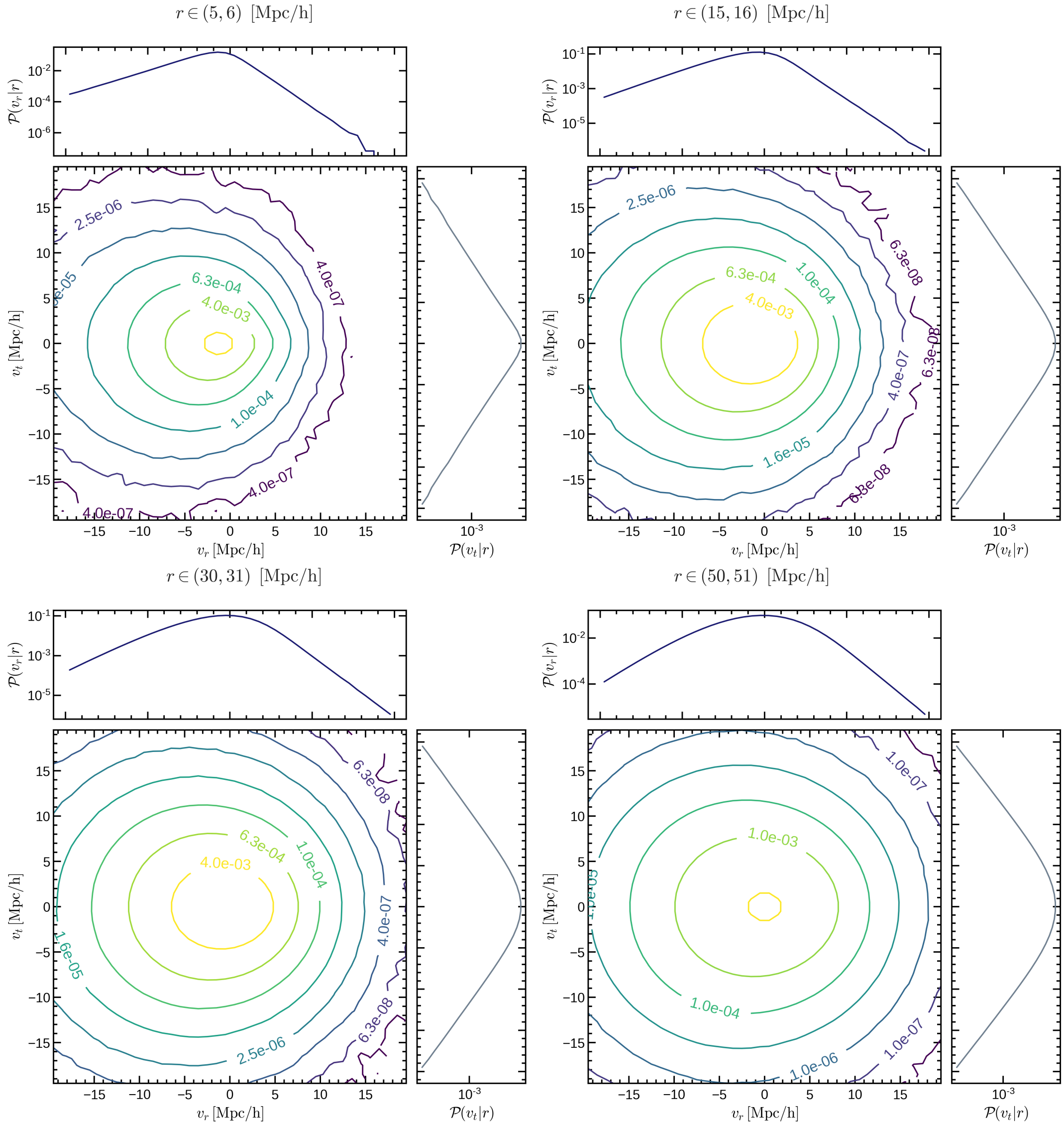


Figure 3.3: The mean joint probability distribution of the radial and transverse pairwise velocities of dark matter halos measured in N-body simulations. The marginal distributions are shown on the sides. At small pair separations, infall towards larger structures produces a large skewness in the radial component, and the mean turns more negative. At large pair separations, the distributions of the two components are symmetric but still show heavy tails.

non-local bias in Eulerian space (Baldauf et al., 2012; Chan et al., 2012).

- *Convolutional Lagrangian perturbation theory (CLPT)* – Wang et al. (2014) extended the formalism of Carlson et al. (2013), to include predictions for the lowest-order pairwise velocity moments. The Lagrangian approach formulates the problem in terms of initial positions and displacement field, where the latter fully specifies the motion of the cosmological fluid. Instead of expanding the fluid equations in terms of the linear density field, the expansion is performed on the displacement field that gives the mapping between initial Lagrangian positions and final Eulerian positions. To describe the Lagrangian bias functional, $\delta_h = F[\delta]$, the authors include three free parameters, b_1, b_2, and b_s, which we fit to the real-space two-point correlation function. The first two of these bias parameters, b_1 and b_2, are the first and second derivatives of the Lagrangian bias function with respect to a long-wavelength density contrast, δ^{L}, whereas b_s encodes the dependence of the bias on a long-wavelength tidal tensor. The variance of the pairwise velocities is, however, not accurately reproduced by CLPT: a constant shift needs to be added to describe the variance on linear scales. Interestingly, this constant offset is the same for both the radial and transverse components, as one would expect from the effect of virial motions. Including the growth factor, CLPT requires five parameters to describe clustering in redshift space.

- *Convolutional Lagrangian effective field theory (CLEFT)* – Carrasco et al. (2012) developed an analytical effective field theory to capture the effects of very small scales on large-scale observables. Vlah et al. (2016) used this idea, together with CLPT, to predict the lowest-order velocity moments that enter the Gaussian streaming model. They found that predictions for the mean pairwise velocity were greatly improved compared to CLPT, especially the derivative, which ultimately controls the accuracy of the redshift space quadrupole. Moreover, it was shown that in the context of effective field theory, the constant shift in Wang et al. (2014) was identified as one of the effective parameters to describe the effect of small scales. Increased accuracy comes at the expense of requiring more free parameters, the effective field theory counter-terms. There are two extra parameters, one for the real space correlation function and the other for the mean pairwise velocity. Therefore, the simplest CLEFT has seven parameters.

The top two panels of Fig. 3.4 compare the predictions for the two lowest-order moments of the three different methods[1]. The symbols show measurements from simulations. In the upper panel we show the mean of the radial pairwise velocity. The two extra EFT counter-terms extend the agreement of CLPT with N-body simulations from scales of $\sim 60\,h^{-1}$Mpc down to $\sim 20\,h^{-1}$Mpc. For the radial and transverse components of the variance, shown in the second panel, the CLPT and CLEFT predictions are qualitatively similar. The reason for this is that the EFT counter-term is very close to a constant shift in the variance, which is already included in CLPT to match N-body simulation results. The moment predicted with the lowest accuracy is the radial component of the variance, where per cent-level predictions are limited to scales above $40\,h^{-1}$Mpc. The radial component of the variance, $c_{20} = m_{20} - m_{10}^2$, has a contribution from the mean pairwise velocity and will also be affected by errors in modelling non-linear infall.

The bottom two panels of Fig. 3.4 show the simulation measurements of the third- and fourth-order moments (symbols). Perturbation theory predictions for moments higher than the second have only been obtained for the third order moment using CLPT in Uhlemann et al. (2015). However, the authors found that it fails to capture the non-Gaussian effects encoded in the skewness for scales below $100\,h^{-1}$Mpc. Since third and fourth order moments only play a role on the accuracy of the redshift space correlation function on small scales, it is extremely difficult to produce accurate enough predictions to unlock access to the cosmological information contained on those scales. To improve these predictions, we plan to explore both effective field theory extensions to CLPT, and the use of emulators for the moments on small scales.

In Fig. 3.5, we also show simulation measurements (symbols) of the real-space two-point correlation function of dark matter halos, together with the predictions using both CLPT (dashed line) and CLEFT (dash-dotted line). The CLEFT prediction is accurate over a broad range of scales – it gives per cent-accuracy results on scales between 15 and $70\,h^{-1}$Mpc – at the expense of only one extra free parameter. For more details on the accuracy of the different perturbation theory models, we refer the reader to Appendix A.3.

3.2.3 Fitting the pairwise velocity distribution

To infer the parameters of the Gaussian and ST distributions that best fit the simulation measurements, we use two different methods,

[1] Perturbation theory predictions have been obtained using the publicly available code github/CLEFT_GSM.

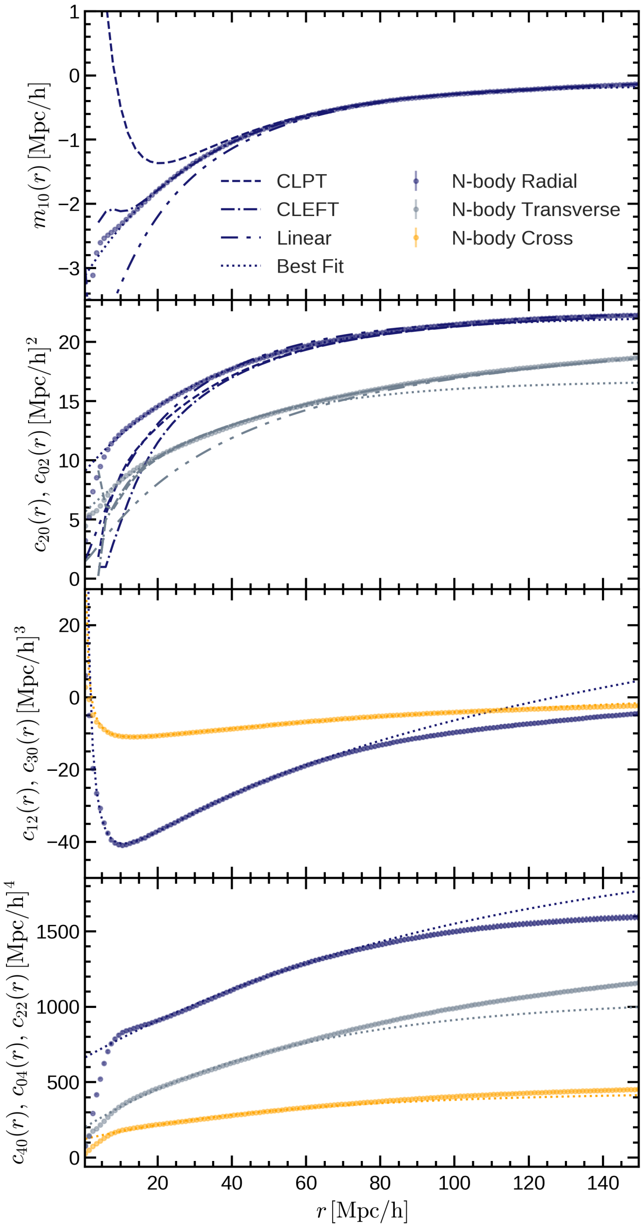


Figure 3.4: The four lowest order moments of the radial and transverse pairwise velocity distributions of dark matter halos. In each panel we show the mean measurements from the simulations, together with errorbars showing one standard deviation (note these are too small to be seen). We also show the different perturbation theory predictions for the two lowest order moments. Linear theory is shown in dotted-dashed-dashed lines, CLPT in dashed lines, and CLEFT in dashed dotted lines. Finally, we show the best-fitting curves as dotted lines, which are used to show the accuracy of the Taylor expansion in Section 3.3.2. Note the best-fitting curves have been fitted to the moments on scales smaller than $60\,h^{-1}$Mpc.

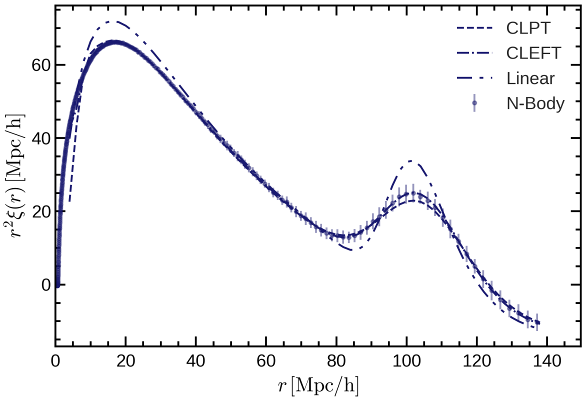


Figure 3.5: The real space correlation function measured from the simulations for dark matter halos, compared with predictions from CLPT and CLEFT. Both these perturbation theory predictions use a Lagrangian prescription for the bias, and have been computed by simultaneously fitting the correlation function and the two lowest order pairwise velocity moments. For more details on the fitting see Appendix A.3.

- *Maximum likelihood estimation*, found by maximising the probability that the model reproduces the simulation measurements. This is equivalent to a least-χ^2 fit to the simulation measurements using the given Gaussian or ST distribution function when the errors are approximately Gaussian distributed. We refer to this method as 'ML' occasionally in this chapter. The χ^2 we minimise is,

$$\chi^2(\mathbf{r}) = \sum_{v_\parallel} \frac{\Big(\mathcal{P}_{\rm measured}(v_\parallel|\mathbf{r}) - \mathcal{P}_{\rm model}(v_\parallel|\mathbf{r})\Big)^2}{\mathcal{P}_{\rm model}(v_\parallel|\mathbf{r})}, \tag{3.2.4}$$

  where $\mathcal{P}_{\rm model}$ is either a Gaussian or an ST distribution.

- *Method of moments*, that uses the analytical relation between the parameters of the distribution and its lowest-order moments to convert the moments estimated from the simulation measurements into distribution parameters.

If the distribution measured from the simulation and the fitted distribution are the same, both methods are equivalent. However, this is not the case when the fitted distribution is an approximation to the simulation results or when noise is present.

In Fig. 3.2 we show the best-fitting distributions for both the Gaussian and the ST models using these two approaches. For the Gaussian case the conversion between moments and parameters is trivial, while for the ST model we have used the relations given in Appendix A.1 to obtain the model parameters given the four lowest-order moments. In this figure we can see that even for large pair separations the Gaussian approximation is inaccurate, where the method of moments and the maximum likelihood estimation produce slightly different results, both being poor approximations.

The ST model, however, is flexible enough to represent the varying degrees of skewness and kurtosis over a broad range of pair separations when using the method of moments. At large separations, the maximum likelihood estimate and the method of moments produce similar distributions. Nonetheless, on small scales the tails of the distribution are mis-estimated by the maximum likelihood method.

For a more detailed comparison of the different models around the peak of the distribution see Fig. A.1.

3.2.4 The redshift space correlation function

In this subsection we use the Gaussian and ST models of the pairwise velocity distribution with the streaming model (Eq. 3.1.2) to predict redshift space clustering. We will focus on the mapping between real and redshift spaces, and show that using the more flexible ST distribution for pairwise velocity leads to more accurate predictions of the higher order redshift-space multipoles than are obtained with the simpler GSM model. For this reason, we measure all the real-space quantities from the simulation, including the real-space halo two-point correlation function and the pairwise velocity distribution moments, as inputs to reproduce the redshift space clustering by using Eq. 3.1.2. The impact of the accuracy of the modelling of the individual ingredients of the streaming model will be studied in a later section.

The pairwise velocity distribution has been measured in the range $0 < r_{\|}/[\,h^{-1}\mathrm{Mpc}] < 70$ and $0 < r_{\perp}/[\,h^{-1}\mathrm{Mpc}] < 50$ in bins spaced by 0.5 h^{-1}Mpc. To perform the streaming model integration in Eq. 3.1.2 we have used the Simpsons rule, with a linear interpolation of the real space correlation function and the pairwise velocity distribution.

Due to the difficulty of analysing two dimensional plots, together with the complex covariance matrix between the different measurements for $\xi^S(s_\perp, s_\|)$, it is common to decompose the redshift space correlation function into multipole moments using its Legendre expansion (Hamilton, 1998),

$$\xi(s,\mu) = \sum_\ell \xi_\ell(\mu) L_\ell(\mu), \tag{3.2.5}$$

where ℓ is the order of the multipole and $L_\ell(\mu)$ is the Legendre polynomial at the ℓ-th order, which depends on the angular coordinate $\mu = \cos\theta$. The redshift space correlation function is symmetric in μ, so only even values of ℓ give a non-zero contribution. Inverting Eq. (3.2.5), we find that the multipole moments are given by

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1} \xi(s,\mu) L_\ell(\mu) d\mu. \tag{3.2.6}$$

The three lowest multipoles are denoted as monopole ($\ell = 0$), quadrupole ($\ell = 2$) and hexadecapole ($\ell = 4$). Recent cosmological analyses are based mainly on the monopole and quadrupole moments, however the cosmological information carried by the hexadecapole has also been shown to be important (Taruya et al., 2011).

We show these three multipole moments predicted by the different models, as well as the measurements from the simulations, in Fig. 3.6. In the lower subpanels of each panel we show the relative differences between the model predictions and the simulation results in units of the standard deviation (σ) (middle subpanel) calculated using the 15 simulation realisations each of which has a volume of 8 $(h^{-1}\mathrm{Gpc})^3$, and the relative percent error in the lowest subpanel. The yellow horizontal shaded bands represent the $\pm 1\sigma$ ranges on the multipoles.

Surprisingly, the two Gaussian distributions that we found by using the method of moments and the maximum likelihood estimate yield multipoles that can be more than five standard deviations away from each other. Furthermore, the Gaussian distribution obtained using the method of moments reproduces the three multipoles within one standard deviation for scales larger than approximately $30\,h^{-1}$Mpc, although it gives a very poor fit to the pairwise velocity distribution on these scales; cf. Fig. 3.2.

Regarding the ST model, although the maximum likelihood ST lies closer than the Gaussian method of moments to the pairwise velocity measured in the simulation (see Fig. 3.2), it gives a biased result for the multipoles. On the other hand, the ST model found by the method of moments is able to reproduce the correct clustering down to scales of around $10\,h^{-1}$Mpc.

The Edgeworth model does improve the predictions of the multipoles compared to the Gaussian Streaming Model, however to extend its validity to even smaller scales we need to also add fourth order moments.

As a result of the large simulated volume, the error bars on the monopole and quadrupole on scales below $20\,h^{-1}$Mpc are extremely small, meaning that the one sigma deviations for the monopole and quadrupole (the yellow horizontal bands in the lower subpanels of Fig. 3.6) are within one per cent of the mean measurement up to $20\,h^{-1}$Mpc.

Finally, although the measurement of the hexadecapole is itself very noisy, the ST model is within one standard deviation for scales larger than around $10\,h^{-1}$Mpc, whilst the Gaussian

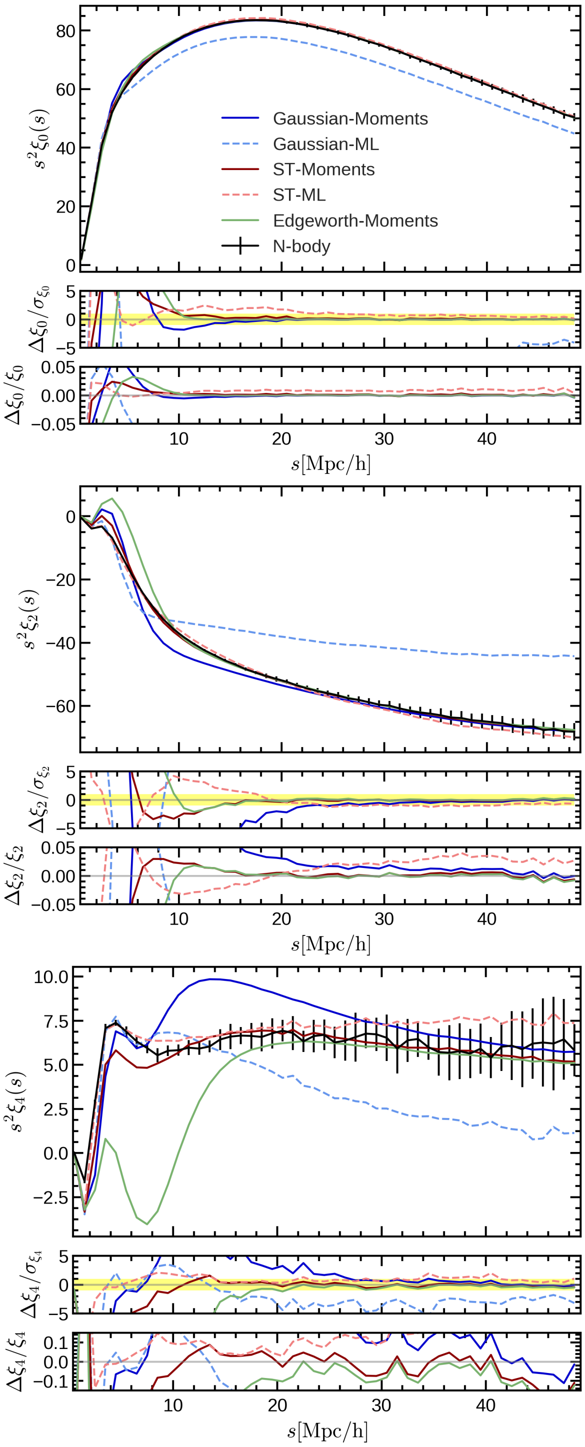


Figure 3.6: Comparison of the accuracy of the different models for reproducing the multipoles of the redshift space correlation function. In the upper sub-panels the multipole directly measured from the simulation is shown together with the model predictions. In the lower sub-panels the deviation between the model and the simulation in units of the variance calculated across the different independent simulations are shown. The yellow bands show the 1σ deviation.

model on those scales is already more than five sigma away from the measurement from the simulations.

To sum up, we have found that the use of the method of moments is critical to accurately reproduce the clustering on quasi-linear scales. The accuracy of the Gaussian streaming model we obtain is consistent with previous findings (Reid & White, 2011; Wang et al., 2014; Bianchi et al., 2016): the prediction is within the measurement errors from the simulations for scales larger than $30\,h^{-1}$Mpc. However, the model prediction rapidly diverges from the simulation results on smaller scales. On the contrary, the ST model is able to reproduce the redshift space clustering very accurately on scales down to $10\,h^{-1}$Mpc, by introducing a pairwise velocity distribution that incorporates the skewness and kurtosis of the pairwise velocity PDF.

On the other hand, we need to understand why the Gaussian model reproduces the clustering on scales above $30\,h^{-1}$Mpc more accurately than the ST distribution obtained through the ML method, even though the latter is a better description of the pairwise velocity distribution on those scales, as shown in Fig. 3.2. To this end, we will study the behaviour of the integrand of Eq. 3.1.2 in more detail in the next section.

3.3 The importance of the moments for accurate clustering predictions

In this section we show how the accuracy of the streaming model on quasi-linear scales is directly related to the lowest order moments of the pairwise velocity distribution. We start by studying how well the different models reproduce the streaming model integrand.

3.3.1 Lessons from the streaming model integrand

We show the integrand of Eq. 3.1.2, for a few pair separation vectors, in Fig. 3.7. In broad terms the integrand is the outcome of a competition between the probability of finding a pair of haloes at a given separation, i.e. the two-point correlation function, and the probability that the pair has the necessary relative velocity to move from real space position $\mathbf{r}$ to redshift space position $\mathbf{s}$. Whilst the first quantity is evaluated as $\xi\left(\sqrt{s_\perp^2 + r_\parallel^2}\right)$ for fixed $s_\perp$, and therefore peaks at $r_\parallel \sim 0$, the latter is evaluated as $\mathcal{P}(v_\parallel = s_\parallel - r_\parallel)$, and peaks around its mean, close to $v_\parallel \approx 0$ ($r_\parallel \approx s_\parallel$) for large pair separations.

The effect of this competition can be seen in Fig. 3.7. For large pair separations, e.g., as shown by the middle and bottom panels, the real-space correlation function is small, $\xi\left(\sqrt{s_\perp^2 + r_\parallel^2}\right) \ll 1$, so that the integrand is dominated by $\mathcal{P}(v_\parallel = s_\parallel - r_\parallel)$ and has a peak at $r_\parallel \approx s_\parallel$. On the other hand, for small pair separations (the top panel), ξ is no longer negligible and the integrand acquires a second, albeit smaller peak around $r_\parallel \approx 0$.

As for the different streaming models, the Gaussian one obtained through the method of moments systematically shifts the main peak of the integrand from its true position, and makes it wider. Although this seems to be a poorer estimate of the integrand than the Gaussian model obtained through maximising the likelihood, which is consistent with what Fig. 3.2 suggests, it predicts the clustering multipoles with a precision that is one order of magnitude higher after integrating, as can be seen on the resulting redshift space correlation function annotated on Fig. 3.7. This same effect is present on all scales larger than $s \approx 30\,h^{-1}$Mpc.

More interestingly, both the ST moments and the ST ML methods give visually much better predictions for the integrand than the Gaussian moments method, which is a consequence of the pairwise velocity distribution being non-Gaussian for all pair separations. However, as shown in the previous section, after integration we find that the Gaussian model yields a comparable accuracy to the non-Gaussian ST model on scales larger than $30\,h^{-1}$Mpc for the monopole and quadrupole. This coincidental behaviour has been noted previously by Kuruvilla & Porciani (2018). Taking the middle panel of Fig. 3.7 as an example, the "errors" of the integration in Eq. 3.1.2, $\Delta\xi^S$, defined as the difference between the integration of the model curve and the integration using the simulation results (black dots), for the four streaming models considered here, are shown in the figure labels. We note that the Gaussian moments method gives a slightly smaller error than the ST Moments at the particular pair separation shown. As the former underestimates the integrand for $14 \lesssim r_\parallel/(\,h^{-1}\mathrm{Mpc}) \lesssim 20$ and $r_\parallel \gtrsim 26\ h^{-1}$Mpc, and overestimates it in other regimes, this seems to suggest that a precise cancellation of the errors from different $r_\parallel$ intervals takes place, which makes the final integration result accurate. However, this cancellation of errors happens for all larger pair separations $s > 30\ h^{-1}$Mpc. In the next subsection, we will show that this is a consequence of the integration being sensitive only to the moments of the pairwise velocity distribution. In particular, for large pair separations it is the two lowest-order moments which dominate the outcome of the integral Eq. 3.1.2, while higher order moments only become important on scales smaller than $30\,h^{-1}$Mpc (Fig. 3.6).

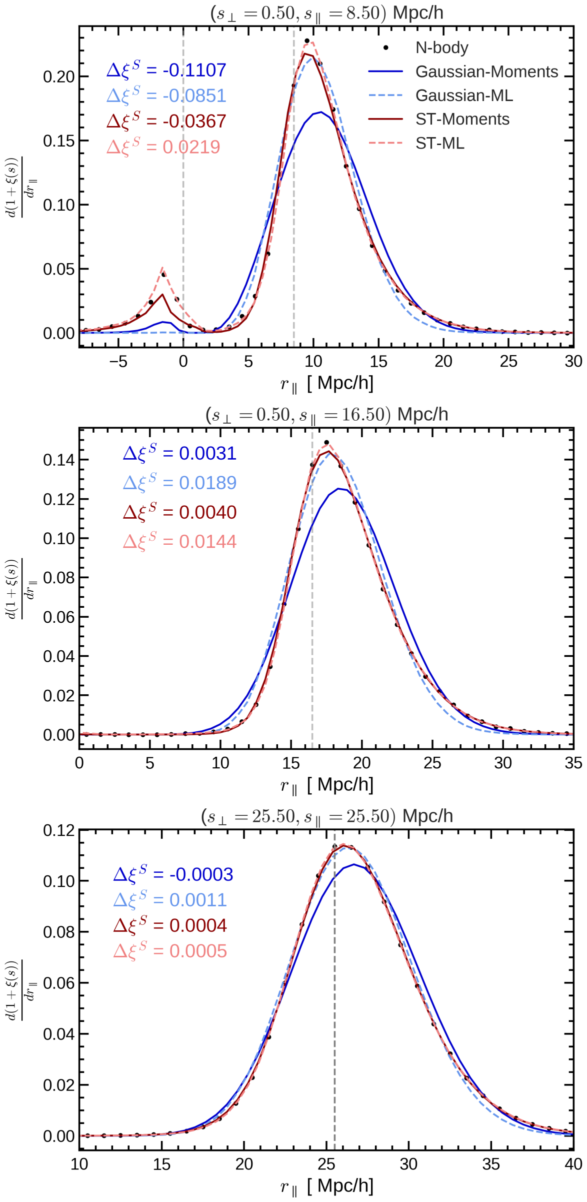


Figure 3.7: Integrand of Eq. 3.1.2 shown for different redshift space pair separations. At small pair separations and for small μ (top panel) we find two peaks situated at $r_\parallel = 0$ and $r_\parallel = s_\parallel$, marked by the grey vertical dashed lines. For larger μ (middle and bottom panels), the second peak dominates since the correlation function decays rapidly at large separations. The result of the integral for the different models minus the integral obtained using the pairwise velocity distribution measured from the simulations, $\Delta\xi(s)$, is plotted with different models shown by the different colours and line styles as shown by the legend in the top panel.

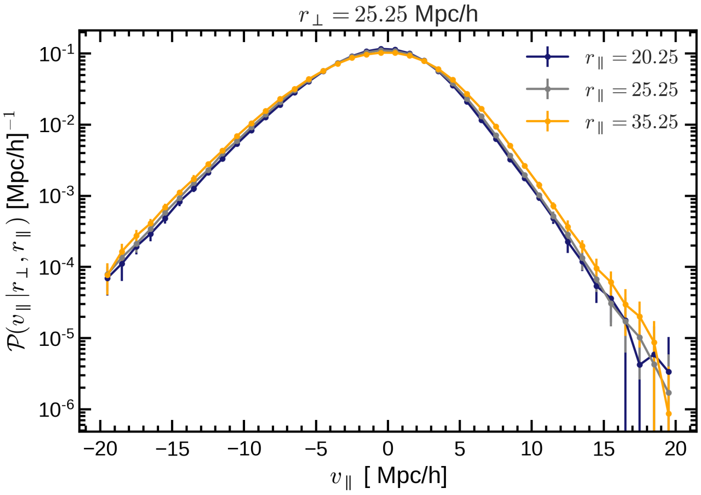


Figure 3.8: The dependence of the pairwise velocity distribution on $r_\parallel$, for fixed $r_\perp = 25.25\,h^{-1}$Mpc and for three values of $r_\parallel$ over a range of $15\,h^{-1}$Mpc $(20.25, 25.25$ and $35.25\,h^{-1}$Mpc). The distribution has only a weak dependence on $r_\parallel$, which is why the Taylor expansion described in the text works.

3.3.2 The importance of the moments on quasi-linear scales

The integration in the streaming model, Eq. 3.1.2, is different from taking the expectation value of $1 + \xi^R(r)$ since the pairwise velocity distribution $\mathcal{P}(v_\parallel|\mathbf{r})$ is different for different $r_\parallel$ values, rather than a fixed probability distribution function $\mathcal{P}(v_\parallel)$. However, $\mathcal{P}(v_\parallel|\mathbf{r})$ is a slowly varying function of pair separation $r_\parallel$, for $r_\parallel \gtrsim 15\,h^{-1}$Mpc, as can be seen in Fig. 3.8. The outcome of the streaming model integral for the values $(s_\perp = r_\perp = 25.5, s_\parallel = 25.5)$, shown in the bottom panel of Fig. 3.7, is dominated by contributions from the pairwise velocity distribution in the range $20\ h^{-1}$Mpc $< r_\parallel < 35\ h^{-1}$Mpc, which is the range of values shown in Fig. 3.8. The same features is found at other separations larger than about $10\ h^{-1}$Mpc.

Therefore, we can Taylor expand the integrand around its peak at $r_\parallel = s_\parallel$ as follows,

$$\Big(1 + \xi^R(r)\Big)\,\mathcal{P}(v_\parallel|\mathbf{r}) \approx \Big(1 + \xi^R(s)\Big)\,\mathcal{P}(v_\parallel|\mathbf{s}) + \\ + \sum_n \frac{1}{n!}(r_\parallel - s_\parallel)^n \frac{d^n}{ds_\parallel^n}\Big((1 + \xi^R(s))\mathcal{P}(v_\parallel|\mathbf{s})\Big)\,. \tag{3.3.1}$$

This expansion was already used by Fisher (1995a), Scoccimarro (2004) and Bianchi et al. (2015) to obtain the Kaiser limit of the streaming model. Here, we will show that this is still accurate on quasi-linear scales.

Inserting Eq. 3.3.1 into Eq. 3.1.2, we find that the derivatives with respect to $s_\parallel$ can be taken out of the integral over $r_\parallel$, together with the real space correlation function, and therefore after integration we are left with the derivatives of the moments through the dependency of $v_\parallel$ on $r_\parallel$. For the lowest order term on the right-hand side of Eq. 3.3.1, we find after a change of variables $v_\parallel = s_\parallel - r_\parallel$,

$$\int_{-\infty}^{\infty} dv_\parallel \left(1 + \xi^R(s)\right) \mathcal{P}(v_\parallel|\mathbf{s}) = \left(1 + \xi^R(s)\right), \tag{3.3.2}$$

whilst for the higher order terms,

$$\begin{aligned} \int_{-\infty}^{\infty} dv_\parallel (-1)^n \sum_n \frac{1}{n!} v_\parallel^n \frac{d^n}{ds_\parallel^n} \left((1 + \xi^R(s))\mathcal{P}(v_\parallel|\mathbf{s})\right) = \\ \sum_n \frac{(-1)^n}{n!} \frac{d^n}{ds_\parallel^n} \left((1 + \xi^R(s)) \int_{-\infty}^{\infty} dv_\parallel v_\parallel^n \mathcal{P}(v_\parallel|\mathbf{s})\right) = \\ \sum_n \frac{(-1)^n}{n!} \frac{d^n}{ds_\parallel^n} \left((1 + \xi^R(s)) m_n(\mathbf{s})\right), \end{aligned} \tag{3.3.3}$$

where m_n denotes the n-th order moment about the origin of the pairwise velocity distribution, which is related to the central moments through,

$$m_n = \sum_{k=0}^{n} \binom{n}{k} c_k m_1^{n-k}. \tag{3.3.4}$$

As a result, an approximation to the streaming model is given by,

$$\xi^S(s_\perp, s_\parallel) \approx \xi^R(s) + \sum_n \frac{(-1)^n}{n!} \frac{d^n}{ds_\parallel^n} \left((1 + \xi^R(s)) m_n(\mathbf{s})\right), \tag{3.3.5}$$

where the integral of Eq. 3.1.2 has now been replaced by derivatives of the pairwise velocity moments evaluated at the redshift space position, $\mathbf{s}$. Consequently, for large pair separations, where the above approximation works well, the exact shape of the pairwise velocity distribution does not affect the clustering, and it is only the moments of the distribution that influence the redshift space correlation function. This explains why the Gaussian moments model works so well in Fig. 3.6 while the Gaussian ML model, which describes the integrand better, fails to reproduce the multipoles.

We can use Eq. 3.3.5 to obtain analytical predictions for the redshift space clustering based on the moments. Up to first order terms the resulting expression is simply

$$\begin{aligned} \xi^{(1)}(s, \mu) \approx \xi^R(s) - \frac{d\xi^R(s)}{ds} m_{10}(s)\mu^2 \\ - \left(1 + \xi^R(s)\right) \left(\frac{m_{10}(s)}{s}(1 - \mu^2) + \frac{dm_{10}(s)}{ds}\mu^2\right), \end{aligned} \tag{3.3.6}$$

where $m_{10} = m_{10}(s)$, defined by Eq. 3.2.2, denotes the radial mean infall.

Interestingly, we can use Eq. 3.3.6 to derive the linear order two-point correlation function for any two tracers, including the void-matter cross-correlation. It has been argued in Nadathur & Percival (2019) that it is not correct to use the standard streaming model result for the galaxy correlation in the void-galaxy case. However, regardless of how the void-matter pairwise velocities are distributed, we can use the Taylor expansion in Eq. 3.3.6, together with the linear theory prediction for the mean pairwise velocity, to derive the linear void-matter two-point correlation function in Nadathur & Percival (2019).

The linear void-matter mean pairwise velocity (Peebles, 1980a; Sheth et al., 2001; Nadathur & Percival, 2019) is given by

$$v_{vm} = -\frac{1}{3} f \Delta(r) r \tag{3.3.7}$$

where $\Delta(r)$ is the average mass density contrast within radius r of the void centre,

$$\Delta_r = \frac{3}{r^2} \int_0^r \delta(y) y^2 \mathrm{d}y. \tag{3.3.8}$$

Substituting Eq. 3.3.7 into Eq. 3.3.6 we find

$$\begin{aligned} \xi_{vm}(s,\mu) \approx & \xi^R_{vm}(s) + \frac{f}{3}\Delta(s)\left(1+\xi^R_{vm}(s)\right) \\ & + f\mu^2\left[\delta(s)-\Delta(s)\right]\left(1+\xi^R_{vm}(s)\right) + \frac{f\mu^2}{3} s \frac{d\xi^R_{vm}}{ds}\Delta(s), \end{aligned} \tag{3.3.9}$$

which is exactly equation (21) in Nadathur & Percival (2019). Therefore, the linear void-matter cross-correlation can be derived from the streaming model.

Going back to Eq. 3.3.5, we compute the first order multipoles

$$\xi_0^{(1)}(s) \approx \xi^R - \frac{1}{3}\frac{d\xi^R}{ds} m_{10} - \frac{1}{3}\left(1+\xi^R\right)\left(2\frac{m_{10}}{s} + \frac{dm_{10}}{ds}\right), \tag{3.3.10}$$

$$\xi_2^{(1)}(s) \approx -\frac{2}{3}\frac{d\xi^R}{ds} m_{10} + \frac{2}{3}\left(1+\xi^R\right)\left(\frac{m_{10}}{s} - \frac{dm_{10}}{ds}\right), \tag{3.3.11}$$

$$\xi_4^{(1)}(s) \approx 0. \tag{3.3.12}$$

Adding second order terms, we find

$$\begin{aligned} \xi^{(2)}(s,\mu) \approx & \frac{1}{2}\left(\frac{d^2\xi(s)}{ds_\parallel^2} m_2(s_\parallel, s_\perp) + (1+\xi(s))\frac{d^2 m_2(s_\parallel, s_\perp)}{ds_\parallel^2}\right) \\ & + \frac{d\xi(s)}{ds_\parallel}\frac{dm_2(s_\parallel, s_\perp)}{ds_\parallel}. \end{aligned} \tag{3.3.13}$$

The analytical result for the second-order multipoles includes many more terms than its first-order equivalent. Therefore, instead of calculating the resulting multipoles analytically, we take numerical derivatives of moments higher than one in the remainder of this work.

In what follows we address the question of how accurate the expansion Eq. 3.3.5 is on small scales, and how the different moments affect the clustering multipoles. The expansion turns out to give accurate predictions for the multipoles even on scales of about 10 h^{-1}Mpc, with the advantage of replacing the integral in Eq. 3.1.2, which sums up the contributions of the pairwise velocity distributions on different scales, with a derivative at the scale s under consideration. Therefore, it converts the non-local relationships between the redshift and real space correlation functions with the pairwise velocity PDF, into a local relation between the redshift space correlation and the derivatives of the pairwise velocity moments and real space correlation function.

In the next subsection, we will test the accuracy of the expansion by comparing it to a full streaming model in which we assume the pairwise velocity distribution is either Gaussian or ST.

3.3.3 The range of validity of the streaming model expansion

Assessing the exactitude of the Taylor expansion on different scales is not straightforward, since including higher order terms involves higher order derivatives of the velocity moments and the real space correlation function.

As one can see from Fig. 3.4, on small scales the velocity moments as functions of pair separation measured from the simulation are not smooth and their high-order derivatives can be noisy, which will affect the accuracy of the analytical predictions. Given that our main objective in this subsection is to test the validity of the expansion method, to eliminate the impact of such noise, we fit the ingredients of the Taylor expansion of the streaming model. The fits to the moments are shown as dotted lines in Fig. 3.4. For the real space correlation function, we fit with a simple power law. We then take the fitted curves as the "truth", and compare the predictions of the Taylor expansion with two different streaming models in which we can convert the fitted moments into pairwise velocity distributions. These include a Gaussian streaming model, to demonstrate why the GSM gives accurate predictions on quasi-linear scales where the pairwise velocity distribution is highly non-Gaussian, and an ST model, to show the effect of skewness and kurtosis in improving the accuracy of the expansion. The Gaussian distribution has the correct two lowest central moments, whilst the ST matches the four lowest moments.

We shall not compare the expansion to the simulation measurements directly, since the analytical fitting formulae to the measured moments already induce at least per cent-level modifications to the multipoles on small scales. However, this exercise is realistic enough (both the real space correlation function and the moments have been fitted to the N-body simulation results) to demonstrate up to which scale the Taylor expansion method can be used.

Compared with the full streaming model, the Taylor expansion makes minimal assumptions regarding the pairwise velocity distribution, since it only uses the moments, and removes the integration over all pair separations in Eq. 3.1.2. For the Gaussian case, the results of expanding Eq. 3.3.5 up to $n = 4$ are shown as coloured lines in Fig. 3.9. Although for a Gaussian distribution the odd central moments vanish, the odd moments about the origin get contributions from lower order even central moments, as expressed in Eq. 3.3.4. Therefore, even orders of the expansion do contribute. Similarly, while odd central moments higher than the second order vanish for a Gaussian distribution, m_n is nonzero for $n > 2$. The full streaming model predictions using the integral in Eq. 3.1.2 and a Gaussian pairwise velocity distribution are shown as black sold lines.

In Fig. 3.9 we see how including only terms up to $n = 2$ we can reproduce the monopole to within 1% down to $10\,h^{-1}$Mpc. To achieve a comparable accuracy for the quadrupole, however, we need to add higher order moments up to $n = 4$. For the hexadecapole, we expect the Taylor expansion to be less accurate, because it is more strongly affected by the finger-of-God effect, which originates from the very small and nonlinear scales on which the Taylor expansion breaks down. This is confirmed by the fact that in the lower panel of Fig. 3.9 there are larger differences between the coloured and black lines. Nevertheless, we find that for the Gaussian model, the hexadecapole predicted by the expansion up to fourth order is accurate to within 3% down to $15\,h^{-1}$Mpc.

Finally, since the ST distribution reproduces the measured line-of-sight velocity distribution with a higher accuracy than a Gaussian distribution (Fig. 3.2), we also demonstrate the effect of higher order moments on the Taylor expansion using an ST model for $\mathcal{P}$. In Fig. 3.10 we show both the third and fourth order moment expansion assuming a Gaussian distribution, with zero skewness and fixed kurtosis shown as dashed-dotted lines, and the fully non-Gaussian moments. Although for the monopole, non-Gaussianity does not play an important role, adding the skewness extends the 1% agreement in the quadrupole from scales of around $30\,h^{-1}$Mpc down to $20\,h^{-1}$Mpc. The effect of the fourth order moment, kurtosis, is important to extend close agreement even further to about $10\,h^{-1}$Mpc. These results are consistent with the findings in Fig. 3.6, where we find that the ST model improves the agreement of the quadrupole in the range 10-30 h^{-1}Mpc.

Note that for the hexadecapole (shown in the bottom panel of Fig. 3.10) the Taylor

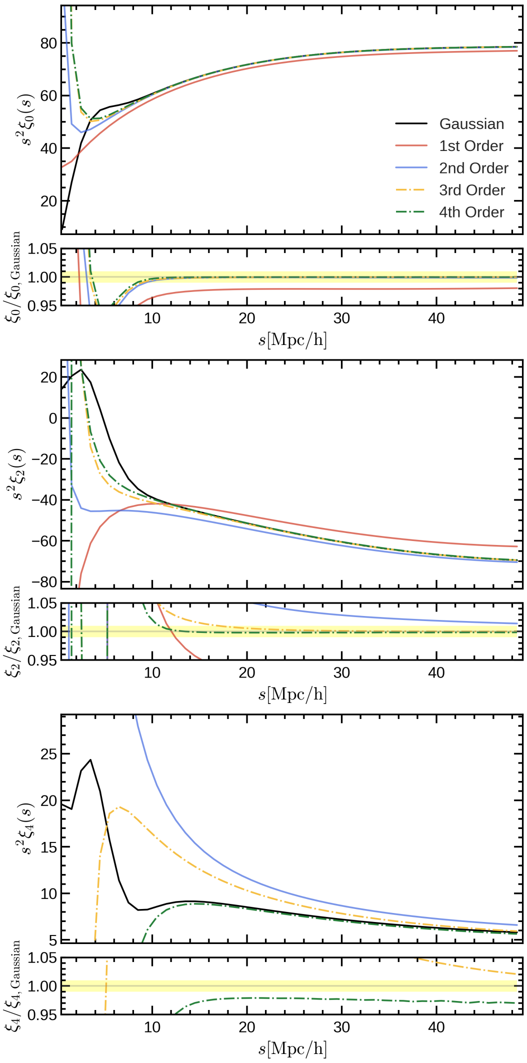


Figure 3.9: Accuracy of the Taylor expansion up to 4-th order, assuming the pairwise velocity distribution is Gaussian, compared to the full streaming model under the same assumptions. Note all the real space ingredients to the streaming model, the real space correlation function and the two lowest order pairwise velocity moments, are analytical functions fitted to the simulation measurements. The yellow shaded region shows one per cent level agreement between the Taylor expansion and the full Streaming Model. The monopole achieves an accuracy better than the one per cent on scales above 10 h^{-1}Mpc when the expansion is truncated at second order. For the quadrupole to achieve a similar accuracy, we need to retain up to fourth order terms.

