HELSINKI INSTITUTE OF PHYSICS INTERNAL REPORT SERIES

HIP-2022-05

Two-Point Correlation Function as a Cosmological Probe

Valtteri Lindholm

Helsinki Institute of Physics
University of Helsinki
Finland

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, in Auditorium E204, Physicum building, on the 17th of February, 2023 at 13 o'clock.

Helsinki 2023

ISBN 978-951-51-1309-2 (print)
ISBN 978-951-51-1310-8 (pdf)
ISSN 1455-0563 (print)
ISSN 2814-9459 (online)
http://ethesis.helsinki.fi
Unigrafia
Helsinki 2023

Äidille.

Acknowledgements

First and foremost, I would like to thank my supervisors, Elina Keihänen and Hannu Kurki-Suonio, for their support and expertise. They, along with all the other people who have worked in our group, have created a safe and constructive environment to work and learn in. I also want to thank Alexis Finoguenov, who introduced me to the art of galaxy clusters and guided me through the writing process of my first first author paper. The pre-examiners of the thesis, Wojciech Hellwing and Carlo Schimd, provided many useful comments and insights. I also wish to acknowledge the Jenny and Antti Wihuri Foundation who financially supported my work throughout the years 2015–2017. This thesis would not have been possible without my family and friends, for whom I am thankful that I turned out curious and kind. Lastly, I want to thank Riikka for being there for the good times and the bad times.

Abstract

The near-future of observational cosmology lies in extensive galaxy surveys that will map the dynamic evolution of the Universe for large part of its history. An example of such a survey is the European Space Agencys (ESA) Euclid mission. It is a space telescope scheduled to launch in 2023. The Euclid wide survey will cover 15 000 square degrees, which is more than one third of the sky. It is expected to observe photometrically roughly 1.5 billion galaxies for the purposes of weak lensing analysis. In addition, 50 million galaxies will be observed spectroscopically, which will allow accurate determination of their three-dimensional distribution. The spectroscopic measurements will reach above a redshift of 2 – covering over 75% of the history of the Universe – and photometric measurements to even higher redshifts.

An important aspect of measuring the galaxy distribution is that its evolution reflects the expansion history of the Universe. It is thus a powerful tool for understanding the processes that drive the expansion. To be able to compare the evolution of the distribution with predictions from, for example, models of dark energy, we need to extract various statistics of the galaxy distribution. An important quantity is the two-point correlation function, which is a measure of amount of clustering at different spatial scales. Estimators for the two-point correlation functions rely on counting pairs of objects, usually galaxies or their clusters. The large number of galaxies Euclid is going to observe will make the clustering analysis a non-trivial computational task. The situation becomes even more challenging when one needs to estimate the covariance of the two-point correlation function estimates. This is usually done by measuring clustering of mock galaxy samples. To reach the parameter accuracy Euclid is aiming for, the number of samples needed is of the order of many thousands, thus in principle increasing the cost of a two-point correlation function estimate by the same factor. It is therefore of paramount importance to perform this analysis step as efficiently as possible.

Two of the three publications included in this thesis deal with this need for efficiency. A central part of methods introduced in them is the so-called random catalog. It is a counterpart to the measured galaxy catalog that is used to model the survey geometry and various selection effects. We show that it is possible to use the random sample in clever ways to significantly speed up the computation with negligible loss of accuracy. We present one method for a single two-point correlation function estimate and another one for its covariance. We show that in a situation representative of a spectroscopic galaxy sample we can reach speed-up by a factor of order of ten – for each method. The combined factor of hundred could already be decisive in whether the analysis is computationally feasible or not.

In the end, the computational efficiency is only a tool for reaching more and more precise scientific results. The third paper in this thesis offers an example of a clustering-based cosmological analysis. Within the current theoretical understanding of the large scale structure of the Universe it is possible to obtain relations between the clustering of clusters of galaxies and the masses of their hosting dark matter halos. Generally, the more massive the halos, the more strongly clustered they are. In the paper we study this connection in detail using an X-ray selected, already-published sample of galaxy clusters. We find that the cosmological predictions for the clustering of clusters gives results that are compatible with the observations and the cosmological parameters inferred from the measurements are consistent with those obtained from the cluster mass function and the Cosmic Microwave Background.