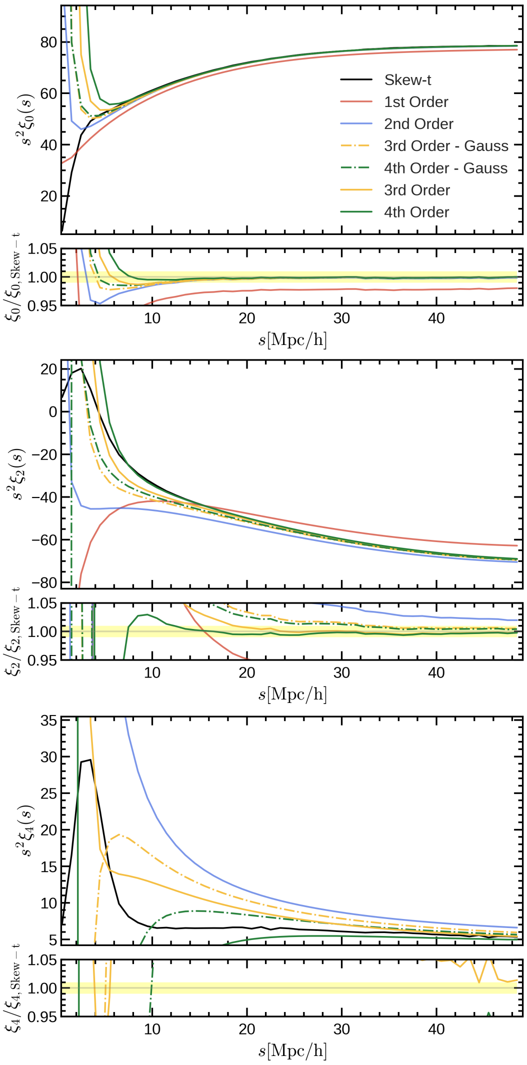


Figure 3.10: Same as in Fig. 3.9, but assuming the pairwise velocity distribution follows a Skew-t distribution. Compared with a Gaussian distribution, which has the correct first and second order moments, the Skew-t also matches the skewness and kurtosis. For comparison, we show the Taylor expansion found assuming the distribution is Gaussian and the fully non-Gaussian result, in which we include skewness and kurtosis. The effect of the skewness can therefore be seen as the difference between the orange dashed and orange solid lines, whilst the effect of the kurtosis is given by the difference between the green dashed and green solid lines. We find that the effects of the skewness and kurtosis are particularly important for the quadrupole on small scales.

expansion method introduces a substantial fractional error of $> 5\%$ on all scales, even if the fourth-order corrections are included. This is not surprising because now the assumed true model – in which the pairwise velocity satisfies an ST distribution – is more complicated, and because the absolute value of the hexadecapole is much closer to zero which tends to magnify the relative error. Nevertheless, we still observe that including higher-order terms brings the expansion prediction closer to the correct answer. In Fig. 3.6, the multipoles have been numerically calculated using the full Streaming model of Eq. (3.1.2), rather than the Taylor expansion, under the assumption of the pairwise velocity PDF being either Gaussian or ST.

3.4 Sensitivity to the real space streaming model ingredients

In this section, we study the effects of varying the various ingredients – the real space quantities – needed to predict the redshift space clustering through the ST streaming model. This will indicate to us what precision is required for each ingredient in order to make the final prediction of the multipoles accurate.

In Fig. 3.11, we show the effects on the multipoles of varying the real space correlation function (the first row), the mean pairwise velocity (the second row) and its variance (both in the radial and transverse components; the last two rows). We have studied the impact of two types of variations: a constant change by $\pm 5\%$, or a fractional change that increases towards smaller pair distances as $1/r$, to emulate the fact that perturbation theory predictions worsen towards small scales. The latter gradual change is tuned to vary the given function by $\pm 5\%$ on scales of $5\,h^{-1}$Mpc and by $\pm 1\%$ on the scale of $30\,h^{-1}$Mpc. In this way we can compare the effect of a varying slope due to uncertainties in our predictions, which we know is important since derivatives of the moments appear in the Taylor expansion Eq. 3.3.5. The fractional changes to the predicted redshift-space monopole, quadrupole and hexadecapole (from left to right) are respectively shown as orange and blue shaded regions for the constant and scale-dependent changes.

Varying the real space correlation function by 5% produces approximately the same fractional change in the monopole. Since the 0-th order contribution to the quadrupole and hexadecapole is zero, both the constant and the scale-dependent variations in the real space correlation function produce a sub-per cent effect on the quadrupole and hexadecapole on scales larger than $20\,h^{-1}$Mpc.

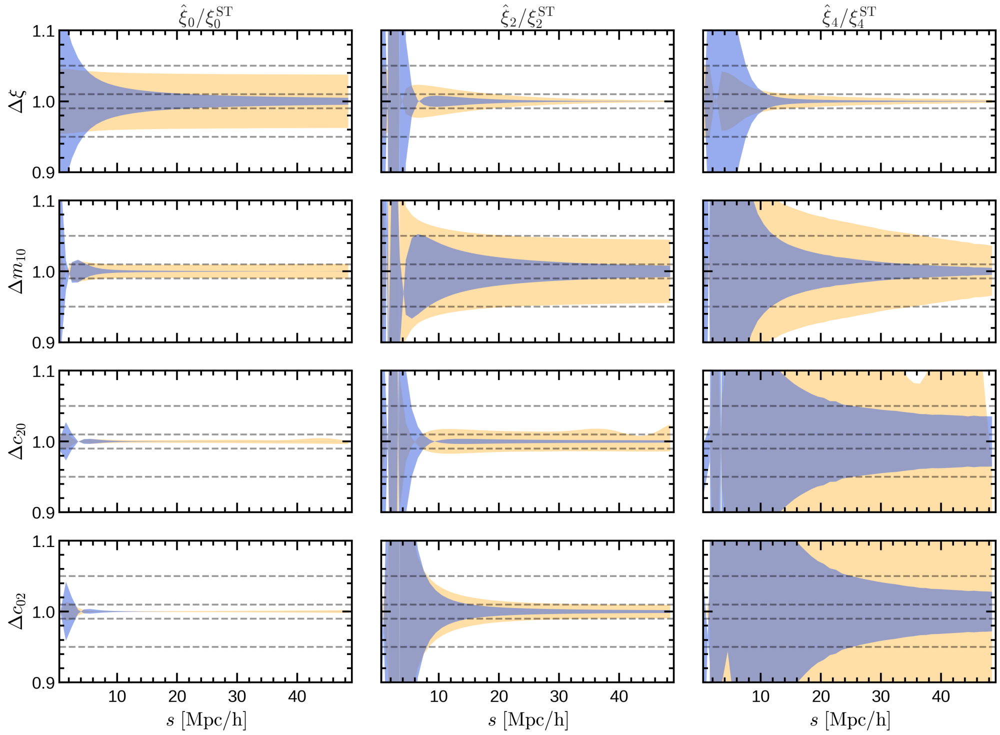


Figure 3.11: The fractional variation in the monopole, quadrupole and hexadecapole after modifying the real space ingredients of the streaming model. In each row we show the effect of varying: the real space correlation function, the mean pairwise velocity, the radial variance of the pairwise velocity and the transverse variance of the pairwise velocity. Orange contours show the effect of varying each of these quantities by ±5%, while the blue contours vary them by a percentage that depends on scale and increases with $1/r$. On small scales, where perturbation theory predictions degrade, we vary each of the ingredients by a larger percentage. The variation is tuned to produce a 5% change on scales of $5\,h^{-1}$Mpc and a 1% one on scales of $30\,h^{-1}$Mpc. Gray dashed lines determine 5 and 1 per cent deviations from the true model.

The mean pairwise velocity has a stronger effect on both the quadrupole and hexadecapole. The importance of getting the slope of the mean right, found in the Taylor expansion, can also be seen in the blue contours. A change of 1% above $30\,h^{-1}$Mpc produces a slightly larger effect on the quadrupole. The monopole, in contrast, is much less sensitive to variations of the mean pairwise velocity, and the effect is at the sub per cent level at $s \gtrsim 10\,h^{-1}$Mpc for both variation scenarios.

On the other hand, the hexadecapole is most sensitive to the radial and transverse standard deviations. Changes of 5% can produce a change that is twice as large in the hexadecapole.

Regarding the third-order moments, we found in the last section (Fig. 3.10) that the skewness has at most a per cent-level effect on the monopole and quadrupole on scales below $30\,h^{-1}$Mpc. Since its effect is very small, we do not show the equivalent in Fig. 3.11. We find that varying the third order radial and transverse moments by 50% introduces modifications smaller than 5% on the quadrupole on small scales.

Finally, the effect of fourth order terms is important on scales below $20\,h^{-1}$Mpc. However, we have already shown in the previous section that setting the fourth order moments to zero, by assuming Gaussianity, also gives only a few percentage level corrections to the quadrupole on small scales (see difference between orange dashed line and solid green line in Fig. 3.10).

Therefore, even on small scales, we need to predict most accurately the lower order moments: the mean and the standard deviation, particularly the latter, if we want to utilise information contained in the hexadecapole. We can afford to have a larger margin of error on the predictions of the higher order moments, and still extend the validity up to scales of around $10\,h^{-1}$Mpc.

3.5 Conclusions and Discussion

The new generation of surveys (Amendola et al., 2013; Takada et al., 2014; Levi et al., 2019; de Jong et al., 2018) is going to measure redshift space clustering of galaxies with unprecedented precision. To translate the high accuracy of these measurements into tighter constraints on the cosmological parameters or on possible deviations from general relativity, we need to improve our theoretical models of redshift space distortions (RSD). Within the streaming model of RSD, we need to: i) improve the mapping from real to redshift space, i.e., by developing the modelling of the pairwise velocity distribution including its higher order moments, ii) increase the accuracy of the predictions of the ingredients of the streaming model – the real space correlation function and the pairwise velocity moments – for given cosmological parameters. Here, we have focused on the first of these aspects, but we have also briefly analysed the effects of the second.

In N-body simulations, where the fully non-linear evolution of collisionless particles is solved, we observe that the distribution of the pairwise velocities of dark matter halos is skewed towards negative velocities, and has broader tails than a Gaussian. Therefore, models that use Gaussian distributions do not give an accurate description of the pairwise velocities. We have introduced an extension to the Gaussian streaming model by using the Skew-T probability distribution for the pairwise velocity. The parameters of this distribution can be tuned to match the four lowest-order velocity moments measured from simulations. The ST model describes the simulation measurement of the pairwise velocity distribution significantly better than a simple Gaussian.

We compare two different methods to find the best-fitting parameters of the pairwise velocity distribution: maximum likelihood estimation and the method of moments. Although the results of both approaches seem to describe the measured velocity distribution equally well on large scales, they give very different results for the redshift space clustering once inserted into the streaming model. Using the method of moments is crucial for describing all multipoles, including the small scales. Even though the Gaussian distribution gives a very poor fit to the measured pairwise velocities distribution, it can reproduce the true multipoles on quasi-linear scales within the small statistical errors of our simulations when we tune it to have the two lowest-order moments extracted from the simulations. On the other hand, the best-fit Gaussian found by maximising the likelihood gives results that are more than five standard deviations away from the simulation measurement.

The ST model, also using the method of moments, gives predictions for the redshift space multipoles (monopole, quadrupole and hexadecapole) that are within the small statistical sampling variance errors (driven by the simulation volume) down to about $10\,h^{-1}$Mpc. On such small scales, the Gaussian streaming model gives predictions that are more than five standard deviations away from the mean measurement from simulations. Therefore, the ST model extends the validity of the streaming model from $30\,h^{-1}$Mpc to $10\,h^{-1}$Mpc, and gives a more accurate description of the hexadecapole, which has so far not been used in analyses that rely on the Gaussian streaming model (e.g., Satpathy et al., 2017; Zarrouk et al., 2018), due to its poor accuracy.

We have used a Taylor expansion of the integrand to show why the Gaussian streaming model can reproduce the clustering on quasi-linear scales within the error bars of the simulation measurement, despite giving a poor description of the pairwise velocity distribution. At $s \gtrsim 30\,h^{-1}$Mpc, only the first and second order moments, the mean and the standard deviation, of the pairwise velocity distribution, are crucial for determining the monopole and quadrupole of the two-point correlation function in redshift space.

We have also shown that the Taylor expansion can describe the non-Gaussian ST streaming model down to smaller scales, of about $10\,h^{-1}$Mpc, when expanded up to fourth order. The main advantage of the Taylor expansion is that it replaces the integral of the pairwise velocities over all scales by a derivative of the moments at the scale under consideration. It therefore makes no assumptions about the details of the underlying velocity distribution, and can give analytical predictions for the monopole and quadrupole. However, it cannot reproduce the hexadecapole as accurately as the full ST streaming model integral, Eq. 3.1.2, nor is it as accurate on smaller scales, $s \lesssim 15\,h^{-1}$Mpc.

The Taylor expansion could be particularly useful to measure the velocity moments from the observed redshift space multipoles, as was already proposed by Bianchi et al. (2015), along the line of previous measurements of the pairwise velocity dispersion (Li et al., 2006; Loveday et al., 2018). The main difficulty to measure the pairwise distribution from observations lies in the pair distance dependence of the moments, imprinted by gravity. We would need to develop analytical formulae to summarise the pair distance dependence in a small set of parameters, that are valid independently of the underlying model of gravity. These parameters could then be inferred from observations of redshift space clustering, by running a Monte Carlo Markov Chain. The direct measurement of the moments could be a complementary test of gravity to the growth rate, and it would utilise more information of the full scale dependence of different gravity models.

Finally, we qualitatively analysed the effects of inaccurate knowledge of the real space correlation function or of the velocity moments on the predictions of the redshift multipoles. As expected, the monopole is mainly determined by the real space correlation function. We have shown that perturbation theory based CLEFT method (Vlah et al., 2016) produces per cent-level-accuracy predictions of the real space correlation function. However, to obtain per cent-level accurate predictions for both the monopole and quadrupole, we also need similar accuracy for the mean pairwise velocity and its slope. Fitting the CLEFT predictions, with five free parameters, we were only able to obtain predictions accurate at the per cent level for the mean on scales above $35\,h^{-1}$Mpc. On the other hand, the hexadecapole is very sensitive to the variance of pairwise velocities, for which CLEFT is only accurate to one per cent above scales of about $45\,h^{-1}$Mpc. Therefore, future efforts to utilise the information content in the hexadecapole will have to obtain more accurate theoretical prescriptions for the variance. Per cent level errors on the prediction of the variance become even larger errors on the hexadecapole. On scales smaller than $30\,h^{-1}$Mpc, we also need predictions for the skewness and kurtosis of pairwise velocities. However, these do not need to be as accurate: per cent errors on the skewness and kurtosis have negligible impact on the multipoles.

To summarise, we have developed a streaming model based on the Skew-t distribution of pairwise velocities, that accurately describes redshift space clustering on scales larger than $10\,h^{-1}$Mpc, given the first four moments of the pairwise velocity distribution are known. In order to improve constraints on the growth rate by using the ST model, we need to improve the theoretical predictions of the real space two-point correlation function and the pairwise velocity moments dependency on the cosmological parameters. This will be the focus of the next chapter. Moreover, we have here described the motions of halo centres. However, to model the motions of galaxies we will need to include the so-called one halo contribution representing the internal motion of galaxies inside their host halos. When the internal motion of galaxies is virialized, the mean infall velocity is negligible compared with the random motion and the internal velocity distribution is isotropic and close to Gaussian distributed.

Chapter 4

Simulation-based models for real space clustering

In the previous chapter, we showed how the mapping from real to redshift space could be accurately modelled if the real space two-point correlation function and the four lowest order moments of the pairwise velocity distribution were known. In this chapter we focus on obtaining accurate predictions for the first item, the real space two-point correlation function, whilst the pairwise velocity moments are the subject of ongoing work.

To obtain fully non-linear predictions for the properties of the large-scale structure and recover all the cosmological information contained in the small-scale clustering, we must resort to N-body simulations (Kuhlen et al., 2012). N-body simulations have been widely used as cosmic laboratories to test the precision and robustness of analytical methods for the large-scale structure (e.g., Carlson et al. 2009; Vlah et al. 2015; Cuesta-Lazaro et al. 2020), together with the effects of systematic errors in our measurements. Over the past decade, advances in computing have allowed us to produce a large enough number of dark matter only N-body simulations covering a significant fraction of the cosmological parameter space, which allows us to use the simulations themselves as predictive models that directly constrain the cosmological parameters. The simulations must be large enough to reduce sample variance, and have high enough resolution to resolve the tracers that will be surveyed.

Moreover, in order to compare the outcomes of dark matter only simulations to the observed distribution of galaxies we have to model the connection between dark matter halos and galaxies (see Section 1.2.1 for an introduction to the topic). Uncertainties in the galaxy-halo connection can limit the amount of information that we can extract from small scale clustering. We would like to use flexible models that can reproduce clustering in different scenarios of galaxy formation, whilst still being able to recover cosmological information after marginalising over the free parameters of the galaxy-halo connection model. Here, we use the empirical model of the halo occupation distribution (HOD) (Benson et al., 2000; Zheng et al., 2005), based on estimating the probability that a given halo hosts a galaxy. See Section 1.2.1 for more details on different models of the galaxy-halo connection.

Over the past few years, several studies (Zhai et al., 2019b; Lange et al., 2019; Kobayashi et al., 2020; Miyatake et al., 2021) have shown how N-body simulations can be leveraged to extract small scale information. Solving the inverse problem, estimating the posterior over the cosmological parameters given the observed clustering, would require the order of $\mathcal{O}(10^6)$ N-body simulations to perform Bayesian inference with Markov Chain Monte Carlo. Therefore, most studies rely on modelling the dependence of the two-point correlation function on cosmology with surrogate models that are trained on a small set of $\mathcal{O}(100)$ N-body simulations (Zhai et al., 2019b; Lange et al., 2019; Kobayashi et al., 2020). The surrogate models are orders of magnitude faster than the original N-body simulations and can then be used to sample the posterior of cosmological parameters.

For instance, Kobayashi et al. (2020) developed an N-body version of the halo model for the galaxy power spectrum by training a neural network to reproduce the dark matter halo clustering properties in Fourier space. Zhai et al. (2019b) and Yuan et al. (2022a) followed a different route by emulating galaxy clustering as both a function of cosmology and galaxy-halo connection parameters with Gaussian processes (Rasmussen & Williams, 2005). Alternatively, Lange et al. (2019) developed the so-called cosmological evidence modelling (CEM) method. Lange et al. (2019) used N-body simulations to compute the evidence of the data as a function of cosmology after marginalising over the HOD parameters, which can then be used to sample the posterior distribution of the cosmological parameters. In this way, the authors do not have to account for the errors introduced by the surrogate model. However, this approach does not yield joint constraints on the galaxy-halo connection and cosmological parameters, since the HOD parameters are marginalised over.

Simulation-based methods currently produce the tightest constraints on the parameter combination $f\sigma_8$ (Lange et al., 2021; Kobayashi et al., 2022; Yuan et al., 2022a; Zhai et al., 2022) when confronted with observations. Interestingly, all studies find values for $f\sigma_8$ that are lower than those obtained from the CMB. The current challenge for emulator-based approaches is to both make sure that theoretical predictions are on a par with the statistical

Simulation Suite	Code	$L_{\rm box}$ [h^{-1}Gpc]	$N_{\rm part}$	$M_{\rm part}$	Halo Finder	Reference
DarkQuest HR	GADGET2	1	2048^3	1.02×10^{10}	Rockstar	Nishimichi et al. (2019)
DarkQuest LR	GADGET2	2	2048^3	8.158×10^{10}	Rockstar	Nishimichi et al. (2019)
AbacusSummit Base	ABACUS	2	6912^3	2.1×10^9	CompaSO	Maksimova et al. (2021b)
Aemulus	GADGET2	1.05	1400^3	3.51×10^{10}	Rockstar	DeRose et al. (2019b)

Table 4.1: Comparison of the characteristics of the DarkQuest suite of simulations and those others used to train clustering emulators in the literature. The mass of dark matter particles $M_{\rm part}$ has units of $(\Omega_{\rm m}/0.3)h^{-1}M_\odot$

errors expected from future surveys, and that the modelling of how galaxies populate dark matter halos does not introduce biases into the analysis from small-scale clustering.

In this chapter, we present a surrogate model for real space clustering that will be combined with pairwise velocity moment emulators to predict redshift space clustering. We will then be able to combine constraints from clustering measurements and estimates of peculiar velocities, obtained through either the kinetic Sunyaev-Zeldovich effect (Sunyaev & Zeldovich, 1980) (see Calafut et al. (2021) for a recent measurement) or through peculiar velocity surveys (Dupuy et al., 2019), to obtain more precise constraints on the cosmological parameters. Peculiar velocity surveys and redshift space distortions have been shown to be specially complementary to test gravity theories (Kim & Linder, 2020).

In this chapter, we focus on modelling small-scale galaxy clustering in real space, improving the emulators presented in Nishimichi et al. (2019) in terms of both accuracy and speed. We show how a combination of neural networks trained using the predictions of N-body simulations and the halo model can produce extremely accurate predictions for the clustering of galaxies over a wide range of pair separations $0.01 < r < 150\ h^{-1}$ Mpc, as opposed to the range $r < 30\ h^{-1}$ Mpc, covered by previous emulators in configuration space (Zhai et al., 2019b; Kobayashi et al., 2020). This allows us to compute the likelihood using the full shape of the two-point correlation function, spanning the behaviour of the one- and two-halo terms. Finally, we demonstrate the limitations of the current implementation of the halo model to recover unbiased constraints when an assembly bias signal (Wechsler & Tinker, 2018) is present in the data to be analysed.

4.1 The Dark Quest simulation suite

Here, we briefly describe Dark Quest, a suite of cosmological N-body simulations designed to build emulators of sumary statistics. A detailed description can be found in Nishimichi et al. (2019).

4.1.1 N-body simulations

The Dark Quest simulations were performed with 2048^3 dark matter particles in $1\,h^{-1}$ Gpc (hereafter high-resolution runs, denoted HR) or $2\,h^{-1}$ Gpc (low-resolution runs, labelled LR) side-length boxes, using the Gadget2 N-body solver (Springel, 2005). The mass resolutions of the HR and LR runs are 1.02×10^{10} and $8.16 \times 10^{10}(\Omega_\mathrm{m}/0.3)\,h^{-1}\,M_\odot$, respectively.

In Table 4.1, we show a comparison of the specifications of Dark Quest with those of other simulation suites that have been used to train clustering emulators in the literature (Zhai et al., 2019b; Lange et al., 2019; Kobayashi et al., 2020; Miyatake et al., 2021). Dark Quest, used int his work, has a higher resolution and a larger box size than Aemulus, but a lower resolution than AbacusSummit. In the future, it will be important to demonstrate the impact of differences in N-body codes (e.g. Grove et al. 2022), halo finders (e.g. Gómez et al. 2022), and resolution on the cosmological parameters inferred using simulation-based methods.

The initial conditions were generated using second-order Lagrangian perturbation theory (2LPT, Crocce et al. (2006)) and the redshift at which to generate the initial conditions was chosen depending on the cosmology and resolution (Nishimichi et al., 2019), with $z_\mathrm{init} \approx 59$ and 29 adopted for the fiducial HR and LR simulations respectively. Each simulation used different random number seeds to generate the initial conditions.

The cosmologies used in the simulations cover 101 flat geometry wCDM models, as shown in Fig. 4.1. In wCDM, the equation of state (EoS) for dark energy is parameterised through the value of w, also known as the EoS parameter of dark energy, $p_\mathrm{de} = w\rho_\mathrm{de}$, whose value is $w = -1$ in ΛCDM. Here, w is assumed to be constant.

The set of cosmological parameters is defined using optimal maximin distance sliced Latin hypercube designs (Ba et al., 2015), which enable efficient sampling from the six-dimensional parameter space,

$$\mathcal{C} = \left\{\omega_\mathrm{b}, \omega_\mathrm{c}, \Omega_\mathrm{de}, \ln\left(10^{10} A_\mathrm{s}\right), n_\mathrm{s}, w\right\}, \tag{4.1.1}$$

where $\omega_\mathrm{b} \equiv \Omega_\mathrm{b} h^2$ and $\omega_\mathrm{c} \equiv \Omega_\mathrm{c} h^2$ are the physical density parameters of baryons and cold dark matter, respectively. The total matter density is the sum of the contributions from baryons, cold dark matter, and non-relativistic neutrinos:

$$\Omega_\mathrm{m} = \Omega_\mathrm{b} + \Omega_\mathrm{c} + \Omega_\nu, \tag{4.1.2}$$

where the physical density of neutrinos is fixed in the Dark Quest simulations as $\omega_\nu \equiv$

$\Omega_\nu h^2 \equiv 0.00064$, corresponding to 0.06 eV for the total mass of the three mass eigenstates. For given values of $\omega_\mathrm{b}, \omega_\mathrm{c}$ and the density parameter for dark energy Ω_de, the Hubble constant is derived from spatial flatness, that is,

$$\Omega_\mathrm{m} h^2 = \omega_\mathrm{b} + \omega_\mathrm{c} + \omega_\nu, \tag{4.1.3}$$

$$\Omega_\mathrm{m} + \Omega_\mathrm{de} = 1. \tag{4.1.4}$$

A_s and n_s are the amplitude and slope of the primordial curvature power spectrum normalised at $0.05\,\mathrm{Mpc}^{-1}$. The range of parameters explored is

$$\begin{aligned}
&0.0211375 < \omega_\mathrm{b} < 0.0233625,\\
&0.10782 < \omega_\mathrm{c} < 0.13178,\\
&0.54752 < \Omega_\mathrm{de} < 0.82128,\\
&2.4752 < \ln\left(10^{10} A_\mathrm{s}\right) < 3.7128,\\
&0.916275 < n_\mathrm{s} < 1.012725,\\
&-1.2 < w < -0.8,
\end{aligned} \tag{4.1.5}$$

which is centred on the fiducial best fitting ΛCDM model to the Planck 2015 data alone (Planck Collaboration et al., 2016b): $\omega_\mathrm{b} = 0.02225, \omega_\mathrm{c} = 0.1198, \Omega_\mathrm{de} = 0.6844, \ln\left(10^{10} A_\mathrm{s}\right) = 3.094, n_\mathrm{s} = 0.9645$ and $w = -1$. Fig. 4.1 shows a two-dimensional representation of the parameter space.

These parameter ranges correspond to the ranges of $(\pm 5\%, \pm 10\%, \pm 20\%, \pm 20\%, \pm 5\%)$ for the parameters $(\omega_\mathrm{b}, \omega_\mathrm{c}, \Omega_\mathrm{de}, \ln(10^{10} A_\mathrm{s}), n_\mathrm{s})$, respectively. These ranges were chosen to cover a parameter space that extends well beyond the constraints from the 2015 Planck data for a flat-ΛCDM model, for which the corresponding 68% intervals are $(0.72\%, 1.25\%, 1.33\%, 1.10\%, 0.51\%)$. Therefore, the Dark Quest simulations cover roughly up to a $\sim 10\,\sigma$ range around the central best-fitting model to the Planck 2015 data. However, for the dark energy EoS parameter, w, a different approach was taken. Since Planck data alone cannot place a stringent constraint on w, and also, assuming that wCDM significantly loosens the constraints on the other parameters, we chose a strategy that is not strictly consistent for the six parameters. Instead, we tried to cover a much wider range for w (ie, $w = -1.019^{+0.075}_{-0.08}$ at 95% CL).

The simulation outputs were stored at 21 redshifts: 1.48, 1.35, 1.23, 1.12, 1.02, 0.932, 0.846, 0.765, 0.689, 0.617, 0.549, 0.484, 0.422, 0.363, 0.306, 0.251, 0.198, 0.147, 0.0967, 0.0478, and 0. These redshifts are evenly spaced in the linear growth factor for the fiducial Planck

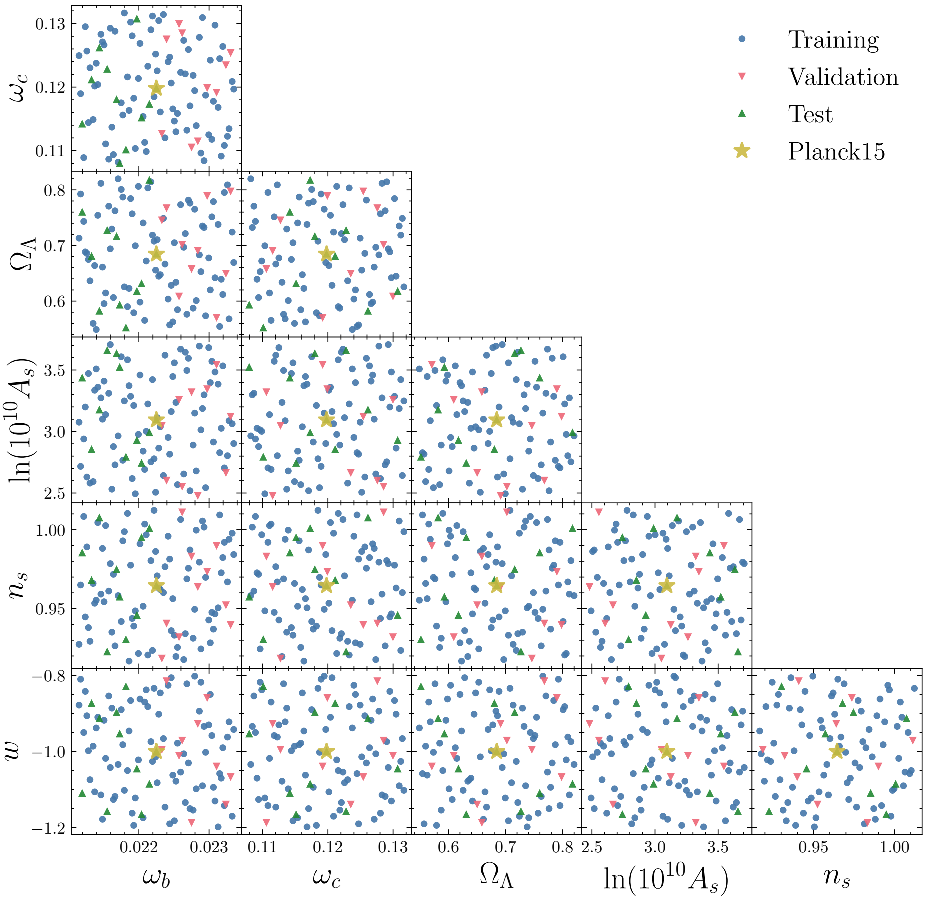


Figure 4.1: Corner plot representation of the 101 wCDM cosmologies covered by the Dark Quest simulation suite. We show the cosmologies chosen as training, test and validation sets, together with the best fitting fiducial cosmology to the 2015 Planck data, using different symbols, as indicated by the key.

cosmology.

4.1.2 Halo catalogues

The identification of halos is of crucial importance, since the central premise of our method is to emulate dark matter halo properties, which can be robustly measured from N-body simulations. Appendix E of the Dark Quest paper (Nishimichi et al., 2019) provides comprehensive convergence tests of halo properties such as halo mass, the halo mass functions, and halo autocorrelation functions, with respect to the choice of halo finder, halo substructure separation, central/satellite split criterion, etc. In this section, we briefly review the main definitions that will be used throughout.

The halo catalogues used here were identified using Rockstar (Behroozi et al., 2013b), a friends-of-friends (FOF) halo finder that operates in six-dimensional phase space. The halo centre is defined as the centre of mass position of the "core particles", a subset of member particles in the inner part of the halo. M_{200m} is adopted as the halo mass definition in Dark Quest, which is the mass enclosed within R_{200m}, the radius within which the average density is 200 times the mean mass density $\bar{\rho}_{m0}$. This definition of halo mass includes all simulation particles within a radius of R_{200m} from the halo centre, including gravitationally unbound ones. When the separation between the centres of different halos is within R_{200m} of any other halo, the most massive halo is marked as a central halo and the other halo(s) as a satellite halo(s). Only central halos with mass $M_{200m} \geq 10^{12}\, h^{-1} M_{\odot}$ are used in our analysis.

4.2 From dark matter halos to galaxies

As in Nishimichi et al. (2019) and Kobayashi et al. (2020) we use the halo model to express the galaxy two-point correlation function in terms of dark matter halo properties. This allows us to make theoretical predictions for different galaxy samples, including cross-correlations of two different tracers, such as the ones that would be used in a multitracer analysis (McDonald & Seljak, 2009), or the cross-correlation between clusters and galaxies. Moreover, a halo model implementation allows us to model the halo-galaxy connection analytically, which means that the accuracy of the results will not be worsened by emulator inaccuracies. As a downside, complex models of the halo-galaxy connection such as environment-based assembly bias may be harder to implement.

The halo model assumes that galaxies occupy dark matter halos, and therefore that the two-point galaxy correlation function can be split into contributions from galaxy pairs that inhabit the same dark matter halo, and pairs in which each member occupies a different dark mater halo (these terms will be referred to as the one and two halo terms, respectively):

$$\xi_{\rm gg}(r) = \xi_{\rm gg}^{\rm 1h}(r) + \xi_{\rm gg}^{\rm 2h}(r). \tag{4.2.1}$$

The one and two halo terms can be further split into correlations between two types of galaxies: centrals and satellites. Central galaxies are positioned at the minimum of the potential well of the dark matter halo and move with the halo's centre of mass velocity. Satellite galaxies orbit within the dark matter halo with virialised velocities. We assume that the distribution of satellite galaxies is given by an NFW profile, $u_{\rm NFW}(r|c(M))$ (Navarro et al., 1997). This approximation has been tested against hydrodynamical simulations, finding it valid for galaxies selected by number density (Bose et al., 2019). The NFW profile is defined by one parameter: the concentration of the halo, c, which varies with halo mass, redshift, and cosmological parameters (Ludlow et al., 2016; Diemer & Joyce, 2019). Here, we use the median concentration-mass relation $c(M)$ from Diemer & Joyce (2019).

Regarding the galaxy-halo connection, we use the halo occupation distribution (HOD) (Zheng et al., 2005) to model the number of galaxies in a given halo as a function of halo mass. The occupation of central galaxies is parameterized as a Bernoulli distribution, whereas that of satellites is assumed to be Poisson distributed. Both distributions are described by their mean parameters

$$\langle N_{\rm g}\rangle(M) = \langle N_{\rm c}\rangle(M) + \langle N_{\rm s}\rangle(M). \tag{4.2.2}$$

We parameterize the mean galaxy numbers as in Zheng et al. (2005) by introducing the following HOD parameters

$$\mathcal{G} = \{M_{\rm min}, \sigma_{\log M}, M_1, \kappa, \alpha\}, \tag{4.2.3}$$

where $M_{\rm min}, \sigma_{\log M}$, and M_1, κ, α define the occupation of the centrals and satellites, respectively.

We describe the mean number of central galaxies for a given halo as

$$\langle N_{\rm c}\rangle(M|\mathcal{G}) = \frac{1}{2}\left(1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right), \tag{4.2.4}$$

where erf(x) is the error function. The mean occupation number of satellite galaxies is defined as

$$\begin{aligned}\langle N_{\rm s}\rangle(M|\mathcal{G}) &= \langle N_{\rm c}\rangle(M|\mathcal{G})\,\lambda_{\rm s}(M|\mathcal{G})\\ &= \langle N_{\rm c}\rangle(M)\left(\frac{M-\kappa M_{\rm min}}{M_1}\right)^{\alpha}.\end{aligned}\tag{4.2.5}$$

The empirical HOD model that we use is extremely simple. One of the simplifying assumptions is that galaxy occupation depends solely on the mass of the dark matter halo. Although dark matter halo mass correlates strongly with clustering, we know that dark matter halos experience different assembly histories even at a fixed halo mass, which can affect their clustering (Gao et al., 2005b; Gao & White, 2007b). These different assembly histories influence secondary properties of halos, and this might, in turn, affect the formation of galaxies and hence the galactic content of halos of a given mass. These effects together – the variations in halo clustering and galactic content with halo mass and a second halo property – are known as galaxy assembly bias (see Wechsler & Tinker 2018 for a recent review on the galaxy-halo connection and assembly bias). The question we will address in Section 4.4.3, is whether the simplified version of the galaxy-halo connection used here is flexible enough to recover unbiased constraints on the cosmological parameters.

Given these assumptions, we can express the two-point galaxy correlation function in terms of dark matter halo properties. To simplify the calculations, we further split the one and two halo terms into correlations of central and satellite galaxies

$$\xi_{\rm gg}(r) = \xi^{\rm 1h}_{\rm ss}(r) + 2\xi^{\rm 1h}_{\rm cs}(r) + \xi^{\rm 2h}_{\rm cc}(r) + 2\xi^{\rm 2h}_{\rm cs}(r) + \xi^{\rm 2h}_{\rm ss}(r).\tag{4.2.6}$$

In the equations below, we highlight the emulated quantities in blue, such as the halo mass functions, $\mathrm{d}n/\mathrm{d}M$, and halo auto correlation functions, $\xi_{\rm hh}(r)$, following the convention used in Miyatake et al. (2020). Terms involving both centrals and satellites lead to the convolution of the halo profiles and the halo two-point correlation function. It is therefore simpler to compute these terms in Fourier space, where convolutions in coordinate space become simple products, and then apply an inverse Fourier transform to the result. Therefore, we compute

$$P^{\rm 1h}_{\rm ss}(k) = \frac{1}{\bar{n}_{\rm g}^2}\int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M}(M)\,\langle N_{\rm c}\rangle(M)\lambda^2_{\rm s}(M)u_{\rm NFW}(k|M,c(M))^2,\tag{4.2.7}$$

where $u_{\rm NFW}(k|M,c(M))$ is the Fourier transform of the truncated NFW profile (see Eq. (81) in Cooray & Sheth 2002).

The cross-correlation between centrals and satellites that occupy the same halo is given by

$$P^{1\mathrm{h}}_{\mathrm{cs}}(k) = \frac{1}{\bar{n}_{\mathrm{g}}^2} \int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M}(M) \left\langle N_{\mathrm{c}} \right\rangle (M) \lambda_{\mathrm{s}}(M) u_{\mathrm{NFW}}(k|M, c(M)), \tag{4.2.8}$$

where $\mathrm{d}n/\mathrm{d}M(M)$ is the halo mass function defined as the comoving number density of halos for a given halo mass, and $\bar{n}_{\mathrm{g}}$ is the galaxy number density that we obtain by integrating the halo mass function weighted by the halo occupation

$$\bar{n}_{\mathrm{g}} = \int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M} \left(\left\langle N_{\mathrm{c}} \right\rangle (M) + \left\langle N_{\mathrm{s}} \right\rangle (M) \right). \tag{4.2.9}$$

Meanwhile, the different two-halo terms will result in weighted averages of the dark matter halo two point correlation function and convolutions with NFW profiles when satellite correlators are involved

$$\begin{aligned} P^{2\mathrm{h}}_{\mathrm{cs}}(k) =& \frac{1}{\bar{n}_{\mathrm{g}}^2} \int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M}(M) \left\langle N_{\mathrm{c}} \right\rangle (M) \\ & \int \mathrm{d}M' \frac{\mathrm{d}n}{\mathrm{d}M}(M') \left\langle N_{\mathrm{c}} \right\rangle (M') \lambda_{\mathrm{s}}(M') \\ & P_{\mathrm{hh}}(k|M, M') u_{\mathrm{NFW}}(k|c(M')), \end{aligned} \tag{4.2.10}$$

$$\begin{aligned} P^{2\mathrm{h}}_{\mathrm{ss}}(k) =& \frac{1}{\bar{n}_{\mathrm{g}}^2} \int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M}(M) \left\langle N_{\mathrm{c}} \right\rangle (M) \lambda_{\mathrm{s}}(M) \\ & \int \mathrm{d}M' \frac{\mathrm{d}n}{\mathrm{d}M}(M') \left\langle N_{\mathrm{c}} \right\rangle (M') \lambda_{\mathrm{s}}(M') \\ & P_{\mathrm{hh}}(k|M, M') u_{\mathrm{NFW}}(k|c(M')) u_{\mathrm{NFW}}(k|c(M)). \end{aligned} \tag{4.2.11}$$

We avoid the Fourier transform when computing central-central terms

$$\begin{aligned} \xi^{2\mathrm{h}}_{\mathrm{cc}}(r) =& \frac{1}{\bar{n}_{\mathrm{g}}^2} \int \mathrm{d}M \frac{\mathrm{d}n}{\mathrm{d}M}(M) \left\langle N_{\mathrm{c}} \right\rangle (M) \\ & \int \mathrm{d}M' \frac{\mathrm{d}n}{\mathrm{d}M}(M') \left\langle N_{\mathrm{c}} \right\rangle (M') \xi_{\mathrm{hh}}(r|M, M'). \end{aligned} \tag{4.2.12}$$

In the next section, we show how we can use neural networks to emulate the two statistics shown in blue that vary with cosmological parameters: $\mathrm{d}n/\mathrm{d}M$ and ξ_{hh}.

4.2.1 The best of both universes: combining simulations of different resolutions

Although the high-resolution (HR) simulations can resolve halos of lower masses than their low-resolution (LR) counterparts, their smaller box size results in a larger sample-variance noise than in the LR boxes.

The halo model approach outlined above allows us to calibrate the halo autocorrelation function using the LR simulations, to reduce sample variance when using measurements from one realisation, while calibrating the halo mass function with the HR simulations to ensure an accurate estimate of the halo mass function for low mass halos. In this section, we examine the impact of combining the halo mass function of HR simulations with the halo correlation function measured in LR simulations.[1]

In Fig. 4.2, we show a comparison of a mock LOWZ-like catalogue obtained from the 25 realisations of the fiducial cosmology for the HR simulations, to the result of Eq. 4.2.6 when i) we combine the halo mass function from HR simulations, with the halo two-point correlation function estimated from one of the HR boxes (solid blue line), ii) estimate both the halo mass function and halo two-point correlation function from the LR simulations (dashed red), and iii) measure the halo mass function in the HR simulation, and the halo auto-correlation from the LR simulation. Fig. 4.2 shows that combining clustering measurements from low-resolution simulations with a halo mass function measured in the HR simulation does not introduce any biases and reduces the sample-variance noise.

4.3 Neural Network emulators for dark matter halo properties

Nishimichi et al. (2019) fitted both the halo mass function and the halo autocorrelation function measured from the N-body simulations using a combination of principal component analysis (PCA), to reduce the dimensionality of the data vector, and Gaussian processes (GP), to fit the dependence of the principal component coefficients on cosmology. Here, we show how dimensionality reduction can be avoided by using neural network emulators, leading to increased accuracy in the prediction of halo properties.

Fully connected neural networks approximate a function f such that

$$\mathbf{y} = f(\mathbf{x}|\boldsymbol{\theta}), \tag{4.3.1}$$

where $\mathbf{x}$ represents the features of the data set, $\mathbf{y}$ the desired outputs, and $\boldsymbol{\theta}$ the network-free parameters, also called trainable parameters. The optimal function f is defined by the set of values θ that minimise the loss function (the form of which is discussed below). The loss function provides a measure of the model's performance when evaluated on the data set.

[1] Note we could also have extended the mass resolution of the LR halo catalogues, using a scheme like the introduced by Ramakrishnan & Velmani (2021) or Armijo et al. (2022).

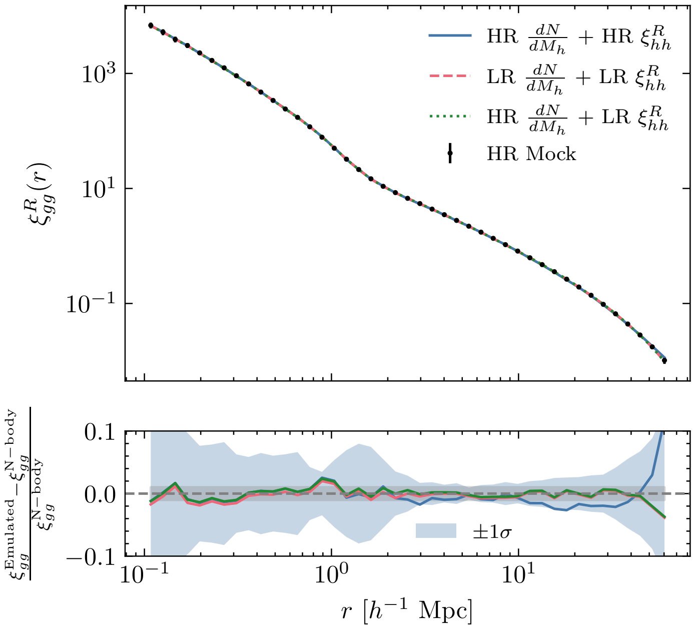


Figure 4.2: We show $\xi^R_{\rm gg}$ obtained by populating the 25 realizations of the fiducial cosmology on the HR simulations with mock LOWZ galaxies, compared to the result of Eq. 4.2.6 when either: i) both $\mathrm{d}N/\mathrm{d}M_{\rm h}$ and $\xi^R_{\rm hh}$ are measured on the HR simulations (in blue), ii) both $\mathrm{d}n/\mathrm{d}M_{\rm h}$ and $\xi^R_{\rm hh}$ are measured on the LR simulations (in red) and iii) $\mathrm{d}n/\mathrm{d}M_{\rm h}$ is obtained from the HR simulations and $\xi^R_{\rm hh}$ from the larger boxsize LR ones (in green). The fractional difference plot in the lower panel shows that the sample variance in the blue line based on the correlation function measured from one HR box is greatly reduced by replacing it with LR simulations without introducing bias. Blue shaded denote the standard deviation of the 25 realizations of the HR simulations. the gray shaded regions denotes 1% errors.

ReLU (Rectified Linear Unit; Agarap 2018) is the most commonly used activation function in current neural networks used to add non-linearities in the mapping between inputs and outputs, and is defined as

$$\mathrm{ReLU}(x) = \max(0, x), \tag{4.3.2}$$

where x is the output of the previous layer of the neural network. Note that ReLU activations are not differentiable at zero. Here, however, we are interested in functions that are differentiable with respect to their inputs and, in particular, with respect to the cosmological parameters (since these derivatives could be used to accelerate parameter inference through Hamiltonian Monte Carlo techniques, e.g. Duane et al. 1987, or to accelerate Fisher forecasts). Therefore, throughout, we use Gaussian error linear units (GELUs) as activation functions instead (Hendrycks & Gimpel, 2016):

$$\mathrm{GELU}(x) = 0.5x \left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{4.3.3}$$

To find the optimal parameters, θ, which reproduce the statistics measured from the N-body simulations, we minimise the L1 norm loss function

$$\mathcal{L} = \frac{1}{N} \sum_{i=0}^{N} |y^i_{\mathrm{true}} - y^i_{\mathrm{predicted}}|, \tag{4.3.4}$$

using the Adam optimiser (Kingma & Ba, 2014). The L1 loss reduces the importance given to outlier errors compared to the use of the mean squared error (also known as the L2 norm). We will refer to the value of Eq. 4.3.4 evaluated in the training and validation dataset as training and validation loss, respectively.

Moreover, we avoid fine-tuning the value of the learning rate by using a learning rate scheduler that reduces the learning rate by a factor of 10 every time the validation loss does not improve after 20 epochs. We also stop training the model when the validation loss does not improve after 100 epochs. This iterative reduction of the learning rate allows the model to quickly learn the broad characteristics of the data and then reduce the errors by adopting a smaller learning rate. The initial learning rate is always set to 0.015.

In the following subsections, we demonstrate the precision of fully connected networks in reproducing the real-space correlation function and the halo mass function obtained from the DARK QUEST simulations.

4.3.1 Real space correlation function

Measurement

The details of the halo correlation function measurements are introduced in Nishimichi et al. (2019). Here, we present only a summary of the most important aspects.