Included publications

This thesis consists of three research articles and an introduction that provides relevant background for the articles and summarizes the most important results. The articles are:

I E. Keihänen, H. Kurki-Suonio, V. Lindholm, et al. (Nov. 2019). "Estimating the galaxy two-point correlation function using a split random catalog". In: *Astronomy & Astrophysics* 631, A73

II E. Keihänen, V. Lindholm, P. Monaco, et al. (Oct. 2022). "Euclid: Fast two-point correlation function covariance through linear construction". In: *Astronomy & Astrophysics* 666, A129

III V. Lindholm, A. Finoguenov, J. Comparat, et al. (Feb. 2021). "Clustering of CODEX clusters". In: *Astronomy & Astrophysics* 646, A8

Author's contribution

Paper I: This paper deepens the understanding of the bias and variance of a popular two-point correlation function estimator and introduces a method for increasing its computational efficiency. Using an existing code I computed the two-point correlation functions used to validate the method and the theoretical results presented in the paper.
Paper II: In this paper we study a method for increasing the computational efficiency of estimating the covariance matrix of the two-point correlation function estimate produced with the method presented in Paper I. I computed the two-point correlation functions on which all the numerical results in the paper are based. I wrote the code to construct the covariance matrices and did all the numerical analysis, except for Table 3. In addition to the numerical work I did some analytical calculations related to the pair count covariances. I also wrote parts of Sec. 4 and produced all the figures.
Paper III: This paper studies the connection between the masses of galaxy clusters and their spatial clustering. I wrote an analysis pipeline around already-existing tools and ran all the analysis starting from the raw cluster catalog. I produced all the numerical results and figures and wrote the paper, except for Sec. 2.

Contents

Chapter 1

Introduction

Throughout the 20th century physical cosmology matured from a field of speculation into a precision science, where theoretical models of the origin and the evolution of the Universe provide testable predictions and parameters can be constrained observationally down to a percent level at best. Important milestones in this development were the discoveries of the Cosmic Microwave Background (CMB) in 1964 (Penzias and Wilson, 1965) and its temperature fluctuations by NASA's Cosmic Background Explorer (COBE) satellite in 1992 (Smoot et al., 1992). The CMB temperature fluctuations offer a snapshot of the Universe in its infancy, roughly at the age of 380 000 years. The small variations in the temperature correspond to density fluctuations in the primordial plasma. The overdense regions act as seeds that gradually accumulate matter through gravitational pull and eventually form structures that we observe today, such as galaxies and their clusters.

After COBE the temperature fluctuations have been studied extensively by various telescopes. Important contributions were made by, for example, the balloon-born experiment BOOMERANG (Balloon Observations Of Millimetric Extragalactic Radiation ANd Geophysics, Netterfield et al., 2002) that greatly improved on resolution of COBE and measured the CMB polarization signal. The first all-sky maps that had the necessary resolution for the temperature fluctuations to be fully exploited were produced by NASA's Wilkinson Microwave Anisotropy Probe (WMAP, Spergel et al., 2003; Spergel et al., 2007; Komatsu et al., 2011; Bennett et al., 2013). These measurements shrank the volume of the allowed cosmological parameter space by orders of magnitude. Currently, the CMB temperature fluctuations have been measured practically exhaustively by ESA's Planck satellite (Planck Collaboration, 2020a). There is still room for improvement in the CMB polarization measurements, both with ground-based telescopes (Abazajian et al., 2016) and satellite missions, such as LiteBIRD (JAXA, selected) (Hazumi et al., 2020) and CORE (ESA, proposal) (Delabrouille et al., 2018).