First, noisy measurements of $\xi(r|M,M')$ are avoided by instead measuring ξ as a function of halo number density, n, and switching from differential to cumulative mass limits. We then use the halo mass function to translate predictions as a function of number density into predictions as a function of differential mass through the relation

$$\begin{aligned}\xi(r|n(m),n(m')) &= \frac{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \xi(r|M,M') \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')}{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')} \\ &= \frac{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \xi(r|M,M') \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')}{n(M)n(M')},\end{aligned} \tag{4.3.5}$$

which can be inverted to obtain

$$\begin{aligned}\xi(r|M,M') &= \frac{\frac{\partial^2}{\partial m \partial m'}\left[n(m)n(m')\xi(r|n(m),n(m'))\right]}{\frac{\mathrm{d}n}{\mathrm{d}M}(M)\frac{\mathrm{d}n}{\mathrm{d}M}(M')} \\ &= \frac{\partial^2}{\partial n \partial n'}\left[n(m)n(m')\xi(r|n(m)n(m'))\right].\end{aligned} \tag{4.3.6}$$

Measurements are made in 8 logarithmically spaced bins in number density over the range $n_\mathrm{h} = [10^{-6}, 10^{-2.5}]\ (h^{-1}\mathrm{Mpc})^{-3}$. In total, there are 36 independent combinations for two halo samples with different number densities. The pair separation r is split into 40 logarithmically spaced bins from 0.01 to 5 h^{-1} Mpc and 75 linear bins from 5 to 150 h^{-1} Mpc, and over the 21 simulation snapshots spanning from $z = 1.48$ to $z = 0$.

In total, the data set is made up of 80 cosmologies in the training set, 10 in the validation set and 10 in the test set, each with its corresponding 21 snapshots and 36 number density bins.

On large scales, we can reduce cosmic variance by using the propagator-based prescription of Crocce & Scoccimarro (2006). For Gaussian initial conditions, the propagator can be expressed as the ratio of the cross-power spectrum between the density field at the initial conditions and the nonlinear field at the redshift of interest, to the linear power spectrum. This calculation was originally performed for the matter density, but can be extended to the halo density field. The propagator quantifies how much of the memory of the initial conditions is preserved in the final nonlinear density field. The propagator describes the smearing of BAO feature due to large-scale random flow. One can straightforwardly generalize this approach to any tracer. This function also describes the linear bias factor in the large-scale limit. The advantage of using the propagator is that a large fraction of sample-variance error is cancelled when the ratio between the two spectra is taken. In addition, it is known that the k dependence of the propagator is simple. A Gaussian-like parameterized function is sufficient to model this accurately (see Nishimichi et al. 2019 for more details).

We have slightly updated the implementation of this idea here. In Nishimichi et al. (2019), to evaluate the correlation function, both the directly emulated correlation function (for small separations) and the propagator-based model (for large separations), in which the propagator is also emulated, are computed and then stitched together to cover a wide range of separations. This requires us to build two separate emulators and both of them must be used when evaluating the correlation function. Here, instead, we now work at the data level: for each simulation box, we construct a data vector that combines the two methods. We refined the stitching scheme to yield a smoother transition between the two regimes (Nishimichi et al. in prep.). Now, our neural-network emulator learns this new datavector, to which the propagator trick has already been applied.

Emulation

We train a fully connected neural network, f, to perform the following mapping

$$\log_{10}\left(\xi^{\rm R}_{\rm hh}(r)\right) = f(\mathcal{C}, \log_{10}(n_1), \log_{10}(n_2), z), \tag{4.3.7}$$

where n_1 and n_2 denote the number densities of each halo sample, z is the redshift and $\mathcal{C}$ represents the set of cosmological parameters in Eq. 4.1.1.

The input to the neural network has been standardised to facilitate training (such that its mean is 0 and standard deviation is 1). The output of the neural network is the logarithm of the correlation function $\log_{10}(\xi_{\rm hh})$, which is also standardised by

$$\log_{10}\left(\xi^{\rm R}_{\rm hh}(r)\right) \rightarrow \frac{\log_{10}\left(\xi^{\rm R}_{\rm hh}(r)\right) - \left\langle \log_{10}\left(\xi^{\rm R}_{\rm hh}(r)\right)\right\rangle}{\sqrt{\mathrm{Var}(\log_{10}(\xi^{\rm R}_{\rm hh}(r)))}}, \tag{4.3.8}$$

where $\left\langle \log_{10}\left(\xi^{\rm R}_{\rm gg}(r)\right)\right\rangle$ and $\mathrm{Var}\left(\log_{10}\left(\xi^{\rm R}_{\rm gg}(r)\right)\right)$ are the mean and variance of all correlation functions, estimated from the training set.

The output of the neural network is all the values of the correlation function evaluated for the pair-separation vector, r. Interestingly, when fitting the neural network with r as input, the model tends to overfit the data and converges to a less accurate overall model, while

Statistic	Batch size	Activation	$N_{\rm hidden}$	Resolution
$\xi_{\rm hh}$	5000	GELU	1024, 512, 512	LR
$\frac{\mathrm{d}n}{\mathrm{d}M}$	5000	GELU	1024, 512, 512	HR

Table 4.2: Summary of the best performing set of hyperparameters for the neural network emulators used to predict halo properties. The last column indicates the simulation resolution from which the quantity listed in the first column is measured.

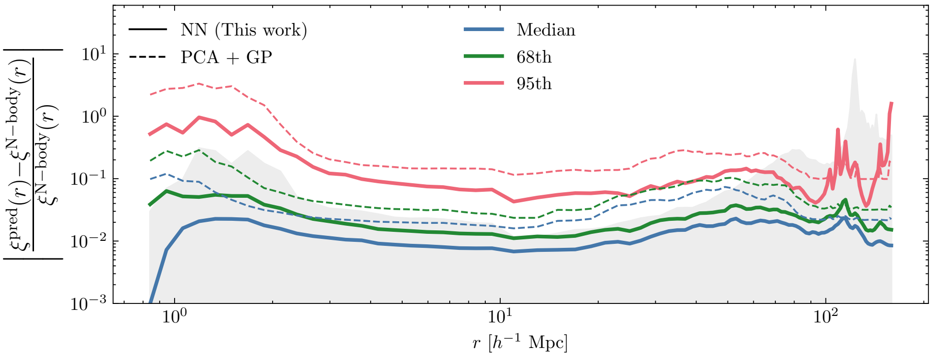


Figure 4.3: Comparison of the absolute fractional errors of the neural network emulator for the halo real space two point correlation function, with the Gaussian process + PCA approach presented in Nishimichi et al. (2019). We only include test set data, but for all redshifts and halo number densities. The grey shading shows the variance estimated from the simulations using the 15 realisations of the fiducial Planck cosmology, $\sigma_{\xi_{\rm fiducial}}/\xi_{\rm fiducial}$.

combining all pair separations shares the weights of the neural network across the values of r and reduces the level of overfitting.

We summarise the best-fitting hyperparameters of the neural network in Table 4.2.

In Fig. 4.3, we show the performance of the neural network as a function of pair separation compared to that found in Nishimichi et al. (2019). Fig. 4.3 shows the absolute errors estimated in the test set, as a function of pair separation r. Number densities and redshifts have been averaged.

The median absolute errors are lower than 2% throughout the entire scale range, a factor of 4 smaller than the upper limit of Nishimichi et al. (2019), while 68% had errors smaller than 6%, which is a factor of 5 smaller. We further compare the variance of the emulator errors (68th percentile fractional residuals) to the variance in the simulations themselves (grey solid background). This comparison shows that the emulator is already performing at a level similar to the variance in the simulations over the full-scale range. Note also that we cannot accurately estimate the model accuracy below the level of sample variance in the simulations, given that we only compare the accuracy of the model against one N-body realisation for each cosmology in the test set.

4.3.2 Halo mass function

Measurement

As explained earlier, we used the HR simulations to model the halo mass function. To do this, we first create a histogram of the number of halos in 80 logarithmically spaced bins in halo mass over the range of 10^{12} to $10^{16}\,h^{-1}\,M_{\odot}$. Following Nishimichi et al. (2019), we apply a correction to individual halo masses to account for systematics due to the finite number of particles. The corrected mass is given by (e.g. Warren et al. 2006):

$$\tilde{M} = (1 + N_{\mathrm{p}}^{-0.55})M, \tag{4.3.9}$$

where N_{p} is the number of simulation particles contained in the halo. The raw histogram is rather noisy, especially at the high-mass tail due to the small number of halos per bin. To produce a smooth mass function, we fit the data points using the functional form employed in Tinker et al. (2008). In doing so, we fix the parameter "b" in the formula, which controls the low mass behaviour, to the original value in Tinker et al. (2008) and allow the other three parameters to vary freely. We weight the bins according to the Poisson noise, which is more important at high masses, and the mass-determination accuracy, which is sensitive to the number of particles in the halo

$$\frac{\Delta N_{\mathrm{h}}}{N_{\mathrm{h}}} = \frac{1}{\sqrt{N_{\mathrm{h}}}} + \frac{1}{N_{\mathrm{p}}}. \tag{4.3.10}$$

The uncertanties in the fitted parameters are propagated to the smooth model prediction to obtain the expectation value, as well as the uncertanties of the estimated halo number counts in each mass bin.

Emulation

As in the case of the halo two-point correlation function, we train the model on the logarithm of the halo mass function to reduce the dynamic range of the observable. In this case, the mapping we obtain is

$$\log_{10}\left(\frac{\mathrm{d}n}{\mathrm{d}M}\right) = f(\mathcal{C}, z). \tag{4.3.11}$$

As before, we standardise inputs and outputs before training the model.

In Fig. 4.4, we compare the N-body measurements from the 10 test cosmologies with the emulator predictions at $z = 0$. The emulator achieves subpercent accuracy for halo masses smaller than $10^{14}\,h^{-1}\,M_{\odot}$, with the error increasing for larger halo masses. Estimating the

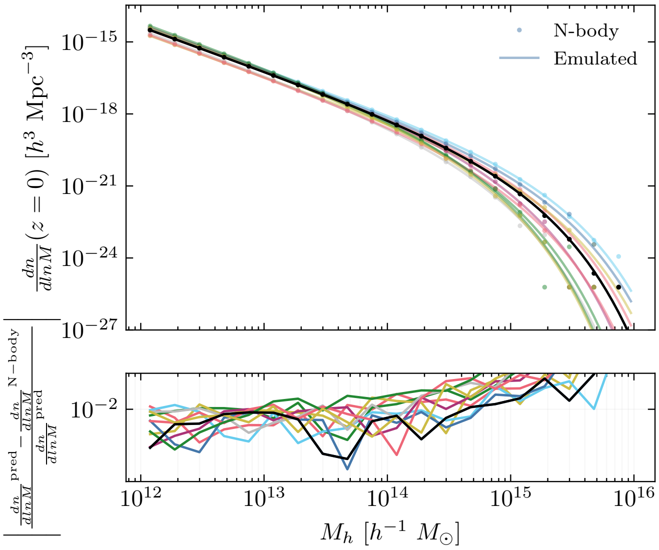


Figure 4.4: N-body measurements (points) and emulator predictions (lines) for the halo mass function at $z = 0$ in the 10 test set cosmologies. The lower panel shows the absolute fractional errors as a function of halo mass. The fiducial Planck cosmology is shown in black.

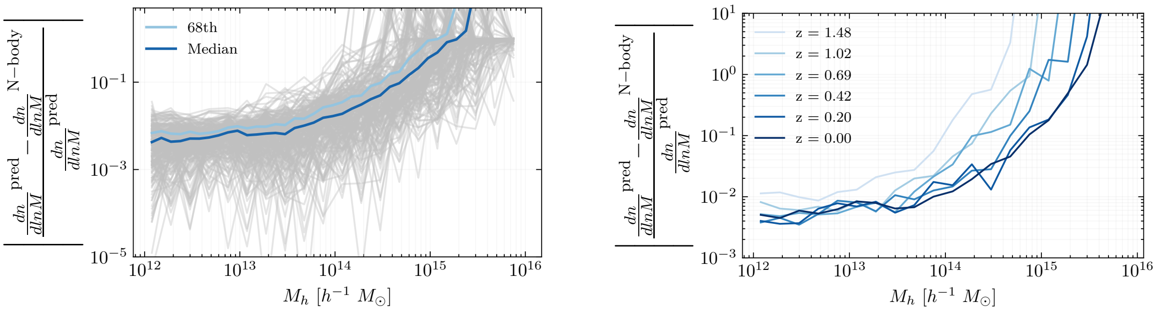


Figure 4.5: Absolute fractional errors on the halo mass function emulator predictions as a function of halo mass. The left panel shows the result for each test set sample (the 10 set cosmologies evaluated at the 21 different redshifts) as a gray line, along with the median (dark blue line) and 68th percentile range (light blue line) of the absolute fractional errors. The right panel shows the median absolute error as a function of halo mass, with different lines showing different redshifts, as indicated by the legend.

error is, however, challenging for halo masses larger than 10^{14} h^{-1} $M_\odot$ due to the large Poisson noise that affects the measurenents caused by the small number of cluster-size halos in the simulations.

In Fig. 4.5, we evaluate the overall accuracy of the halo mass function emulator at all redshifts (left panel) and as a function of the redshift (right panel). We find that the median emulator error for all redshifts is below 1 per cent for halo masses smaller than $10^{13.5}$ h^{-1} $M_\odot$, and increases rapidly to values larger than 10 per cent for the most massive halos ($M_{\rm h} > 10^{15}$ h^{-1} $M_\odot$). The right panel of Fig. 4.5 shows that the accuracy of the emulator degrades slightly at the highest redshifts considered ($z = 1.48$).

4.3.3 Galaxy clustering

We now assess the impact that inaccuracies in halo emulators have on galaxy clustering predictions. To do so, we populate the 10 test and 10 validation LR simulations with mock galaxies. We populate each cosmology at four different snapshots (z=0.1,0.25,0.5 and 0.75) and 5 different galaxy number densities, logarithmically spaced between $\log\left(\bar{n}_{\mathrm{gal}}/(h^{-1}\mathrm{Mpc})^{-3}\right) = -3.7$ and $\log\left(\bar{n}_{\mathrm{gal}}/(h^{-1}\mathrm{Mpc})^{-3}\right) = -4.3$. Note that halo property emulators cannot estimate galaxy clustering for arbitrary number densities, given that the lowest halo mass resolved by the Dark Quest simulations is $10^{12}\, h^{-1}\, M_{\odot}$.

For each combination of cosmology, redshift, and number density, we randomly sampled the HOD parameters from the ranges

$$\begin{aligned}
\sigma_{\log M} &\in [0.1, 0.8] \\
\alpha_{\mathrm{sat}} &\in [0.5, 1.] \\
\kappa &\in [0.1, 0.8] \\
\log M_1 &\in [13.5, 14.5]\,.
\end{aligned}$$

The remaining HOD parameter, $\log M_{\mathrm{min}}$, is fixed by the given galaxy number density. In total, we built a diverse sample of 400 HOD mocks with varying cosmology, HOD parameters, and redshift, to test the performance of the emulator.

Fig. 4.6 shows the emulator predictions for 20 HOD mocks at fixed redshift ($z = 0.25$), each of the curves is generated from a different set of cosmological parameters in the test and validation sets. Comparing the mock HOD catalogues with the emulator predictions, we find that the median error of the emulator is below 3 per cent on scales smaller than 50 h^{-1} Mpc, as shown in Fig. 4.7. Furthermore, the 68th percentile interval of the error increases only by 1 per cent with respect to the median. There is a small increase (≈ 1 per cent) in the error in the transition from one-to-two-halo term that occurs between 1 and $2\, h^{-1}$ Mpc. On large scales, the variance of the measurements is large, making it difficult to accurately determine the error of the emulator.

Fig. B.2 shows the performance of the emulator as a function of the galaxy number density and redshift. In both cases, the emulator shows similar levels of performance and therefore does not show any bias.

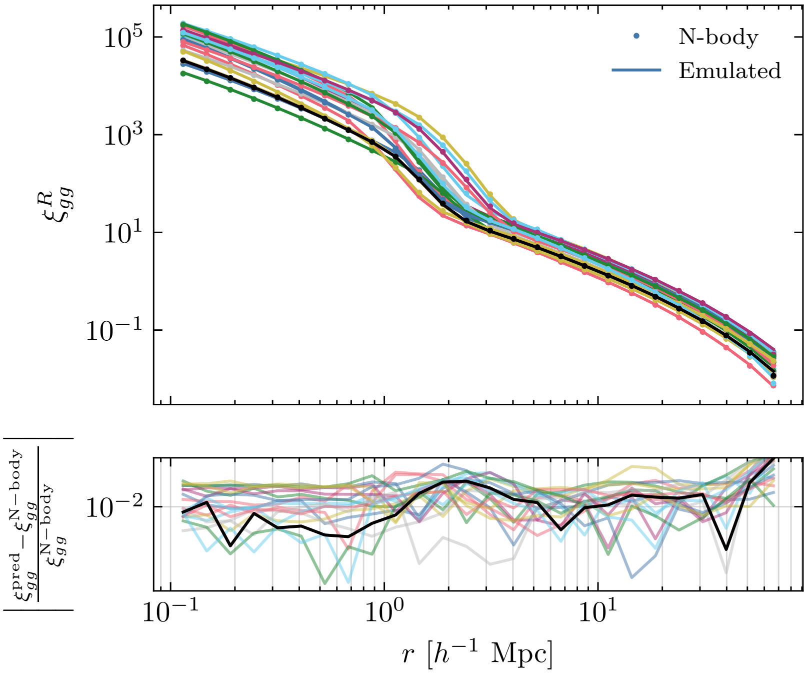


Figure 4.6: Emulator predictions for a subset of the 400 HOD mocks generated to test the accuracy of galaxy clustering. We show only those at $z = 0.25$. Planck cosmology is shown in black. The top panel shows all measurements from the 20 HOD catalogues and the corresponding emulator prediction. On the bottom pannel, we show the absolute error of the emulator as a function of scale.

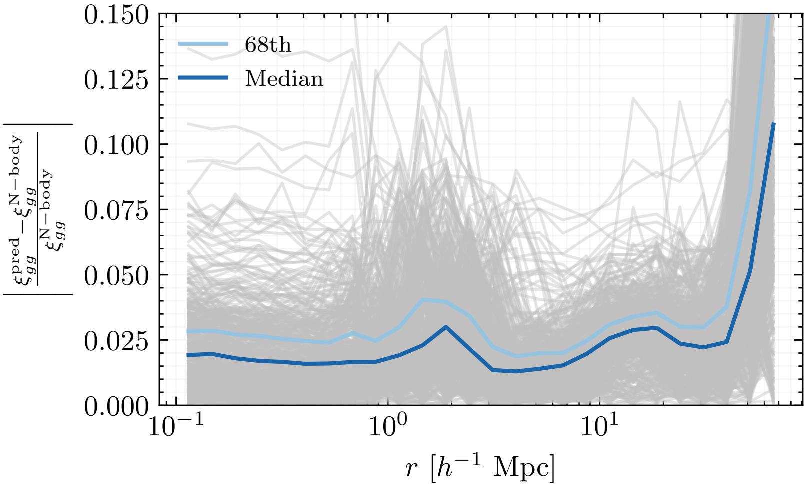


Figure 4.7: We show the absolute error of the emulator as a function of scale for each of the 400 HOD mocks generated to test the accuracy of galaxy clustering predictions for different cosmologies, redshifts, and galaxy number densities. The light and dark blue lines show the 68th credible interval and the median of the absolute errors.

	$\bar{z}$	$\bar{n}_{\rm g}$ $[(h^{-1}{\rm Mpc})^{-3}]$	$\log M_{\rm min}$ $[h^{-1}M_{\odot}]$	$\sigma_{\log M}$	$\log M_1$ $[h^{-1}M_{\odot}]$	κ	$\alpha_{\rm sat}$
Fiducial	0.251	2.174×10^{-4}	13.62	0.6915	14.42	0.51	0.9168
Min prior	-	-	12	0.1	12	0.01	0.5
Max prior	-	-	14.5	1	16	3	3

Table 4.3: The fiducial values and priors of the parameters for mock galaxy surveys that resemble the LOWZ galaxy sample.

4.4 The inverse problem: From correlations to cosmology

Here, we show how the galaxy two-point correlation function emulator is able to recover the cosmological parameters from mock simulated galaxies, first using the same HOD prescription as the one implemented in our theoretical model within the 68% credible interval for all parameters.

It should be emphasised that we focus on the three-dimensional two-point correlation of galaxies in real space, which is not directly observable in galaxy surveys. What we observe is the redshift space two-point correlation function of galaxies, which will be the subject of future work. However, it is important to show that the emulator is capable of recovering the parameters of interest for a mock dataset and to study the potential biases that might arise from adopting a too simplistic HOD model. We will also examine the scale dependence of the cosmological information content, which will, in turn, be important in determining the information content in redshift space.

We generated mock galaxy catalogues for LOWZ SDSS-like galaxies based on the fiducial Planck cosmology of the Dark Quest HR simulations, following Kobayashi et al. (2020). See Table 4.3 for the characterisation of the mock sample.

We use nested sampling, in particular the implementation of pymultinest (Buchner et al., 2014), to obtain samples from the posterior distribution. The posterior is defined as

$$p(\theta|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\theta)p(\theta), \tag{4.4.1}$$

where θ are the parameters to be estimated, $p(\theta|\mathcal{D})$ is the posterior distribution of the parameters given the data, $\mathcal{L}(\mathcal{D}|\theta)$ describes the likelihood of the data given the parameters, and $p(\theta)$ is the prior distribution of the model parameters.

We used a combination of the real space two-point correlation function and galaxy number density as our data vector and assumed that the likelihood follows a Gaussian distribution.

Therefore, we compute the log-likelihood, $\ell(\mathcal{D}|\theta)$, (up to a normalisation factor) as follows

$$\begin{aligned}\ell(\mathcal{D}|\theta) = -\frac{1}{2}\sum_{r_i,r_j}[\xi^s(r_i) - \xi^s(r_i|\theta)] \times C^{-1}(\xi^s(r_i), \xi^s(r_j)) \\ \times [\xi^s(r_j) - \xi^s(r_j|\theta)] + \frac{(n_{\mathrm{g}}^s - n_{\mathrm{g}}^s(\theta))^2}{\sigma_{n_{\mathrm{g}}}^2},\end{aligned} \tag{4.4.2}$$

where $\xi^s(r_i)$ denotes the two-point correlation function of the data for sample s, and $\xi^s(r_i|\theta)$ is the prediction of the theoretical model where θ denotes the model parameters, i.e. cosmological and HOD ($\mathcal{C} + \mathcal{G}$), C is the data covariance matrix, n_{g}^s is the galaxy number density estimated from the data, $n_g^s(\theta)$ the theoretical prediction, and $\sigma_{n_{\mathrm{g}}}$ the estimated error of the data that we fix to a nominal value of 5 per cent. The galaxy number density depends both on the HOD parameters and on cosmology, as seen in Eq. (4.2.9). See Appendix B.2 for a description of how the covariance matrix is estimated from N-body simulations.

Unless otherwise stated we will use the entire range of scales on which the emulator was trained, $0.1\,h^{-1}\,\mathrm{Mpc} \leq r \leq 150\,h^{-1}\,\mathrm{Mpc}$, to perform inference. Furthermore, although we vary the cosmological parameters $\mathcal{C} = \{\Omega_\Lambda, \ln A_{\mathrm{s}}, \omega_{\mathrm{c}}\}$, we show constraints on the derived parameters most commonly used $\mathcal{C} = \{\Omega_{\mathrm{m}}, \sigma_8, h\}$. The priors on the cosmological parameters are chosen to be uniform within the range of the sampled latin hyper-cube (Eq. 4.1.5); the priors on the HOD parameters are also chosen to be uniform with the ranges shown in Table 4.3.

4.4.1 Fiducial constraints

Here, we show that the emulator is capable of recovering the fiducial parameters of the mock catalogue within the 68% confidence interval for all parameters. The resulting 2-D posterior distributions are shown in blue in Fig. 4.8.

In the same figure, we also show the resulting constraints when the HOD parameters are fixed to their fiducial values (green) and the constraints on the HOD parameters when the cosmological parameters are fixed to their fiducial values (red).

Although taking either of these two steps in a real analysis would underestimate the error on the estimated parameter values, and most likely bias them, this is a useful exercise to determine how much more one could learn by combining the two-point correlation function with other statistics that can constrain the HOD parameters more accurately. For example, Hahn & Villaescusa-Navarro (2021) demonstrated how using the bispectrum could help us to improve constraints on both the cosmological and HOD parameters, by breaking degeneracies between them. Other probes, such as galaxy-galaxy weak lensing (More et al., 2015) can also be used to infer the HOD parameters. Fig. 4.8 shows that the constraints on Ω_m and σ_8 could be significantly improved by breaking the degeneracies with the HOD parameters.

On the other hand, it is mostly the mass scales M_min and M_1 that are better constrained by galaxy clustering when fixing the cosmological parameters. The remaining satellite parameters α and κ do not improve significantly by fixing cosmology. This is probably due to the fact that LOWZ galaxies have a low fraction of satellites, compared with other galaxy selections, and therefore their galaxy two-point correlation function is not very sensitive to these two satellite occupation parameters.

Fig. B.3 shows the effect of removing the number density constraint from the likelihood. As previously found in Miyatake et al. (2020), the constraints on cosmological parameters are not strongly affected by the number density term. However, the HOD parameters are sensitive to this change, with the parameters that influence the number of centrals becoming much more poorly constrained when the number density is not used.

4.4.2 The complementary role of small scales

Here, we study how the constraints vary as a function of the minimum scale included in the likelihood evaluation. This is a test of the performance of our model and its accuracy on small scales, and serves to illustrate the usefulness of small scales in reducing the errors on the recovered parameters. We show the results of this test in Fig. 4.9.

The small-scale information mainly constrains the fluctuation amplitude, σ_8, as shown in the upper panel of Fig. 4.9. From $r_\mathrm{min} = 1\,h^{-1}\,\mathrm{Mpc}$ to $r_\mathrm{min} = 5\,h^{-1}\,\mathrm{Mpc}$, the errorbars on σ_8 increase by a factor of 2.

In the same figure, we also show how the constraints on cosmological parameters would change if we fixed the HOD parameters. Interestingly, the Ω_m constraints would also be improved by including small-scale information by about a factor of 2 if there were no degeneracies with the HOD parameters. The constraints on h are dominated by the BAO scale and therefore do not change noticeably when smaller scales are included or the HOD parameters are fixed.

In the bottom panel of Fig. 4.9, we show the opposite effect, that of excluding large-scale information. The BAO scale has a very small effect on the recovered value of σ_8, whereas it dominates the constraints on the cosmological parameters Ω_m and h, after marginalising

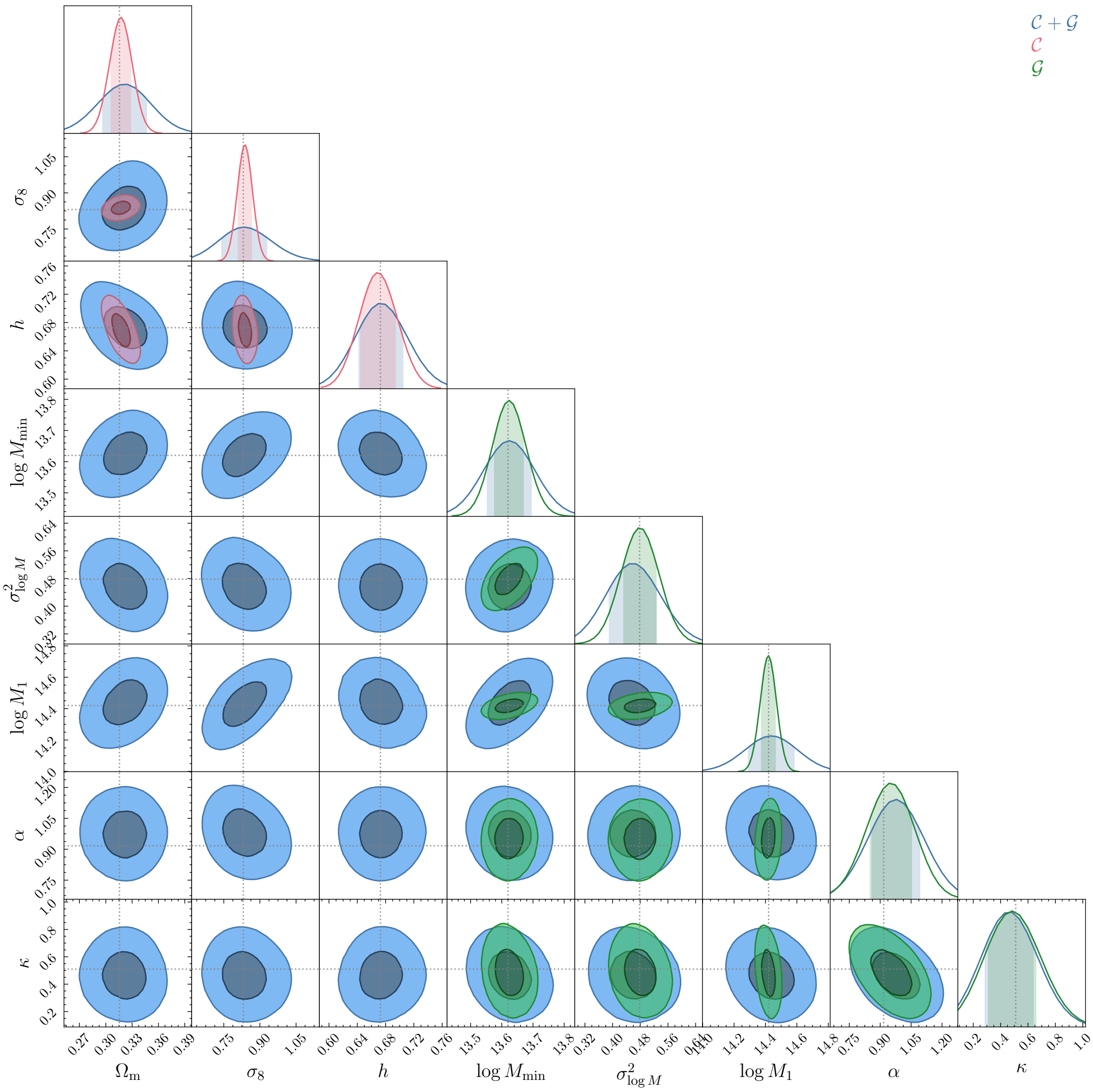


Figure 4.8: This plot shows that the emulator can recover the true cosmological and HOD parameters within the confidence intervals. We show the posteriors which result when varying both cosmology and HOD parameters ($\mathcal{C}$ and $\mathcal{G}$) (blue, labelled "$\mathcal{C}+\mathcal{G}$") and the cosmological constraints found when the HOD parameters ($\mathcal{C}$) are set to their fiducial values (red, labelled "$\mathcal{C}$"). The constraints on the HOD parameters ($\mathcal{G}$) obtained by fixing the cosmological parameters to their fiducial values are shown in green (labelled "$\mathcal{G}$"). The true values that generated the simulated data are shown by the dotted gray lines.

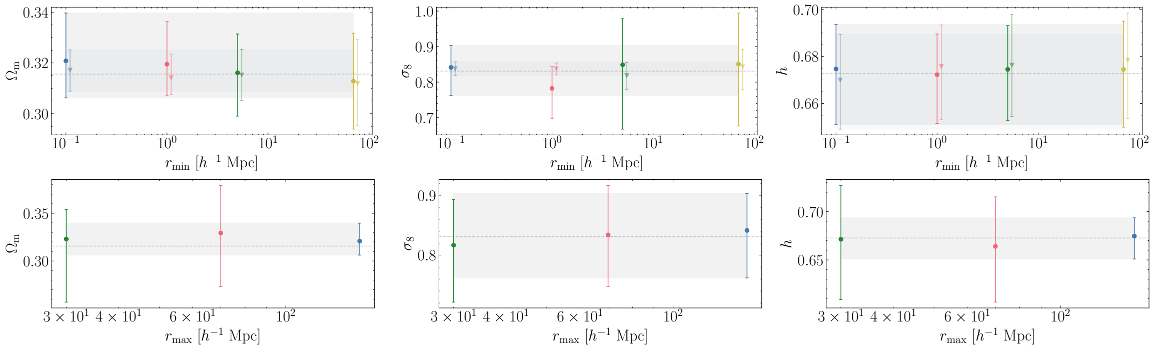


Figure 4.9: We show the estimated maximum likelihood parameters, together with their estimated uncertainties, for varying minimum and maximum pair separation scales used in the analysis. In the top panel we show both the cosmological constraints obtained when marginalizing over the HOD parameters (circles) and when fixing the HOD parameters to their fiducial values (triangles). This shows that the constraints on the cosmological parameters improve as more non-linear scales are included for all parameters but h, whose constraints are dominated by the BAO information.

over the HOD parameters. Most emulators (Zhai et al., 2019b; Yuan et al., 2022a) focus on scales smaller than $30\,h^{-1}$ Mpc, and therefore lose constraining power on Ω_m and h.

4.4.3 The consequences of ignoring assembly bias

We now test whether the halo-connection model used here is flexible enough to obtain unbiased cosmological constraints when modelling the clustering of a sample known to contain assembly bias. Although dark matter halo mass correlates strongly with galaxy clustering, we know that dark matter halos experience different assembly histories even at fixed halo mass, and can display different clustering. These different assembly histories influence secondary properties of halos, and this, in turn, might also affect the formation of galaxies, and hence result in different galactic contents for halos of the same mass.

These effects are known as *halo* and *galaxy assembly bias.* Although these two effects share the words assembly bias, they refer to different effects

- *Halo assembly bias* refers to differences in the clustering of dark matter halos at a fixed halo mass. These differences depend on the choice of secondary halo properties, which usually correlate with the formation history of the halo, such as halo concentration or substructure fraction.

- *Galaxy assembly bias* refers to differences in the number of galaxies within dark matter halos at a fixed halo mass, which in turn may depend on secondary halo properties.

Galaxy clustering is shaped by both of these effects. On one hand, halo assembly bias implies that, at fixed halo mass, grouping dark matter halos by a secondary property results in a different clustering signal. On the other hand, the way galaxies occupy dark matter halos might depend on properties other than mass. The combination of both effects determines how strongly galaxy clustering depends on secondary dark-matter halo properties, and therefore how important it is to model this dependency in order to obtain unbiased cosmological constraints.

Here, we want to test how assembly bias affects our constraints when we include effects similar to those observed in hydrodynamical simulations (Hadzhiyska et al., 2021) and semi-analytical models of galaxy formation (Zehavi et al., 2018; Xu et al., 2021; Jiménez et al., 2021) in our mock galaxy catalogues. In this way, we can assess whether the halo model is flexible enough to recover unbiased constraints from realistic galaxy mocks when including small-scale information.

In particular, we implement the assembly bias model based on environment introduced in Xu et al. (2021). The authors showed that the smoothed matter density can account for most of the assembly bias signal observed in a semi-analytic galaxy formation model. This is in agreement with other studies using hydrodynamical simulations (Hadzhiyska et al., 2021). Note however that this environmental assembly bias effect has not been found in observational data yet in SDSS-like survey volumes (Abbas & Sheth, 2007; Paranjape et al., 2018)

To create mock galaxy catalogues with an environment-based assembly bias signal, we first determine the local density around each halo. We compute the dark matter density field smoothed with a Gaussian filter over a scale of $2.5\,h^{-1}$Mpc, by first measuring the counts-in-cells dark matter particle density on a 512^3 grid and then multiplying with a Gaussian kernel in Fourier space. The matter overdensity value at the position of each halo is found by interpolating over the 3D grid. Finally, we rank the overdensity values of the halos at fixed halo mass and normalise them to be between 0 and 1. We have computed the ranks inside 50 logarithmically spaced halo mass bins in the range $12 < \log_{10}\left[M_\mathrm{h}/(h^{-1}M_\odot)\right] < 16$. These ranks, $\delta_{2.5}^\mathrm{rank}$, are then normalised between 0 and 1 in each halo mass bin.

Once we have determined the ranked environment density around each halo, we assign galaxies to dark matter halos through equations Eq. (4.2.4) and Eq. (4.2.5), modifying the values of $\log M_\mathrm{min}$ and $\log M_1$ with the rank of the halo's overdensity value

$$\log_{10} M_\mathrm{min}(\delta_{2.5}^\mathrm{rank}) = \log_{10} M_\mathrm{min}^0 + B_\mathrm{cen} \times \left(\delta_{2.5}^\mathrm{rank} - 0.5\right), \tag{4.4.3}$$

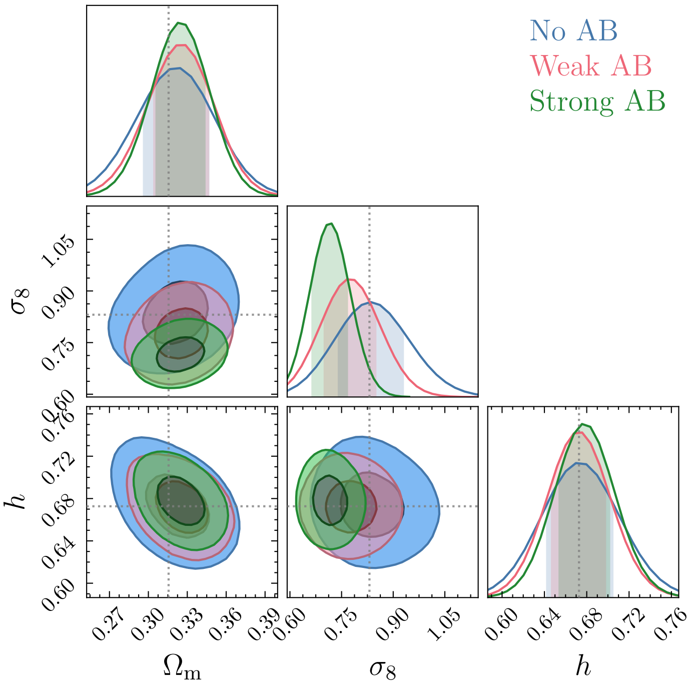


Figure 4.10: Constraints obtained when fitting mock catalogues that include the environment-based assembly bias model presented in Xu et al. (2021) with our halo model emulator, which ignores the effect of assembly bias. The cosmological parameters Ω_m and h can still be recovered within the estimated confidence intervals, since they are mainly constrained by the BAO peak, whereas σ_8 shows a small bias towards smaller values in both the weak and strong assembly bias scenarios.

$$\log_{10} M_1(\delta_{2.5}^{\mathrm{rank}}) = \log_{10} M_1^0 + B_{\mathrm{sat}} \times \left(\delta_{2.5}^{\mathrm{rank}} - 0.5\right), \tag{4.4.4}$$

where B_cen and B_sat are the central and satellite assembly bias parameters that control the strength of the effect. Since more galaxies will form in overdense regions, the values of B_cen and B_sat will be negative.

To explore the possible biases that ignoring assembly bias may introduce in the estimated cosmological parameters, we study two scenarios: i) a weak assembly bias effect with values $B_\mathrm{cen} = -0.1$ and $B_\mathrm{sat} = -0.2$, and ii) a strong one with values $B_\mathrm{cen} = -0.2$ and $B_\mathrm{sat} = -0.4$. The weak assembly bias parameters have been chosen to mimic the level of assembly bias signal found in Xu et al. (2021) for a sample with a galaxy number density of $n_\mathrm{gal} = 0.01$ $(h^{-1}\mathrm{Mpc})^{-3}$. In Fig. B.4, we show that the weak scenario produces changes in the two-point correlation function of up to 10 per cent compared with the case with no assmebly bias, while the strong case increases the clustering by up to 20 per cent.

Fig. 4.10 shows the constraints obtained using our model (which ignores assembly bias) to fit the clustering measured from the mock galaxy samples described above, with weak and strong assembly bias. In both the weak and strong assembly bias scenarios, we can robustly recover the cosmological parameters Ω_m and h since they are mostly determined by the BAO scale. However, σ_8 is biased towards smaller values in both scenarios. In the strong assembly bias case, this shift is more than $1-\sigma$ away from its true value. However, we note that the strong assembly bias scenario is unrealistic for a LOWZ-like sample of galaxies (Yuan et al., 2022a).

Fig. B.5 shows the full 2D posterior, including the HOD parameters that have shifted in the expected direction. Intuitively, the environment assembly bias effect leads to more galaxies forming in overdense regions (thus, the assembly bias parameters are negative). The left hand side of Fig. B.4 shows that higher number densities in the assembly bias mocks correspond to a higher mean number of galaxies, that could be effectively reproduced by lowering M_min.

Fig. 4.11 shows how the constraints on σ_8 change as we vary the minimum scale included in the determination of the likelihood. If we restrict the analysis to scales larger than 10 h^{-1} Mpc, the halo model recovers unbiased cosmological constraints by biasing the HOD parameters. However, on scales smaller than 10 h^{-1} Mpc, when the constraining power on σ_8 doubles, lowering the mass of halos that host a central cannot mimic the effects shown in Fig.B.4, and σ_8 needs to be lowered to describe the changes around the one to two halo term transition.

We can monitor the evidence of the model to detect whether the halo-galaxy connection model has been mispecified. The evidence is defined as

$$P(\mathcal{D}) = \int \mathrm{d}\theta \, \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta), \tag{4.4.5}$$

and can be interpreted as the likelihood of the data given the model. The values of the evidence estimated by nested sampling are 20.87 for mocks without assembly bias, 18.34 for those with a weak assembly bias signal, and 16.37 for those with a strong assembly bias effect.

Given the importance of unbiased constraints on σ_8 to resolve the $\sigma_8 - S_8$ tension, we will work on adding environment-based assembly bias to our emulator for its application to DESI Y1 data.

4.4.4 Comparison with Lagrangian Perturbation Theory

In this section, we compare the emulator constraints with those obtained by 1-loop Lagrangian perturbation theory (Chen et al., 2020, 2021) using the publicly available code VELOCILEP-

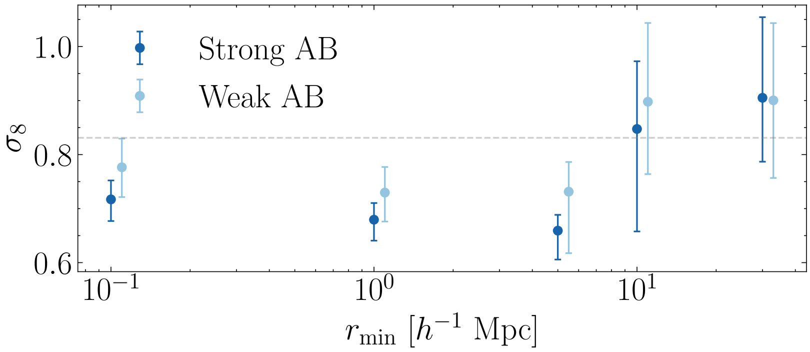


Figure 4.11: Inferred values of σ_8 and their estimated uncertainties as a function of the minimum scale, $r_{\rm min}$, used in the likelihood analysis. This plot shows the systematic introduced by assembly bias can only be removed by excluding the small scale information.

tors[2]. We fit the bias parameters b_1, b_2, and b_s, together with the cosmological parameters. Since we are only looking at the real space correlation function and not at velocity statistics, we do not include the one-loop effective field theory counter-terms in the analysis.

In Fig. 4.12, we show how the emulator can obtain constraints similar to LPT when analysed over the same scale range, even after marginalising the halo-galaxy connection parameters, which are in total 6 free parameters (compared to only 3 for LPT). The LPT predictions are slightly biased in σ_8, this is due to the strong degeneracy between b_1 and σ_8 that is accentuated in real space. In such a situation, the 1-D marginalized posterior for σ_8 can depend strongly on the prior or the parameterisation of the nuisance parameters, potentially leading to a biased estimate (Sugiyama et al., 2020). The biased estimate of σ_8 tends to be alleviated by including more information, e.g., redshift space distortions. As shown in Fig. 4.12, including small scale information does allow the emulator to constrain the parameters more accurately.

4.5 Discussion and Conclusion

We show that after marginalizing over uncertainties in the galaxy-halo connection parameters, an emulator of the real space correlation function based on the halo model can obtain tighter constraints on the cosmological parameters than Lagrangian Perturbation Theory (LPT) given that the latter cannot extract the additional information contained in small scale galaxy clustering.

The treatment of galaxy bias in both approaches is very different. On the one hand,

[2] https://github.com/sfschen/velocileptors

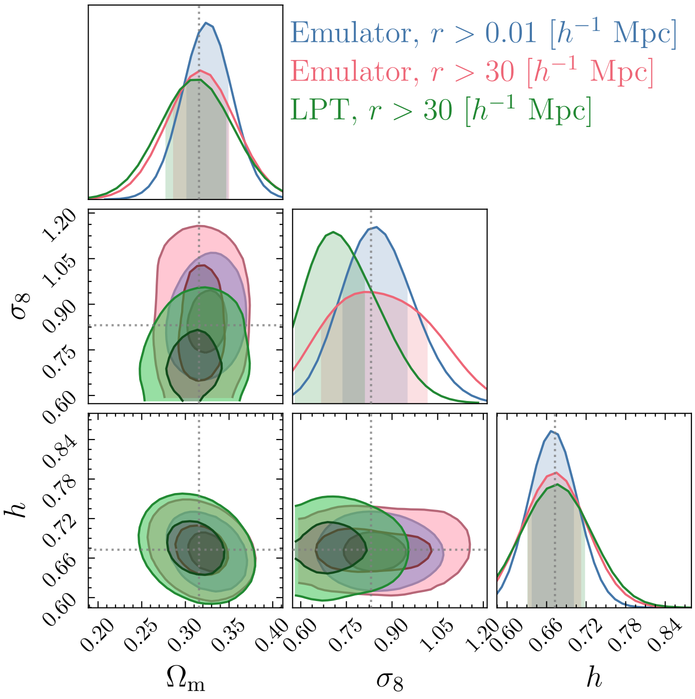


Figure 4.12: Comparison of the constraints obtained by the emulator based model, using the whole range of scales or only quasi-linear scales, with the 1 loop Perturbation Theory model presented in Chen et al. (2020, 2021) on quasi-linear scales.

the bias treatment of LPT is based on expanding the galaxy number density perturbation $\delta_\mathrm{g}(\boldsymbol{x})$, in terms of all local operators that are relevant at a given order in perturbation theory (Desjacques et al., 2018b). The free coefficients that accompany each operator are called bias parameters, and these are the ones that need to be fitted to the data. The flexibility of this bias expansion to be able to reproduce the observed clustering in different galaxy-halo connection models will be determined by the operators included and their degeneracies. On the other hand, the HOD approach implemented in this chapter is restricted by the assumption one makes about the halo properties that determine halo clustering and galaxy occupations. More work is needed to determine the robustness of both approaches against uncertainties in the model connecting halos to galaxies. In the future, we plan to compare the constraints obtained with both models using large hydrodynamic simulations or semi-analytic models of galaxy formation.

Regarding the emulation approach, we have combined an emulator trained in halo properties with an analytical prescription of how galaxies populate halos, as already done by Nishimichi et al. (2019). Most other emulators, however, are trained on HOD catalogues built on N-body simulations (Zhai et al., 2019b; Yuan et al., 2022a). Our approach has advantages and disadvantages. In particular, the halo model allows us to reduce emulator errors through an analytical galaxy-halo connection, which also simplifies the task for the emulator that only needs to learn the dependency of halo clustering on cosmological parameters. Moreover, the analytical model allows us to compute different observables, such as the galaxy-cluster cross-correlation function or a multitracer two-point correlation function. Obtaining cosmological information from small scales through these observables will be the subject of future work. It also allows us to combine emulators trained on simulations with different resolutions to reduce cosmic variance on large scales and perform an analysis using the full-shape of the correlation function.

Regarding the disadvantages of our approach, extending the halo model approach to arbitrary statistics could potentially be difficult. The emulation of statistics such as the bispectrum, would be simplified if one were to follow the procedure outlined in Zhai et al. (2019b) and Yuan et al. (2022a). Moreover, more work needs to be done in order to go beyond the vanilla HOD model used in this work to introduce effects such as the environment-based assembly bias shown in Section 4.4.3. In the future, we plan to introduce a correction based on binning the halo two-point correlation function in terms of halo environment.

We have shown that including environment-based assembly bias in the model is important to avoid biased constraints on σ_8. This is especially relevant given the $f\sigma_8$ tension. Previously, Kobayashi et al. (2022) and Miyatake et al. (2020) had performed tests similar to the one presented in Sec. 4.4.3 to emulators based also on the halo model. Kobayashi et al. (2022) studied the effect that ignoring concentration-based assembly bias would have on the cosmological parameters inferred when emulating the redshift space power spectrum through the halo model. They found that although the mock galaxies show $10 - 20$ per cent higher amplitudes than the mocks without assembly bias, they can still recover unbiased cosmological constraints through a change in the HOD parameters. In contrast, Miyatake et al. (2020) found that the same effects of assembly bias would introduce biases in Ω_m and σ_8 when the data vector is a combination of the projected two-point correlation function of galaxies and galaxy-galaxy lensing. In this case, the fact that one can use galaxy-galaxy lensing to accurately determine the scaling of halo bias with halo mass restricts the flexibility of the HOD model, which is not able to adapt the parameters in such a way that unbiased constraints can be recovered.

We have here explored an assembly bias model inspired by semi-analytic methods of galaxy formation and hydrodynamical simulations. In fact, these studies find that the magnitude of concentration-based assembly bias is small. Ignoring environment-based assembly bias in the theory model, we find that the halo model is not flexible enough to obtain unbiased cosmological constraints already when the effect of assembly bias only impacts clustering by about 10%. Moreover, we find that including the BAO scale allows us to obtain robust constraints on Ω_m.

To summarise, we have

- Presented a neural network which models the full-shape galaxy clustering in real space based on the halo model, which is more accurate and faster than previously published Gaussian process emulators Nishimichi et al. (2019), when trained on the same dataset. The method presented here can produce a galaxy correlation function in less than 300 ms.

- Shown that small scale galaxy clustering ($r < 5\ h^{-1}\,\mathrm{Mpc}$) in real space improves the constraints on σ_8 by a factor of 2, whereas marginalising over the HOD parameters erases the information contained on small scales for Ω_m.

- Shown that a halo model that ignores effects of environment-based assembly bias similar to those observed in hydrodynamic simulations and semianalytic models of galaxy formation could introduce bias in the inferred σ_8, while the BAO peak ensures that we can recover Ω_m and h robustly.

- Found that the above-mentioned bias in the value of inferred σ_8 disappears when analysing scales larger than $10\ h^{-1}\,\mathrm{Mpc}$.

Currently, we are working on analogous neural network emulators of the pairwise velocity moments that will be used to i) perform the real to redshift space mapping to predict the cosmological dependence of redshift-space galaxy clustering, and ii) constrain observations of the peculiar velocity field.

In the future, we also plan to use the neural network emulators on DESI Y1 data to constrain the cosmological parameters. This requires that the models be trained on simulations with lower particle mass so that they can reach the high galaxy number densities that DESI will measure. For this, a new simulation campaign, Dark Quest II., is currently ongoing to cover a wider mass range (down to a few $10^{11}\,h^{-1}M_\odot$) in an extended cosmological model space including massive neutrinos, time-varying dark energy equation-of-state parameter and spatial curvature using a newly developed fast N-body code (Nishimichi *et al.* in prep.).

Chapter 5

The information content of environment dependent clustering

Extracting the relevant information from complex and high-dimensional datasets is challenging; one needs a mapping between a potentially noisy high-dimensional space, such as the three-dimensional density field, to a reduced set of parameters that define a particular theory, such as the cosmological parameters. We can devise different techniques by thinking about the properties of the data, like its symmetries or signal-to-noise ratio. In Chapter 4, we have shown how the correlation function can be used to constrain the cosmological parameters from three dimensional galaxy maps. The two-point correlation function fully describes a Gaussian random field statistically, and the symmetries of the density field allow us to simplify its description as a function of scale only.

However, as discussed in Section 1.1.3, gravitational evolution introduces non-Gaussianities in an initially Gaussian random field and therefore deems the two-point function sub-optimal for the task of constraining cosmology. In this chapter, we will present a study of an alternative summary statistic: environment-dependent clustering (Abbas & Sheth, 2007; Tinker, 2007; Paillas et al., 2021; Bonnaire et al., 2022).

Splitting the galaxy field into different density bins naturally captures the non-Gaussian nature of the PDF. In this work, we perform a Fisher analysis to quantify the precision with which density-split (DS) clustering (Paillas et al., 2021) can constrain the value of cosmological parameters in a $\nu\Lambda$CDM model. We study how different definitions of environmental density can affect the constraints of DS and compare them with the results of the standard two-point correlation function (2PCF). In particular, we compare the information content of DS when the environments are defined in either real or redshift space. Furthermore, in previous studies (Paillas et al., 2021), several limiting assumptions had to be made to model the clustering of DS analytically. To overcome this problem and estimate the full information content of DS, we use the Quijote suite of N-body simulations (Villaescusa-Navarro et al., 2020).

5.1 The Quijote Simulations

The Quijote project (Villaescusa-Navarro et al., 2020) is a suite of 44 100 full N-body simulations constructed to quantify the information content of cosmological observables. The simulations span a wide range of values around the fiducial cosmology, which is set to a matter density parameter of $\Omega_m = 0.3175$, a baryon density of $\Omega_b = 0.049$, a dimensionless Hubble constant of $h = 0.6711$, a spectral index of $n_s = 0.9624$, an amplitude of density fluctuations of $\sigma_8 = 0.834$, a neutrino mass of $M_\nu = 0.0\,\mathrm{eV}$, and a dark energy equation of state of $w = -1$. The fiducial cosmological parameters are in good agreement with the latest Planck constraints (Planck Collaboration et al., 2020a). There are 15 000 realisations of the fiducial cosmology that can be used to calculate covariance matrices, as well as 500 realisations of paired simulations where only one cosmological parameter is changed at a time, which can be used to estimate derivatives numerically. The specifications of these simulations are listed in Table 5.1.

The halo catalogues in each simulation are generated using a Friends of Friends algorithm (Davis et al., 1985), with the linking length parameter set to b = 0.2. Throughout, we select haloes at redshift $z = 0.0$ by imposing a minimum halo mass cut of $M_{\min} = 3.2 \times 10^{13}\,h^{-1}\mathrm{M}_\odot$. Future surveys, such as DESI, will be able to sample galaxies living in haloes of much lower masses. Therefore, the constraints shown in this chapter do not serve as a forecast for future surveys but rather serve as a comparison between two-point statistics and DS.

Adopting a fixed mass cut can modify the bias of the halo samples with respect to the underlying matter distribution, which in turn affects the measured clustering statistics. To disentangle this effect from those coming from variations in cosmological parameters, we also build halo catalogues where we impose mass cuts of $3.1 \times 10^{13}\,h^{-1}\mathrm{M}_\odot$ and $3.3 \times 10^{13}\,h^{-1}\mathrm{M}_\odot$, so that we can compute derivatives of the data vectors with respect to this mass cut and marginalise over this dependence.

Name	$\Omega_{\rm m}$	$\Omega_{\rm b}$	h	n_s	σ_8	M_ν	realizations
Fiducial	0.3175	0.049	0.6711	0.9624	0.834	0.0	15000
Fiducial (ZA)	0.3175	0.049	0.6711	0.9624	0.834	0.0	500
$\Omega_{\rm m}^+$	0.3275	0.049	0.6711	0.9624	0.834	0.0	500
$\Omega_{\rm m}^-$	0.3075	0.049	0.6711	0.9624	0.834	0.0	500
$\Omega_{\rm b}^+$	0.3175	0.051	0.6711	0.9624	0.834	0.0	500
$\Omega_{\rm b}^-$	0.3175	0.047	0.6711	0.9624	0.834	0.0	500
h^+	0.3175	0.049	0.6911	0.9624	0.834	0.0	500
h^-	0.3175	0.049	0.6511	0.9624	0.834	0.0	500
n_s^+	0.3175	0.049	0.6711	0.9824	0.834	0.0	500
n_s^-	0.3175	0.049	0.6711	0.9424	0.834	0.0	500
σ_8^+	0.3175	0.049	0.6711	0.9624	0.849	0.0	500
σ_8^-	0.3175	0.049	0.6711	0.9624	0.819	0.0	500
$\mathrm{M}_\nu{}^+$	0.3175	0.049	0.6711	0.9624	0.834	0.1	500
$\mathrm{M}_\nu{}^{++}$	0.3175	0.049	0.6711	0.9624	0.834	0.2	500
$\mathrm{M}_\nu{}^{+++}$	0.3175	0.049	0.6711	0.9624	0.834	0.4	500

Table 5.1: Characteristics of the Quijote simulations suite that are used in this work. Each row corresponds to a set of simulations with a varying cosmological parameter. The simulations are set to span a grid of cosmologies ready for numerically estimating derivatives with respect to cosmological parameters.

5.2 Density-split clustering

The DS clustering method (Paillas et al., 2021) consists of grouping a collection of random points according to the local galaxy density around them and then extracting cosmological information from the clustering statistics that characterise each environment. This information would be averaged out in the two-point correlation function.

We apply the DS algorithm to the halo catalogues of Quijote simulations using our publicly available code[1]. In Fig. 5.1 we show a sketch of the density split pipeline, which can be summarised as follows:

1. Generate a set of $N_{\rm random}$ random points that cover the sample volume and measure the integrated halo number density contrast $\Delta(R_s)$ in spheres of radius R_s around each random point.

2. Classify the random points into five density bins, or *quintiles*, based on the densities measured from the previous step. By definition, each quantile will have the same number of points. We find that five quantiles are a good compromise between distinguishing different environments and reducing the shot noise that a higher number of quantiles would introduce. In Fig. 5.2 we show the random points that were classified as the least

[1] https://github.com/epaillas/density-split-rsd

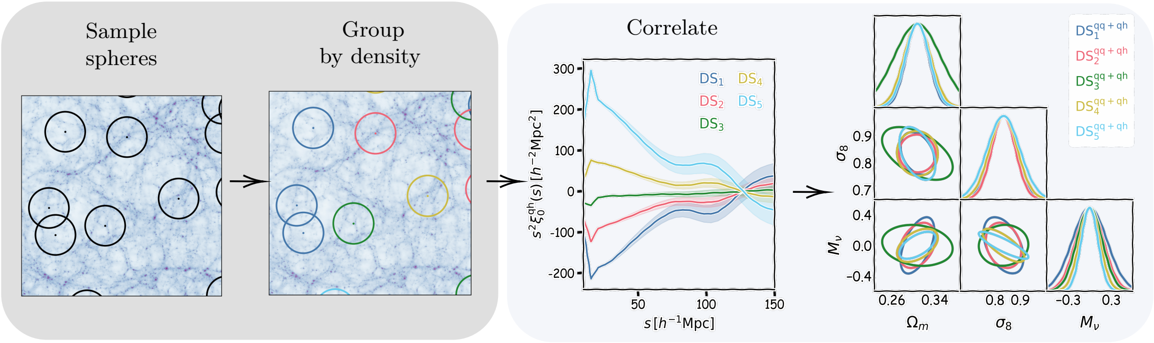


Figure 5.1: Summary of the density split clustering pipeline. First, we sample random spheres throughout the simulation box. Then we compute the environment density around each sphere within a smoothing scale and group them into 5 quintiles. Using the grouped spheres, we compute the cross-correlation between the random centres in each quintile and the all the redshift space dark matter halos, and the auto-correlation of the random points in each quintile. Finally, we estimate the constraints on the cosmological parameters through a Fisher analysis.

(DS_1) and most dense (DS_5) environments in a slice of the Quijote simulations. It can be seen that the DS_1 points correspond to regions that would normally be denoted as voids, while the DS_5 points correspond to nodes of the cosmic web.

3. Measure the multipole moments of the cross-correlation functions between the points in each quantile and the redshift-space halo field, as well as the autocorrelation function of the points in each quintile. The use of autocorrelations is an addition that was not previously considered in Paillas et al. (2021). In what follows, we denote autocorrelations of the i-th quintile as $\mathrm{DS}_i^{\mathrm{qq}}$ and cross-correlations between the i-th quintile and the redshift-space halo field as $\mathrm{DS}_i^{\mathrm{qh}}$.

4. Use changes in measured multipoles with cosmology to estimate constraints on the parameters of the $\nu\Lambda$CDM model through a Fisher analysis.

The multipole moments are defined as

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1} \mathrm{d}\mu\,\xi(\mathrm{s},\mu)\mathrm{P}_\ell(\mu), \tag{5.2.1}$$

where $\mu = \cos\theta$, $P_\ell(\mu)$ is the Legendre Polynomial, $\ell = 0, 2$ for monopole and quadrupole, respectively, and $\xi(s,\mu)$ denotes either the cross-correlations between quintiles and the halo field in redshift space, or autocorrelations of quintiles.

We have run tests with different choices of N_{random}, and have found that the clustering measurements converge when this number is set to five times the number of haloes in each simulation, when using five density bins. This guarantees that the number of centres in each quintile is the same as the number of halos. Therefore, we set $N_{\mathrm{random}} = 5N_{\mathrm{haloes}}$

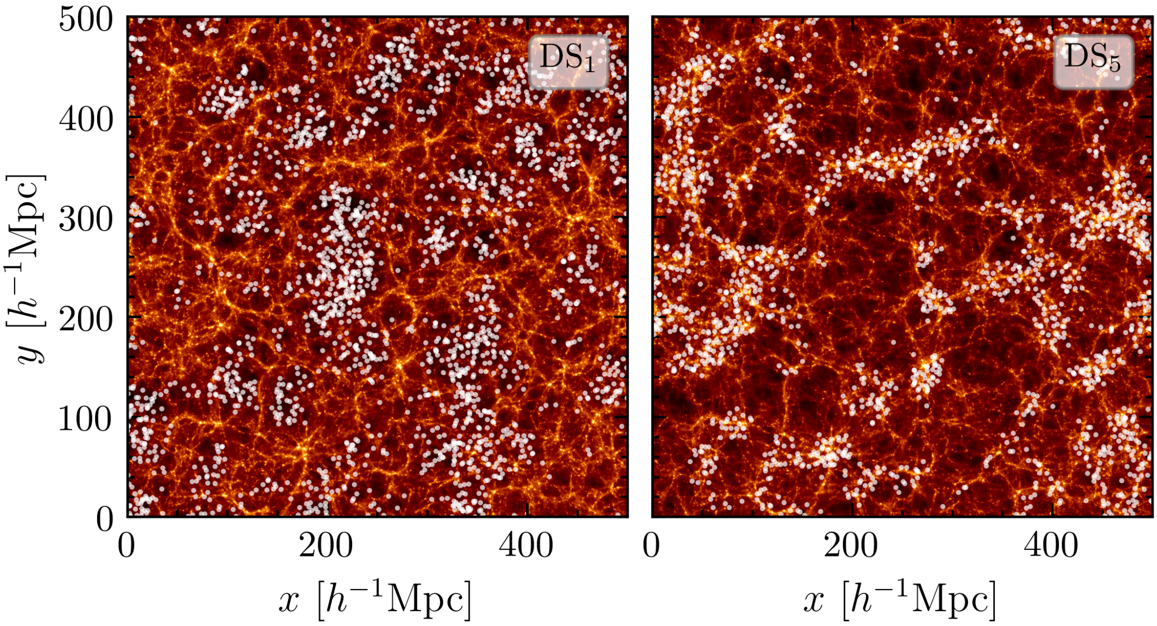


Figure 5.2: The positions of the DS$_1$ and DS$_5$ density-split quintiles (white circles) in a region of $500 \times 500 \times 50\,(h^{-1}\text{Mpc})^3$ from one of fiducial Quijote simulations at $z = 0$. The colourmap shows the projected dark matter density within the same region. DS$_1$ centres populate the most underdense environments of the cosmic web, whereas DS$_5$ centres cluster on high density environments.

throughout the rest of this work. We set the default smoothing radius R_s to $20\,h^{-1}$Mpc, which is well above the mean halo separation in the simulations but still sufficiently small to capture non-Gaussianities in the density PDF.

Estimation of the halo density around random points in step (i) can be performed in real or redshift space. Paillas et al. (2021) showed that, from a theoretical point of view, it is easier to model the real to redshift space mapping when quintiles are defined in real space. However, in observations, we only have direct access to the redshift-space galaxy field. A similar problem is found in void-galaxy cross-correlation studies (Nadathur et al., 2019) where reconstruction algorithms (Nadathur et al., 2018) are commonly used to detect voids in real space. However, the reconstruction step also introduces additional complexity when estimating the likelihood of the data given the cosmological parameters since the reconstructed data depend on some of the parameters being fitted (such as the growth rate of structure, f, or the linear galaxy bias). Moreover, reconstruction algorithms are not perfect and might introduce biases in the estimates of real space quantities that would impact the inference on cosmological parameters. This will be particularly relevant when including small scales in the analysis, where the signal-

[1] The projected dark matter density has been estimated using the DTFE public software (https://github.com/MariusCautun/DTFE).

to-noise ratio is largest. Here, we compare both the definitions of the density split and the resulting constraints.

The autocorrelation and cross-correlation functions of each density environment are calculated using PYCORR[2], which is a wrapper around CorrFunc (Sinha & Garrison, 2020). We used 28 radial bins within $10 < s < 150\,h^{-1}$Mpc, and 240 μ bins from -1 to 1 for the calculation of redshift-space multipoles. We restrict ourselves to using the monopole and quadrupole moments of the correlation functions. In principle, valuable information could also be contained in the hexadecapole, but its statistical uncertainty for the samples resolved by the Quijote simulations is too large to be included in this analysis. We also measure the multipoles from the halo 2PCF with the same binning scheme for comparison.

5.2.1 The impact of identifying density environments in real or redshift space

For observational data, we can only access the redshift space positions of galaxies. However, as in void-galaxy cross-correlation studies, their real space positions can be estimated using reconstruction algorithms (Nadathur et al., 2018). In this section, we examine the key differences between density splits identified in real (r-split), redshift (z-split), or reconstructed space (recon-split), and we will later use the Fisher formalism to determine the impact that split identification has on cosmological constraints.

First, we compare the real and redshift splits using the same set of random centres. This allows us to make a one-to-one comparison of real and redshift space environments. In Fig. 5.3, we show the joint distribution of overdensities estimated using either the real space positions of the halos, $\Delta^{\rm R}$, or the redshift space positions, $\Delta^{\rm S}$. The contours are slightly tilted; underdense (overdense) regions appear more underdense (overdense) in redshift space. In underdense regions, outflows of matter will produce deeper density contrasts in redshift space, whereas in overdense regions, coherent infall of matter will tend to produce denser environment estimates.

On the right hand side of Fig. 5.3, we show the percentage of random points that belong to a given quintile in real and redshift space. When the density split is performed in redshift space, a substantial fraction of each quintile consists of misclassified points, which would have been part of a different quintile based on their true (real-space) density. This misclassification

[2] https://github.com/cosmodesi/pycorr

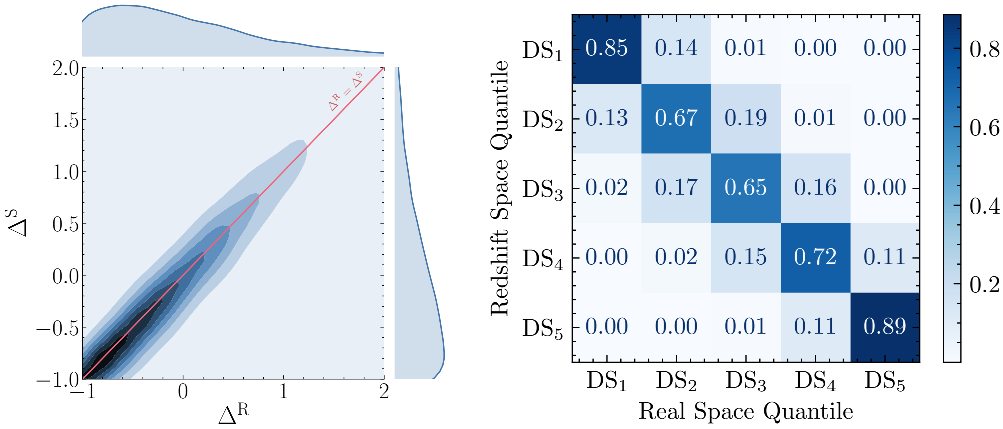


Figure 5.3: On the left, we show the joint distribution of overdensities, Δ, when identified either in real space Δ^{R}, or in redshift space, Δ^{S}. In underdense regions, redshift space densities tend to appear slightly more underdense whereas overdense regions also appear more overdense. On the right hand side, we show the percent of centres in real space that have been identified as split i but appear as split j in redshift space.

mostly shifts points from one quintile to its nearest neighbour(s), and larger shifts are rare. We will now focus on the effect that this has on the multipoles of autocorrelations and cross-correlations.

Fig. 5.4 shows the multipoles of DS cross-correlation ($\mathrm{DS}_i^{\mathrm{qh}}$) between points in a quintile and the halos' redshift space positions, and autocorrelation ($\mathrm{DS}_i^{\mathrm{qq}}$) functions of random points within the same quintile, when the overdensities are estimated from the real-space positions of halos (r-split) or from their redshift-space positions (z-split).

Quintile Autocorrelations

For the autocorrelations, shown on the right-hand side of Fig. 5.4, the monopole is very similar in both the identified real-space and redshift-space splits. In both cases, the largest signal is found for the overdense regions, DS$_5$, closely followed by the underdense regions, DS$_1$. Although DS$_1$, DS$_2$ and DS$_3$ are expected to have a negative tracer bias due to their underdense nature, all monopoles are positive since the bias enters squared in the mapping from matter to tracer autocorrelation functions, i.e. $\xi_{\mathrm{tracer}} = b^2 \xi_{\mathrm{matter}}$. Both DS$_1$ and DS$_5$ show a significant enhancement in clustering on a scale of approximately $100\,h^{-1}$Mpc corresponding to the acoustic scale set by the Baryon Acoustic Oscillations (BAO).

The quadrupole, on the other hand, is completely different for the real and redshift space identification scenarios. It is compatible with zero for splits identified in real space, whereas it is always negative for splits done in estimated redshift-space densities. In the r-split

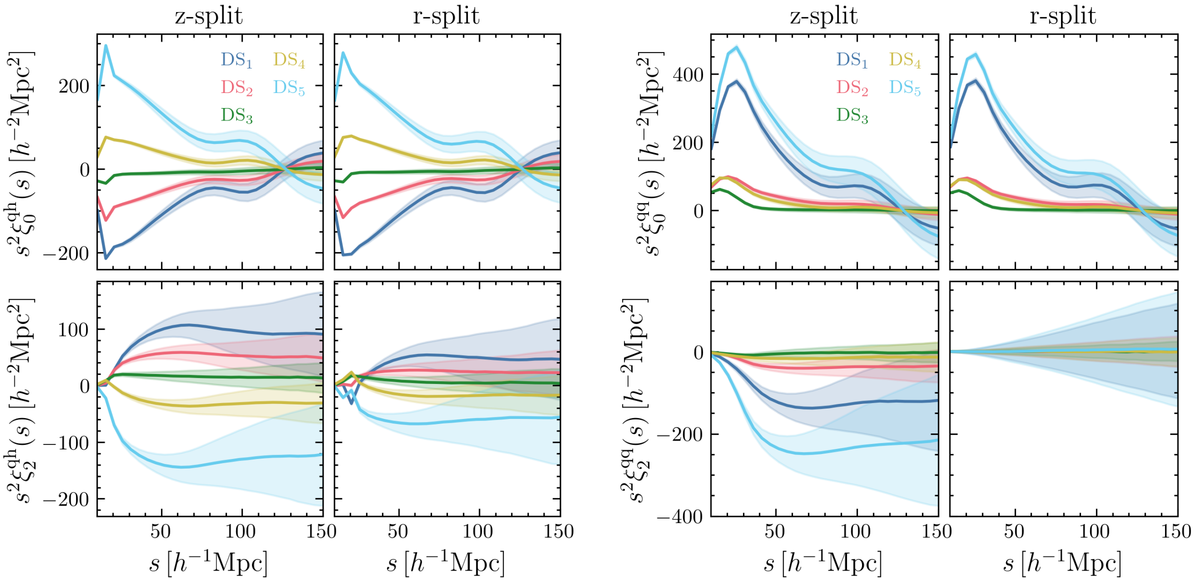


Figure 5.4: Multipoles of the DS-halo cross-correlation functions (left panel) and DS autocorrelation functions (right panel). The subpanels compare the cases when the quintiles are defined in redshift or real space (left and right sub-panels, respectively). Error bars represent the standard deviation associated to a $(1\,h^{-1}\mathrm{Gpc})^3$ volume, estimated from multiple mock realizations of the fiducial cosmology.

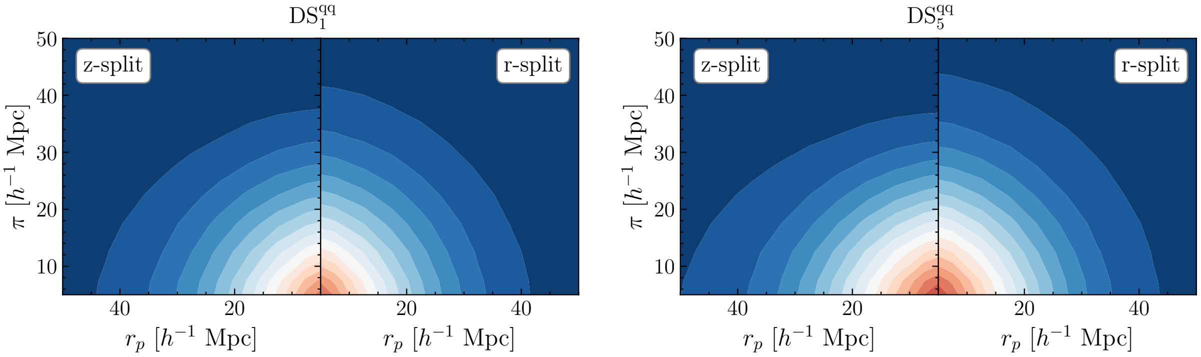


Figure 5.5: Two-dimensional auto-correlation functions for the two extreme density splits DS1 (left) and DS5 (right), when identified in real (r-split) and redshift space (z-split). Overall, the two point correlation functions appear squashed along the line of sight when the quintiles are identified in redshift space.

scenario, where density splits are performed in real space, there is no preferred direction, and so statistical isotropy dictates a quadrupole signal consistent with zero. When estimating densities in redshift space, peculiar velocities along the line of sight introduce a direction-dependent distortion to the estimated density field, which creates a redshift-space distortion (RSD) anisotropy in the distribution of the DS centres themselves. Generally, a non-linear transformation of a tracer density field performed in redshift space must have an additional RSD-induced bias (Seljak, 2012; Chuang et al., 2017).

This negative quadrupole corresponds to a squashing of the two-point autocorrelation functions along the line of sight, as can be seen in Fig. 5.5. The origin of this squashing is related to the misidentification of the random centres shown in Fig. 5.3. Let us imagine that we focus on a fixed position within the simulation. If the local density at this position is high, random points around this region will be classified as DS_5 in real space. In redshift space, while it is likely that these points will also be classified as DS_5, there is a chance that the local density environment will look less dense than in real space which could cause this point to fall within DS_4 instead. This can be seen on the right-hand side of Fig. 5.3, where approximately 11% of the points identified as DS_5 in real space have been identified as DS_4 in redshift space. This will most likely happen at the boundaries of the overdense regions due to coherent infall motions. On the other hand, the same percentage of random centres have been misidentified as DS_5 while belonging to DS_4 in real space. These are more likely to be found within overdense regions. Both effects contribute to squashing the autocorrelation function; in Appendix C.1, we explicitly show how these misidentifications contribute to the quadrupole by decomposing it into the contributions from the correctly identified and misidentified centres.

For the underdense regions, DS_1, Fig. 5.5 shows the same squashing pattern as in the overdense case. It might, however, seem counterintuitive if one thinks of the bulk motions of the centres themselves. The bulk motions of DS_1 centres can be characterised by their pairwise velocity, which we can estimate through the pair conservation equation (Peebles, 1980a; Sheth et al., 2001)

$$v_{12}(r) = -\frac{2}{3}\frac{\beta a H r \bar{\xi}(r)}{(1+\xi(r))} \tag{5.2.2}$$

where $\beta = \frac{f}{b}$ and $\bar{\xi}(r)$ is the spherically averaged quintile autocorrelation. For a negatively biased sample such as DS_1, the pairwise velocity will be positive. Therefore, DS_1 centres are, on average, moving away from each other.

The centres are, however, not moving from real to redshift space. Instead of moving the centres themselves, we sample a set of random points in either real and redshift space, and then classify these. When sampling the DS_1 centres, we still sample the inner regions of the voids that would be moved outwards in the picture of moving centres.

Quintile crosscorrelations with dark matter haloes

On the left-hand side of Fig. 5.4, we show the multipoles resulting from cross-correlating the random centres in each quantile with the halos' redshift space positions. On the left column, we show the cross-correlation with centres identified in redshift space, whilst on the right we show the same cross-correlation when centres are identified in real space. In both cases, the halo positions are in redshift space.

The monopole, which appears to be largely unaffected by the density split definition, shows a wide range in amplitudes at small scales, going from the most underdense regions in DS_1, having density contrasts close to -1, to the overdense environments of DS_5, which correspond to cluster-like environments with density contrasts around 2. These amplitudes also reflect the non-Gaussian nature of the density PDF: DS_1 regions are always constrained from below, as voids cannot be emptier than empty ($\delta = -1$). However, the densities in DS_5 can go well beyond 1, breaking the symmetry of the distribution. At large scales, the monopole moments slowly converge towards the mean density. In a Gaussian random field, the splits would be perfectly symmetric (i.e. $\mathrm{DS}_i^{\mathrm{qh}} = \mathrm{DS}_{6-i}^{\mathrm{qh}}$); deviations from it are a signature of non-Gaussianity in the density field, see Appendix C.3 for a comparison between the Quijote simulations and Gaussian random fields.

On large scales, around $100\,h^{-1}$Mpc, we can distinguish the signal coming from baryon acoustic oscillations for all density quintiles, both for the cross-correlation and autocorrelation functions.

Regarding the quadrupole moment of the cross-correlations, they show features that can be very different between the two identification scenarios. On large scales, where the two cases behave qualitatively similarly, we see positive amplitudes in DS_1, DS_2 and DS_3, while negative amplitudes are observed in DS_4 and DS_5. According to our convention for the redshift-space multipoles (Eq. 5.2.1), a negative (positive) quadrupole for overdensities (underdensities) means that the distribution of haloes around these quantiles appears to be flattened along the line of sight. We also observe that the amplitudes of the quadrupoles for DS_1 and DS_5 are larger in z-split than in r-split. This is again a consequence of the misidentification of quintiles and the additional anisotropy that the redshift-space definition of quintiles introduces.

For the redshift-space identification scenario, the quadrupoles maintain their sign across the whole scale range. However, for the real-space identification, we see an abrupt change from positive to negative amplitudes for DS1. This transition, which translates to an apparent elongation of the underdensities along the line of sight, has also been observed in the void-galaxy cross-correlation function (Nadathur et al., 2020; Woodfinden et al., 2022), and can be driven by the coherent outflow of galaxies from voids (see Cai et al., 2016; Nadathur & Percival, 2019, for a more in-depth discussion about the physical interpretation of this feature).

5.2.2 Reconstructing real-space positions

Nadathur et al. (2019) proposed to detect voids after reconstructing the approximate real-space galaxy positions by removing the effects of large-scale velocity flows from the redshift-space positions. The reconstruction algorithm is similar to that used in Baryon Acoustic Oscillation (BAO) analyses (Padmanabhan et al., 2012; Bautista et al., 2018; Chen et al., 2022), but is used only to remove the RSD. This is motivated by the theoretical challenges that arise from modelling the clustering around cosmic voids when these are identified from redshift-space galaxy catalogues. By using a density-field reconstruction algorithm, they were able to move galaxies back to their approximate real-space positions, which can then be used to identify voids. Here, we use the same method to remove RSD from the redshift-space Quijote halo catalogues and then identify the DS quintiles in the reconstructed catalogues.

Let us place ourselves in a Lagrangian framework, in which the Eulerian position $\vec{x}$ at time t can be described in terms of the initial Lagrangian position $\vec{q}$ and a non-linear displacement field $\vec{\Psi}(\vec{q}, t)$:

$$\vec{x}(\vec{q}, t) = \vec{q} + \vec{\Psi}(\vec{q}, t)\,. \tag{5.2.3}$$

The halo overdensity field $\delta_h(\vec{x}, t)$, can be related to the displacement field by (Nusser & Davis, 1994)

$$\nabla \cdot \vec{\Psi} + \frac{f}{b}\nabla \cdot (\vec{\Psi} \cdot \hat{r})\hat{r} = -\frac{\delta_h}{b}\,, \tag{5.2.4}$$

where b is the linear bias of the halo sample. The full solution to Eq. 5.2.4 includes contributions to the velocity flow coming from galaxy peculiar velocities at the corresponding redshift, as well as additional non-linear evolution that can be traced back to earlier epochs. In BAO analyses (e.g. Alam et al., 2017), in an attempt to undo all effects of non-linear clustering to sharpen the BAO feature to the best extent possible, galaxy or halo positions are shifted by $-\vec{\Psi}$ using the full displacement field. In our analysis, we are only concerned with removing the RSD coming from halo peculiar velocities at a certain epoch, so the part of the solution we are interested in is

$$\vec{\Psi}_{\mathrm{RSD}} = -f(\vec{\Psi} \cdot \hat{r})\hat{r}\,. \tag{5.2.5}$$

Shifting the redshift-space halo positions by $-\vec{\Psi}_{\mathrm{RSD}}$, we obtain a pseudo real-space halo catalogue that can be used to define the DS quintiles.

Several reconstruction implementations have been introduced in the literature. Here, we use the Iterative FFT Particle Reconstruction code implemented in PYRECON[3], which solves Eq. 5.2.4 by using an iterative fast Fourier transform procedure (Burden et al., 2015). This is the same algorithm that was applied to reconstruct the galaxy field in the eBOSS cosmological analysis (Bautista et al., 2018). Eq. 5.2.4 shows that reconstruction is sensitive to the ratio of the linear growth rate of structure f and the linear bias parameter b. We estimate the value of f from the cosmology of the fiducial Quijote simulation as $f = \Omega_m(z)^{0.55} = 0.532$. We estimate the linear halo bias taking the square root of the ratio between the halo and the matter power spectrum, which yields a value of $b = 1.7$ on large scales. The FFT procedure operates on the density field on a regular grid, which we set to have a size of 512^3. The density field δ_h is smoothed with a Gaussian kernel of width R_s^{recon} to reduce the sensitivity to small-scale density modes, for which Eq. 5.2.4 becomes inaccurate. We adopt $R_s^{\mathrm{recon}} = 10\,h^{-1}\mathrm{Mpc}$, in line with Nadathur et al. (2020) for easier comparison.

We show the multipoles obtained when splitting the density field using the reconstructed real-space positions of halos (recon-split) in Fig. 5.6, where we also compare against the real-space identification scenario (r-split). Qualitatively, we find that the recon-split multipoles closely follow the key features observed in the r-split multipoles: i) the quadrupole of the autocorrelation functions being consistent with zero, ii) the smaller amplitudes of the cross-correlation functions' quadrupole with respect to the z-split case, and iii) the transition from a positive to negative quadrupole for the DS_1 cross-correlation function. Although the recon-split monopole is always within 1-σ of the r-split monopole, both for auto and cross-correlation functions, the recon-split quadrupole of the two most extreme quintiles (DS_1 and DS_5) is biased with respect to the real-space identification at scales below $\sim 30\,h^{-1}\mathrm{Mpc}$, close to the smoothing scale. In the next section, we show that this offset in the quadrupole signal can potentially lead to biased constraints of the cosmological parameters if small scales are included in the analysis.

[3] https://github.com/cosmodesi/pyrecon

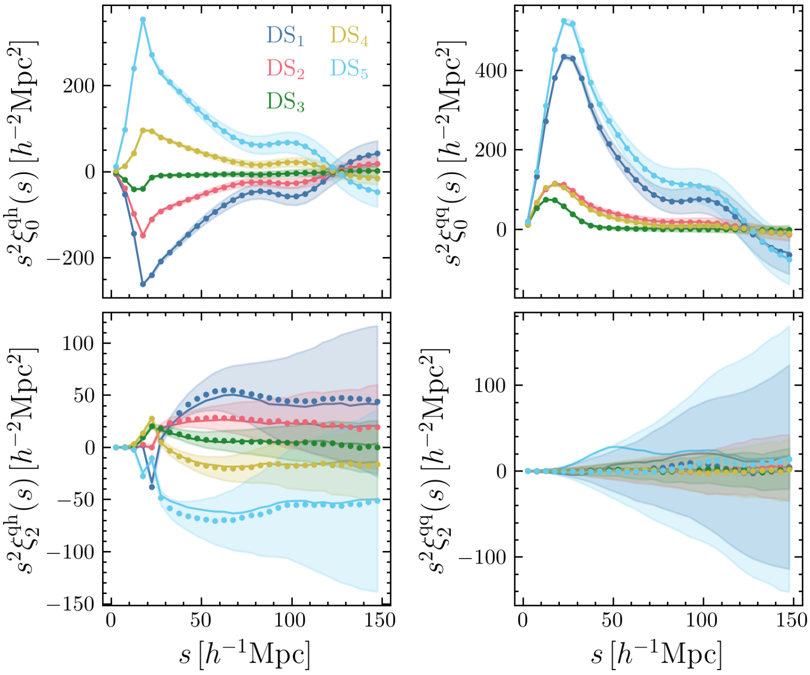


Figure 5.6: Comparison of multipoles when the densities are identified in either real (dots) or reconstructed halo positions (lines). Error bars represent the standard deviation associated to a $(1\,h^{-1}{\rm Gpc})^3$ volume, estimated from multiple mock realizations of the fiducial cosmology.

5.3 Fisher formalism

We quantify the information content of the summary statistics using the Fisher formalism (Fisher, 1935; Tegmark et al., 1997; Tegmark, 1997) described in Section 5.3.

In particular, if the likelihood follows a multivariate Gaussian distribution, we can compute the expectation value in the calculation of the Fisher matrix (Eq. 5.3.2) analytically, finding

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{Tr}\left[C^{-1}\frac{\partial C}{\partial\theta_i}C^{-1}\frac{\partial C}{\partial\theta_j} + C^{-1}\left(\frac{\partial s}{\partial\theta_i}\frac{\partial s}{\partial\theta_j}^{\top} + \frac{\partial s}{\partial\theta_i}^{\top}\frac{\partial s}{\partial\theta_j}\right)\right], \tag{5.3.1}$$

where C is the covariance matrix associated with the data vector $\boldsymbol{s}$. As shown in (Carron, J., 2013), the first term in Eq. 5.3.1 artificially adds information that was already included in the second term through the derivative of the mean vector. In what follows, we neglect this term to rather produce a conservative estimate of the information content and compute the Fisher matrix as

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = \frac{\partial s}{\partial\theta_i}\mathbf{C}^{-1}\frac{\partial s}{\partial\theta_j}^{\top}. \tag{5.3.2}$$

In Appendix C.2, we show that the likelihood for DS statistics is indeed very close to a multivariate Gaussian. Non-Gaussianities in the likelihood could lead to artificially tight bounds on the cosmological parameters when using the Fisher matrix formalism described by Eq. 5.3.2 (Park et al., 2022).

For most of the cosmological parameters, the derivatives can be numerically approximated as

$$\frac{\partial s}{\partial \theta} \simeq \frac{\boldsymbol{s}(\theta + \mathrm{d}\theta) - \boldsymbol{s}(\theta - \mathrm{d}\theta)}{2\mathrm{d}\theta} \, . \tag{5.3.3}$$

Eq. 5.3.3 cannot be used to estimate derivatives with respect to M_ν, as the neutrino mass cannot be negative. In that case, we instead approximate it as follows:

$$\frac{\partial s}{\partial M_\nu} \simeq \frac{s(4\mathrm{d}M_\nu) - 12s(2\mathrm{d}M_\nu) + 32s(\mathrm{d}M_\nu) - 21s(M_\nu = 0)}{12\mathrm{d}M_\nu} \, . \tag{5.3.4}$$

While the initial conditions for most simulations in our sample were generated using 2LPT, the simulations with non-zero neutrino mass were initialised using the Zeldovich approximation (ZA). For a consistent estimation of the derivatives, the $\mathrm{M}_\nu = 0$ data vector in Eq. 5.3.4 is measured from simulations of the fiducial cosmology that were also run with ZA initial conditions.

We calculate derivatives of the redshift-space 2PCF and DS multipoles on each of the 500 realisations of the paired simulations along three different lines of sight (taken to be the x, y and z axes of the simulations), which effectively gives us 1500 realisations over which we take the average (Smith et al., 2020). Fig. 5.7 shows an example of these derivatives for the matter density parameter, Ω_m. Each quintile shows a distinct sensitivity to Ω_m as a function of scale. The largest contribution comes from small scales, where we expect the density field to deviate the most from a Gaussian distribution. The auto- and cross-correlation functions also show different scale dependencies, which, as we will corroborate later, highlight the importance of combining these two sets of statistics to maximise the cosmological constraining power.

We estimate the covariance matrix from the multiple realisations of the fiducial cosmology as

$$\mathbf{C} = \frac{1}{n_{\mathrm{mocks}} - 1} \sum_{k=1}^{n_{\mathrm{mocks}}} \left(\boldsymbol{s}_k - \overline{\boldsymbol{s}}\right)\left(\boldsymbol{s}_k - \overline{\boldsymbol{s}}\right) \, , \tag{5.3.5}$$

where $n_{\mathrm{mocks}} = 7000$ and $\overline{\boldsymbol{s}}$ is the mean data vector averaged over all the realisations. In Appendix C.4 we show that the inferred errors on the parameters converge when using these numbers of realisations for the calculation of the derivatives and covariance.

To obtain the parameter constraints, two matrix inversions must be performed: the inversion of the covariance matrix in Eq. 5.3.2, and that of the Fisher matrix in Eq. 2.1.6. Although the estimator of the covariance matrix (Eq. 5.3.5) is unbiased, these two inversions

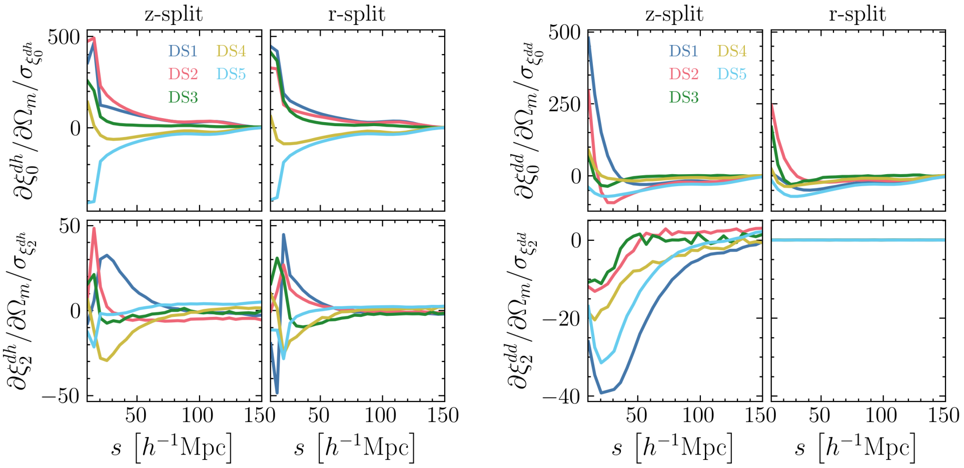


Figure 5.7: (left) derivatives of the DS-halo cross-correlation multipoles with respect to Ω_m, expressed in units of the variance of the multipoles. The upper and lower rows in each panel show derivatives of the monopole and quadrupole moments, respectively, while the left and right columns compare results when the quintiles are defined in redshift or real space. (right) same as the other panel, but showing the DS autocorrelation functions.

lead to biased constraints on the parameters. To account for this, we apply the Hartlap (Hartlap et al., 2007) correction to the covariance matrix

$$\hat{\mathbf{C}}^{-1} = \frac{N_{\mathrm{mocks}} - N_{\mathrm{bins}} - 2}{N_{\mathrm{mocks}} - 1}\mathbf{C}^{-1} \tag{5.3.6}$$

where N_{mocks} is the number of mocks used to estimate the covariance and N_{bins} is the number of bins of the data vector.

Fig. 5.8 shows the correlation matrix associated with this covariance for the DS and 2PCF data vectors. For DS, the covariance includes contributions from the monopole and quadrupole moments of the auto and cross-correlation functions for each for the DS quintiles. Since we use 30 radial bins in the range $10 < s < 150h^{-1}$Mpc, this results in a 600×600 matrix. For the 2PCF, we have a 60×60 matrix resulting from the contributions from the monopole and quadrupole.

5.4 Information content of density-split clustering

5.4.1 Identifying environments

The first step of the DS algorithm described in Sect. 5.2 consists of estimating the halo density in spheres of radius R_s centred around random points, which is then used to calculate the density PDF and define the DS quantiles. The density PDF itself depends on cosmology,

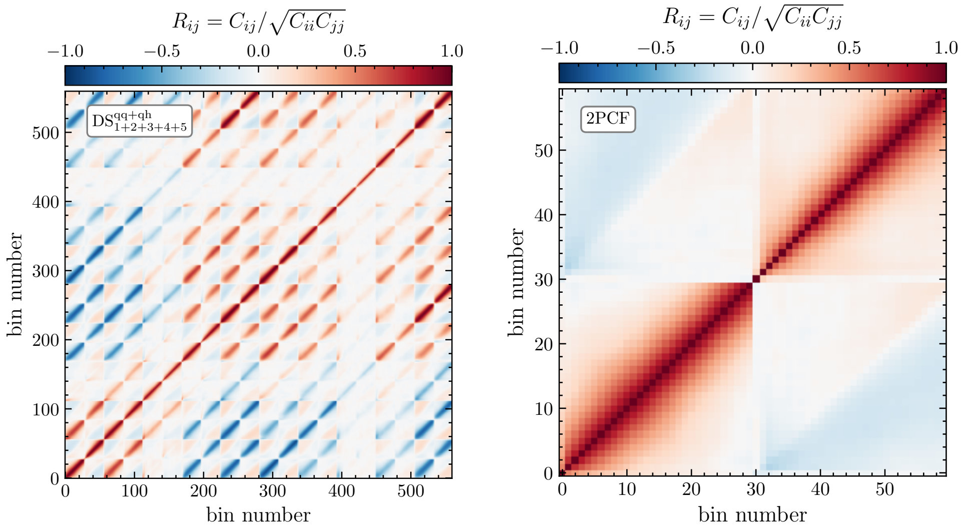


Figure 5.8: Correlation matrices of the DS and 2PCF data vectors, which include contributions from the monopole and quadrupole moments of the redshift-space correlation functions.

which is the main source of information used in methods such as counts-in-cells statistics (Uhlemann et al., 2020). We also expect DS to be sensitive to this information, as any changes in the density PDF will translate into changes in the average density in each quintile, which then propagates into changes in the observed multipoles.