However, in the years to come the focus is going to shift more and more from CMB to large galaxy surveys. Such surveys have already been conducted, notable examples being the Sloan Digital Sky Survey (SDSS) (York et al., 2000; Abazajian et al., 2003; Abazajian et al., 2004; Tegmark et al., 2004) and the Two-degree-Field Galaxy Redshift Survey (2dFGRS) (Percival et al., 2001; Cole et al., 2005). In the near future the data quality and volume in the field will be significantly increased by next generation surveys such as the Legacy Survey of Space and Time (LSST) (Ivezić et al., 2019), the Dark Energy Spectroscopic Instrument (DESI) survey (DESI Collaboration, 2016), Nancy Grace Roman Space Telescope (Akeson et al., 2019) and the Euclid satellite (Laureijs et al., 2011).

Whereas the CMB measurements most directly tell us about the initial conditions and the early stages of the Universe, galaxy surveys offer direct observational information of a much larger span of the history of the Universe. They are thus well-suited for studying the dynamic evolution of the Universe that could shed light on, for example, the problem of dark energy. In the context of galaxy surveys the expansion history of the Universe is mapped by the time evolution of the large-scale distribution of matter. The two most important probes for this evolution are galaxy clustering and weak gravitational lensing. The former has a more intuitive connection to the overall matter distribution; the distribution of galaxies traces the underlying matter distribution. The situation is, however, more convoluted. The correspondence between the matter density and the galaxy number

density is not one-to-one and the complex nature of galaxy formation obscures the exact relation between the two distributions. Gravitational lensing, on the other hand, is directly caused by the total mass between the observer and the source. The effect of the gravitational lensing on observed galaxy shapes is known as the cosmic shear and it has already been studied by various surveys, such as the ones conducted at the Canada France Hawaii Telescope (CFHT) (Van Waerbeke et al., 2000; Semboloni et al., 2006; Hoekstra et al., 2006; Heymans et al., 2012), the Cerro Tololo Inter-American Observatory (CTIO) lensing survey (Jarvis et al., 2006), the Red-Sequence Cluster Survey (RCS, combines data from CFHT and CTIO) (Hoekstra et al., 2002) and the Kilo Degree Survey (KiDS) (Hildebrandt et al., 2017). Compared to galaxy clustering the cosmic shear analysis is theoretically more involved but in principle a direct measurement of the matter distribution. The problem is that the effect is extremely weak and to be fully exploited requires large samples of accurate galaxy shape measurements.

Such a sample will be provided by the Euclid survey, which, of the near-future galaxy surveys, is the most important regarding my PhD work. It is ESA's next large cosmology mission, scheduled to launch in 2023. The satellite will observe with two instruments: the visible imager VIS (Cropper et al., 2016) and the near-infrared spectrophotometric instrument NISP (Costille et al., 2018). The Euclid wide survey will cover 15 000 square degrees, which is more than one third of the sky. It is expected to observe photometrically roughly 1.5 billion galaxies for the purposes of weak lensing analysis. In addition, 50 million galaxies will be observed spectroscopically. This will allow studying the three-dimensional clustering of galaxies accurately over 75% of the lifetime of the Universe, up to redshifts of ~ 2, and the photometric sample is expected to reach even higher redshifts. Both probes are vital to the success of the Euclid mission but this thesis mostly concerns galaxy clustering.

An important quantity in characterizing the galaxy clustering is the two-point correlation function, which is a statistical measure of clustering at different length scales. The clustering analysis using two-point correlation functions has been the context of most of my PhD work. Since the beginning of this work I have been a member of the Euclid consortium and Papers I & II originate from our group's work on the Euclid analysis pipeline. Estimators for the two-point correlation functions rely on counting pairs of objects (galaxies or their clusters). The large number of galaxies Euclid is going to observe will make the clustering analysis a non-trivial computational task. The situation becomes even more challenging when one needs to estimate the covariance of the two-point correlation function estimates. This is usually done by measuring clustering of mock galaxy samples. To reach the parameter accuracy Euclid is aiming for the number of samples needed is of the order of many thousands, thus in principle increasing the cost of a two-point correlation function estimate by the same factor. It is therefore of paramount importance to perform this analysis step as efficiently as possible. Papers I & II deal with this need for efficiency.