Fig. 5.9 illustrates this by showing how the average density per quintile responds to changes in the cosmological parameters. Increasing Ω_m makes DS_1, DS_2, DS_3, and DS_4 denser, while the opposite occurs for DS_5. On one hand, given that we have fixed the minimum halo mass, increasing Ω_m will increase the number of halos above this threshold. For the densest quintile, DS_5, the increased merger rate could reduce the number of halos in a given sphere. On the other hand, when all other parameters are kept fixed, the effect of raising Ω_m is to reduce the amplitude of the galaxy or halo power spectrum (Kobayashi et al., 2020) by reducing the halo bias with respect to the underlying matter distribution, which brings the density of the quantiles slightly closer to the cosmic average. Changing σ_8 produces a similar effect on the quintiles, which is again related to an increase in the number of halos above the mass threshold and a reduced halo bias for larger σ_8 values (see Fig. C.6 in the Appendix).

The effect of varying the neutrino mass goes in the opposite direction. Having a non-zero neutrino mass lowers the density from DS_1 to DS_3 and boosts the density in DS_5. This effect is very similar to that of decreasing Ω_m, since increasing the mass of neutrinos reduces the amount of cold dark matter. This is consistent with the picture that neutrinos, which do not cluster below their free-streaming scale, reduce the growth of cold dark matter perturbations.

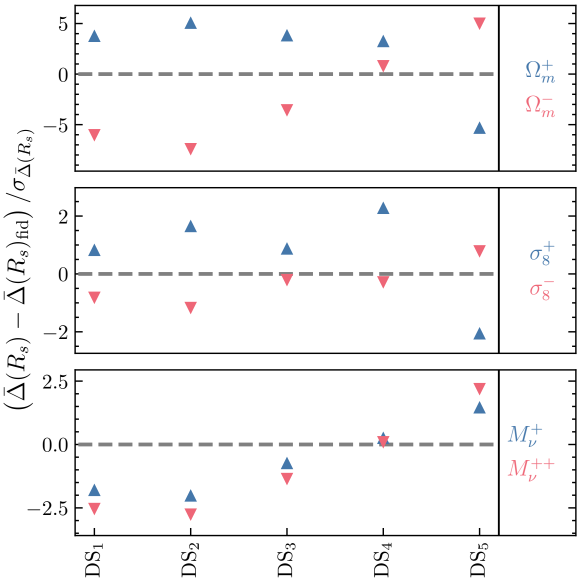


Figure 5.9: Response of the average density of each DS quintile to changes in cosmology. The vertical axis shows the difference in average density that is produced when we change Ω_m, σ_8 or M_ν, in units of the 1-σ errors of the density. The horizontal axis shows results for each quintile separately.

Although massive haloes can still form in the peaks of the density field and be resolved in Quijote, haloes forming in shallower regions of the density field will not reach masses above our selection threshold. The overall effect is an increased halo bias with respect to the fiducial case with $M_\nu = 0$ (Kreisch et al., 2019), which in turn makes the voids emptier and the clusters denser. This can be corroborated by looking at how the increase in neutrino mass increases the amplitude of the halo-halo power spectrum in Fig. C.6.

5.4.2 Comparing the information content of density-split RSD to two point statistics

In this section, we present the constraints obtained on the cosmological parameters through Eq. 2.1.6 and Eq. 5.3.2. Unless stated otherwise, the DS constraints we show correspond to the z-split scenario, i.e., when density quantiles are defined in terms of the redshift-space overdensities.

Modelling either the real-space or redshift-space-identified quintiles analytically would be challenging. In fact, previous studies (Paillas et al., 2021) have only modelled the real-to-redshift mapping. However, the Fisher formalism allows us to estimate the entire information content from direct measurements in N-body simulations.

In Fig. 5.10 we compare the constraints obtained by combining the DS auto and cross-correlation functions of all quintiles, $\mathrm{DS}^{\rm qq+qh}_{1+2+3+4+5}$, against the halo 2PCF, using multipoles within the scale range $10 < s < 150\,h^{-1}$Mpc. We limit the measurements to scales larger than $10\,h^{-1}$Mpc since we are only analysing central halos, whose behaviour is very different from that of galaxies on small scales, and because on these scales the effects of baryonic physics would be negligible.

Fig. 5.10 shows how DS can break some key parameter degeneracies that result when analysing two-point statistics, such as the one between Ω_m and σ_8, or that of n_s and σ_8. In particular, when we combine the information from all quintiles, the degeneracy between M_ν and the other parameters is significantly reduced. The standard halo 2PCF suffers from the well-known degeneracy found between σ_8 and M_ν, which limits its constraining power. Although the individual quintiles $\mathrm{DS}_1^{\mathrm{qq+qh}}$ and $\mathrm{DS}_5^{\mathrm{qq+qh}}$ also exhibit this degeneracy to some extent, the combined DS dataset is able to reduce it due to the different sensitivity of each density environment to these parameters. Overall, $\mathrm{DS}_{1+2+3+4+5}^{\mathrm{qq+qh}}$ increases the constraining power with respect to the halo 2PCF by a factor of approximately ×5, ×8, ×3, ×4, ×6, and ×6 for Ω_m, M_ν, Ω_b, h, n_s, and σ_8, respectively.

In Fig. 5.11 we show the individual contribution of each quintile to the parameter constraints. Interestingly, we find that DS_1 produces the weakest constraints for the sum of neturino masses after maginilizing over all other parameters. On the other hand, it produces the tightest unmarginalized constraints. One expects underdense regions to be more senstive to the properties of neutrinos, since their free streaming motions imply that the ratio of neutrino density to that of dark matter is higher in void regions than in overdensities.

Moreover, most quintiles individually produce tighter constraints than the 2PCF, except DS_3 and Ω_m. We show the equivalent of Fig. 5.11 for quintiles identified in real space in Fig. C.7.

Fig. 5.12 compares the information content of density-split clustering when the overdensities are identified in redshift (z-split) or real space (r-split). The combined constraints on the cosmological parameters are shown in Table 5.2. The real space identification of quintiles consistently produces better parameter constraints, especially for the parameters Ω_m and σ_8. When quintiles are identified in redshift space, some cosmological information is lost by the blurring of the density-split quintiles.

However, additional information is obtained through autocorrelations when these are identified in redshift space. This can be seen in Fig. 5.12; while the additional information contained in the autocorrelations is small for the r-split scenario, it has a large impact in improving the constraints for density split centres identified in redshift space. This additional

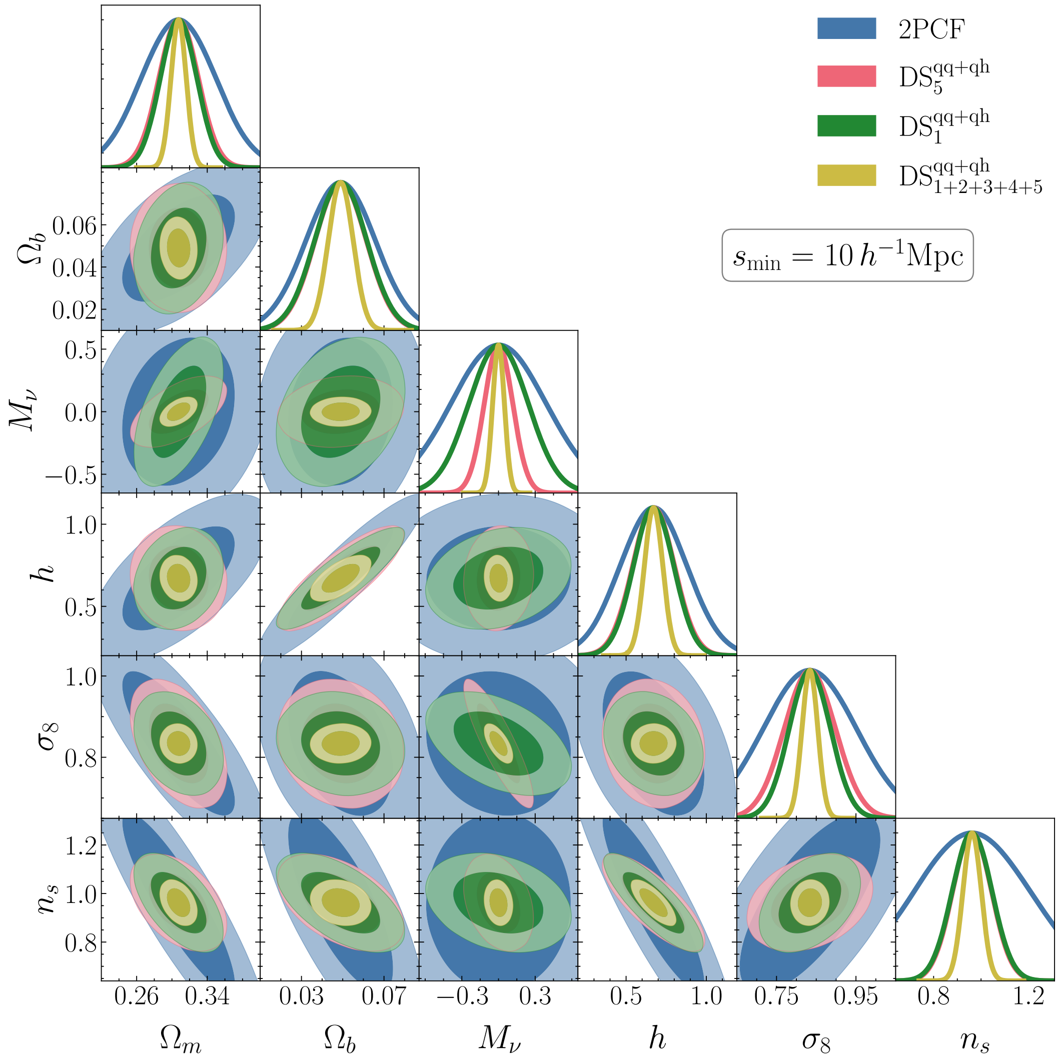


Figure 5.10: Fisher forecasts for constraints on the $\nu\Lambda$CDM model parameters from the 2PCF multipoles (blue) and density-split clustering using only voids (DS_1, red), clusters (DS_5, green) or the combination of all quintiles (yellow).

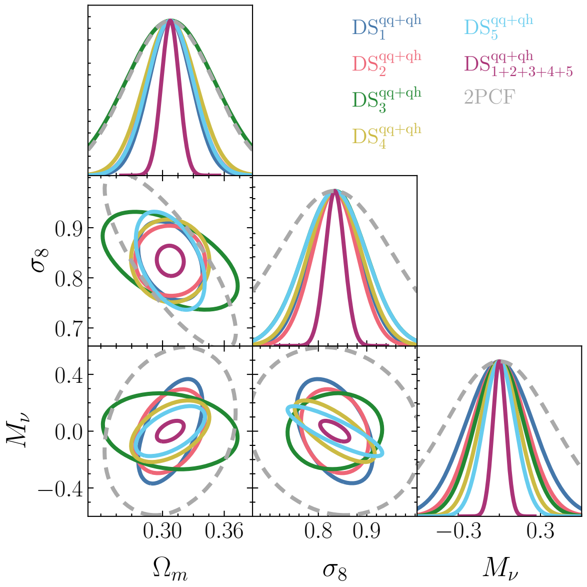


Figure 5.11: Constraints on the cosmological parameters from individual and combined DS quintiles identified in redshift space (solid). The constraints from the two-point correlation function are shown by the grey, dashed contours for comparison.

information comes mainly from the quadrupole of quintile autocorrelations, given that the monopole changes very little between r-split and z-split.

Finally, Fig. 5.13 shows the resulting constraints on each parameter as a function of the minimum scale $s_{\min}$ used in the analysis. It demonstrates how, even on large scales (i.e. BAO peak), splitting the density field into different quantiles allows us to extract more information on the cosmological parameters than the two-point correlation function. This would happen even in the case of a Gaussian random field, if one compares the information content above a minimum scale $s_{\min}$. In Appendix C.3, we show a qualitative explanation for the fact that the additional information of denstiy split statistics comes from the environment definition, which uses information from small scales that have been removed from the two-point correlation function when comparing their information content.

5.4.3 Biases introduced by reconstruction errors

The goal of this section is to determine whether reconstruction algorithms introduce a bias in the cosmological parameters estimated by density split measurements if we were to model them assuming that the post-reconstruction results can recover the real-space data vectors exactly. The reconstruction algorithm has been described in Sect. 5.2.2.

We estimate the bias in the inferred cosmological parameters introduced by an inaccurate reconstruction algorithm using the Fisher matrix (Huterer & Takada, 2005)

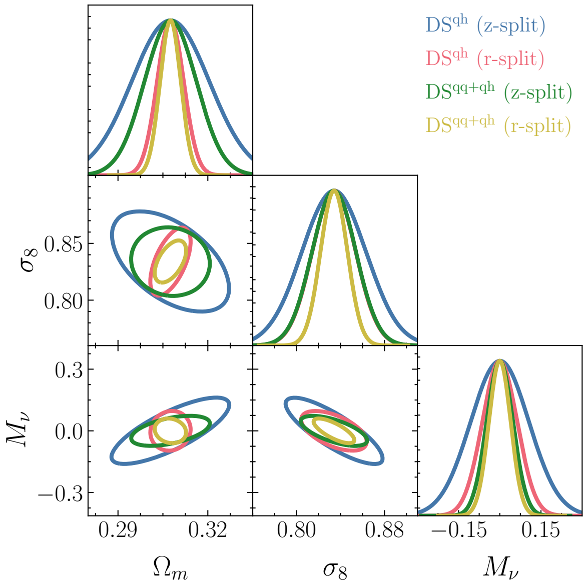


Figure 5.12: We compare the constraints obtained through cross-correlations of the density split centres and the entire halo field, DS$^{\rm qh}$, to those obtained from the combination of cross-correlations and auto-correlations of the density split centres, DS$^{\rm qq+qh}$. We show both results for density split centres identified in real space (r-split), and density split centres identified in redshift space (z-split). This figure demonstrates that quintile autocorrelations, DS$^{\rm qq}$, have a bigger impact in redshift identified quintiles than they do in real identified ones.

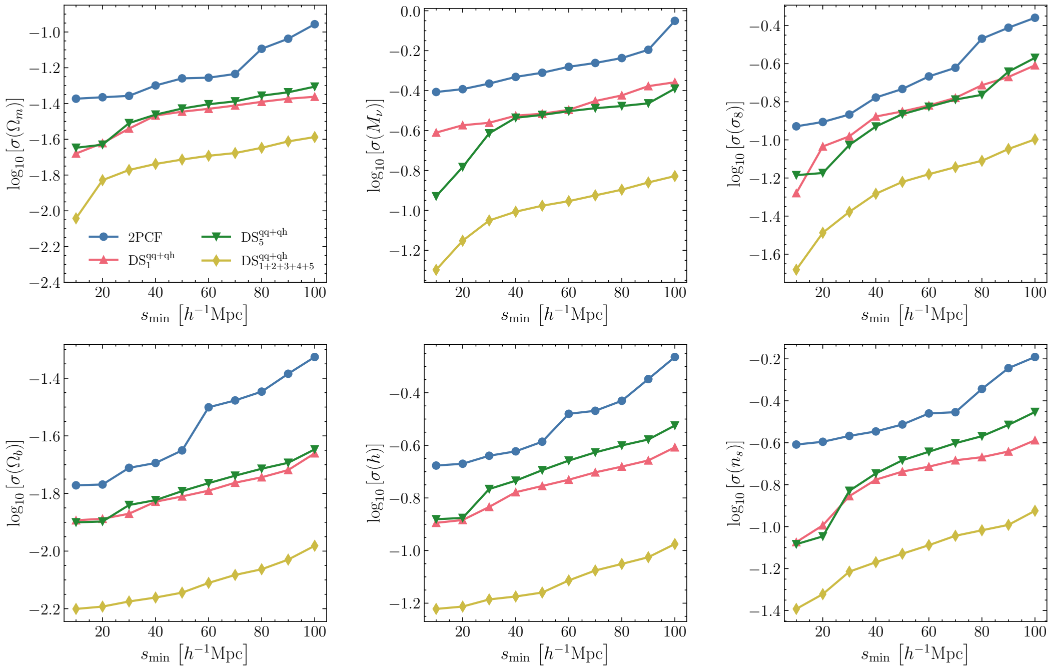


Figure 5.13: Constraints on the cosmological parameters from DS and the 2PCF, as a function of the minimum scale used to calculate the Fisher matrix. We also include the individual constraints obtained through the two extreme quintiles, DS1 and DS5.

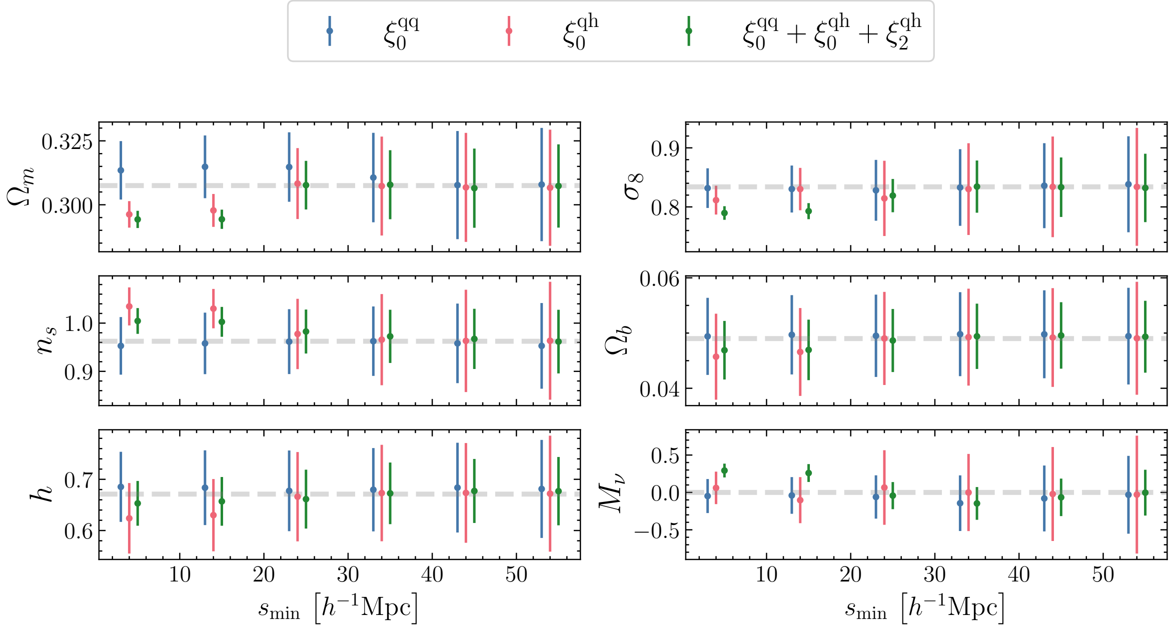


Figure 5.14: Bias in the cosmological parameters introduced by systematic errors caused by reconstructing the halo's real space positions, computed using Eq. 5.4.1. The true value of the parameters is shown on a gray dashed line. We show the bias introduced by each of the statistics used to infer the cosmological parameters: i) ξ_0^{qq}, the monopole of the quintile autocorrelation, ii) ξ_0^{qh}, the monopole of the cross-correlations between quintiles and halos, and iii) $\xi_0^{\mathrm{qq}}+\xi_0^{\mathrm{qh}}+\xi_2^{\mathrm{qh}}$, the combination of all the above with the quadrupole of the cross-correlations between quintiles and halos.

$$\begin{aligned}\delta\theta_\alpha &= \langle \theta^{\mathrm{recon}} \rangle - \langle \theta^{\mathrm{r-split}} \rangle \\ &= \sum_\beta \mathcal{F}^{-1}_{\alpha\beta} \sum_{ij} \left[s_i^{\mathrm{recon}} - s_i^{\mathrm{r-split}} \right] C^{-1}_{ij} \frac{\partial s_j^{\mathrm{r-split}}}{\partial \theta_\beta}\end{aligned} \tag{5.4.1}$$

where s^{recon} is the data vector obtained using the reconstructed halo positions and $s^{\mathrm{r-split}}$ is obtained through the true real space positions of the halos. We note that the bias we quantify here is associated with the statistical errors of a $(1\,h^{-1}\mathrm{Gpc})^3$ volume.

In Fig. 5.14, we show the potential biases in the estimated cosmological parameters caused by systematic errors in reconstructing the halo's real space positions, as a function of the minimum scale considered in the analysis. M$_\nu$, Ω_m and σ_8 are the parameters that are most affected by errors in the reconstructed halo positions. In particular, biases are found when including the monopole and quadrupole of cross-correlations between quintiles and the halo field, $\xi_{0,2}^{\mathrm{dh}}$ on scales smaller than the smoothing radius. In Fig. 5.6, we have shown that the errors introduced by reconstruction mostly affect the quadrupole of cross-correlations. Using only the monopole of quintile autocorrelations, ξ_0^{dd}, one can obtain unbiased constraints on the cosmological parameters using the full range of scales. However, the constraining power of

Statistic	Scales	Redshifts	Ω_m	M_ν	Ω_b	h	n_s	σ_8	Reference
$\mathrm{DS}^{\mathrm{qq+qh}}_{1+2+3+4+5}$ (z-split)	$10 < r < 150$	$z = 0$	±0.0087	±0.0484	±0.0060	±0.0576	±0.0389	±0.0200	This work
$\mathrm{DS}^{\mathrm{qq+qh}}_{1+2+3+4+5}$ (r-split)	$10 < r < 150$	$z = 0$	±0.0033	±0.0396	±0.0054	±0.0445	±0.0280	±0.0119	This work
Halo 2PCF	$10 < r < 150$	$z = 0$	±0.0422	±0.3907	±0.0169	±0.2099	±0.2456	±0.1175	This work
$\mathrm{B}_0(k)$	$k < 0.5$	$z = 0$	±0.011	±0.054	±0.004	±0.039	±0.034	±0.014	Hahn et al. (2020a)
kNN	$10 < r < 40$	$z = 0, 0.5$	±0.0111	±0.0925	±0.0029	±0.0273	±0.0206	±0.0108	Banerjee & Abel (2021)
MST(d,l,b,s)	$k < 0.5$	$z = 0$	±0.036	±0.23	±0.0083	±0.073	±0.065	±0.067	Naidoo et al. (2022)

Table 5.2: Comparison to Fisher forecasts for different summary statistics also based on the halo field.

autocorrelations on small scales is smaller than that of cross-correlations with the halo field, and therefore we would lose more information than if we were to estimate the overdensity around random centres directly in redshift space.

We note that the results presented in this section apply to a particular choice of reconstruction algorithm, which has been described in Sect. 5.2.2. Other algorithms (e.g., White, 2015; Wang et al., 2020) may lead to different constraints on the parameters, although a thorough comparison of different reconstruction techniques is beyond the scope of this manuscript.

As described in Sect. 5.2.2, reconstruction also smooths the density field below a given scale, which is a free parameter in the algorithm. In our analysis, this scale was set to $R_s^{\mathrm{recon}} = 10\,h^{-1}\mathrm{Mpc}$. We do not expect reconstruction to work below $R_s^{\mathrm{recon}} = 10\,h^{-1}\mathrm{Mpc}$, where the clustering information has been washed out, and consequently, the removal of RSD may be inaccurate. Future surveys, such as DESI-BGS (Zarrouk et al., 2022), are expected to reach much higher tracer number densities than those probed by Quijote, and the range of scales at which reconstruction is reliable may differ. We plan to study this in further detail in future work.

5.5 Discussion and conclusions

In this work, we have studied the cosmological information of density-split clustering (DS, Paillas et al., 2021) in the context of the $\nu\Lambda$CDM model. This method consists in characterising the clustering of biased tracers as a function of environmental density, exploiting the sensitivity of each environment (density quintiles) to the cosmological parameters. The density field at small scales is highly non-Gaussian due to nonlinear gravitational evolution, and therefore the power spectrum or the two-point correlation function (2PCF), which are measures of the variance of the density field, become incomplete descriptions of the galaxy distribution. DS is able to capture the missing information through a collection of correlation functions that are conditioned on environmental density, which naturally captures the non-Gaussian nature of the PDF.

We quantify the information content of DS through the Fisher matrix, estimated numerically from the halo catalogues of the Quijote suite of simulations (Villaescusa-Navarro et al., 2020). We have found that DS improves the constraints on each cosmological parameter by factors between 3 and 8, when compared to the standard halo two-point correlation function.

In Paillas et al. (2021), it was already shown that the cross-correlations between galaxies and DS quintiles could improve the constraints on the growth rate of structure by 30 per cent over the 2PCF function analysis if the Gaussian streaming model (Peebles, 1980b; Fisher, 1995b) was used to model the real to redshift space mapping. However, the analytical model presented in Paillas et al. (2021) relied on measurements of cross-correlation functions of real space galaxy catalogues from ΛCDM simulations, and their cosmological dependence was ignored in the analysis. This limits the amount of information that can be extracted to that of the real-to-redshift-space mapping. Here, we have shown for the first time that if we can model the full cosmological dependence of DS using N-body simulations, we can obtain much tighter constraints.

Moreover, we have presented the autocorrelations of the DS quintiles for the first time and have shown that they are also a valuable source of cosmological information, in addition to the DS cross-correlation functions. In particular, the quintile autocorrelations can recover some of the cosmological information that is lost when performing the density split in redshift space. Introducing them in the likelihood analysis will therefore be useful to avoid the use of reconstruction techniques and to analyse directly the redshift space identified multipoles.

The Quijote simulations have allowed us to explore the sensitivity of DS clustering to different cosmological parameters, such as the sum of neutrino masses M_ν. The combination of all DS quintiles places a constraint of $\sigma_{\mathrm{M}_\nu} = 0.0483$ for a $(1\,h^{-1}\mathrm{Gpc})^3$ volume, assuming that we can model the DS multipoles down to a scale of $10\,h^{-1}\mathrm{Mpc}$, which results in a factor of 8 improvement respect to the two-point correlation function constraints. DS also improves constraints by factors of 5, 3, 4, 6, and 6 for Ω_m, Ω_b, h, n_s, and σ_8, respectively. Our constraints are conservative, since the number density of resolved dark matter halos in the Quijote simulations is much lower than that expected in future galaxy surveys.

Our results are in line with forecasts from other summary statistics that aim at extracting non-Gaussian information from density fields. A natural approach is to include higher-order correlation functions or polyspectra. Hahn et al. (2020a) found that the redshift-space halo bispectrum provides tighter constraints on the cosmological parameters of $\nu\Lambda$CDM, compared to the halo power spectrum. In particular, the bispectrum is five times better at constraining the sum of neutrino masses M_ν, assuming that the bispectrum can be modelled up to $k_{\mathrm{max}} = 0.5\,h/\mathrm{Mpc}$. Including even higher-order correlations might tighten the cosmological constraints; however, even the full hierarchy of polyspectra may fail to contain all statistical information; see Carron (2011) for an example using log-normal fields. Moreover, the signal-to-noise ratio of higher-order moments decreases with the order of the correlators, and the computational complexity of higher-order statistics increases with the order of function chosen. Therefore, it is important to develop alternative statistics to the hierarchy of moments.

Most alternative summary statistics exploit the environmental dependence of clustering, but differ on the particular definition of environment. Massara et al. (2022) showed that the marked power spectrum of the galaxy field can improve the constraints over the standard power spectrum by a factor of 3-6 for the $\nu\Lambda$CDM parameters. In their method, galaxies are weighted or "marked" with a function that depends on local density. Marks can be chosen so that low-density regions are upweighted, which increases the sensitivity of the clustering to certain regions of the parameter space. As opposed to density split, where random centres are used to estimate environment densities, marked correlations use the positions of tracers to determine environment densities, and therefore the tracer marked power spectra might not have access to lowest density regions in the matter field.

Uhlemann et al. (2020) showed that the one-point probability distribution function of counts-in-cells statistics provides particularly powerful constraints for Ω_m, σ_8 and M_ν. They highlight the importance of combining information from different redshift bins in order to maximise information gain, which is something we have not explored in this work but could potentially be promising for DS. Moreover, given the low number density of our halo catalogues, we have not explored the additional information that the PDF might bring to DS statistics. We plan to study how complementary these two statistics are in future work.

Banerjee & Abel (2021) used the k-nearest-neighbour (NN) distributions of haloes as a way to constrain cosmology. Validating their method with the Quijote halo catalogues, they found that the NN cumulative distribution functions improve the constraints on the cosmological parameters by roughly a factor of 4, using the scale range $10 < s < 40\,h^{-1}\mathrm{Mpc}$ and two redshift slices $z = 0, 0.5$.

Alternatively, one could also detect the positions in the cosmic web of tracers of different environments and use their statistics to constrain cosmology. For example, Kreisch et al. (2021) looked at the constraining power of cosmic void statistics, finding that the void size function, the void autocorrelation, and the void-halo cross-correlation functions provide tight constraints on M_ν on their own. Furthermore, Bonnaire et al. (2022) used the eigenvalues of the tidal tensor to segment the cosmic web into nodes, filaments, walls, and voids, and used them to compute their respective power spectra in real space. Here, we have shown that cross-correlations between the halo field and the different environments add additional cosmological information to that of the autocorrelations (see Fig. 5.12). Although the environment here is defined differently from Bonnaire et al. (2022), we expect that similar gains could be achieved through the introduction of cross-correlations using their environment definition. Moreover, Bonnaire et al. (2022) assumed that the real space positions of tracers were known when identifying environments, but did not analyse the impact that identifying environments in redshift space could have on the resulting cosmological constraints.

Table 5.2 summarises the constraining power of different summary statistics found using the dark matter halos of the Quijote suite of simulations. We do not include studies based on the dark matter field, since a one-to-one comparison would not be possible. It shows how DS can obtain state-of-the-art constraints on the cosmological parameters Ω_m, M_ν, and n_s while still obtaining competitive constraints on the remaining parameters. Rather than advocating for a particular summary statistic, we highlight the possibility of complementing these different probes, exploiting the degeneracy-breaking power that each of them has to offer.

We have shown that the DS clustering statistics depend on whether the density environments are defined in real or redshift space. Real-space identified quintiles yield better constraints for all cosmological parameters, in particular Ω_m and σ_8, and indeed in Paillas et al. (2021) it was shown that if one has access to the real-space galaxy positions to identify the quintiles in this way, it is possible to model the real to redshift space mapping of the DS cross-correlation functions analytically using the Gaussian streaming model down to $\sim 15\,\mathrm{h}^{-1}\mathrm{Mpc}$. However, galaxy catalogues in real space are not immediately available in observations, and one would have to rely on reconstruction algorithms to approximately remove RSD from galaxies (Nadathur et al., 2019). But, as shown in Sec. 5.4.3, reconstruction algorithms could potentially introduce systematic errors in the inferred cosmological parameters when including small-scale information, which would then need to be added to the total error budget.

When presenting the main cosmological constraints of our analysis, we have put aside the complications related to the theoretical modelling and implicitly assumed that we have access to a model that can perfectly match the measurements down to $10\,h^{-1}$Mpc. An analytical prediction of how the multipoles of DS statistics change with cosmology is a challenging task. We plan to work on a simulation-based model to allow for a comparison between simulations and data, which will be presented in future work. This framework could potentially allow us to directly emulate the redshift-space DS multipoles, without the need for reconstruction. Moreover, we have focused here on DS statistics for dark matter halos, but we will work on simulation-based models for the DS statistics of galaxies. We expect DS to set tight constraints on environment-based assembly bias (Xu et al., 2021).

We note that since the different samples obtained through DS are expected to share the same sample variance, they can also use sample variance cancellation techniques such as those proposed in McDonald & Seljak (2008) and Seljak (2008). In fact, part of the gain in S/N we obtained over the standard 2PCF analysis might be related to this effect. However, sample variance cancellation can only meaningfully contribute to the S/N if the shot noise contribution is small, which is not the case for the Quijote simulations. Nevertheless, DS might be a promising analysis technique to exploit sample variance cancellation in future high density sample like DESI-BGS.

It has also been shown that zero-biased tracers might be a promising way to achieve optimal constraints on primordial non-Gaussianity (Castorina et al., 2018). Since it is basically impossible to obtain zero biased tracers through colour or magnitude cuts, DS again might provide a useful tool for such studies.

Relativistic effects can only be analyzed in the cross-correlation of differently biased tracers with the signal itself being proportional to the difference in galaxy bias (Yoo, 2010; Bonvin & Durrer, 2011; Challinor & Lewis, 2011). DS might prove useful for such studies, given the large range in galaxy bias, accessible with this technique.

Ongoing and upcoming large-area surveys, such as DESI (DESI Collaboration et al., 2016a), Euclid (Laureijs et al., 2011), and Roman Space Telescope (Green et al., 2012), will offer unprecedented statistical precision for galaxy clustering. A vast amount of information from these Stage-IV experiments will be available in the mildly nonlinear regime, where the density field is non-Gaussian. Methods that can grant access to higher-order statistical information beyond two-point statistics, such as DS, will thus play a key role in extracting cosmological information that cannot be readily accessed with the power spectrum. This will require percent-level precision from the modelling side, and ensuring that the models can circumvent the observational systematic effects that will be inherent to these datasets.

Chapter 6

Computational methods across disciplines

So far we have shown examples of how machine learning, Bayesian statistics, and high performance computing can push the frontiers of what is known about the Universe and help us detect signatures of new physics.

Although these techniques have been shown to have successful application in astrophysics, they were not initially developed for this field. Warren S. McCulloch, a neuroscientist, and Walter Pitts, a logician, first proposed the computational model for a neural network. In McCulloch & Pitts (1943), the authors attempted to understand the functioning of the human brain and its ability to produce complex patterns by connecting basic cell units. Monte Carlo methods (Metropolis & Ulam, 1949), which make Bayesian inference for cosmology computationally feasible, were invented by Stanislaw Ulam and John von Neumann and developed in the area of nuclear physics.

Currently, the connection between different scientific fields is stronger than ever. On one hand, the development of sophisticated simulations capable of replicating complex real-world phenomena means that we need to develop tools to contrast simulations with data in an interpretable manner. On the other hand, the collection of large datasets whose patterns might contain the answers to the most pressing scientific questions calls for developments at the intersection of high-performance computing and machine learning. Examples of how such developments have led to exciting progress in science and engineering include predicting the outcome of protein folding (Jumper et al., 2021) and learning to represent languages for applications in natural language processing (Wolf et al., 2020). Past and future advances in cosmology are and will be the result of a collective endeavour that spans across disciplines.

During my Ph.D., I have had the opportunity to work on research in different areas of science, ranging from medical imaging and epidemiology to natural language processing. In this chapter, I will briefly describe the results in two of these areas: medical imaging and epidemiology. In both cases, we used similar statistical and computational techniques to those shown earlier in this Thesis.

6.1 XNet: A neural network for medical X-Ray imaging segmentation

X-Ray image enhancement, along with many other medical image processing applications, requires segmentation of images into bone, soft tissue, and open beam regions. In Fig. 6.1, we show an example input/output pair for image segmentation. The input is the greyscale X-Ray image of the body part, whilst the output is a segmented map in which each pixel is classified as belonging to either bone, soft tissue, or open beam regions (i.e., nothing lies in the path of the X-ray). The classic image processing methods (Pakin et al., 2003) developed to solve this problem rely on a complex system of classical image processing techniques and require tuning the hyperparameter of the models for each class of body parts.

In Bullock et al. (2019), we instead developed a machine learning approach that presents an end-to-end solution resulting in robust and efficient inference. Since medical institutions frequently do not have the resources to process and label the large number of X-ray images usually needed for neural network training, we designed an end-to-end solution for small datasets while achieving state-of-the-art results. Our dataset is composed of only 150 labelled X-ray images, compromising 19 body parts in an imbalanced way. Given that the data set size is small, we artificially augment the training images with the two-fold purpose of creating a larger dataset to avoid overfitting, and balancing the different body part classes through augmented oversampling.

We present a neural network architecture for X-ray image segmentation based on an encoder-decoder style architecture commonly used in image segmentation (Badrinarayanan et al., 2015). The different components of this are described below.

Encoder The encoder consists of a series of convolutional layers, for feature extraction, and max pooling layers to downsample the input image. Max pooling is a pooling operation that selects the maximum value in each patch of a feature map, keeping only the most salient features. Breaking up the downsampling into multiple stages allows for varying levels of

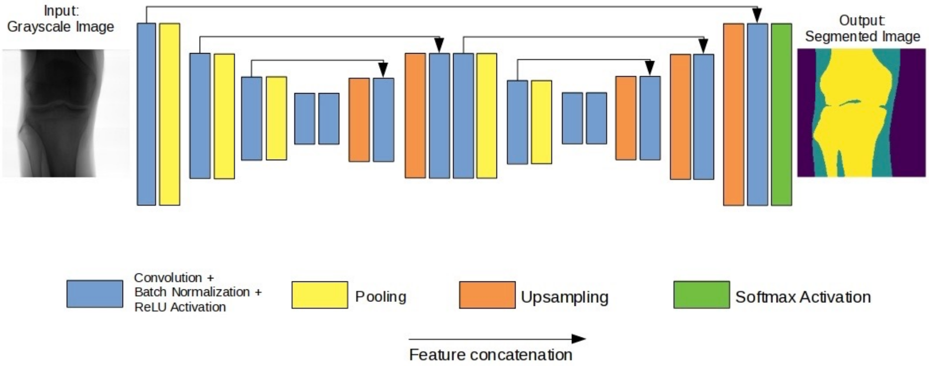


Figure 6.1: Visualisation of the XNet architecture including an example input image, left, and output segmented mask, right. Feature concatenation of same dimension layers helps to avoid losing fine-grained detail. Softmax activation function provides final pixel-wise classification.

feature extraction, with increasingly global features learnt through the convolutional layers at each pooling stage.

Decoder After feature extraction, the decoder performs upsampling (a transposed convolution) of the image that has been passed through pooling to generate a segmented mask of the same dimension as the input image. Similarly to the encoder, using a multistage upsampling process with convolutional layers in between allows for varying degrees of fine-grain feature reconstruction during upsampling, thus producing dense feature maps.

Due to the small size of our images, we aim to avoid large serial downsampling of the input image compared to many other networks, particularly those used in image classification. We avoid this, since performing a greater number of downsamplings in series can be detrimental to accurate boundary-level detail, particularly around smaller structures. However, downsampling allows us to learn features that are invariant to small distortions of the image, such as shifts, and is therefore important to include in the network.

We present an architecture which incorporates a comparable, or greater, number of downsampling stages for feature extraction as other segmentation networks, whilst avoiding overly reducing image resolution. This is achieved by using two encoder-decoder modules in succession, whilst storing encoder feature maps and using them during the creation of the dense feature maps in the decoders (as can be seen in Fig. 6.1).

Our implementation produces an overall accuracy of 92%, averaged across all pixels. The open beam region is both the most prevalent in the dataset and the easiest to classify. Therefore, the accuracy in all classes is not necessarily the most effective metric to measure the performance of the model. Alternatively, we use the F1 score (Taha & Hanbury, 2015) which is the harmonic mean of precision and recall, and therefore it is a better metric to describe the performance of a model on an unbalanced dataset. We find an F1 score of 0.92. These results surpass classical image processing techniques, such as clustering and entropy-based methods, while improving the output of existing neural networks used for segmentation in non-medical contexts.

6.2 Agent based simulations for epidemiology

The spread of SARS-CoV-2 in populations with little or no immunological resistance has caused considerable disruption to health care systems and a large number of fatalities around the globe since 2020. The evaluation of policies which aim to mitigate the impact of this and other epidemics on the health of individuals relies on a detailed understanding of the spread of the disease and requires both short term operational forecasts and longer term strategic resource planning.

There are various modelling approaches that aim to provide insight into the spread of an epidemic. They range from analytic models, formulated through differential equations, which reduce numerous aspects of the society–virus–disease interaction to a small set of parameters, to purely data-driven parameterisations which inherently rely on a probability density that has been fitted to the current and past state of the system in an often untraceable way. As a complement to analytic models, agent-based models (ABMs) (Bonabeau, 2002) focus on the interactions of individuals and groups of individuals in complex social networks. They are able to capture social mixing by modelling direct contacts between individuals belonging to different sub-populations, as well as the geographic and demographic heterogeneity of the populations.

In essence, ABMs can record transmission chains between individuals. Perhaps most importantly, for an epidemic, ABMs are also able to capture individual behavioural adjustments, which can change as agents interact with the larger complex system. Such models also provide the flexibility to experiment with different policies and practises based on realistic changes in the model structure, such as the inclusion of new treatments, changes in social behaviour, and restrictions on movement.

However, capturing the behaviour of individuals, their activities, and social networks requires much more detailed data inputs and greater computing power than analytical approaches. In addition, the complexity of ABMs, as well as the sometimes strong effects of stochasticity, can make the process of fitting models to make predictions more challenging.

We will show how a similar emulation process and Bayesian analysis developed for cosmology can be used in the context of epidemiological simulations for parameter inference and uncertainty quantification.

6.2.1 The JUNE model

We developed JUNE (Aylett-Bullock et al., 2021a), a generalisable modular framework to simulate the spread of infectious diseases using fine-grained geographic and demographic information, and with a strong focus on detailed simulation of policy interventions. Individuals in JUNE follow detailed spatio-temporal activity profiles that are informed by the available data, including time surveys, geographic, and movement data. JUNE simulates, simultaneously, the full population of a country in its spatio-temporal setting, and how a disease spreads through the population mediated by contacts between individuals. The main cost for this level of detail in the model is increased computational load; in fact, models such as JUNE would probably not have been possible prior to the 2010s, as they use what would have been a prohibitive amount of computing power at the time.

The JUNE framework is built on four interconnected layers: population, interactions, disease and policy. The layers and their interfaces are illustrated in Fig. 6.2.

JUNE models the transmission of an infection from the infecting individual, i, to the susceptible individual, s, in a probabilistic way. The probability of infection in a social setting within a group of people g, in a location L, depends on several factors:

- the number, N_i, of infectious people $i \in g$ present,
- the infectiousness of the infectors, i, at time t, $I_i(t)$,
- the susceptibility, ψ_s, of the potential infectee, s,
- the exposure time interval, $[t, t+\Delta t]$, during which the group, g, is at the same location,
- the number of possible contacts, $\chi_{si}^{(L)}$, and the proportion of physical contacts, $\phi_{si}^{(L)}$, at location L,
- and the overall intensity, $\beta^{(L,g)}$, of group contacts at location L.

Most of these "ingredients" depend on the time, t, of contact. For example, the number of contacts, $\chi_{si}^{(L)}$, and the proportion of physical contacts, $\phi_{si}^{(L)}$, and the overall contact intensity,

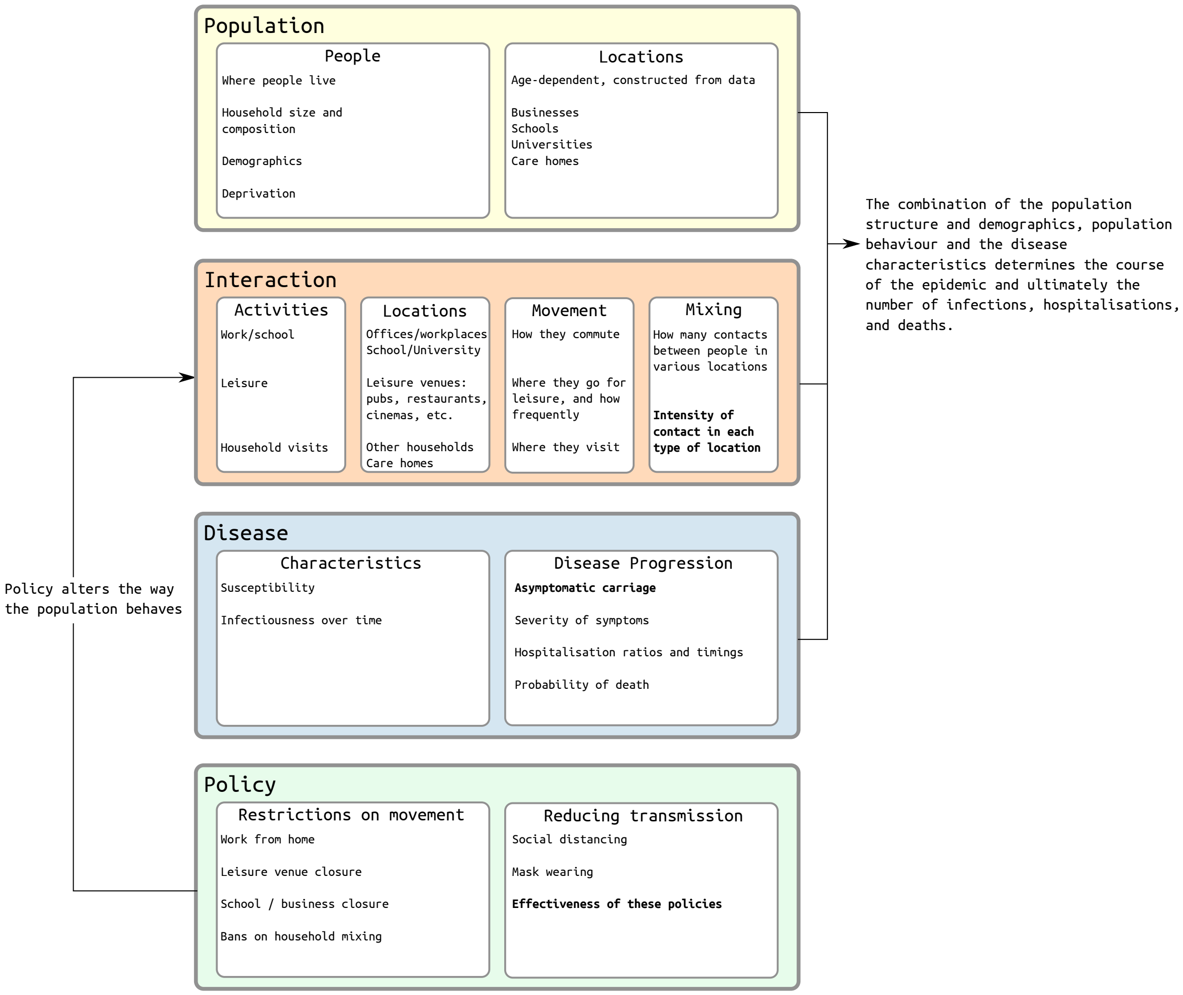

Population
People
Where people live
Household size and composition
Demographics
Deprivation
Locations
Age-dependent, constructed from data
Businesses
Schools
Universities
Care homes
The combination of the population structure and demographics, population behaviour and the disease characteristics determines the course of the epidemic and ultimately the number of infections, hospitalisations, and deaths.
Interaction
Activities
Work/school
Leisure
Household visits
Locations
Offices/workplaces
School/University
Leisure venues: pubs, restaurants, cinemas, etc.
Other households
Care homes
Movement
How they commute
Where they go for leisure, and how frequently
Where they visit
Mixing
How many contacts between people in various locations
Intensity of contact in each type of location
Policy alters the way the population behaves
Disease
Characteristics
Susceptibility
Infectiousness over time
Disease Progression
Asymptomatic carriage
Severity of symptoms
Hospitalisation ratios and timings
Probability of death
Policy
Restrictions on movement
Work from home
Leisure venue closure
School / business closure
Bans on household mixing
Reducing transmission
Social distancing
Mask wearing
Effectiveness of these policies

Figure 6.2: Overview of the structure of `JUNE`. Fitted parameters are shown in bold.

$\beta^{(L,g)}$, will change with the implementation of social distancing policies. To simplify the notation, we introduce a combined contact intensity for a group g with size N_g at location L,

$$\beta_{si}^{(L,g)}(t) = \beta^{(L,g)} \cdot \frac{\chi_{si}^{(L)}(t)}{N_g} \left\{1 + \phi_{si}^{(L)}(t)\left[\alpha(t) - 1\right]\right\}, \tag{6.2.1}$$

where the ratio χ/N_g provides a simple parameterisation of the probability that s is in contact with another individual in the group and $\alpha(t)$ describes the relative impact of close physical contacts. Both the factor $\alpha(t)$, which we assume to be the same for all locations, and the location- and group-specific contact intensities, $\beta^{(L,g)}$, are taken from fits to the data that will be explained in the following section.

In constructing an infection probability for a susceptible individual, s, we make several assumptions. First, we model the probability of being infected as a Poisson process. In keeping with the probabilistic process, the argument of the Poissonian is given by a sum over individual pairs of infectious individuals with the susceptible person, implying a simple superposition of individual infectiousness. The underlying individual transmission probabilities are written as the product of the susceptibility of the susceptible individual, the infectiousness of the infected person, and the intensity of contact, all integrated over the time interval in which the interaction occurs. Integration over time ensures that the transmission probability increases with exposure time. Therefore, we arrive at the transmission probability, i.e. the probability for s to be infected as:

$$\bar{\mathcal{P}}_s(t,\, t+\Delta t) = 1 - \exp\left[-\psi_s \sum_{i\in g} \int_t^{t+\Delta t} \beta_{si}^{(L,g)}(t')\mathcal{I}_i(t')\mathrm{d}t'\right] . \tag{6.2.2}$$

In the actual implementation, we approximate the integral over time with a simple product

$$\int_t^{t+\Delta t} \beta_{si}^{(L,g)}(t')\mathcal{I}_i(t')\mathrm{d}t' \;\longrightarrow\; \beta_{si}^{(L,g)}(t)\mathcal{I}_i(t)\Delta t\,. \tag{6.2.3}$$

6.2.2 Modelling the spread of Covid-19 in England

As a first application of JUNE, we modelled the spread of Covid-19 in England. In this context, JUNE uses census, household composition, and workplace data to ensure that each of the 53 million people in England is assigned a specific and identifiable location at any time. Their activities, health, age, and other demographic attributes are then modelled at a fine-grained geographical level, which helps to ensure that the local heterogeneity in population and movement characteristics is well recovered. A full description of the virtual twin of England can be found in Aylett-Bullock et al. (2021a).

The challenge faced when developing a model like JUNE is to calibrate its large number of free parameters (18 for the Covid-19 spread model in England) and the general uncertainty of the analysis. Different regions of parameter space might be compatible with the noisy aggregated datasets one is fitting, which implies that a point estimate of the parameters might bias the model forecasts. A full exploration of the parameter space is necessary to estimate the uncertainty of the model outputs.

In Vernon et al. (2022) we solve this problem with the introduction of a fast surrogate model that emulates the number of hospital admissions and deaths in each region as a function of the 18 free parameters of the model. This approach is extremely similar to the one developed in Chapter 4. The fast surrogate model can then be used to obtain a set of input parameters with a high likelihood when compared to the data. In Fig. 6.3, we show the results for 14 of these high-likelihood parameter sets.

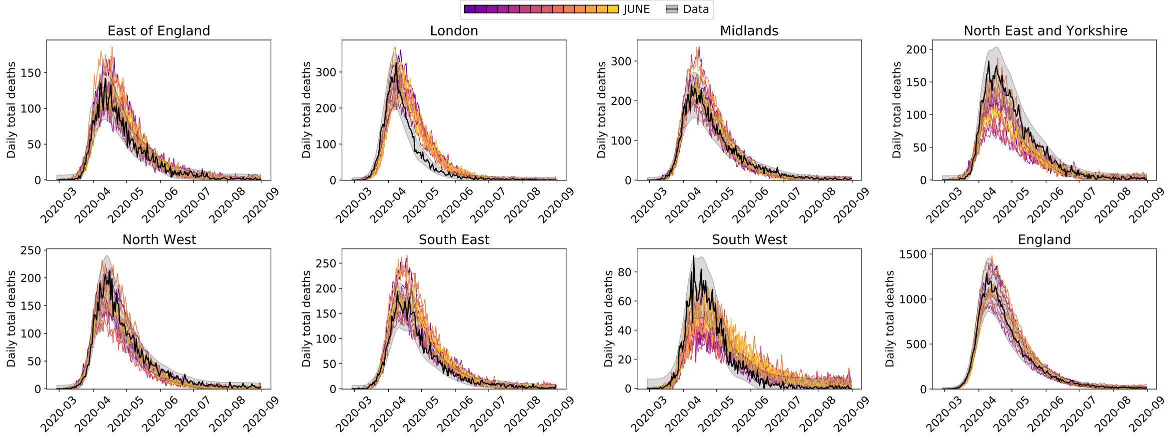


Figure 6.3: Daily hospital deaths for each region in England, and for England overall, as labelled in the title of each panel, for 14 realisations of JUNE. Data is shown in black, with 3 standard deviation error bands.

The level of detail included in JUNE allows us to study how different subgroups of the population were affected by the virus. Although everyone is, in principle, equally susceptible to the virus, Covid-19 has hit harder those in less privileged socioeconomic groups. Since JUNE can answer questions at any spatial level and for any demographic, we can compare the prevalence of antibodies in the JUNE population with that found in the REACT2 study during the first wave of England (Ward et al., 2020). In addition to regional differences, Ward et al. (2020) found that the prevalence of Covid-19 is a function of age, ethnicity, the deprivation quintile and the size of the household. Fig. 6.4 shows that JUNE reproduces these trends, demonstrating that we can use JUNE to understand and assess the effect of different policies to an unprecedented degree of precision.

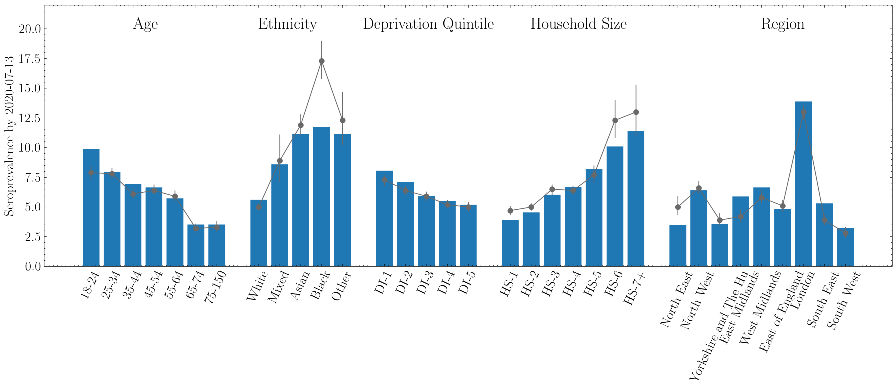


Figure 6.4: Comparison of the predicted seroprevalence of JUNE after the first wave, with data from Ward et al. (2020). Data are shown as grey errorbars, whereas simulation results are shown as blue solid bars.

6.2.3 Modelling the spread of Covid-19 in refugee settlements: Cox's Bazar

The spread of infectious diseases such as Covid-19 presents many challenges to healthcare systems and infrastructures around the world, exacerbating inequalities and leaving the most vulnerable populations the most affected. Given their high population density and limited available infrastructure, refugee and internally displaced person (IDP) settlements can be particularly susceptible to spreading the disease. In Aylett-Bullock et al. (2021b), we have adapted the JUNE framework described in the previous section to simulate the spread of disease in refugee and IDP settlements under various non-pharmaceutical intervention strategies. The model is informed by data on geography, demographics, comorbidities, physical infrastructure, and other parameters obtained from real-world observations and previous literature. The development and testing of this approach focus on the Cox's Bazar refugee settlement in Bangladesh, although our model is designed to be generalisable to other informal settings.

An important difference between the population of England and that of Cox's Bazar is the presence of comorbidities at younger ages. An individual's response to Covid-19 and other diseases can depend on the presence of diseases such as diabetes, heart conditions, and conditions causing immune suppression. Specifically, we allow the probability of following one of the disease trajectories to depend on the comorbidity status, together with age and sex.

Furthermore, given the incompleteness of the testing and case reporting data in the Cox's Bazar settlement, we cannot use the data to inform us about the most plausible model parameters and perform a complete uncertainty quantification analysis, as we did in the case of England. Therefore, we focus primarily on analysing the efficacy of the intervention by comparing the relative magnitudes of the infection curves between various implementation conditions.

In particular, we use our model to assess interventions that were deemed the most important by public health officials operating in the settlement according to an assessment of short- and medium-term needs, including feasibility and timeliness. All interventions were compared with a baseline scenario which includes current policy decisions, such as closing certain venues and changes in the probability with which people perform certain tasks.

We first examined the effectiveness of self-isolation. In many countries, those with symptoms that are not yet severe enough to require hospitalisation are encouraged to stay home and self-quarantine. In the case of settlements such as Cox's Bazar, the density and living conditions of the residents mean that it is not possible to avoid contact with family in the home environment, and people often have to leave their shelter to use facilities such as hand pumps and latrines. In an attempt to better allow the isolation of symptomatic individuals, public health officials in the settlement established isolation and treatment facilities to house those who tested positive for Covid-19 but did not require hospitalisation.

Our findings suggest that encouraging self-isolation at home of mild to severe symptomatic patients, as opposed to the isolation of all positive cases in purpose-built isolation and treatment centres, does not increase the risk of secondary infection, meaning that the centres can be used to provide hospital support to the most intense cases of Covid-19.

Second, we studied the effectiveness of mask wearing, finding that mask wearing in all communal indoor areas can be effective in dampening viral spread, even with low mask efficacy and compliance rates.

Finally, we modelled the effects of re-opening learning centres in the settlement under various mitigation strategies. For example, a combination of mask wearing in the classroom, halving attendance regularity to allow physical distancing, and better ventilation can almost completely mitigate the increased risk of infection that keeping the learning centres open may cause.

These modelling efforts are being incorporated into decision-making processes to inform future planning, and more exercises should be performed in similar geographies to help protect those most vulnerable.

Chapter 7

Conclusions and Future Work

7.1 Summary

The large scale structure (LSS) of the Universe as traced by three-dimensional galaxy maps carries a wealth of information, which can be used to constrain theories of gravity. In particular, we can use the clustering properties of the LSS to address some of the most pressing questions opened up by the standard cosmological model, such as "What drives the accelerated expansion of the universe?" and "What is the dark matter?".

Ongoing and future surveys, such as the Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016b), the Subaru Prime Focus Spectrograph (PFS) Takada et al. (2014), and the space-based mission Euclid (Laureijs et al., 2011) will provide LSS maps of unprecedented statistical precision. The challenge cosmologists face now is to develop statistical methods that are i) accurate enough to match the precision of the data and ii) optimal regarding the amount of information extracted, so that we can extract all of the valuable information on gravity and cosmology contained in the LSS. Overcoming these challenges would help us reduce the uncertainties on the estimated cosmological parameters, which would in turn determine if the observed tensions among different values inferred for the cosmological parameters (see Section 1) are the result of systematics, statistical bad luck, or even the imprint of new physics that is yet to be discovered.

In this Thesis, we worked on these two challenges by: i) developing simulation-based methods which yield predictions of summary statistics at percent-level accuracy over a wider range of scales than previous work, and ii) showing how the constraints obtained from the two-point functions could be enhanced through the dependence of galaxy clustering on environment.

7.1.1 The mapping between real and redshift space

Although the statistical precision of data on small scales is higher than that on large scales, most studies that rely on perturbation theory (e.g. Chen et al., 2021) to model the dependence of two-point functions on cosmology restrict their analysis to pair separations larger than ≈ 30 h^{-1}Mpc. On smaller scales, perturbation theory models break down rapidly and their use introduces biases into the inferred cosmological parameters.

To obtain fully non-linear predictions for the properties of the large-scale structure and recover all of the cosmological information contained in the small-scale clustering, we must resort to N-body simulations (Kuhlen et al., 2012). N-body simulations have been widely used as cosmic laboratories to test the precision and robustness of analytical methods for characterising the large-scale structure (e.g. Carlson et al. 2009), together with the effects of systematic errors in our measurements.

In Chapter 3, we showed an application of N-body simulations in this spirit. Our contributions were:

- We extended the Gaussian streaming model of redshift space distortions with an analytical description of the pairwise velocity distribution based on the four lowest-order velocity moments of the PDF, the Skew-T distribution (ST). Our model describes the pair separation dependence of the PDF significantly better than a simple Gaussian.

- Using the description of the pairwise velocity distribution to map pairs of galaxies from real to redshift space, we find that the ST model can reproduce the multipoles of halo autocorrelaitons to within 2% on scales larger than 1 h^{-1} Mpc for the monopole, 4% for the quadrupole on scales larger than 5 h^{-1} Mpc, and 10% for the hexadecapole on scales larger tan 8 h^{-1} Mpc. On scales smaller than around 5 h^{-1} Mpc, the contribution of satellite galaxies would dominate the multipoles of the correlation function, whose effects have been ignored in this work.

- We showed that a Taylor expansion of the streaming model can give an accurate description of the full non-Gaussian streaming model down to about 10 h^{-1} Mpc for the monopole and quadrupole moments, when expanded up to the fourth order.

- We demonstrated the importance of modelling the mean pairwise velocity to better than one per cent accuracy to obtain similar levels of precision for the monopole and quadrupole moments. Fitting CLEFT perturbation theory estimates, with five free parameters, we could only obtain accurate predictions at the percent level on scales greater than 35 h^{-1} Mpc. Therefore, more accurate models of the mean pairwise velocity of dark matter halos are needed to achieve the targets of future surveys.

7.1.2 Emulating summary statistics measured from N-body simulations

Given that accurate predictions for the ingredients of the streaming model, the real space two-point correlation function, and the lowest four-order velocity moments are needed to accurately reproduce the real to redshift space mapping on small scales, we worked on developing simulation-based models of these summary statistics. In Chapter 4, we presented a real space two-point correlation function emulator.

Over the past decade, advances in computing power and algorithms have allowed us to produce a large enough number of dark matter only N-body simulations covering a significant fraction of the cosmological parameter space. This allows us to use the simulations themselves as predictive models that directly constrain the cosmological parameters by leveraging machine learning techniques. However, in order to compare the outcomes of dark matter only simulations to the observed distribution of galaxies, we have to model the connection between dark matter halos and galaxies (see Wechsler & Tinker 2018 for a review on this topic).

Uncertainties in the galaxy-halo connection can limit the amount of information that we can extract from small scale clustering. We would like to use flexible models that can reproduce clustering in different scenarios of galaxy formation, whilst still being able to recover cosmological information after marginalising over the free parameters of the galaxy-halo connection model. In Chapter 4, we used the empirical model of the halo occupation distribution (HOD) (Benson et al., 2000; Zheng et al., 2005), which describes the probability that a given halo hosts a galaxy based on the mass of the halo.

In summary, in Chapter 4 we:

- Presented a neural network that models the full-shape galaxy clustering in real space based on the halo model, which is more accurate and faster than previously published Gaussian process emulators (Nishimichi et al., 2019), when trained on the same dataset.
- Showed that small scale galaxy clustering ($r < 5\ h^{-1}$ Mpc) in real space improves the constraints on σ_8 by a factor of 2, while marginalising over the HOD parameters erases the information contained on small scales for Ω_m.

- Showed that a halo model that ignores effects of environment-based assembly bias similar to those observed in hydrodynamic simulations and semi-analytical models of galaxy formation could introduce bias into the inferred σ_8, while the BAO peak ensures that we can recover Ω_m and h robustly.
- Found that the bias mentioned above in the value of the inferred σ_8 disappears when analysing scales larger than 10 $h^{-1}\,$Mpc.

7.1.3 Extracting more cosmological information: density-split clustering

If the galaxy field were a Gaussian random field, its two-point statistics (the power spectrum or the two-point correlation function) would be complete summaries of the 3-D maps. But while the density field at high redshift is indeed close to Gaussian over a wide range of scales, nonlinear gravitational evolution produces non-Gaussianity, and limits the scales on which the field is Gaussian to progressively larger scales at later times. Given that the mass overdensity δ is bounded at low values by -1, since a region of the universe cannot have a negative density, the distribution of δ values must develop skewness as the density contrast grows. Finding alternative summary statistics to supplement the constraints obtained from the two-point functions is currently an active area of research (see, for instance, studies on the bispectrum Hahn et al. 2020b and the scattering transform Valogiannis & Dvorkin 2022). In Chapter 5, we demonstrate how environment-dependent clustering can provide an effective and interpretable method for extracting non-Gaussian information from galaxy surveys.

In particular, we show that:

- Splitting the dark matter density field into quintiles of varying local density can improve constraints on the cosmological parameters of $\nu\Lambda$CDM by factors between 3 and 8, depending on the parameter. In particular, density split improves the constraints on the sum of neutrino masses by a factor of 8×, and by factors of 5×, 3×, 4×, 6×, and 6× for Ω_m, Ω_b, h, n_s, and σ_8, respectively.
- Density split clustering statistics depend on whether the environment density is defined in real or redshift space. Real-space identified quintiles yield better constraints for all cosmological parameters, in particular Ω_m and σ_8.
- However, galaxy catalogues in real space are not immediately available in observations, and one would have to rely on reconstruction algorithms to approximately remove RSD

from galaxies (Nadathur et al., 2019). But, as shown in Sect. 5.4.3, reconstruction algorithms could potentially introduce systematic errors in the inferred cosmological parameters when including small-scale information, which would then need to be added to the total error budget.

- Quintile autocorrelations can recover some of the cosmological information that is lost when performing the density split in redshift space as opposed to real space. Introducing these into the likelihood analysis will therefore be useful to avoid the use of reconstruction techniques and to analyse directly the redshift space-identified multipoles.

7.2 Future work

The next few years will be exciting times in cosmology. Not only will we have access to the largest and most precise three-dimensional maps of the universe that will allow us to set very stringent constraints on ΛCDM theories, but also new statistical methods will allow us to unlock the potential to discover new physics.

In the following, we outline some interesting extensions to the work presented in this thesis.

7.2.1 Simulation-based summary statistics with machine learning

Simulation-based models of summary statistics will complement the constraints obtained using perturbation theory techniques by modelling the non-linearities present in small scale clustering. The work presented here focused on modelling the real-space correlation function of galaxies, but we are currently working on extending our approach to the real to redshift space mapping through a similar simulation-based model of the pairwise velocity distribution.

Moreover, the Fisher analysis presented in Chapter 5 demonstrated the potential of environment dependent clustering to constrain $\nu\Lambda$CDM. To realise this constraining power with DESI Y1 data, we will need to develop accurate simulation-based methods for density split statistics. In particular, we would like to leverage the environment dependence to set stringent constraints on assembly bias models. As shown in Chapter 5 , the surrogate model could directly aim at reproducing redshift space identified splits, which will allow us to avoid the introduction of reconstruction algorithms when analysing data from galaxy surveys.

A common challenge that simulation-based models of summary statistics face is that of attaining sub per cent level precision. A particular case is when modelling summary statistics other than two-point correlations functions, since these might have a higher dimensionality and more complex dependencies on the cosmological parameters. Modelling summary statistics on non-linear scales based on a set of $\mathcal{O}(100)$ N-body simulations to the required accuracy is a challenge for current emulation techniques.

In Lange et al. (2019), the authors showed that one could directly model the evidence for a particular summary statistic, after marginalising the halo-galaxy connection in an exact manner. In this way, they overcame the limitation of percent accurate predictions for the summary statistic as a function of the halo-galaxy connection parameters. Moreover, by modelling the evidence, a single number, instead of multi-dimensional observables, they showed that they could use simple multi-dimensional Gaussian functions to model the dependency of the evidence with cosmology.

However, since Lange et al. marginalised over the halo-galaxy connection, their approach does not allow one to obtain constraints on galaxy formation physics. To overcome this limitation, we would like to work on a method that emulates the likelihood function of a given observable. This could either be done by evaluating an analytical likelihood over a set of N-body simulations populated with mock galaxies, or by learning the likelihood directly from N-body simulations through neural networks (Glöckler et al., 2022). While the first option would be preferred for summary statistics such as the two-point correlation function, where the Gaussian likelihood is a good approximation to the one measured in N-body simulations, the second would be needed for alternative summary statistics whose likelihood might significantly deviate from Gaussian.

In addition, we would like to combine the surrogate model for likelihood evaluation with variational inference (VI) techniques (Blei et al., 2017; Glöckler et al., 2022) to estimate the posterior of cosmological and galaxy-halo connection parameters without having to produce costly MCMC samples. Variational inference is a family of machine learning techniques that turns the classical inference problem into an optimisation one by defining a set of flexible probability densities and then optimising a distance measure between the true posterior and the set of flexible densities. Commonly used densities are mixture of Gaussians and normalizing flows (Rezende & Mohamed, 2015).

Variational inference would allow for the introduction of complex relations in the connection between halos and galaxies, which would otherwise introduce a prohibitively large number of free parameters to be constrained through MCMC chains. Additionally, these models can be trained sequentially to increase their accuracy while reducing the number of simulations used.

Finally, yet importantly, we will have to make sure that halo-galaxy connection models are on one hand flexible enough to encompass all plausible galaxy formation scenarios, whilst allowing for precise estimates of the cosmological parameters. It will therefore be crucial to test the robustness of simulation-based methods with large hydrodynamical simulations and semi-analtyic models of galaxy formation.

7.2.2 Testing gravity with simulation-based methods

Although simulation-based methods increase our constraining power, they also reduce the hypothesis space that we can test. This is especially relevant for cosmology, since N-body simulations are extremely costly, particularly for testing gravity theories. The standard cosmological model, ΛCDM, assumes that general relativity is the correct theory of gravity on cosmological scales. However, despite the success of the theory, the true nature of dark energy and dark matter remains unknown. Therefore, it is important to test well-motivated alternative gravity theories, also referred to as modified gravity theories (Clifton et al., 2012; Joyce et al., 2015).

Previous simulation-based models of the galaxy two-point correlation function have introduced linear scaling parameters to account for departures from general relativity (e.g., Zhai et al. (2022)). However, this assumption is extremely simplifying and might bias constraints on deviations from GR. Therefore, it is important to develop simulation-based techniques for alternative gravity models.

For this purpose, fast N-body codes (Ruan et al., 2022) are being developed for modified gravity theories. These will then be able to produce large suites of N-body simulations (such as Arnold et al. (2021)) that can be used to train emulators and constrain gravity with unprecedented precision.

7.2.3 Machine learning the optimal summary of the Universe

By constructing accurate emulators for observables with complementary information content to that of two-point functions, we can improve current constraints on cosmological parameters and gravity. Nevertheless, there is no guarantee that all the summaries that we design

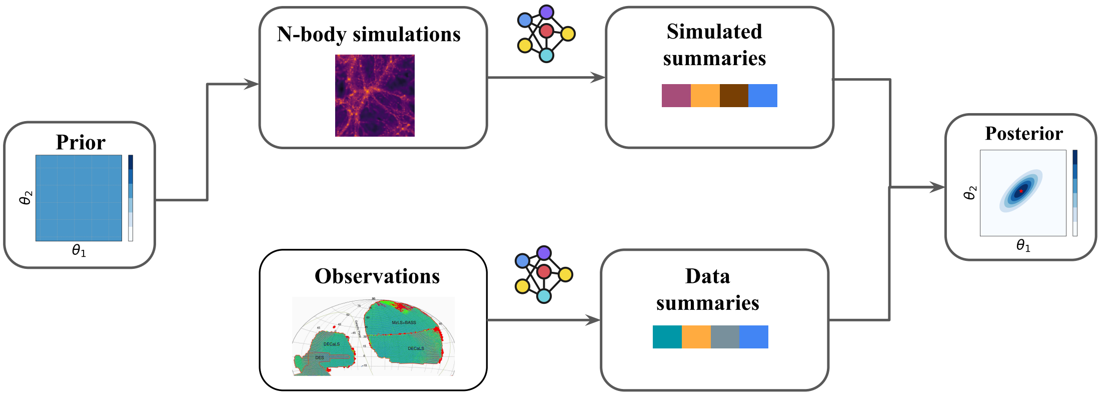


Figure 7.1: Sketch of a deep learning model to turn priors on cosmological parameters into posteriors given the data.

combined would exhaust the information content of 3-D galaxy maps. Instead, we could harness machine learning techniques to design summary statistics that perform inference in an optimal manner.

Although ML methods for cosmological parameter inference have been developed (e.g. Ntampaka et al. (2020)), these have not yet revolutionised the way inference is done in cosmology, nor have they been applied to galaxy maps to produce estimates of the cosmological parameters. We attribute this to:

- *Lack of uncertainty estimates.* Since deviations from CDM are likely to be small, a cautious application of statistics is important to separate a true discovery from a spurious one. To date, a reliable estimate of the posterior over the cosmological parameters inferred from 3-D galaxy maps has yet to be done.

- *Lack of interpretability.* Understanding where the theory fails to reproduce the data is as important as knowing whether it does fail, in order to inform future modelling. There are currently no general methods to untangle neural network output as a function of interpretable quantities, such as separating contributions from different scales or environmental densities.

- *Lack of robustness.* Training and testing our methods on numerical simulations may lead us to overfit the simulations used for training. This would imply that either we lose constraining power or, even worse, we might end up with overconfident predictions. This is especially relevant when we account for the uncertainty introduced by our poor understanding of baryonic effects.

Overcoming these problems could unlock the potential to discover new physics with machine learning, and working on the development of robust and interpretable summary statistics would help us do so. In Fig. 7.1, we show a sketch of a deep learning model that constrains cosmology through the learnt summary statistics.

It will also be important to account for inductive biases, assumptions inherent to a learning algorithm that are independent of the data, to build an ML framework tailored to cosmology. For instance, the Universe is spatially homogeneous and isotropic on the largest scales, and these symmetries should be preserved in the summary statistics. Instead of relying on our model to learn symmetries from data, we can construct rotational and translational invariant architectures (Bronstein et al., 2021). Adapting neural network architectures to cosmological datasets will not only make models learn more efficiently but also ease the interpretability of their outcomes.

It is possible that deviations from ΛCDM already hide in current datasets, and we are blinded by our lack of suitable inference techniques. In the coming years, machine learning will allow us to extract and quantify the wealth of information contained across the entire cosmic web.

Appendix A

Appendix: The real to redshift mapping on small scales

A.1 Method of moments for the ST distribution

The four parameters of the ST distribution (v_c, w, α, ν) are determined by the first four order moments. To simplify the relation between moments and parameters, we introduce,

$$b_\nu = \left(\frac{\nu}{\pi}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma(\nu/2)}, \tag{A.1.1}$$

$$\delta = \frac{\alpha}{\sqrt{(1+\alpha^2)}}, \tag{A.1.2}$$

The moments can be shown to be,

$$m_1 = v_c + w\delta b_\nu, \tag{A.1.3}$$

$$c_2 = w^2 \left(\frac{\nu}{\nu-2} - \delta^2 b_\nu^2\right), \tag{A.1.4}$$

$$\gamma_1 = \frac{c_3}{c_2^{3/2}} = \delta b_\nu \left(\frac{\nu(3-\delta^2)}{\nu-3} - \frac{3\nu}{\nu-2} + 2\delta^2 b_\nu^2\right) \left(\frac{\nu}{\nu-2} - \delta^2 b_\nu^2\right)^{-\frac{3}{2}}, \tag{A.1.5}$$

$$\begin{aligned}\gamma_2 =& \frac{c_4}{c_2^2} = \left(\frac{3\nu^2}{(\nu-2)(\nu-4)} - \frac{4\delta^2 b_\nu^2 \nu(3-\delta^2)}{\nu-3} - \frac{6\delta^2 b_\nu^2 \nu}{\nu-2} - 3\delta^4 b_\nu^4\right) \\ &\left(\frac{\nu}{\nu-2} - \delta^2 b_\nu^2\right)^{-2}.\end{aligned} \tag{A.1.6}$$

The parameters α and ν are obtained from the last two equations that determine the skewness and kurtosis of the distribution, these form a system of non-linearly coupled equations that we solve numerically. The remaining two parameters, v_c and w, can then directly be obtained from the equation for the mean and the variance.

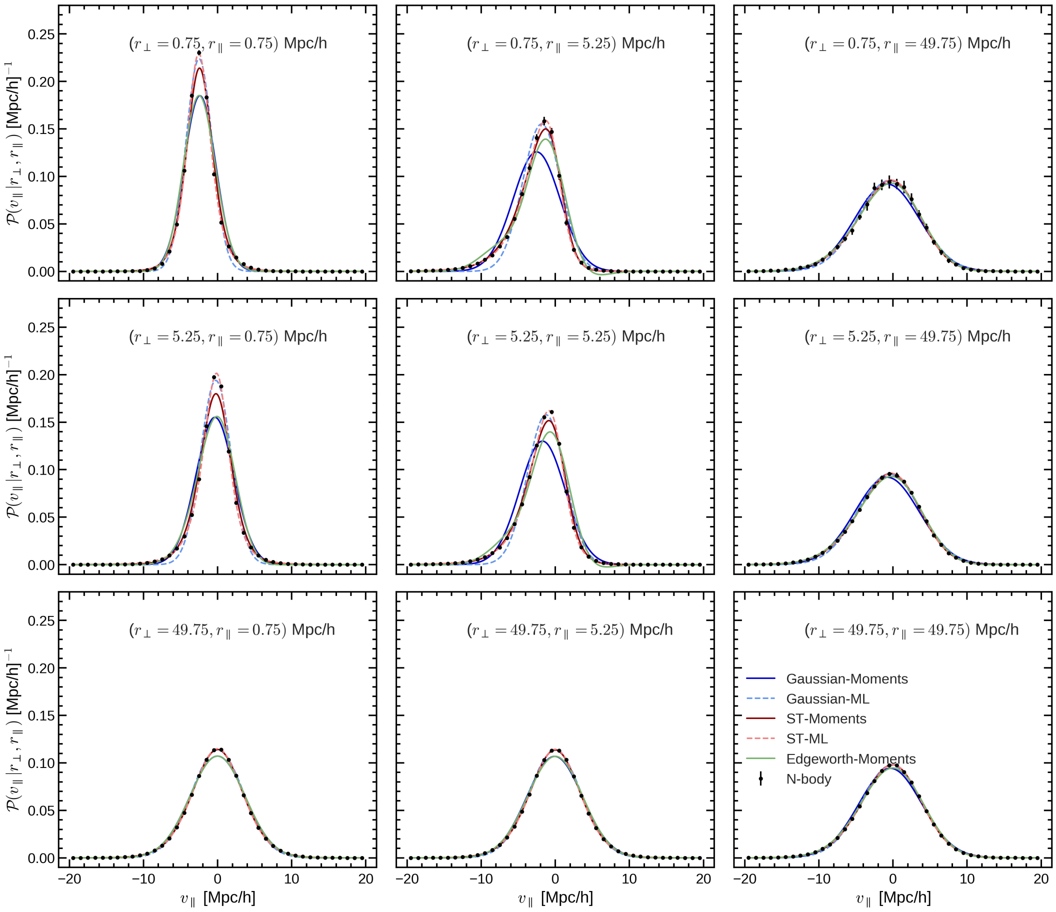


Figure A.1: Linear scale representation of the pairwise velocity distribution to highlight the behaviour of the PDF close to its peak. The models shown are the same one as in Fig. 3.2. Note that the Edgeworth expansion predicts negative probabilities for certain pair separations such as $(r_\perp = 0.75, r_\parallel = 5.25)\, h^{-1}$Mpc and $(r_\perp = 5.25, r_\parallel = 5.25)\, h^{-1}$Mpc, where the skewness is more pronounced. Moreover, an Edgeworth expansion behaves very differently from a Taylor expansion since it produces an asymptotic expansion, and therefore adding more terms does not guarantee convergence. See Sellentin et al. (2017) for an interesting discussion on the Edgeworth expansion and its applications to cosmology. In the application of the Edgeworth expansion to the pairwise velocity distribution, we see that it does not reproduce the N-body measurements as well as the ST distribution does with only one extra parameter.

A.2 Zoom in distributions

A.3 Perturbation Theory results in detail

In this Appendix, we show a detailed summary of the state-of-the-art CLPT and CLEFT perturbation theory predictions for the Gaussian Streaming Model ingredients. Note that we show the predictions for real space statistics, since we want to separately analyse the accuracy of perturbation theory predicting the ingredients of the Streaming Model, and the assumption of a Gaussian pairwise velocity distribution.

The free parameters are found by maximising the combined Gaussian likelihood that the

	b_1	b_2	b_s	α_ξ	α_v	$\sigma_{\rm FoG}$
CLPT	0.29 ± 0.01	-1.63 ± 0.31	1.80 ± 0.38	-	-	-17.35 ± 0.15
CLEFT	0.30 ± 0.02	-1.69 ± 0.26	2.16 ± 0.37	-39.58 ± 16.32	90.30 ± 73.68	-17.45 ± 0.28

Table A.1: Perturbation theory parameters for both CLEFT and CLPT. Note that b_1, b_2, and b_s are obtained by expanding the bias function in Lagrangian space. We show the maximum likelihood estimate and errors representing 1-sigma deviations in the posterior distribution of the given parameter.

simulation measurements are most probable under the given theory,

$$\log(\mathcal{L}) = \log(\mathcal{L}_\xi) + \log(\mathcal{L}_{m_{10}}) + \log(\mathcal{L}_{c_{20}}) + \log(\mathcal{L}_{c_{02}}), \tag{A.3.1}$$

where the individual likelihoods are given by,

$$\log(\mathcal{L}_y) = -\frac{1}{2}\sum_i \frac{(y_{i,\rm measured} - y_{i,\rm model})^2}{\sigma_i^2}. \tag{A.3.2}$$

where y is the mean simulation measurement across the 15 independent simulations, and σ its standard deviation. Note that the covariance matrix is assumed to be diagonal, which means that the parameter uncertainties obtained from the fit will be underpredicted. While this assumption will also affect the values of the best-fit parameters in detail, we do not expect this to have a qualitative impact on the relative agreement between the model predictions and data, which is our main objective here. We maximise the likelihood in the pair separation range 15 h^{-1}Mpc $< r <$ 150 h^{-1}Mpc and the resulting mean parameter values are shown in Table A.1. We find a value for the second order Lagrangian bias b_2 that is in good agreement with previous measurements (Lazeyras et al., 2016), whereas the tidal bias is rather different from its local Lagrangian value ($b_s = 0$), which is in contrast with other analyses in the literature (Lazeyras & Schmidt, 2018; Abidi & Baldauf, 2018). We also note that the EFT parameters are the least constrained by our measurements, which is to be expected as they only have an impact on the small-scale regime.

In Fig A.2 we show a detailed comparison of the best-fit model predictions for the two methods. The second counter-term introduced in CLEFT improves notably the prediction for the mean pairwise velocities on scales between 20 h^{-1}Mpc and 60 h^{-1}Mpc. Regarding the second order moments, the predictions for m_{20} are similar for CLPT and CLEFT, however, since $c_{20} = m_{20} - m_{10}^2$, the variance of the radial component is influenced by the predictions of the mean. Conincidentally, the error made by CLPT in the mean improves the agreement with the variance of the radial component (dotted blue line in the lowest panel).

Finally, we show the redshift space monopole and quadrupole in Fig. A.3, obtained by

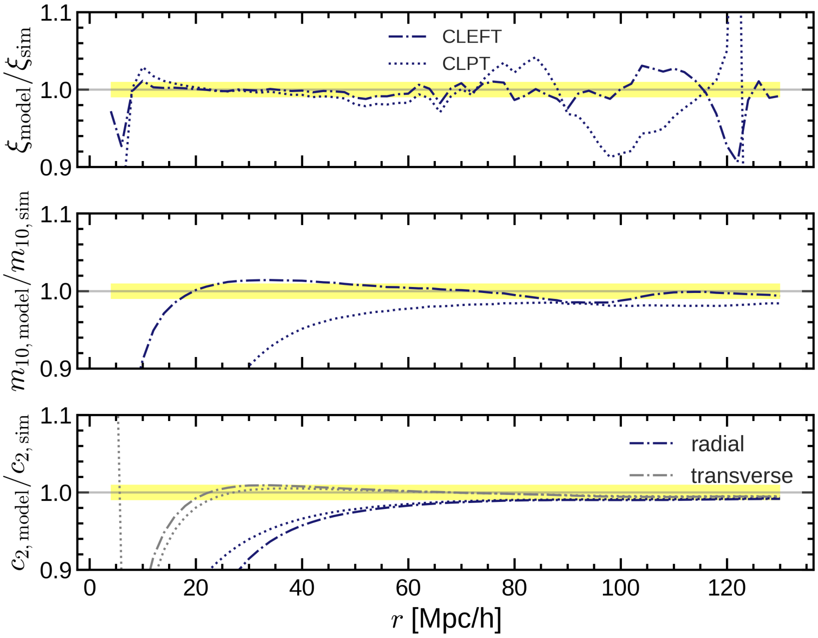


Figure A.2: Detailed comparison of the different predictions for the Gaussian Streaming Model ingredients made by CLPT and CLEFT. The top panel shows the ratio of the predicted two-point correlation function to the measurement in the simulation, for both CLPT (dotted) and CLEFT (dotted-dashed line). The solid yellow bands marks the one per cent agreement. The middle and bottom panels show the same comparison for the mean radial velocity, and the second order radial and transverse moments.

combining these predictions with the Gaussian Streaming Model. The CLEFT predictions of the monopole and quadrupole are more accurate than those from CLPT, mainly due to the increased accuracy in estimating the mean pairwise velocity, which is consistent with our findings in Sec. 3.4. As shown in Sec. 3.2.4, on scales smaller than $30\,h^{-1}$Mpc it is necessary to include higher order moments to further improve the accuracy of the predictions. A more detailed comparison of these different models applied to mock catalogues that mimic actual data at different redshifts and different halo mass ranges will be the subject of future work.

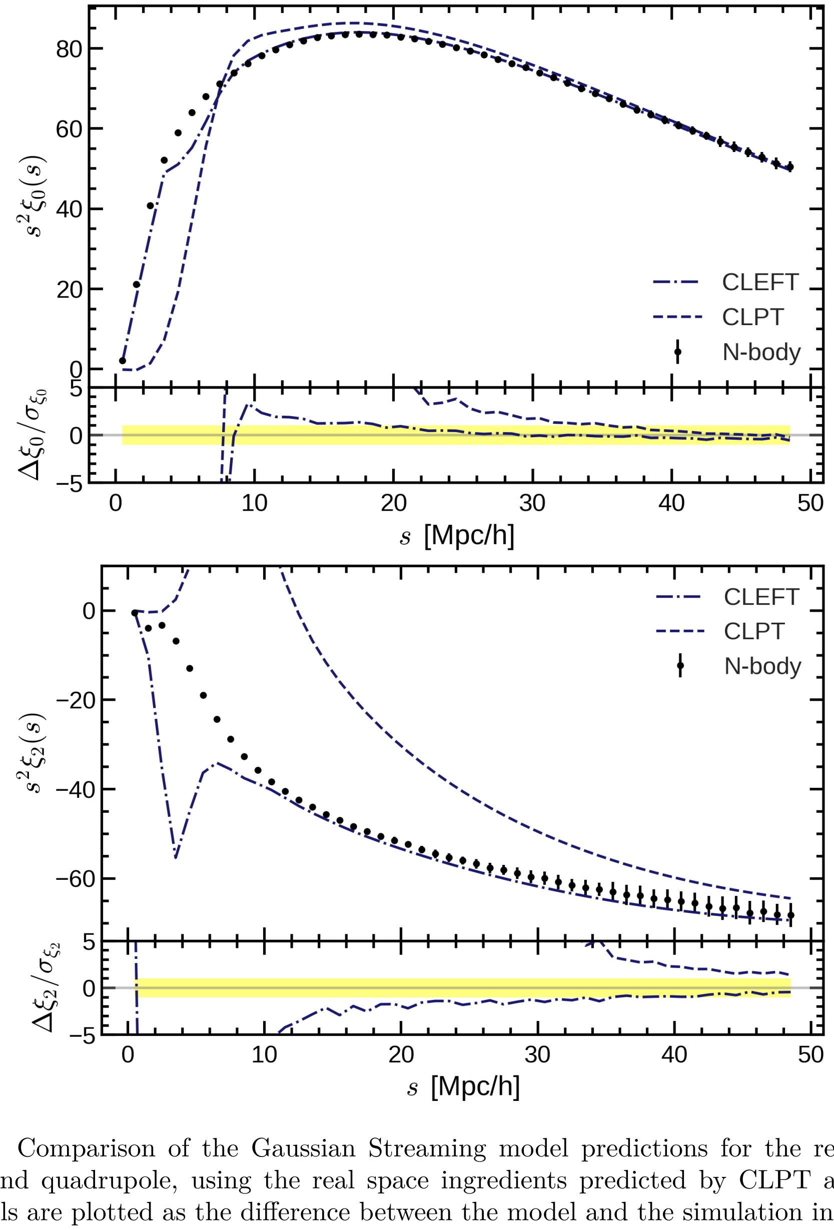


Figure A.3: Comparison of the Gaussian Streaming model predictions for the redshift space monopole and quadrupole, using the real space ingredients predicted by CLPT and CLEFT. The residuals are plotted as the difference between the model and the simulation in units of the variance calculated across the different independent simulations. The yellow bands show the 1σ deviation.

Appendix B

Appendix: Simulation-based models for real space clustering

B.1 Evaluation of the emulators as a function of redshift and number density

In this appendix we show detailled evaluations of the halo auto-correlation emulator (Fig. B.1) and the galaxy auto-correlation emulator (Fig. B.2).

For halo auto-correlations, we find that the emulator accuracy decreases for lower number densities, which are more affected by shot noise, whereas it decreases for high redshifts ($z = 1.5$).

For galaxy auto-correlations we do not find any substantial biases for redshift and galaxy number density.

B.2 Estimating the covariance matrix

In Section 4.4, we used an estimate of the covariance matrix to obtain the posterior of cosmological parameters given a mock data vector. The covariance matrix was estimated from a set of 1600 N-body simulations part of the AbacusSummit suite (Maksimova et al., 2021b). These are high resolution small boxsize simulations ($L_{\mathrm{box}} = 500\ h^{-1}\,\mathrm{Mpc}$).

Given the small boxsize of the simulations, we re-scale the covariance by a factor of $0.5^3/0.67$ to estimate the expected errors for a LOWZ-like sample, whose effective volume is

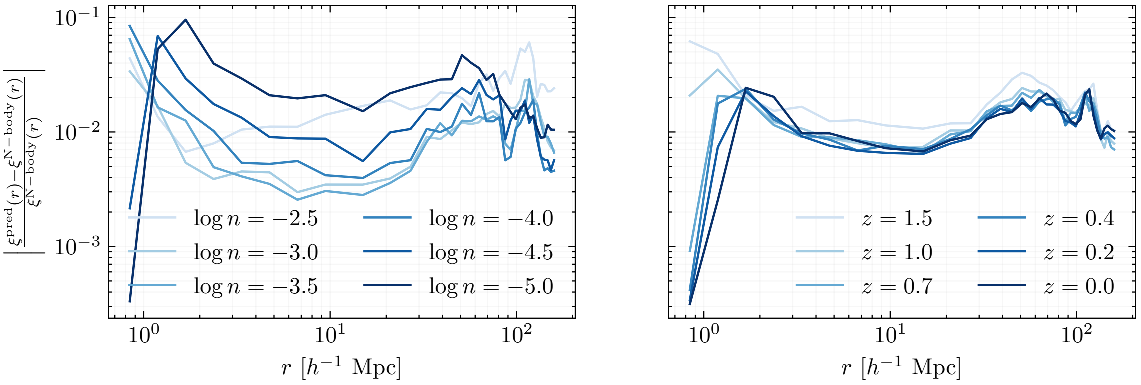


Figure B.1: Median absolute errors of the halo two-point correlation function as a function of number density (left), averaged over redshift and test set cosmologies, and as a function of redshift (right), averaged over number density and test set cosmologies.

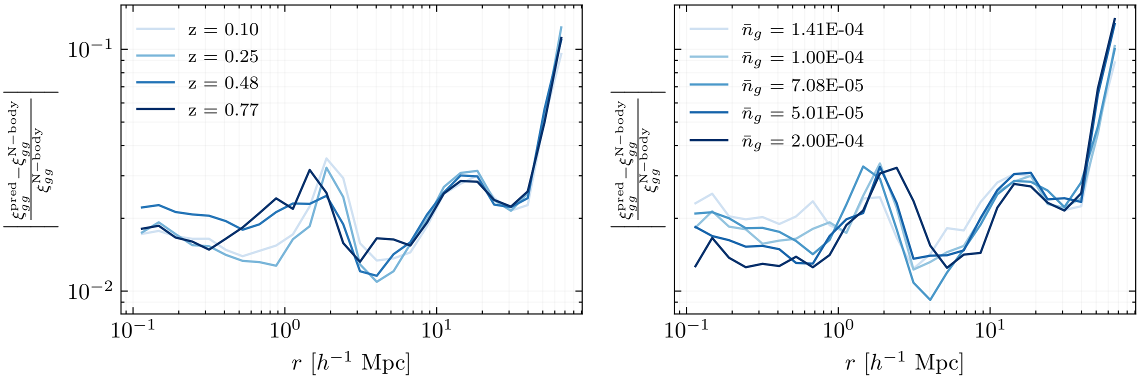


Figure B.2: Median absolute errors of the galaxy two-point correlation function as a function of number redshift (left), averaged over galaxy number density and test set cosmologies, and as a function of galaxy number density (right), averaged over number redshift and test set cosmologies. In both cases the emulator accuracy does not show noticeable biases.