A central part of the methods introduced in these papers is the so-called random catalog. It is a counterpart to the measured galaxy catalog and is used for modeling the survey geometry and various selection effects. In these papers we show that it is possible to use the random catalog in clever ways to significantly speed up the computations with negligible loss of accuracy. In Paper I we present one method for a single two-point correlation function estimate and in Paper II another one for computing the covariance from a large set of mock catalogs.

Main goal of a cosmological survey is not algorithm development but the scientific results, such as parameter constraints, and improving the computational efficiency of an analysis pipeline is just a way to increase the accuracy of the results. Paper III is an example of a cosmological analysis that uses two-point correlation function measurements. We analyse the distribution of galaxy clusters but as far as their clustering is concerned they are in principle not any different from individual galaxies. In the paper we use data combined from the ground based SDSS galaxy survey and an X-ray satellite (ROentgen SATellite, ROSAT, Voges et al., 1999) mission but the data is similar in nature than what Euclid will produce. Also the theoretical framework and analysis methods are closely related to the Euclid survey.

The thesis is organized as follows. Chapter 2 introduces the theoretical framework for the analysis of the large scale structure of the Universe. First, I present the homogeneous and isotropic model for the global cosmic evolution, then the linear perturbation theory for studying the evolution of small density fluctuations and finally I discuss the formation of stable structures through gravitational collapse. Chapter 3 deals with statistics of the these collapsed objects. I first discuss

the statistics of peaks of matter density fields, which are thought to be seeds of the collapsed halos. Then I present some models for predicting mass distribution and clustering properties of these halos from a cosmological model. This predicting power allows one to test cosmological models by observing the number density and spatial distribution of clusters of galaxies as a function of their mass. In Chapter 4 I discuss the two-point correlation functions. I give an overview of the theoretical framework underlying the predicted and observed two-point correlation functions. After this, I discuss algorithms for estimating the two-point correlation function and its covariance, and explain how we have made these more efficient. Here I also summarize results from Papers I & II. Lastly, I present some two-point correlation function results we obtained from the aforementioned galaxy cluster sample in Paper III. In Chapter 5 I offer some concluding remarks.

Chapter 2

Cosmological perturbations

I will work with the natural units in which speed of light and the reduced Planck's constant equal unity.

2.1 Background cosmology

An important cosmological observation is that the Universe is expanding. Distant galaxies seem to be receding from us with a velocity v that is proportional to their distance d

$$v = H_0 d. \tag{2.1}$$

The proportionality constant is called the Hubble constant. This apparent motion gets a natural explanation from Einstein's general theory of relativity. It describes the gravity as a manifestation of the geometry of a four-dimensional spacetime. The basic degree of freedom is the metric tensor g that defines the geometry of the spacetime. It is the basis for all geometric quantities, such as physical distances. The distances are often expressed in terms of the differential line element

$$\mathrm{d}s^2 = \sum_{\mu,\nu} g_{\mu\nu} \mathrm{d}x^\mu \mathrm{d}x^\nu \,, \tag{2.2}$$

where $\{x^\mu\}$ are some spacetime coordinates and $g_{\mu\nu}$ are the components of the metric tensor within this coordinate system. The differential line element is then integrated to compute finite distances between points.

The components of the metric tensor determine the geometric properties of the spacetime and are in turn determined by the mass-energy content of the universe. This relationship is expressed by the Einstein equation

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} \,. \tag{2.3}$$

Here $G_{\mu\nu}$ is the Einstein tensor, G is Newton's gravitational constant and $T_{\mu\nu}$ is the so-called stress-energy tensor. The Einstein tensor is a non-linear combination of the metric tensor and its first and second derivatives. Thus, Eq. (2.3) is a set of ten (the tensor is symmetric) non-linear second order partial differential equations for the components of the metric tensor and as such impossible to solve analytically, except in some highly symmetrical cases. Luckily, the Universe seems to be spatially homogeneous and isotropic on large enough scales. In the context of general relativity this means that there exists a special time coordinate t, the so-called cosmic time, constant values of which correspond to a homogeneous and isotropic three-dimensional space. In spherical coordinates $(t, r, \vartheta, \varphi)$[1] this homogeneous and isotropic metric takes the form

$$g = \mathrm{diag}\left(-1,\, \frac{a(t)}{1 - Kr^2},\, a(t) r^2,\, a(t) r^2 \sin^2 \vartheta\right) \,. \tag{2.4}$$

[1] Here r is the comoving radial coordinate for which the expansion of the universe has been factored out. Comoving distances coincide with the physical distances (also known as the proper distances) at the time when $a(t) = 1$, often defined to be the present time.

Here K is a constant that parametrizes the spatial curvature (that is, the curvature of space at any given time) and the function $a(t)$ is the scale factor that tells how the physical distances change in time. The metric 2.4 is called the Friedmann-Lemaître-Robertson-Walker (FLRW) metric. Many observations, Suzuki et al. (2012), Planck Collaboration (2020b), and Zhao et al. (2022) for example, suggest that in our universe K is very close to zero, in which case the metric is said to be spatially flat and the homogeneous $t =$ constant three-dimensional slices obey Euclidean geometry. The other two options are closed universe ($K > 0$) and open universe ($K < 0$), in which cases the three-dimensional slices correspond to spherical and hyperbolic geometries, respectively. For the rest of this thesis I will assume that $K = 0$. Often, it is more convenient to define the time coordinate so that the scale factor can be taken as a common factor. This is achieved by defining $\mathrm{d}\eta = \mathrm{d}t/a(t)$. The coordinate η is called the conformal time and in terms of it the flat FLRW metric takes a particularly simple form in Cartesian coordinates: $g = a(\eta)\mathrm{diag}(-1, 1, 1, 1)$.

The assumption of homogeneity and isotropy forces also the stress-energy tensor to take a simple diagonal form

$$T = \mathrm{diag}(p, \rho, \rho, \rho)\,, \tag{2.5}$$

where p and ρ correspond to pressure and energy density of the universe, respectively, and only depend on time. This diagonal form simplifies the Einstein equation into a pair of Friedmann equations for the scale factor

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K^2}{a^2}\,, \tag{2.6}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p)\,. \tag{2.7}$$

Here the dot ($\dot{}$) is used for derivative with respect to time t. The relative rate of change in the scale factor is often expressed as the Hubble parameter $H \coloneqq \dot{a}/a$. The parameter can be also written in terms of the conformal time as $\mathcal{H} \coloneqq a'/a = aH$, where $a' \coloneqq \mathrm{d}a/\mathrm{d}\eta$. For completeness I have included the contribution from the curvature parameter K even though it is assumed to be zero.

What is required to solve these equations is a model for how the pressure and energy density are related. In a homogeneous universe where the energy density falls monotonically it is always possible to express pressure as a function of energy density $p(\rho)$. This dependence is usually parameterized by the so-called the equation of state parameter $w \coloneqq p/\rho$. In general, w is a function of time. Eqs. 2.6 and 2.7 can be combined to obtain

$$\dot{\rho} = -3(\rho + p)H\,. \tag{2.8}$$

Given the relationship between pressure and energy density, Eq. (2.8) expresses how the energy density evolves in time, or equivalently as a function of the scale factor. In terms of w this evolution is given by

$$\mathrm{d}\ln\rho = -3\left[1 + w(a)\right]\mathrm{d}\ln a\,. \tag{2.9}$$

A simple, yet very powerful model for the energy content of the Universe consists of three components:

1. Non-relativistic particles, also called matter, for which $w_\mathrm{m} = 0$ and $\rho_\mathrm{m} \propto a^{-3}$
2. Ultrarelativistic particles, also called radiation, for which $w_\mathrm{r} = 1/3$ and $\rho_\mathrm{r} \propto a^{-4}$
3. Cosmological constant, for which $w_\Lambda = -1$ and $\rho_\Lambda =$ constant

A cosmological constant is the simplest case of an energy component that drives the expansion to accelerate. Such energy components are generally called dark energy. Developing more complex models for dark energy is an active research field but so far observations are consistent with a cosmological constant (see eg. Planck Collaboration, 2020b; DES Collaboration, 2022 and references therein).