$0.67\ (h^{-1}\,\mathrm{Gpc})^3$. We also correct the covariance estimated from the mocks with Eq. 56 in Percival et al. (2021).

B.3 The effect of constraining galaxy number density in the likelihood analysis

In this appendix, we show the effect of removing the galaxy number density term in Eq. 4.4.2.

Fig. B.3 shows that the number density constrain does not change the constraints on cosmological parameters noticeably, whereas it mainly improves those of the HOD parameters. In particular, it breaks the degeneracy between the central occupation parameters, $\log M_{\mathrm{min}}$ and $\sigma_{\log M}$.

B.4 Assembly bias mocks details

Here, we describe here the occupation variations of the environment-based assembly bias mocks used in Section 4.4.3.

Fig. B.4 shows how the mean number of centrals and satellites change as a function of halo mass and halo environment for both the strong and weak assembly bias mocks. At fixed halo mass, halos residing in denser environments will have a higher mean number of galaxies (both centrals and satellites) than those occupying underdense regions.

On the right hand side of Fig. B.4 we also show the ratio of the galaxy two-point correlation function with a strong and weak assembly bias signal to that of the no assembly bias case. The deviations can be as large as 10% for the weak case, and 20% for the strong one.

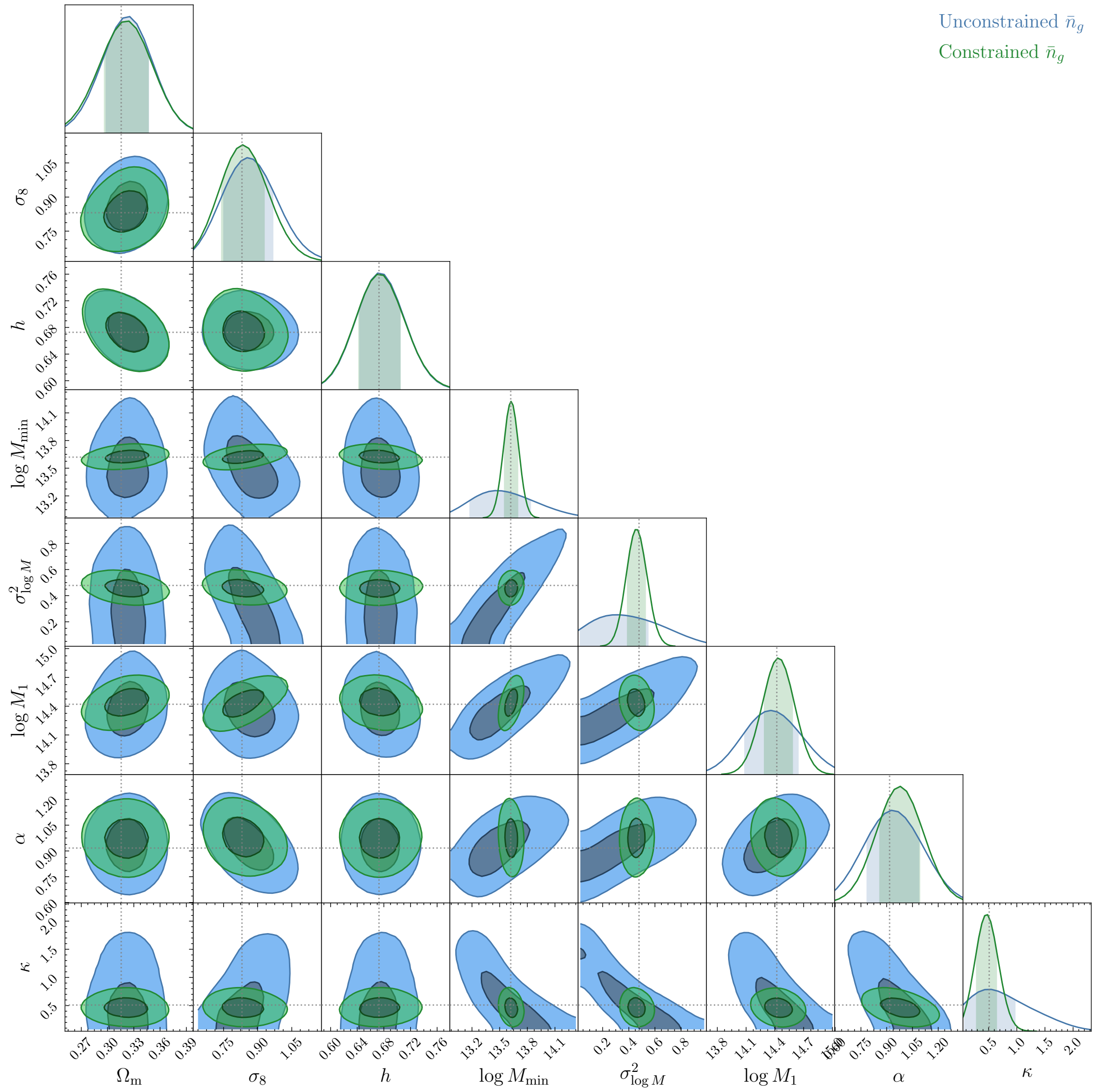


Figure B.3: Comparison of constraints on cosmological and HOD parameters when the galaxy number density is included in the likelihood (Constrained $\bar{n}_g$) and when it isn't (Unconstrained $\bar{n}_g$. Including number density constraints only helps determine the HOD parameters with a higher accuracy.

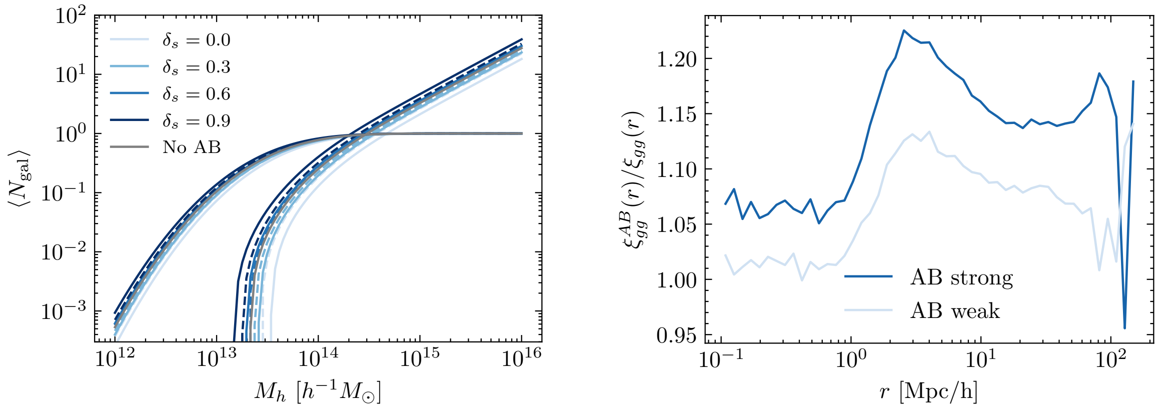


Figure B.4: Details of the assembly bias mocks. On the left, we show the mean number of central and satellite galaxies as a function of halos mass and halo environment for both the strong assembly bias model (solid lines) and the weak assembly bias model (dashed lines). On the right, we show the ratios of the galaxy two-point correlation functions for assembly bias models, and their non-assembly bias counterpart.

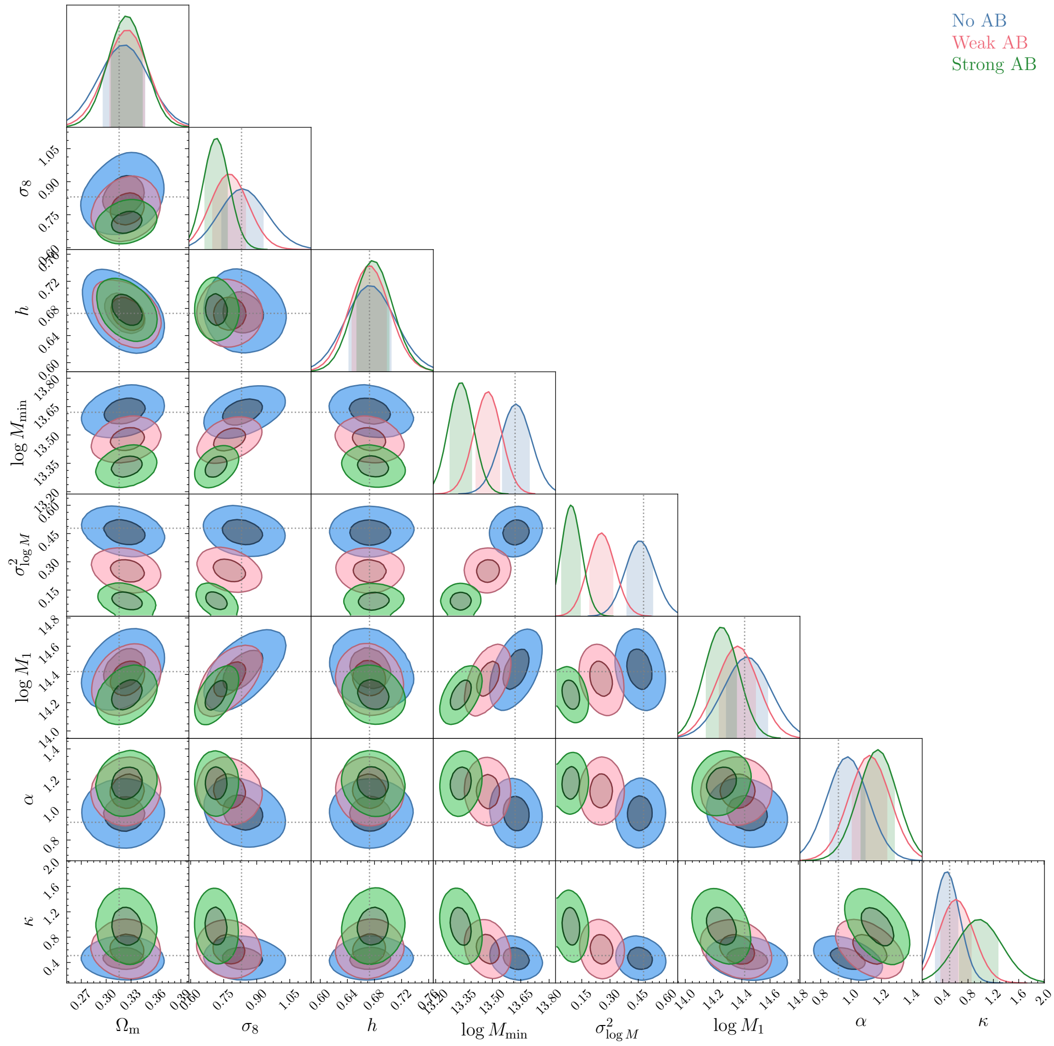


Figure B.5: Full 2D posteriors obtained for data with i) no assembly bias effect, ii) a weak assembly bias signal, and iii) a strong assembly bias signal.

Appendix C

Appendix: The information content of environment dependent clustering

C.1 The impact of mixing quintiles when estimating overdensities in redshift space

In this appendix, we examine the contribution to the quadrupole of quintile autocorrelations into the signal coming from random centres that have been correctly identified in redshift space, and those that have been misidentified.

Let us begin by defining the set of correctly identified random points for DS_i as

$$\mathrm{S} \cap \mathrm{R} = \left\{ \mathbf{x} \in \left(\mathrm{DS}_{\mathrm{i}}^{\mathrm{S}} \cap \mathrm{DS}_{\mathrm{i}}^{\mathrm{R}} \right) \right\}, \tag{C.1.1}$$

where subscript S and R, denote redshift and real space identification respectively. We denot those incorrectly identified as

$$\mathrm{S} \notin \mathrm{R} = \left\{ \mathbf{x} \in \left(\mathrm{DS}_{\mathrm{i}}^{\mathrm{S}} \cap \mathrm{DS}_{\mathrm{i}}^{\mathrm{R}} \right) \right\}. \tag{C.1.2}$$

For a given density split, DS_i, we separate the contribution to the quadrupole from the two sets as

$$\begin{aligned} \xi_2^{\mathrm{qq}} = & \left(\frac{|\mathrm{S} \cap \mathrm{R}|}{N_{\mathrm{random}}} \right)^2 \xi_2^{\mathrm{S} \cap \mathrm{R}} + \left(\frac{|\mathrm{S} \notin \mathrm{R}|}{N_{\mathrm{random}}} \right)^2 \xi_2^{\mathrm{S} \notin \mathrm{R}} \\ & + 2 \frac{|\mathrm{S} \cap \mathrm{R}||\mathrm{S} \notin \mathrm{R}|}{N_{\mathrm{random}}^2} \xi_2^{\mathrm{S} \cap \mathrm{R}, \mathrm{S} \notin \mathrm{R}} \end{aligned} \tag{C.1.3}$$

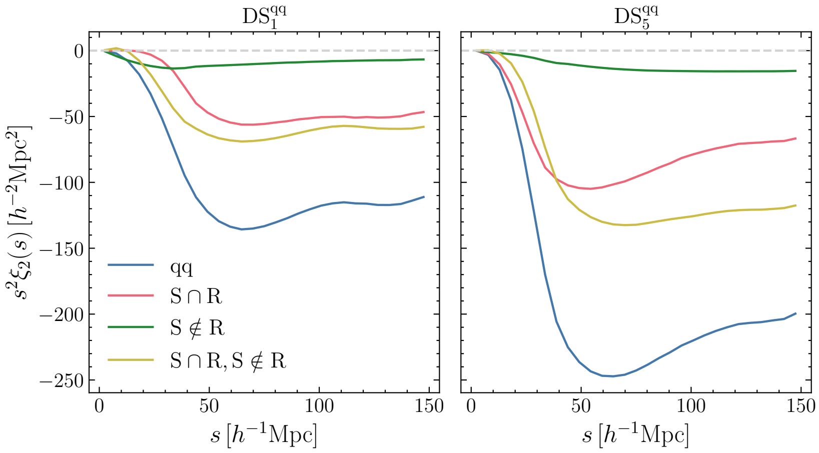


Figure C.1: The contribution from correctly ($\mathrm{S}\cap\mathrm{R}$) and incorrectly classified ($\mathrm{S}\notin\mathrm{R}$) random points to the quadrupole of autocorrelations. We show both the effect for DS_1 (left) and DS_5 (right), estimated for only one realization of the fiducial Quijote simulations.

where $|S\cap R|$ and $|S\notin R|$ are the number of points correctly and incorrectly identified, respectively. The first term in Eq. C.1.3 quantifies the anisotropy resulting from missing random centres that have not been correctly identified, the second term represents the contribution of anisotropies present in the random centres that have been incorrectly added, whereas the last term quantifies the cross-correlation between those centres that have been correctly identified and those that have been added.

Fig. C.1 shows the contribution of each term in Eq. C.1.3. For both DS_1 and DS_5, all terms contribute to the overall squashing of the autocorrelation. For DS_1, points that tend to be correctly classified in redshift space are those inside the void region, whereas those that are missed tend to be at the void boundary. The correctly classified centres, $\mathrm{S}\cap\mathrm{R}$, are therefore more clustered along the line of sight. Moreover, DS_2 points in real space classified as DS_1 in redshift space tend to also be located around void boundaries. Therefore, cross-correlation of these points with the correctly classified ones ($\mathrm{S}\cap\mathrm{R},\mathrm{S}\notin\mathrm{R}$) contributes to the enhanced clustering along the line of sight.

C.2 Assessing the Gaussianity of the density split likelihood

In this section, we check that the likelihood of density split statistics is indeed distributed as multivariate Gaussian following the analysis in Friedrich et al. (2021). We first compute the χ^2 value of the summary statistic measured in each of the fiducial simulations

$$\chi_i^2 = \left(d_i(\mathbf{s}) - \bar{d}(\mathbf{s})\right)^T C^{-1}\left(d_i(\mathbf{s}) - \bar{d}(\mathbf{s})\right), \qquad \text{(C.2.1)}$$

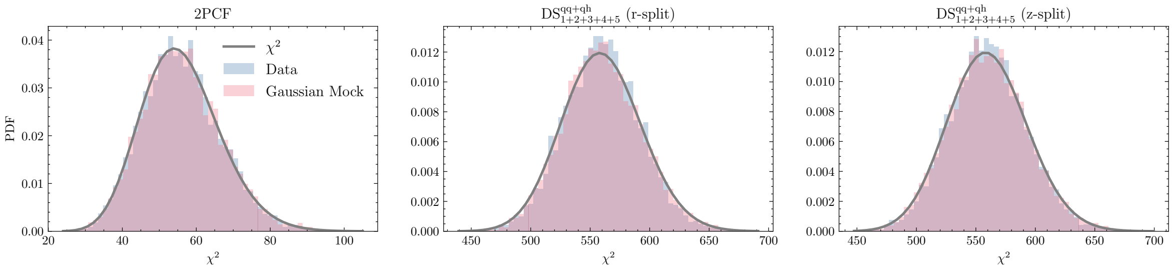


Figure C.2: Gaussianity tests. Left 2PCF, Right DS.

where d_i represents the value of the summary statistic for the i-th fiducial simulation evaluated at the pair separation vector $\mathbf{s}$, $\bar{d}(\mathbf{s})$ is the average of the summary statistic over all fiducial simulations at the pair separation vector $\mathbf{s}$, and C is the covariance matrix estimated from all the fiducial simulations.

If the likelihood of the summary statistic is Gaussian distributed, the χ^2_i values should also follow a χ^2 distribution with degrees of freedom determined by the number of pair-separation bins.

Furthermore, if the likelihood is Gaussian, the distribution of χ^2_i should also be very close to that of sampling from a multivariate Gaussian with a mean given by $\bar{d}$ and the covariance measured from the simulations.

In Figure C.2, we show how both the two-point correlation function and DS statistics χ^2_i computed from the data follow a very similar χ^2 distribution as that of the random samples generated from a multivariate Gaussian.

C.3 Density-split clustering in Gaussian random fields

In Sect. 5.4.2, we showed that density-split clustering leads to improved cosmological constraints when compared against the halo 2PCF. The small-scale halo density field at $z = 0.0$ in Quijote is highly non-Gaussian, and DS, which relies on measurements of correlation functions conditioned by density, is able to extract more information than the standard 2PCF. Here we compare the constraining power of DS and the 2PCF in a Gaussian random field, where the 2PCF, which is a measure of the variance of the field as a function of scale, fully describes its statistical properties. The purpose is to use the Gaussian mocks to develop an intuition for where the constraining power comes from in the case of DS clustering when analysing only large scales.

Starting from primordial power spectra with the same parameters as those described in Table. 5.1, we use MOCK FACTORY[1] to linearly evolve the density field to $z = 0.0$, and then sample a Gaussian random field of particles with a similar tracer bias as the Quijote haloes. We compute the 2PCF and DS correlation functions and estimate the Fisher matrix numerically as described in Sect. 5.3. For simplicity, all measurements are performed in real space, so that all information is contained in the monopole moment of the correlation functions.

Fig. C.3 shows the monopoles of the DS cross-correlation and autocorrelation functions, as computed from the Gaussian mocks. It can be seen that, under this setup, the collection of cross-correlation functions is symmetric around zero, reflecting the Gaussian nature of the density PDF. For the autocorrelation functions, the quintiles with negative bias (DS_1 & DS_2) exactly match those with positive bias (DS_4 & DS_5). In other words, the information from DS_1 & DS_2 is completely degenerate with that of DS_4 & DS_5, while DS_3 does not contribute with any significant information due to its zero mean. In order to avoid including duplicate information in the Fisher matrix, we estimate the DS Fisher matrix using only DS_1 & DS_2. We have verified that we get the same results if we choose to work with DS_4 & DS_5 instead.

Fig. C.4 compares the constraints on Ω_m, σ_8 and h from DS and the 2PCF, using a minimum scale of $s_{\mathrm{min}} = 10\,h^{-1}\mathrm{Mpc}$. It can be seen that DS leads to significantly improved constraints over the 2PCF. This may go against the intuition that DS should not be able to outperform the 2PCF in the Gaussian scenario. However, we need to keep in mind that the DS quintiles are defined in terms of the halo densities in spheres of radius $R_s = 20\,h^{-1}\mathrm{Mpc}$. This makes the DS quintiles sensitive to the smoothed density contrast within R_s, even if the multipoles are truncated at $s_{\mathrm{min}} = 10\,h^{-1}\mathrm{Mpc}$. To account for this, we measure the average density contrast in each DS quintile, $\Delta(R_s)$, and add this information to the 2PCF. It can be seen that the resulting constraints from this combination match much better the constraints from DS.

C.4 Convergence of Fisher forecasts

In Fig. C.5, we demonstrate that the constraints we obtain have converged when we vary the number of simulations used in the analysis to: i) estimate the derivatives respect to the cosmological parameters (left hand panel), and ii) estimate the covariance matrix (right hand panel).

[1] `https://github.com/cosmodesi/mockfactory`

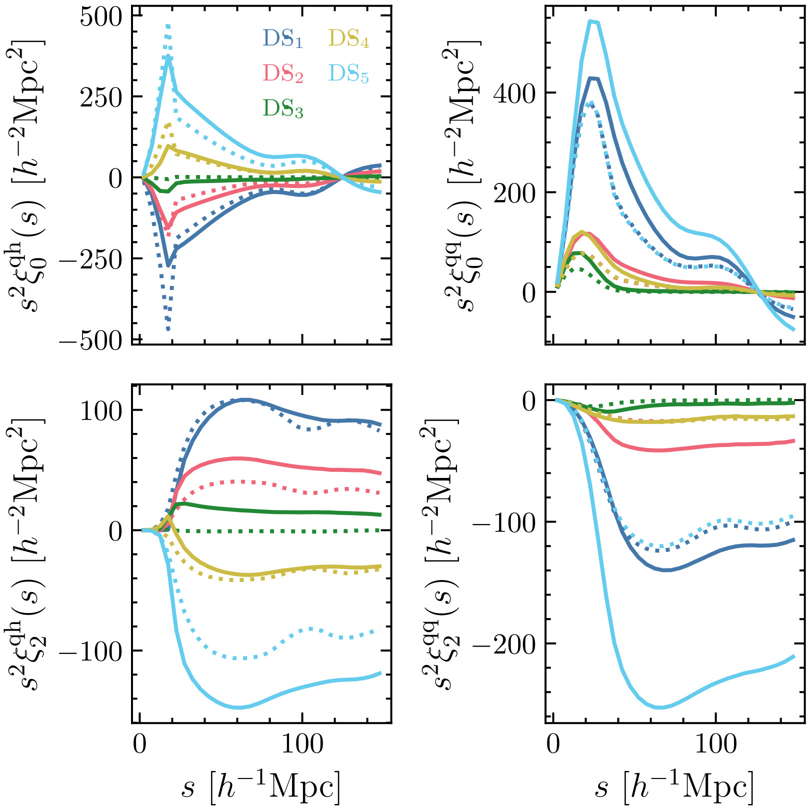


Figure C.3: A comparison of density split clustering measurements in the Quijote simulations (solid lines) and the Gaussian mocks (dotted lines). Note that in the Gaussian mocks, the density splits are symmetric.

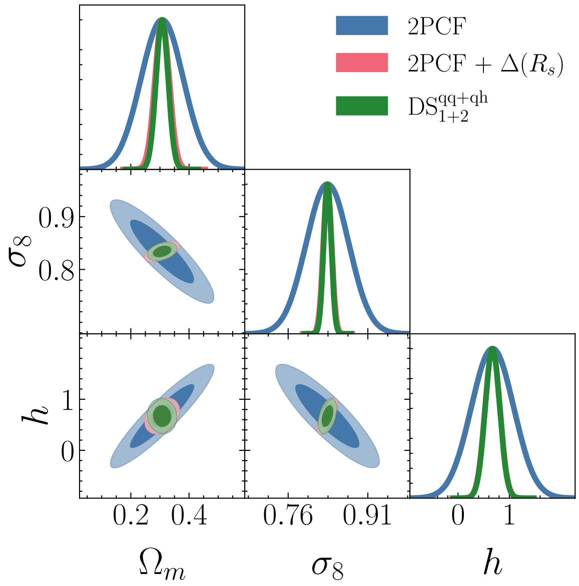


Figure C.4: Cosmological parameter constraints from DS and the 2PCF in a Gaussian random field, using scales down to $s_{\rm min} = 10\,h^{-1}$Mpc. Blue: constraints from the real-space halo 2PCF. Green: constraints from the combination of DS cross-correlation and autocorrelation functions in real space. Red: constraints from the combination of the halo 2PCF and the average density in DS quintiles. Note that DS-related quantities only make use of the first two quintiles, DS$_1$ & DS$_2$, which contain all the information if the density PDF is symmetric.

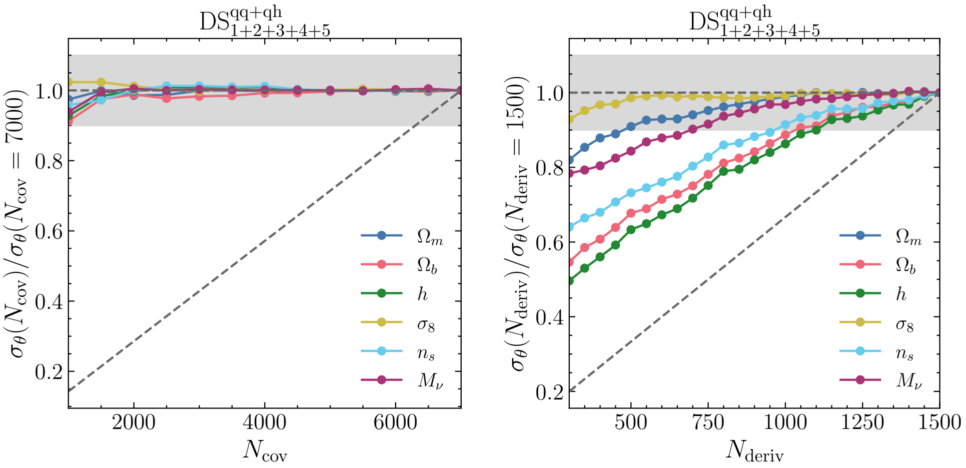


Figure C.5: Convergence of the constraints estimated through the Fisher matrix. The left-hand panel shows the 1-σ errors on each model parameter as a function of the number of mocks used to estimate the covariance matrix, where the errors are normalized by the default case when $N_{\mathrm{cov}} = 7000$. The right-hand panel shows the same convergence tests but for the number of mocks used to estimate the derivatives, where the errors are normalized by the default case $N = 1500$. The grey shaded bands show regions where the agreement is better than 10 per cent.

C.5 Impact of cosmology on the halo power spectrum

In Fig C.6 we show the impact of varying the cosmological parameters in the halo power spectrum measured from the Quijote simulations.

C.6 Density split constraints in r-split

In Fig. C.7, we show the contribution of each density split quintile to constrain the cosmological parameters, when DS quintiles are identified in real space.

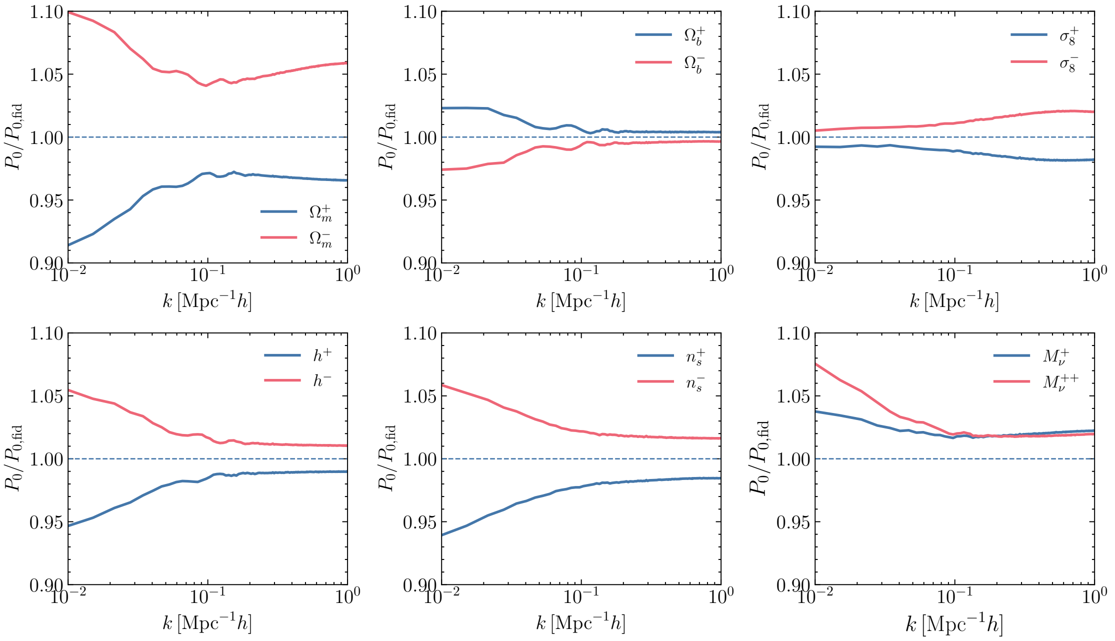


Figure C.6: The impact of changes in cosmological parameters on the halo power spectrum, as measured from the Quijote simulations. Each panel shows the ratio between the power spectrum in each cosmology and the fiducial one.

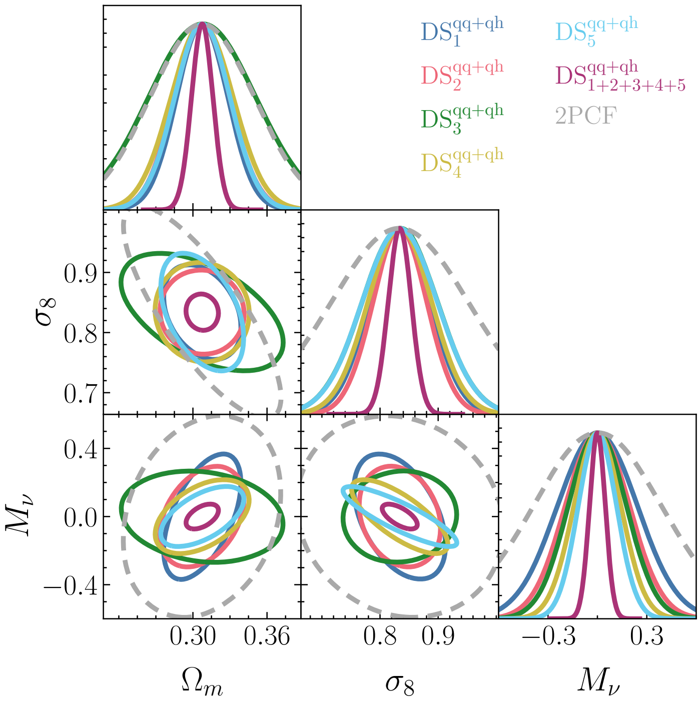


Figure C.7: Constraints on the cosmological parameters from individual and combined DS quintiles identified in real space (solid). The constraints from the two-point correlation function are shown by the grey, dashed contours for comparison.

Bibliography

Abbas, U., Sheth, R.K. *Strong clustering of underdense regions and the environmental dependence of clustering from Gaussian initial conditions*. MNRAS, **378**(2) (2007), 641. astro-ph/0703391.

Abbott, T.M.C., Aguena, M., Alarcon, A., et al. (DES Collaboration). *Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing*. Phys. Rev. D, **105** (2022), 023520. URL https://link.aps.org/doi/10.1103/PhysRevD.105.023520.

Abbott, T.M.C., Aguena, M., Alarcon, A., et al. *Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing*. Phys. Rev. D, **105**(2) (2022), 023520. 2105.13549.

Abidi, M.M., Baldauf, T. *Cubic halo bias in Eulerian and Lagrangian space*. J. Cosmology Astropart. Phys., **2018**(7) (2018), 029. 1802.07622.

Adamek, J., Daverio, D., Durrer, R., et al. *General relativity and cosmic structure formation*. Nature Phys., **12** (2016), 346. 1509.01699.

Adler, R.J. *The Geometry of Random Fields* (1981).

Agarap, A.F. *Deep learning using rectified linear units (relu)*. arXiv preprint arXiv:1803.08375 (2018).

Agarwal, S., Davé, R., Bassett, B.A. *Painting galaxies into dark matter haloes using machine learning*. MNRAS, **478**(3) (2018), 3410. 1712.03255.

Alam, S., Ata, M., Bailey, S., et al. *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample*. MNRAS, **470**(3) (2017), 2617. 1607.03155.

Alcock, C., Paczynski, B. *An evolution free test for non-zero cosmological constant.* Nature, **281** (1979), 358.

Amendola, L., Appleby, S., Bacon, D., et al. *Cosmology and Fundamental Physics with the Euclid Satellite.* Living Reviews in Relativity, **16**(1) (2013), 6. ISSN 1433-8351. URL https://doi.org/10.12942/lrr-2013-6.

Armijo, J., Baugh, C.M., Padilla, N.D., et al. *Making use of sub-resolution haloes in N-body simulations.* MNRAS, **510**(1) (2022), 29. 2111.11321.

Arnold, C., Li, B., Giblin, B., et al. *FORGE – the f(R) gravity cosmic emulator project I: Introduction and matter power spectrum emulator.* arXiv e-prints (2021), arXiv:2109.04984. 2109.04984.

Aylett-Bullock, J., Cuesta-Lazaro, C., Quera-Bofarull, A., et al. *J<span class="smallcaps smallerCapital">une</span>: open-source individual-based epidemiology simulation.* Royal Society Open Science, **8**(7) (2021a), 210506. https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.210506, URL https://royalsocietypublishing.org/doi/abs/10.1098/rsos.210506.

Aylett-Bullock, J., Cuesta-Lazaro, C., Quera-Bofarull, A., et al. *Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox's Bazar settlement.* PLOS Computational Biology, **17**(10) (2021b). URL http://dro.dur.ac.uk/35186/.

Azzalini, A., Capitanio, A. *Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution.* arXiv e-prints (2009), arXiv:0911.2342. 0911.2342.

Ba, S., Myers, W.R., Brenneman, W.A. *Optimal Sliced Latin Hypercube Designs.* Technometrics, **57**(4) (2015), 479. https://doi.org/10.1080/00401706.2014.957867, URL https://doi.org/10.1080/00401706.2014.957867.

Badrinarayanan, V., Kendall, A., Cipolla, R. *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation* (2015). URL https://arxiv.org/abs/1511.00561.

Baldauf, T., Seljak, U., Desjacques, V., et al. *Evidence for quadratic tidal tensor bias from the halo bispectrum.* Phys. Rev. D, **86**(8) (2012), 083540. 1201.4827.

Ballinger, W.E., Peacock, J.A., Heavens, A.F. *Measuring the cosmological constant with redshift surveys.* Mon. Not. Roy. Astron. Soc., **282** (1996), 877. astro-ph/9605017.

Banerjee, A., Abel, T. *Nearest neighbour distributions: New statistical measures for cosmological clustering.* MNRAS, **500**(4) (2021), 5479. 2007.13342.

Bardeen, J.M. *Gauge-invariant cosmological perturbations.* Phys. Rev. D, **22** (1980), 1882. URL https://link.aps.org/doi/10.1103/PhysRevD.22.1882.

Barrera-Hinojosa, C., Li, B. *GRAMSES: a new route to general relativistic N-body simulations in cosmology. Part I. Methodology and code description.* JCAP, **01** (2020), 007. 1905.08890.

Bautista, J.E., Vargas-Magañ a, M., Dawson, K.S., et al. *The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations at Redshift of 0.72 with the DR14 Luminous Red Galaxy Sample.* The Astrophysical Journal, **863**(1) (2018), 110. URL https://doi.org/10.3847%2F1538-4357%2Faacea5.

Behroozi, P., Knebe, A., Pearce, F.R., et al. *Major Mergers Going Notts: Challenges for Modern Halo Finders.* Mon. Not. Roy. Astron. Soc., **454**(3) (2015), 3020. 1506.01405.

Behroozi, P.S., Wechsler, R.H., Wu, H.Y. *The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores.* ApJ, **762**(2) (2013a), 109. 1110.4372.

Behroozi, P.S., Wechsler, R.H., Wu, H.Y. *The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores.* ApJ, **762**(2) (2013b), 109. 1110.4372.

Beisbart, C., Kerscher, M. *Luminosity- and Morphology-dependent Clustering of Galaxies.* ApJ, **545**(1) (2000), 6. astro-ph/0003358.

Benson, A.J., Cole, S., Frenk, C.S., et al. *The nature of galaxy bias and clustering.* Monthly Notices of the Royal Astronomical Society, **311**(4) (2000), 793. URL https://doi.org/10.1046%2Fj.1365-8711.2000.03101.x.

Bernardeau, F., Colombi, S., Gaztanaga, E., et al. *Large scale structure of the universe and cosmological perturbation theory.* Phys. Rept., **367** (2002), 1. astro-ph/0112551.

Bianchi, D., Chiesa, M., Guzzo, L. *Improving the modelling of redshift-space distortions - I. A bivariate Gaussian description for the galaxy pairwise velocity distributions.* MNRAS, **446**(1) (2015), 75. 1407.4753.

Bianchi, D., Percival, W.J., Bel, J. *Improving the modelling of redshift-space distortions- II. A pairwise velocity model covering large and small scales.* MNRAS, **463**(4) (2016), 3783. 1602.02780.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D. *Variational Inference: A Review for Statisticians.* Journal of the American Statistical Association, **112**(518) (2017), 859. URL https://doi.org/10.1080%2F01621459.2017.1285773.

Bonabeau, E. *Agent-based modeling: Methods and techniques for simulating human systems.* Proceedings of the National Academy of Sciences, **99**(suppl_3) (2002), 7280. https://www.pnas.org/doi/pdf/10.1073/pnas.082080899, URL https://www.pnas.org/doi/abs/10.1073/pnas.082080899.

Bonnaire, T., Aghanim, N., Kuruvilla, J., et al. *Cosmology with cosmic web environments.* Astronomy &amp Astrophysics, **661** (2022), A146. URL https://doi.org/10.1051%2F0004-6361%2F202142852.

Bonvin, C., Durrer, R. *What galaxy surveys really measure.* Physical Review D, **84** (2011). Arxiv:1105.5280v3, URL https://arxiv.org/abs/1105.5280.

Bose, B., Koyama, K. *A perturbative approach to the redshift space correlation function: beyond the Standard Model.* J. Cosmology Astropart. Phys., **2017**(8) (2017), 029. 1705.09181.

Bose, S., Eisenstein, D.J., Hernquist, L., et al. *Revealing the galaxy-halo connection in IllustrisTNG.* MNRAS, **490**(4) (2019), 5693. 1905.08799.

Bronstein, M.M., Bruna, J., Cohen, T., et al. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges.* arXiv e-prints (2021), arXiv:2104.13478. 2104.13478.

Buchner, J., Georgakakis, A., Nandra, K., et al. *X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue.* A&A, **564** (2014), A125. 1402.0004.

Bullock, J., Cuesta-Lázaro, C., Quera-Bofarull, A. *XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets.* In B. Gimi, A. Krol, editors, *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10953. International Society for Optics and Photonics, SPIE (2019), pages 453 – 463. URL https://doi.org/10.1117/12.2512451.

Burden, A., Percival, W.J., Howlett, C. *Reconstruction in Fourier space.* Monthly Notices of the Royal Astronomical Society, **453**(1) (2015), 456. URL https://doi.org/10.1093%2Fmnras%2Fstv1581.

Cai, Y.C., Taylor, A., Peacock, J.A., et al. *Redshift-space distortions around voids.* MNRAS, **462**(3) (2016), 2465. 1603.05184.

Calafut, V., Gallardo, P.A., Vavagiakis, E.M., et al. *The Atacama Cosmology Telescope: Detection of the pairwise kinematic Sunyaev-Zel'dovich effect with SDSS DR15 galaxies.* Phys. Rev. D, **104** (2021), 043502. URL https://link.aps.org/doi/10.1103/PhysRevD.104.043502.

Carlson, J., Reid, B., White, M. *Convolution Lagrangian perturbation theory for biased tracers.* MNRAS, **429**(2) (2013), 1674. 1209.0780.

Carlson, J., White, M., Padmanabhan, N. *Critical look at cosmological perturbation theory techniques.* Phys. Rev. D, **80**(4) (2009), 043531. 0905.0479.

Carrasco, J.J.M., Hertzberg, M.P., Senatore, L. *The effective field theory of cosmological large scale structures.* Journal of High Energy Physics, **2012** (2012), 82. 1206.2926.

Carroll, S. *Spacetime and Geometry: An Introduction to General Relativity.* Cambridge University Press (2019). ISBN 9781108775557. URL https://books.google.co.uk/books?id=1XSmDwAAQBAJ.

Carron, J. *ON THE INCOMPLETENESS OF THE MOMENT AND CORRELATION FUNCTION HIERARCHY AS PROBES OF THE LOGNORMAL FIELD.* The Astrophysical Journal, **738**(1) (2011), 86. URL https://doi.org/10.1088%2F0004-637x%2F738%2F1%2F86.

Carron, J. *On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices.* A&A, **551** (2013), A88. URL https://doi.org/10.1051/0004-6361/201220538.

Castorina, E., Feng, Y., Seljak, U., et al. *Primordial non-Gaussianities and zero bias tracers of the Large Scale Structure.* Physical Review Letters, **121** (2018), 101301. Arxiv:1803.11539v1, URL https://arxiv.org/abs/1803.11539.

Challinor, A., Lewis, A. *The linear power spectrum of observed source number counts.* Physical Review D, **84** (2011). Arxiv:1105.5292v2, URL https://arxiv.org/abs/1105.5292.

Chan, K.C., Scoccimarro, R., Sheth, R.K. *Gravity and large-scale nonlocal bias.* Phys. Rev. D, **85**(8) (2012), 083509. 1201.3614.

Charnock, T., Lavaux, G., Wandelt, B.D. *Automatic physical inference with information maximizing neural networks.* Phys. Rev. D, **97**(8) (2018), 083004. 1802.03537.

Chen, S.F., Vlah, Z., Castorina, E., et al. *Redshift-space distortions in Lagrangian perturbation theory.* J. Cosmology Astropart. Phys., **2021**(3) (2021), 100. 2012.04636.

Chen, S.F., Vlah, Z., White, M. *Consistent modeling of velocity statistics and redshift-space distortions in one-loop perturbation theory.* J. Cosmology Astropart. Phys., **2020**(7) (2020), 062. 2005.00523.

Chen, S.F., Vlah, Z., White, M. *A new analysis of galaxy 2-point functions in the BOSS survey, including full-shape information and post-reconstruction BAO.* Journal of Cosmology and Astroparticle Physics, **2022**(02) (2022), 008. URL https://doi.org/10.1088%2F1475-7516%2F2022%2F02%2F008.

Chuang, C.H., Kitaura, F.S., Liang, Y., et al. *Linear redshift space distortions for cosmic voids based on galaxies in redshift space.* Phys. Rev. D, **95** (2017), 063528. URL https://link.aps.org/doi/10.1103/PhysRevD.95.063528.

Clifton, T., Ferreira, P.G., Padilla, A., et al. *Modified gravity and cosmology.* Physics Reports, **513**(1) (2012), 1. ISSN 0370-1573. Modified Gravity and Cosmology, URL https://www.sciencedirect.com/science/article/pii/S0370157312000105.

Commons, W. *File:PowerSpectrumExt.svg — Wikimedia Commons, the free media repository* (2022). [Online; accessed 24-July-2022], URL https://commons.wikimedia.org/w/index.php?title=File:PowerSpectrumExt.svg&oldid=635090300.

Cooray, A., Sheth, R. *Halo models of large scale structure.* Phys. Rep., **372**(1) (2002), 1. astro-ph/0206508.

Cooray, A., Sheth, R. *Halo models of large scale structure.* Physics Reports, **372**(1) (2002), 1. URL https://doi.org/10.1016%2Fs0370-1573%2802%2900276-4.

Cramér, H. *Mathematical methods of statistics*, volume 9 of *Princeton Math. Ser.* Princeton University Press, Princeton, NJ (1946).

Cranmer, M., Melchior, P., Nord, B. *Unsupervised Resource Allocation with Graph Neural Networks.* In *34th Conference on Neural Information Processing Systems* (2021). 2106.09761.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., et al. *Discovering Symbolic Models from Deep Learning with Inductive Biases.* arXiv e-prints (2020), arXiv:2006.11287. 2006.11287.

Crocce, M., Pueblas, S., Scoccimarro, R. *Transients from initial conditions in cosmological simulations.* Monthly Notices of the Royal Astronomical Society, **373**(1) (2006), 369. URL https://doi.org/10.1111%2Fj.1365-2966.2006.11040.x.

Crocce, M., Pueblas, S., Scoccimarro, R. *Transients from initial conditions in cosmological simulations.* MNRAS, **373**(1) (2006), 369. astro-ph/0606505.

Crocce, M., Scoccimarro, R. *Memory of initial conditions in gravitational clustering.* Phys. Rev. D, **73**(6) (2006), 063520. astro-ph/0509419.

Croton, D.J., Gao, L., White, S.D.M. *Halo assembly bias and its effects on galaxy clustering.* MNRAS, **374**(4) (2007), 1303. astro-ph/0605636.

Cuesta-Lazaro, C., Li, B., Eggemeier, A., et al. *Towards a non-Gaussian model of redshift space distortions.* MNRAS, **498**(1) (2020), 1175. ISSN 0035-8711. https://academic.oup.com/mnras/article-pdf/498/1/1175/33731014/staa2249.pdf, URL https://doi.org/10.1093/mnras/staa2249.

Dai, B., Seljak, U. *Translation and Rotation Equivariant Normalizing Flow (TRENF) for Optimal Cosmological Analysis.* arXiv e-prints (2022), arXiv:2202.05282. 2202.05282.

Davis, M., Efstathiou, G., Frenk, C.S., et al. *The evolution of large-scale structure in a universe dominated by cold dark matter.* ApJ, **292** (1985), 371.

de Jong, R.S., Barden, S.C., Bellido-Tirado, O., et al. *4MOST: the 4-metre multi-object spectroscopic telescope project at final design review (Conference Presentation).* In C.J. Evans, L. Simard, H. Takami, editors, *Ground-based and Airborne Instrumentation for Astronomy VII*, volume 10702. International Society for Optics and Photonics, SPIE (2018). URL https://doi.org/10.1117/12.2312012.

DeRose, J., Wechsler, R.H., Tinker, J.L., et al. *The AEMULUS Project. I. Numerical Simulations for Precision Cosmology.* ApJ, **875**(1) (2019a), 69. 1804.05865.