The total energy density is the sum of the individual components and the combined equation of state parameter is determined by the relative contributions of the different components. Even if each of the components has a constant equation of state parameter the total equation of state

changes in time due to evolution of the fraction of the different components. Since the energy density of radiation drops faster than that of matter and cosmological constant, most the expansion history is determined by the two latter components[2]. Most of the matter (84%, Planck Collaboration, 2020b) is so-called dark matter that, apart from gravity, interacts extremely weakly (if at all). However, at the present day the dominating energy component is not matter but the cosmological constant at 69% of the total energy density (Planck Collaboration, 2020b). These two components, the cosmological constant Λ and Cold Dark Matter[3] (CDM) give name to this standard model: ΛCDM.

Energy density of a component i is usually expressed in terms of the density parameter Ω_i which is a fraction of the present day energy density and the so-called critical density (at the present day)

$$\rho_\mathrm{c} \coloneqq \frac{3H_0^2}{8\pi G}\,, \tag{2.10}$$

so that $\Omega_i \coloneqq \rho_i/\rho_\mathrm{c}$. The critical density is simply the energy density in a spatially flat universe that expands with the velocity corresponding to the present-day Hubble parameter H_0 (thus our Universe has a density very close to critical density). Using this definition and the three-component model, Eq. (2.6) takes the form

$$H(a)^2 = H_0^2\left(\Omega_\mathrm{r}a^{-4} + \Omega_\mathrm{m}a^{-3} + \Omega_\Lambda\right)\,. \tag{2.11}$$

This equation can be numerically solved for the evolution of the scale factor and thus determines the expansion history of the Universe.

The scale factor is not directly measurable over cosmological length scales and as such not a useful variable for practical purposes. A directly observable quantity is the redshift of electromagnetic radiation caused by the cosmic expansion

$$z \coloneqq \frac{\lambda_0 - \lambda}{\lambda}\,, \tag{2.12}$$

where λ is the wavelength emitted at the source and λ_0 is the observed wavelength. It is related to the scale factor by

$$1 + z = \frac{a_0}{a}\,, \tag{2.13}$$

where a is the scale factor at the time the radiation was emitted and a_0 at the time it was received (often normalized to be unity).

Eq. 2.11, or a more general expression for $H(z)$, can be used to constrain the cosmological model with the help of a standard ruler. A standard ruler is an object whose physical size as a function of cosmic epoch is known. A way to exploit this knowledge is to relate the angle we observe the object at to its distance from us (in practice the objects redshift). In a curved spacetime the relationship between objects size and the angle it covers is more complicated than in Euclidean geometry. The angle can be parameterized as

$$\vartheta(z) = \frac{s^\mathrm{p}}{d_\mathrm{A}(z)}\,, \tag{2.14}$$

where s^p is the physical size of the object and $d_\mathrm{A}(z)$ is the angular diameter distance to the objects redshift z. In a flat FLRW geometry the angular diameter distance is equal to the proper distance at the time the light left and is given by

$$d_\mathrm{A}(z) = \frac{1}{1+z}\int_0^z \frac{\mathrm{d}z'}{H(z')}\,. \tag{2.15}$$

Thus, by measuring how the angle covered by an object of known size evolves with redshift one can constrain the expansion law $H(z)$. An important standard ruler are the Baryon Acoustic

[2]The CMB temperature directly constrains the present day radiation contribution to be $\sim 0.01\%$ of the total energy density

[3]Cold here means that velocities of the dark matter particles have not been large enough for them to escape the gravity of the first forming structures.

Oscillations (BAO), see for example Aubourg et al. (2015) for a review. These are sound waves in the primordial baryon-photon plasma and the angle corresponding to their wavelength can be accurately observed in the CMB. After the photons decouple from baryons the compressed regions of the wavefront are frozen and act to enhance the galaxy inhomogeneity with the same characteristic wavelength. The enhancement at the comoving scale of $\sim 100h^{-1}\mathrm{Mpc}$[4] can be observed statistically in the distribution of galaxies (see Chapter 4) at different redshifts and provides a measurement of the expansion history.