DeRose, J., Wechsler, R.H., Tinker, J.L., et al. *The AEMULUS Project. I. Numerical Simulations for Precision Cosmology.* ApJ, **875**(1) (2019b), 69. 1804.05865.

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. *The DESI Experiment Part I: Science,Targeting, and Survey Design* (2016a). URL https://arxiv.org/abs/1611.00036.

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. *The DESI Experiment Part I: Science,Targeting, and Survey Design.* arXiv e-prints (2016b), arXiv:1611.00036. 1611.00036.

Desjacques, V., Jeong, D., Schmidt, F. *Large-Scale Galaxy Bias.* Phys. Rept., **733** (2018a), 1. 1611.09787.

Desjacques, V., Jeong, D., Schmidt, F. *Large-scale galaxy bias.* Physics Reports, **733** (2018b), 1. URL https://doi.org/10.1016%2Fj.physrep.2017.12.002.

Di Valentino, E., Mena, O., Pan, S., et al. *In the realm of the Hubble tension-a review of solutions.* Classical and Quantum Gravity, **38**(15) (2021), 153001. 2103.01183.

Diemer, B., Joyce, M. *An Accurate Physical Model for Halo Concentrations.* The Astrophysical Journal, **871**(2) (2019), 168. URL https://doi.org/10.3847%2F1538-4357%2Faafad6.

Duane, S., Kennedy, A.D., Pendleton, B.J., et al. *Hybrid Monte Carlo.* Phys. Lett. B, **195** (1987), 216.

Dupuy, A., Courtois, H.M., Kubik, B. *An estimation of the local growth rate from Cosmicflows peculiar velocities.* MNRAS, **486**(1) (2019), 440. 1901.03530.

Fisher, K.B. *On the Validity of the Streaming Model for the Redshift-Space Correlation Function in the Linear Regime.* ApJ, **448** (1995a), 494. astro-ph/9412081.

Fisher, K.B. *On the Validity of the Streaming Model for the Redshift-Space Correlation Function in the Linear Regime.* ApJ, **448** (1995b), 494. astro-ph/9412081.

Fisher, R.A. *The Logic of Inductive Inference.* Journal of the Royal Statistical Society, **98**(1) (1935), 39. ISSN 09528385. URL http://www.jstor.org/stable/2342435.

Fluri, J., Kacprzak, T., Lucchi, A., et al. *Cosmological constraints with deep learning from KiDS-450 weak lensing maps.* Phys. Rev. D, **100** (2019), 063514. URL https://link.aps.org/doi/10.1103/PhysRevD.100.063514.

Fluri, J., Kacprzak, T., Refregier, A., et al. *Cosmological constraints from noisy convergence maps through deep learning.* Phys. Rev. D, **98**(12) (2018), 123518. 1807.08732.

Fluri, J., Lucchi, A., Kacprzak, T., et al. *Cosmological parameter estimation and inference using deep summaries.* Phys. Rev. D, **104**(12) (2021), 123526. 2107.09002.

Forman, W., Schwarz, J., Jones, C., et al. *X-ray observations of galaxies in the Virgo cluster.* ApJ, **234** (1979), L27.

Freedman, W.L. *Measurements of the Hubble Constant: Tensions in Perspective.* Astrophys. J., **919**(1) (2021), 16. 2106.15656.

Friedrich, O., Andrade-Oliveira, F., Camacho, H., et al. *Dark Energy Survey year 3 results: covariance modelling and its impact on parameter estimation and quality of fit.* Monthly Notices of the Royal Astronomical Society, **508**(3) (2021), 3125. ISSN 0035-8711. https://academic.oup.com/mnras/article-pdf/508/3/3125/40736161/stab2384.pdf, URL https://doi.org/10.1093/mnras/stab2384.

Gao, L., Springel, V., White, S.D.M. *The age dependence of halo clustering.* MNRAS, **363**(1) (2005a), L66. astro-ph/0506510.

Gao, L., Springel, V., White, S.D.M. *The age dependence of halo clustering.* MNRAS, **363**(1) (2005b), L66. astro-ph/0506510.

Gao, L., White, S.D.M. *Assembly bias in the clustering of dark matter haloes.* MNRAS, **377**(1) (2007a), L5. astro-ph/0611921.

Gao, L., White, S.D.M. *Assembly bias in the clustering of dark matter haloes.* MNRAS, **377**(1) (2007b), L5. astro-ph/0611921.

Garrison, L.H., Eisenstein, D.J., Ferrer, D., et al. *Improving Initial Conditions for Cosmological N-Body Simulations.* Mon. Not. Roy. Astron. Soc., **461**(4) (2016), 4125. 1605.02333.

Garrison, L.H., Eisenstein, D.J., Ferrer, D., et al. *The ABACUS cosmological N-body code.* MNRAS, **508**(1) (2021), 575. 2110.11392.

Gil-Marín, H., Percival, W.J., Brownstein, J.R., et al. *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies.* Monthly Notices of the Royal Astronomical Society, **460**(4) (2016), 4188. URL https://doi.org/10.1093%2Fmnras%2Fstw1096.

Glöckler, M., Deistler, M., Macke, J.H. *Variational methods for simulation-based inference.* arXiv e-prints (2022), arXiv:2203.04176. 2203.04176.

Gómez, J.S., Padilla, N.D., Helly, J.C., et al. *Halo merger tree comparison: impact on galaxy formation models.* MNRAS, **510**(4) (2022), 5500. 2106.12664.

Gorski, K. *On the Pattern of Perturbations of the Hubble Flow.* ApJ, **332** (1988), L7.

Green, J., Schechter, P., Baltay, C., et al. *Wide-Field InfraRed Survey Telescope (WFIRST) Final Report* (2012). URL `https://arxiv.org/abs/1208.4012`.

Grove, C., Chuang, C.H., Devi, N.C., et al. *The DESI N-body simulation project – I. Testing the robustness of simulations for the DESI dark time survey.* Monthly Notices of the Royal Astronomical Society, **515**(2) (2022), 1854. ISSN 0035-8711. `https://academic.oup.com/mnras/article-pdf/515/2/1854/45093659/stac1947.pdf`, URL `https://doi.org/10.1093/mnras/stac1947`.

Grove, C., et al. *The DESI N-body Simulation Project I: Testing the Robustness of Simulations for the DESI Dark Time Survey* (2021). `2112.09138`.

Gupta, A., Matilla, J.M.Z., Hsu, D., et al. *Non-Gaussian information from weak lensing data via deep learning.* Phys. Rev. D, **97**(10) (2018), 103515. `1802.01212`.

Guth, A.H. *Inflationary universe: A possible solution to the horizon and flatness problems.* Phys. Rev. D, **23** (1981), 347. URL `https://link.aps.org/doi/10.1103/PhysRevD.23.347`.

Hadzhiyska, B., Liu, S., Somerville, R.S., et al. *Galaxy assembly bias and large-scale distribution: a comparison between IllustrisTNG and a semi-analytic model.* MNRAS, **508**(1) (2021), 698. `2108.00006`.

Hahn, C., Villaescusa-Navarro, F. *Constraining M sub/sub with the bispectrum. Part II. The information content of the galaxy bispectrum monopole.* Journal of Cosmology and Astroparticle Physics, **2021**(04) (2021), 029. URL `https://doi.org/10.1088%2F1475-7516%2F2021%2F04%2F029`.

Hahn, C., Villaescusa-Navarro, F., Castorina, E., et al. *Constraining M_ν with the bispectrum. Part I. Breaking parameter degeneracies.* J. Cosmology Astropart. Phys., **2020**(3) (2020a), 040. `1909.11107`.

Hahn, C., Villaescusa-Navarro, F., Castorina, E., et al. *Constraining M_ν with the bispectrum. Part I. Breaking parameter degeneracies.* J. Cosmology Astropart. Phys., **2020**(3) (2020b), 040. `1909.11107`.

Hamilton, A.J.S. *Linear Redshift Distortions: a Review.* In D. Hamilton, editor, *The Evolving Universe*, volume 231 of *Astrophysics and Space Science Library* (1998), page 185. astro-ph/9708102.

Hartlap, J., Simon, P., Schneider, P. *Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix.* A&A, **464**(1) (2007), 399. astro-ph/0608064.

Hastings, W.K. *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, **57**(1) (1970), 97. http://biomet.oxfordjournals.org/cgi/reprint/57/1/97.pdf, URL http://biomet.oxfordjournals.org/cgi/content/abstract/57/1/97.

He, S., Li, Y., Feng, Y., et al. *Learning to predict the cosmological structure formation.* Proceedings of the National Academy of Science, **116**(28) (2019), 13825. 1811.06533.

Heitmann, K., Lawrence, E., Kwan, J., et al. *THE COYOTE UNIVERSE EXTENDED: PRECISION EMULATION OF THE MATTER POWER SPECTRUM.* The Astrophysical Journal, **780**(1) (2013), 111. URL https://doi.org/10.1088%2F0004-637x%2F780%2F1%2F111.

Hendrycks, D., Gimpel, K. *Gaussian Error Linear Units (GELUs).* arXiv e-prints (2016), arXiv:1606.08415. 1606.08415.

Hey, T., Hey, A., Tansley, S., et al. *The Fourth Paradigm: Data-intensive Scientific Discovery.* Microsoft Research (2009). ISBN 9780982544204. URL https://books.google.co.uk/books?id=oGs_AQAAIAAJ.

Heymans, C., Tröster, T., Asgari, M., et al. *KiDS-1000 Cosmology: Multi-probe weak gravitational lensing and spectroscopic galaxy clustering constraints.* A&A, **646** (2021), A140. 2007.15632.

Huterer, D., Takada, M. *Calibrating the nonlinear matter power spectrum: Requirements for future weak lensing surveys.* Astroparticle Physics, **23**(4) (2005), 369. astro-ph/0412142.

Huterer, D., et al. *Growth of Cosmic Structure: Probing Dark Energy Beyond Expansion.* Astropart. Phys., **63** (2015), 23. 1309.5385.

Jeffrey, N., Lanusse, F., Lahav, O., et al. *Deep learning dark matter map reconstructions from DES SV weak lensing data.* Mon. Not. Roy. Astron. Soc., **492**(4) (2020), 5023. 1908.00543.

Jenkins, A. *Second-order Lagrangian perturbation theory initial conditions for resimulations.* Mon. Not. Roy. Astron. Soc., **403** (2010), 1859. 0910.0258.

Jennings, E., Baugh, C.M., Pascoli, S. *Testing Gravity Using the Growth of Large-scale Structure in the Universe.* ApJ, **727**(1) (2011), L9. 1011.2842.

Jiménez, E., Padilla, N., Contreras, S., et al. *The assembly bias of emission-line galaxies.* MNRAS, **506**(3) (2021), 3155. 2010.08500.

Joudaki, S., Blake, C., Heymans, C., et al. *CFHTLenS revisited: assessing concordance with Planck including astrophysical systematics.* Monthly Notices of the Royal Astronomical Society, **465**(2) (2016), 2033. ISSN 0035-8711. https://academic.oup.com/mnras/article-pdf/465/2/2033/8364676/stw2665.pdf, URL https://doi.org/10.1093/mnras/stw2665.

Joudaki, S., Blake, C., Heymans, C., et al. *CFHTLenS revisited: assessing concordance with Planck including astrophysical systematics.* MNRAS, **465**(2) (2017), 2033. 1601.05786.

Joyce, A., Jain, B., Khoury, J., et al. *Beyond the cosmological standard model.* Physics Reports, **568** (2015), 1. ISSN 0370-1573. Beyond the cosmological standard model, URL https://www.sciencedirect.com/science/article/pii/S0370157314004487.

Jumper, J., Evans, R., Pritzel, A., et al. *Highly accurate protein structure prediction with AlphaFold.* Nature, **596**(7873) (2021), 583.

Juszkiewicz, R., Fisher, K.B., Szapudi, I. *Skewed Exponential Pairwise Velocities from Gaussian Initial Conditions.* The Astrophysical Journal, **504**(1) (1998), L1. URL https://doi.org/10.1086/311558.

Kaiser, N. *Clustering in real space and in redshift space.* MNRAS, **227** (1987), 1.

Kim, A.G., Linder, E.V. *Complementarity of peculiar velocity surveys and redshift space distortions for testing gravity.* Physical Review D, **101**(2) (2020). URL https://doi.org/10.1103%2Fphysrevd.101.023516.

Kingma, D.P., Ba, J. *Adam: A Method for Stochastic Optimization.* arXiv e-prints (2014), arXiv:1412.6980. 1412.6980.

Kitano, H. *Nobel Turing Challenge: creating the engine for scientific discovery.* npj Systems Biology and Applications, **7**(1) (2021), 1.

Kobayashi, Y., Nishimichi, T., Takada, M., et al. *Accurate emulator for the redshift-space power spectrum of dark matter halos and its application to galaxy power spectrum.* Phys. Rev. D, **102** (2020), 063504. URL `https://link.aps.org/doi/10.1103/PhysRevD.102.063504`.

Kobayashi, Y., Nishimichi, T., Takada, M., et al. *Cosmological information content in redshift-space power spectrum of SDSS-like galaxies in the quasinonlinear regime up to* k *=0.3* h *Mpc*$^{-1}$. Phys. Rev. D, **101**(2) (2020), 023510. `1907.08515`.

Kobayashi, Y., Nishimichi, T., Takada, M., et al. *Cosmological information content in redshift-space power spectrum of SDSS-like galaxies in the quasinonlinear regime up to mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML" display="inline"mml:mrowmml:mik/mml:mimml:mo=/mml:momml:mn0.3/mml:mnmml:mtext /mml:mtextmml:mtext /mml:mtextmml:mih/mml:mimml:mtext /mml:mtextmml:mrowmml:msupmml:mrowmml:miMpc/mml:mi/mml:mrowmml:mrowmml:mo- /mml:momml:mn1/mml:mn/mml:mrow/mml:msup/mml:mrow/mml:mrow/mml:math.* Physical Review D, **101**(2) (2020). URL `https://doi.org/10.1103%2Fphysrevd.101.023510`.

Kobayashi, Y., Nishimichi, T., Takada, M., et al. *Full-shape cosmology analysis of the SDSS-III BOSS galaxy power spectrum using an emulator-based halo model: A 5% determination of σ8.* Phys. Rev. D, **105**(8) (2022), 083517. `2110.06969`.

Kreisch, C.D., Pisani, A., Carbone, C., et al. *Massive neutrinos leave fingerprints on cosmic voids.* Monthly Notices of the Royal Astronomical Society, **488**(3) (2019), 4413. URL `https://doi.org/10.1093%2Fmnras%2Fstz1944`.

Kreisch, C.D., Pisani, A., Villaescusa-Navarro, F., et al. *The GIGANTES dataset: precision cosmology from voids in the machine learning era* (2021). URL `https://arxiv.org/abs/2107.02304`.

Kuhlen, M., Vogelsberger, M., Angulo, R. *Numerical Simulations of the Dark Universe: State of the Art and the Next Decade.* Phys. Dark Univ., **1** (2012), 50. `1209.5745`.

Kuruvilla, J., Porciani, C. *On the streaming model for redshift-space distortions.* MNRAS, **479**(2) (2018), 2256. `1710.09379`.

Lange, J.U., Hearin, A.P., Leauthaud, A., et al. *Five per cent measurements of the growth rate from simulation-based modelling of redshift-space clustering in BOSS LOWZ.* Monthly

Notices of the Royal Astronomical Society, **509**(2) (2021), 1779. URL https://doi.org/10.1093%2Fmnras%2Fstab3111.

Lange, J.U., van den Bosch, F.C., Zentner, A.R., et al. *Cosmological Evidence Modelling: a new simulation-based approach to constrain cosmology on non-linear scales.* MNRAS, **490**(2) (2019), 1870. 1909.03107.

Laureijs, R., Amiaux, J., Arduini, S., et al. (Euclid). *Euclid Definition Study Report.* ArXiv e-prints (2011). 1110.3193.

Laureijs, R., Amiaux, J., Arduini, S., et al. *Euclid Definition Study Report.* arXiv e-prints (2011), arXiv:1110.3193. 1110.3193.

Lavin, A., Zenil, H., Paige, B., et al. *Simulation Intelligence: Towards a New Generation of Scientific Methods.* arXiv preprint arXiv:2112.03235 (2021).

Lazeyras, T., Schmidt, F. *Beyond LIMD bias: a measurement of the complete set of third-order halo bias parameters.* J. Cosmology Astropart. Phys., **2018**(9) (2018), 008. 1712.07531.

Lazeyras, T., Wagner, C., Baldauf, T., et al. *Precision measurement of the local bias of dark matter halos.* J. Cosmology Astropart. Phys., **2016**(2) (2016), 018. 1511.01096.

Leclercq, F., Heavens, A. *On the accuracy and precision of correlation functions and field-level inference in cosmology.* MNRAS, **506**(1) (2021), L85. 2103.04158.

Leclercq, F., Pisani, A., Wandelt, B.D. *Cosmology: from theory to data, from data to theory.* arXiv e-prints (2014), arXiv:1403.1260. 1403.1260.

LeCun, Y., Bengio, Y., Hinton, G. *Deep learning.* nature, **521**(7553) (2015), 436.

Lehmann, E.L., Casella, G. *Theory of Point Estimation.* Springer-Verlag, New York, NY, USA, second edition (1998).

Lemos, P., Jeffrey, N., Cranmer, M., et al. *Rediscovering orbital mechanics with machine learning.* arXiv e-prints (2022), arXiv:2202.02306. 2202.02306.

Levi, M.E., et al. (DESI). *The Dark Energy Spectroscopic Instrument (DESI)* (2019). 1907.10688.

Li, C., Jing, Y.P., Kauffmann, G., et al. *The dependence of the pairwise velocity dispersion on galaxy properties.* MNRAS, **368**(1) (2006), 37. astro-ph/0509874.

Lin, Z., Huang, N., Avestruz, C., et al. *DeepSZ: identification of Sunyaev–Zel'dovich galaxy clusters using deep learning*. Mon. Not. Roy. Astron. Soc., **507**(3) (2021), 4149. 2102.13123.

Linde, A., Mezhlumian, A. *Inflation with* $\Omega{\neq}1$. Phys. Rev. D, **52**(12) (1995), 6789. astro-ph/9506017.

Linde, A.D. *A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems*. Physics Letters B, **108**(6) (1982), 389.

Linder, E.V. *Cosmic growth history and expansion history*. Phys. Rev. D, **72**(4) (2005), 043529. astro-ph/0507263.

Loveday, J., Christodoulou, L., Norberg, P., et al. *Galaxy and Mass Assembly (GAMA): small-scale anisotropic galaxy clustering and the pairwise velocity dispersion of galaxies*. MNRAS, **474**(3) (2018), 3435. 1711.05636.

Lovelock, D. *The uniqueness of the Einstein field equations in a four-dimensional space*. Archive for Rational Mechanics and Analysis, **33**(1) (1969), 54.

Ludlow, A.D., Bose, S., Angulo, R.E., et al. *The mass-concentration-redshift relation of cold and warm dark matter haloes*. MNRAS, **460**(2) (2016), 1214. 1601.02624.

Maksimova, N.A., Garrison, L.H., Eisenstein, D.J., et al. *ABACUSSUMMIT: a massive set of high-accuracy, high-resolution N-body simulations*. MNRAS, **508**(3) (2021a), 4017. 2110.11398.

Maksimova, N.A., Garrison, L.H., Eisenstein, D.J., et al. *ABACUSSUMMIT: a massive set of high-accuracy, high-resolution N-body simulations*. MNRAS, **508**(3) (2021b), 4017. 2110.11398.

Massara, E., Villaescusa-Navarro, F., Hahn, C., et al. *Cosmological Information in the Marked Power Spectrum of the Galaxy Field* (2022). URL https://arxiv.org/abs/2206.01709.

Matsubara, T. *Nonlinear perturbation theory with halo bias and redshift-space distortions via the Lagrangian picture*. Phys. Rev. D, **78**(8) (2008), 083519. 0807.1733.

McCulloch, W., Pitts, W. *A Logical Calculus of Ideas Immanent in Nervous Activity*. Bulletin of Mathematical Biophysics, **5** (1943), 127.

McDonald, P., Seljak, U. *How to measure redshift-space distortions without sample variance.* Journal of Cosmology and Astroparticle Physics, **2009** (2008), 007. Arxiv:0810.0323v1, URL https://arxiv.org/abs/0810.0323.

McDonald, P., Seljak, U. *How to evade the sample variance limit on measurements of redshift-space distortions.* Journal of Cosmology and Astroparticle Physics, **2009**(10) (2009), 007. URL https://doi.org/10.1088%2F1475-7516%2F2009%2F10%2F007.

Metropolis, N., Ulam, S. *The Monte Carlo Method.* Journal of the American Statistical Association, **44**(247) (1949), 335. PMID: 18139350, https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310, URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310.

Miyatake, H., Kobayashi, Y., Takada, M., et al. *Cosmological inference from emulator based halo model I: Validation tests with HSC and SDSS mock catalogs.* arXiv e-prints (2020), arXiv:2101.00113. 2101.00113.

Miyatake, H., More, S., Takada, M., et al. *Evidence of Halo Assembly Bias in Massive Clusters.* Phys. Rev. Lett., **116**(4) (2016), 041301. 1506.06135.

Miyatake, H., Sugiyama, S., Takada, M., et al. *Cosmological inference from the emulator based halo model II: Joint analysis of galaxy-galaxy weak lensing and galaxy clustering from HSC-Y1 and SDSS.* arXiv e-prints (2021), arXiv:2111.02419. 2111.02419.

More, S., Miyatake, H., Mandelbaum, R., et al. *THE WEAK LENSING SIGNAL AND THE CLUSTERING OF BOSS GALAXIES. II. ASTROPHYSICAL AND COSMOLOGICAL CONSTRAINTS.* The Astrophysical Journal, **806**(1) (2015), 2. URL https://doi.org/10.1088%2F0004-637x%2F806%2F1%2F2.

Mukhanov, V.F., Feldman, H.A., Brandenberger, R.H. *Theory of cosmological perturbations.* Phys. Rep., **215**(5-6) (1992), 203.

Nadathur, S., Carter, P., Percival, W.J. *A Zeldovich reconstruction method for measuring redshift space distortions using cosmic voids.* Monthly Notices of the Royal Astronomical Society, **482**(2) (2018), 2459. ISSN 0035-8711. https://academic.oup.com/mnras/article-pdf/482/2/2459/26576217/sty2799.pdf, URL https://doi.org/10.1093/mnras/sty2799.

Nadathur, S., Carter, P.M., Percival, W.J., et al. *Beyond BAO: Improving cosmological constraints from BOSS data with measurement of the void-galaxy cross-correlation.* Phys. Rev. D, **100**(2) (2019), 023504. 1904.01030.

Nadathur, S., Carter, P.M., Percival, W.J., et al. *Beyond BAO: Improving cosmological constraints from BOSS data with measurement of the void-galaxy cross-correlation.* Phys. Rev. D, **100**(2) (2019), 023504. 1904.01030.

Nadathur, S., Percival, W.J. *An accurate linear model for redshift space distortions in the void-galaxy correlation function.* MNRAS, **483**(3) (2019), 3472. 1712.07575.

Nadathur, S., Woodfinden, A., Percival, W.J., et al. *The completed SDSS-IV extended baryon oscillation spectroscopic survey: geometry and growth from the anisotropic void-galaxy correlation function in the luminous red galaxy sample.* MNRAS, **499**(3) (2020), 4140. 2008.06060.

Naidoo, K., Massara, E., Lahav, O. *Cosmology and neutrino mass with the minimum spanning tree.* Monthly Notices of the Royal Astronomical Society, **513**(3) (2022), 3596. URL https://doi.org/10.1093%2Fmnras%2Fstac1138.

Navarro, J.F., Frenk, C.S., White, S.D.M. *A Universal Density Profile from Hierarchical Clustering.* ApJ, **490**(2) (1997), 493. astro-ph/9611107.

Nishimichi, T., Takada, M., Takahashi, R., et al. *Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy Clustering.* ApJ, **884**(1) (2019a), 29. 1811.09504.

Nishimichi, T., Takada, M., Takahashi, R., et al. *Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy Clustering.* ApJ, **884**(1) (2019b), 29. 1811.09504.

Nishimichi, T., et al. *Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy Clustering.* Astrophys. J., **884** (2019), 29. 1811.09504.

Ntampaka, M., Eisenstein, D.J., Yuan, S., et al. *A Hybrid Deep Learning Approach to Cosmological Constraints from Galaxy Redshift Surveys.* ApJ, **889**(2) (2020), 151. 1909.10527.

Ntampaka, M., Vikhlinin, A. *The Importance of Being Interpretable: Toward an Understandable Machine Learning Encoder for Galaxy Cluster Cosmology.* Astrophys. J., **926**(1) (2022), 45. 2112.05768.

Nusser, A., Davis, M. *On the Prediction of Velocity Fields from Redshift Space Galaxy Samples.* ApJ, **421** (1994), L1. astro-ph/9309009.

Padmanabhan, N., Xu, X., Eisenstein, D.J., et al. *A 2 per cent distance toiz/i= 0.35 by reconstructing baryon acoustic oscillations – I. Methods and application to the Sloan Digital Sky Survey.* Monthly Notices of the Royal Astronomical Society, **427**(3) (2012), 2132. URL https://doi.org/10.1111%2Fj.1365-2966.2012.21888.x.

Paillas, E., Cai, Y.C., Padilla, N., et al. *Redshift-space distortions with split densities.* Monthly Notices of the Royal Astronomical Society, **505**(4) (2021), 5731–5752. ISSN 1365-2966. URL http://dx.doi.org/10.1093/mnras/stab1654.

Pakin, S.K., Gaborski, R.S., Barski, L.L., et al. *Clustering approach to bone and soft tissue segmentation of digital radiographic images of extremities.* J. Electronic Imaging, **12**(1) (2003), 40. URL https://doi.org/10.1117/1.1526846.

Palmese, A., Graur, O., Annis, J.T., et al. *Gravitational wave cosmology and astrophysics with large spectroscopic galaxy surveys* (2019). URL https://arxiv.org/abs/1903.04730.

Paranjape, A., Hahn, O., Sheth, R.K. *Halo assembly bias and the tidal anisotropy of the local halo environment.* MNRAS, **476**(3) (2018), 3631. 1706.09906.

Paranjape, A., Hahn, O., Sheth, R.K. *Halo assembly bias and the tidal anisotropy of the local halo environment.* Monthly Notices of the Royal Astronomical Society, **476**(3) (2018), 3631. URL https://doi.org/10.1093%2Fmnras%2Fsty496.

Park, C.F., Allys, E., Villaescusa-Navarro, F., et al. *Quantification of high dimensional non-Gaussianities and its implication to Fisher analysis in cosmology.* arXiv e-prints (2022), arXiv:2204.05435. 2204.05435.

Peebles, P.J.E. *The large-scale structure of the universe* (1980a).

Peebles, P.J.E. *The large-scale structure of the universe* (1980b).

Percival, W.J., Friedrich, O., Sellentin, E., et al. *Matching Bayesian and frequentist coverage probabilities when using an approximate data covariance matrix.* Monthly Notices of the Royal Astronomical Society, **510**(3) (2021), 3207. URL https://doi.org/10.1093%2Fmnras%2Fstab3540.

Perlmutter, S., Gabi, S., Goldhaber, G., et al. *Measurements of the Cosmological Parameters Ω and Λ from the First Seven Supernovae at $z >= 0.35$.* ApJ, **483**(2) (1997), 565. astro-ph/9608192.

Philcox, O.H.E., Ivanov, M.M. *BOSS DR12 full-shape cosmology: Λ CDM constraints from the large-scale galaxy power spectrum and bispectrum monopole.* Phys. Rev. D, **105**(4) (2022), 043517. 2112.04515.

Philcox, O.H.E., Ivanov, M.M. *BOSS DR12 full-shape cosmology:* ΛCDM *constraints from the large-scale galaxy power spectrum and bispectrum monopole.* Phys. Rev. D, **105** (2022), 043517. URL https://link.aps.org/doi/10.1103/PhysRevD.105.043517.

Planck Collaboration, Ade, P.A.R., Aghanim, N., et al. *Planck 2015 results. XIII. Cosmological parameters.* A&A, **594** (2016a), A13. 1502.01589.

Planck Collaboration, Ade, P.A.R., Aghanim, N., et al. *Planck 2015 results. XIII. Cosmological parameters.* A&A, **594** (2016b), A13. 1502.01589.

Planck Collaboration, Aghanim, N., Akrami, Y., et al. *Planck 2018 results. VI. Cosmological parameters.* A&A, **641** (2020a), A6. 1807.06209.

Planck Collaboration, Akrami, Y., Arroja, F., et al. *Planck 2018 results. IX. Constraints on primordial non-Gaussianity.* A&A, **641** (2020b), A9. 1905.05697.

Pritchard, J.R., Loeb, A. *21 cm cosmology in the 21st century.* Reports on Progress in Physics, **75**(8) (2012), 086901. 1109.6012.

Ramakrishnan, S., Velmani, P. *Properties beyond mass for unresolved haloes across redshift and cosmology using correlations with local halo environment* (2021). URL https://arxiv.org/abs/2112.15305.

Rao, C.R. *Information and the accuracy attainable in the estimation of statistical parameters.* Bull. Calcutta Math. Soc., **37** (1945), 81. ISSN 0008-0659.

Rasmussen, C.E., Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press (2005). ISBN 026218253X.

Reddick, R.M., Wechsler, R.H., Tinker, J.L., et al. *The Connection between Galaxies and Dark Matter Structures in the Local Universe*. ApJ, **771**(1) (2013), 30. 1207.2160.

Reid, B.A., White, M. *Towards an accurate model of the redshift-space clustering of haloes in the quasi-linear regime*. MNRAS, **417**(3) (2011), 1913. 1105.4165.

Reid, M.J., Pesce, D.W., Riess, A.G. *An Improved Distance to NGC 4258 and its Implications for the Hubble Constant*. Astrophys. J. Lett., **886**(2) (2019), L27. 1908.05625.

Rezende, D.J., Mohamed, S. *Variational Inference with Normalizing Flows* (2015). URL https://arxiv.org/abs/1505.05770.

Riess, A.G., Filippenko, A.V., Challis, P., et al. *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*. AJ, **116** (1998), 1009. astro-ph/9805201.

Riess, A.G., Yuan, W., Macri, L.M., et al. *A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km/s/Mpc Uncertainty from the Hubble Space Telescope and the SH0ES Team*. arXiv e-prints (2021), arXiv:2112.04510. 2112.04510.

Rodriguez, A.C., Kacprzak, T., Lucchi, A., et al. *Fast cosmic web simulations with generative adversarial networks*. Comput. Astrophys. Cosmol., **5** (2018), 4. 1801.09070.

Ruan, C.Z., Hernández-Aguayo, C., Li, B., et al. *Fast full N-body simulations of generic modified gravity: conformal coupling models*. J. Cosmology Astropart. Phys., **2022**(5) (2022), 018. 2110.00328.

Samushia, L., Percival, W.J., Raccanelli, A. *Interpreting large-scale redshift-space distortion measurements*. MNRAS, **420**(3) (2012), 2102. 1102.1014.

Satpathy, S., Alam, S., Ho, S., et al. *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: on the measurement of growth rate using galaxy correlation functions*. MNRAS, **469**(2) (2017), 1369. 1607.03148.

Schaller, M., Gonnet, P., Chalk, A.B.G., et al. *SWIFT: Using Task-Based Parallelism, Fully Asynchronous Communication, and Graph Partition-Based Domain Decomposition for Strong Scaling on More than 100,000 Cores*. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '16. Association for Computing Machinery, New York, NY, USA (2016). ISBN 9781450341264. URL https://doi.org/10.1145/2929908.2929916.

Scoccimarro, R. *Redshift-space distortions, pairwise velocities, and nonlinearities.* Phys. Rev. D, **70** (2004), 083007. URL https://link.aps.org/doi/10.1103/PhysRevD.70.083007.

Sefusatti, E., Crocce, M., Pueblas, S., et al. *Cosmology and the Bispectrum.* Phys. Rev. D, **74** (2006), 023522. astro-ph/0604505.

Seljak, U. *Measuring primordial non-gaussianity without cosmic variance.* Physical Review Letters, **102** (2008). Arxiv:0807.1770v1, URL https://arxiv.org/abs/0807.1770.

Seljak, U. *Bias, redshift space distortions and primordial nongaussianity of nonlinear transformations: application to Ly- forest.* Journal of Cosmology and Astroparticle Physics, **2012**(03) (2012), 004. URL https://doi.org/10.1088%2F1475-7516%2F2012%2F03%2F004.

Sellentin, E., Jaffe, A.H., Heavens, A.F. *On the use of the Edgeworth expansion in cosmology I: how to foresee and evade its pitfalls* (2017). 1709.03452.

Shah, P., Lemos, P., Lahav, O. *A buyer's guide to the Hubble constant.* A&A Rev., **29**(1) (2021), 9. 2109.01161.

Sheth, R.K. *The distribution of pairwise peculiar velocities in the non-linear regime.* MNRAS, **279** (1996), 1310. astro-ph/9511068.

Sheth, R.K., Diaferio, A., Hui, L., et al. *On the streaming motions of haloes and galaxies.* MNRAS, **326**(2) (2001), 463. astro-ph/0010137.

Sheth, R.K., Tormen, G. *Large-scale bias and the peak background split.* MNRAS, **308**(1) (1999), 119. astro-ph/9901122.

Sinha, M., Garrison, L.H. *CORRFUNC - a suite of blazing fast correlation functions on the CPU.* MNRAS, **491**(2) (2020), 3022.

Skilling, J. *Nested sampling for general Bayesian computation.* Bayesian Analysis, **1**(4) (2006), 833 . URL https://doi.org/10.1214/06-BA127.

Smith, A., de Mattia, A., Burtin, E., et al. *Reducing the variance of redshift space distortion measurements from mock galaxy catalogues with different lines of sight.* Monthly Notices of the Royal Astronomical Society, **500**(1) (2020), 259. URL https://doi.org/10.1093%2Fmnras%2Fstaa3244.

Somerville, R.S., Davé, R. *Physical Models of Galaxy Formation in a Cosmological Framework*. ARA&A, **53** (2015), 51. 1412.2712.

Springel, V. *The Cosmological simulation code GADGET-2*. Mon. Not. Roy. Astron. Soc., **364** (2005a), 1105. astro-ph/0505010.

Springel, V. *The cosmological simulation code gadget-2*. Monthly Notices of the Royal Astronomical Society, **364**(4) (2005b), 1105. ISSN 0035-8711. http://oup.prod.sis.lan/mnras/article-pdf/364/4/1105/18657201/364-4-1105.pdf, URL https://doi.org/10.1111/j.1365-2966.2005.09655.x.

Springel, V. *The cosmological simulation code GADGET-2*. MNRAS, **364**(4) (2005), 1105. astro-ph/0505010.

Succi, S., Coveney, P.V. *Big data: the end of the scientific method?* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, **377**(2142) (2019), 20180145. https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2018.0145, URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2018.0145.

Sugiyama, S., Takada, M., Kobayashi, Y., et al. *Validating a minimal galaxy bias method for cosmological parameter inference using HSC-SDSS mock catalogs*. Phys. Rev. D, **102**(8) (2020), 083520. 2008.06873.

Sunyaev, R.A., Zeldovich, Y.B. *The velocity of clusters of galaxies relative to the microwave background. The possibility of its measurement*. MNRAS, **190**(3) (1980), 413. ISSN 0035-8711. https://academic.oup.com/mnras/article-pdf/190/3/413/18223511/mnras190-0413.pdf, URL https://doi.org/10.1093/mnras/190.3.413.

Taha, A.A., Hanbury, A. *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool*. BMC Medical Imaging, **15** (2015), 29. URL http://www.biomedcentral.com/1471-2342/15/29.

Takada, M., Ellis, R.S., Chiba, M., et al. *Extragalactic science, cosmology, and Galactic archaeology with the Subaru Prime Focus Spectrograph*. PASJ, **66**(1) (2014), R1. 1206.0737.

Takada, M., Ellis, R.S., Chiba, M., et al. *Extragalactic science, cosmology, and Galactic archaeology with the Subaru Prime Focus Spectrograph*. Publications of the Astronomical Society of Japan, **66**(1) (2014).

Taruya, A., Nishimichi, T., Saito, S. *Baryon acoustic oscillations in 2D: Modeling redshift-space power spectrum from perturbation theory.* Phys. Rev. D, **82**(6) (2010), 063522. 1006.0699.

Taruya, A., Saito, S., Nishimichi, T. *Forecasting the cosmological constraints with anisotropic baryon acoustic oscillations from multipole expansion.* Phys. Rev. D, **83**(10) (2011), 103527. 1101.4723.

Tegmark, M. *Measuring Cosmological Parameters with Galaxy Surveys.* Physical Review Letters, **79**(20) (1997), 3806. URL https://doi.org/10.1103%2Fphysrevlett.79.3806.

Tegmark, M., Taylor, A.N., Heavens, A.F. *Karhunen-Loeve Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets?* The Astrophysical Journal, **480**(1) (1997), 22. URL https://doi.org/10.1086%2F303939.

Tinker, J., Kravtsov, A.V., Klypin, A., et al. *Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality.* ApJ, **688**(2) (2008), 709. 0803.2706.

Tinker, J.L. *Redshift-space distortions with the halo occupation distribution - II. Analytic model.* MNRAS, **374**(2) (2007), 477. astro-ph/0604217.

Tinker, J.L., Robertson, B.E., Kravtsov, A.V., et al. *The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests.* ApJ, **724**(2) (2010), 878. 1001.3162.

Tinker, J.L., Weinberg, D.H., Zheng, Z., et al. *On the Mass-to-Light Ratio of Large-Scale Structure.* ApJ, **631**(1) (2005), 41. astro-ph/0411777.

Tröster, T., Ferguson, C., Harnois-Déraps, J., et al. *Painting with baryons: augmenting N-body simulations with gas using deep generative models.* Mon. Not. Roy. Astron. Soc., **487**(1) (2019), L24. 1903.12173.

Udrescu, S.M., Tan, A., Feng, J., et al. *AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity.* arXiv e-prints (2020), arXiv:2006.10782. 2006.10782.

Uhlemann, C., Friedrich, O., Villaescusa-Navarro, F., et al. *Fisher for complements: extracting cosmology and neutrino mass from the counts-in-cells PDF.* MNRAS, **495**(4) (2020), 4006. 1911.11158.

Uhlemann, C., Kopp, M., Haugg, T. *Edgeworth streaming model for redshift space distortions.* Phys. Rev. D, **92**(6) (2015), 063004. 1503.08837.

Valogiannis, G., Dvorkin, C. *Towards an optimal estimation of cosmological parameters with the wavelet scattering transform.* Phys. Rev. D, **105**(10) (2022), 103534. 2108.07821.

van den Bosch, F.C., More, S., Cacciato, M., et al. *Cosmological constraints from a combination of galaxy clustering and lensing - I. Theoretical framework.* MNRAS, **430**(2) (2013), 725. 1206.6890.

Vernon, I., Owen, J., Aylett-Bullock, J., et al. *Bayesian Emulation and History Matching of JUNE.* medRxiv (2022). https://www.medrxiv.org/content/early/2022/02/22/2022.02.21.22271249.full.pdf, URL https://www.medrxiv.org/content/early/2022/02/22/2022.02.21.22271249.

Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. *Robust marginalization of baryonic effects for cosmological inference at the field level.* arXiv e-prints (2021), arXiv:2109.10360. 2109.10360.

Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. *The Quijote Simulations.* The Astrophysical Journal Supplement Series, **250**(1) (2020), 2. ISSN 1538-4365. URL http://dx.doi.org/10.3847/1538-4365/ab9d82.

Villaescusa-Navarro, F., et al. *Multifield Cosmology with Artificial Intelligence* (2021). 2109.09747.

Vlah, Z., Castorina, E., White, M. *The Gaussian streaming model and convolution Lagrangian effective field theory.* J. Cosmology Astropart. Phys., **2016**(12) (2016), 007. 1609.02908.

Vlah, Z., Seljak, U.c.v., Baldauf, T. *Lagrangian perturbation theory at one loop order: Successes, failures, and improvements.* Phys. Rev. D, **91** (2015), 023508. URL https://link.aps.org/doi/10.1103/PhysRevD.91.023508.

Vogelsberger, M., Marinacci, F., Torrey, P., et al. *Cosmological Simulations of Galaxy Formation.* Nature Rev. Phys., **2**(1) (2020), 42. 1909.07976.

Voit, E.O. *Perspective: Dimensions of the scientific method.* PLOS Computational Biology, **15**(9) (2019), 1. URL https://doi.org/10.1371/journal.pcbi.1007279.

Wandelt, B. *Gaussian Random Fields in Cosmostatistics* (2012). ISBN 978-1-4614-3507-5.

Wang, L., Reid, B., White, M. *An analytic model for redshift-space distortions.* MNRAS, **437**(1) (2014), 588. 1306.1804.

Wang, Y., Li, B., Cautun, M. *Iterative removal of redshift-space distortions from galaxy clustering.* Monthly Notices of the Royal Astronomical Society, **497**(3) (2020), 3451. URL https://doi.org/10.1093%2Fmnras%2Fstaa2136.

Ward, H., Atchison, C., Whitaker, M., et al. *Antibody prevalence for SARS-CoV-2 following the peak of the pandemic in England: REACT2 study in 100,000 adults.* medRxiv (2020). https://www.medrxiv.org/content/early/2020/08/21/2020.08.12.20173690.full.pdf, URL https://www.medrxiv.org/content/early/2020/08/21/2020.08.12.20173690.

Warren, M.S., Abazajian, K., Holz, D.E., et al. *Precision Determination of the Mass Function of Dark Matter Halos.* ApJ, **646**(2) (2006), 881. astro-ph/0506395.

Wechsler, R.H., Tinker, J.L. *The Connection Between Galaxies and Their Dark Matter Halos.* ARA&A, **56** (2018), 435. 1804.03097.

White, M. *Reconstruction within the Zeldovich approximation.* Monthly Notices of the Royal Astronomical Society, **450**(4) (2015), 3822. URL https://doi.org/10.1093%2Fmnras%2Fstv842.

White, M. *A marked correlation function for constraining modified gravity models.* JCAP, **11** (2016), 057. 1609.08632.

White, S.D.M., Rees, M.J. *Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering.* MNRAS, **183** (1978), 341.

Wolf, T., Debut, L., Sanh, V., et al. *Transformers: State-of-the-Art Natural Language Processing.* In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, Online (2020), pages 38–45. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Woodfinden, A., Nadathur, S., Percival, W.J., et al. *Measurements of cosmic expansion and growth rate of structure from voids in the Sloan Digital Sky Survey between redshift 0.07 and 1.0* (2022). URL https://arxiv.org/abs/2205.06258.

Xu, X., Zehavi, I., Contreras, S. *Dissecting and modelling galaxy assembly bias.* MNRAS, **502**(3) (2021), 3242. 2007.05545.

Xu, X., Zehavi, I., Contreras, S. *Dissecting and modelling galaxy assembly bias.* Monthly

Notices of the Royal Astronomical Society, **502**(3) (2021), 3242. URL https://doi.org/10.1093%2Fmnras%2Fstab100.

Yankelevich, V., Porciani, C. *Cosmological information in the redshift-space bispectrum.* MNRAS, **483**(2) (2019), 2078. 1807.07076.

Yoo, J. *General Relativistic Description of the Observed Galaxy Power Spectrum: Do We Understand What We Measure?* Physical Review D, **82** (2010). Arxiv:1009.3021v1, URL https://arxiv.org/abs/1009.3021.

Yuan, S., Garrison, L.H., Eisenstein, D.J., et al. *Stringent σ_8 constraints from small-scale galaxy clustering using a hybrid MCMC+emulator framework.* arXiv e-prints (2022a), arXiv:2203.11963. 2203.11963.

Yuan, S., Hadzhiyska, B., Bose, S., et al. *Illustrating galaxy-halo connection in the DESI era with ILLUSTRISTNG.* MNRAS, **512**(4) (2022b), 5793. 2202.12911.

Zarrouk, P., Burtin, E., Gil-Marín, H., et al. *The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: measurement of the growth rate of structure from the anisotropic correlation function between redshift 0.8 and 2.2.* Monthly Notices of the Royal Astronomical Society, **477**(2) (2018), 1639. ISSN 0035-8711. http://oup.prod.sis.lan/mnras/article-pdf/477/2/1639/25009821/sty506.pdf, URL https://doi.org/10.1093/mnras/sty506.

Zarrouk, P., Ruiz-Macias, O., Cole, S., et al. *Preliminary clustering properties of the DESI BGS bright targets using DR9 Legacy Imaging Surveys.* MNRAS, **509**(1) (2022), 1478. 2106.13120.

Zehavi, I., Contreras, S., Padilla, N., et al. *The Impact of Assembly Bias on the Galaxy Content of Dark Matter Halos.* ApJ, **853**(1) (2018), 84. 1706.07871.

Zehavi, I., Kerby, S.E., Contreras, S., et al. *On the Prospect of Using the Maximum Circular Velocity of Halos to Encapsulate Assembly Bias in the Galaxy-Halo Connection.* ApJ, **887**(1) (2019), 17. 1907.05424.

Zel'dovich, Y.B. *Gravitational instability: An approximate theory for large density perturbations.* A&A, **5** (1970), 84.

Zhai, Z., Tinker, J.L., Banerjee, A., et al. *The Aemulus Project V: Cosmological constraint* *from small-scale clustering of BOSS galaxies.* arXiv e-prints (2022), arXiv:2203.08999. 2203.08999.

Zhai, Z., Tinker, J.L., Becker, M.R., et al. *The Aemulus Project. III. Emulation of the Galaxy Correlation Function.* ApJ, **874**(1) (2019a), 95. 1804.05867.

Zhai, Z., Tinker, J.L., Becker, M.R., et al. *The Aemulus Project. III. Emulation of the Galaxy Correlation Function.* ApJ, **874**(1) (2019b), 95. 1804.05867.

Zheng, Z., Berlind, A.A., Weinberg, D.H., et al. *Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies.* ApJ, **633**(2) (2005), 791. astro-ph/0408564.

Zhu, G., Zheng, Z., Lin, W.P., et al. *The Dependence of the Occupation of Galaxies on the Halo Formation Time.* ApJ, **639**(1) (2006), L5. astro-ph/0601120.

Zu, Y., Weinberg, D.H. *The redshift-space cluster–galaxy cross-correlation function – I. Modelling galaxy infall on to Millennium simulation clusters and SDSS groups.* Monthly Notices of the Royal Astronomical Society, **431**(4) (2013), 3319. ISSN 0035-8711. http://oup.prod.sis.lan/mnras/article-pdf/431/4/3319/3953195/stt411.pdf, URL https://doi.org/10.1093/mnras/stt411.

Zwicky, F. *Die Rotverschiebung von extragalaktischen Nebeln.* Helvetica Physica Acta, **6** (1933), 110.