2.2 Linear perturbations

Clearly, the Universe is not fully homogeneous but contains various structures, such as galaxies and their clusters. As mentioned, the fully inhomogeneous Einstein equation is extremely difficult to solve. We can, however, consider small perturbations around the homogeneous and isotropic model presented in the previous section. In this perturbative description all quantities are expressed as a sum of the corresponding value in the background model and a small perturbation. Perturbation theory around the FLRW metric is a well-studied subject, see eg. Mukhanov et al. (1992). Here I will just outline the procedure. I denote the background value of a quantity q with overbar $\bar{q}$ and the perturbation with δq. Since the background universe is assumed to be homogeneous and isotropic, all the background quantities only depend on time and the spatial variations are captured by the perturbation part. The smallness of perturbations means that $|\delta q| \ll |\bar{q}|$ at each point and time and we can ignore all the terms that are second or higher order in these perturbations.

In principle we are again dealing with the full Einstein equation and ten components of the metric tensor. However, since the perturbed spacetime is no longer homogeneous or isotropic we have some additional freedom in choosing the coordinate system we use. The requirement for the coordinate system is that it preserves the condition $|\delta q| \ll |\bar{q}|$ for the relevant quantities. This is called gauge freedom and allows us to fix four of the ten metric components. The remaining six can be classified according to how they transform under spatial rotations. The three classes are:

1. Scalar perturbations, which are invariant under rotations and correspond to density perturbations.

2. Vector perturbations, which transform as vector fields and correspond to vorticity perturbations.

3. Tensor perturbations, which transform as rank-2 tensors and correspond to gravitational waves.

To first order all the different classes evolve independent of each other. It can be shown that vector perturbations always decay over time. Gravitational waves can be produced by early Universe processes after which they propagate through space but their amplitude is constrained to be much lower than of the scalar perturbations (Planck Collaboration, 2020b). Moreover, scalar perturbations are the ones that amplify under- and overdensities to eventually form the presently-observed structures in the Universe. Thus, I will only focus on the scalar perturbations. A particularly suitable coordinate system, or gauge, for describing the scalar perturbations is the Conformal Newtonian Gauge, in which the metric takes in terms of the conformal time the form

$$g = a(\eta) \begin{pmatrix} -(1+2\Psi) & 0 \\ 0 & (1-2\Phi)\delta_{ij} \end{pmatrix} . \tag{2.16}$$

Here δ_{ij} is the Kronecker delta and the perturbations Ψ and Φ are called the Bardeen potentials.

In the same way as the metric tensor, the energy tensor is split into a background and a perturbation part and the background part correspond to what was introduced in Sec. 2.1. Likewise, the perturbation part can be decomposed into scalar-, vector- and tensor-like perturbations. The scalar perturbations of the energy tensor couple to scalar perturbations of the metric and so on, hence

[4] $h \coloneqq H_0/(100\mathrm{km/s/Mpc})$

I will only focus on the scalar perturbations in the energy tensor. In the Conformal Newtonian Gauge the scalar perturbations of the energy tensor can be written in the form

$$\delta T = \begin{pmatrix} -\delta\rho & -(\bar{\rho}+\bar{p})v_{,i} \\ (\bar{\rho}+\bar{p})v_{,i} & \delta p\delta_{ij} + \bar{p}(\Pi_{,ij} - \frac{1}{3}\nabla^2\Pi) \end{pmatrix}. \tag{2.17}$$

Here I have introduced notation for partial derivatives: $v_{,1\ldots N} \coloneqq \partial_1 \ldots \partial_N v$. $\delta\rho$ and δp are perturbations of energy density and pressure, respectively. v are velocity perturbations in the cosmic fluid (the background fluid is at rest by the requirement of isotropy but the perturbed fluid can have local flows). Π is the so-called anisotropic stress normalized by the mean pressure $\bar{p}$. It is caused by anisotropic momentum distribution of the underlying particle distribution. Sufficient interactions between the particles tend to isotropize the distribution. The anisotropic stress can be produced when particles decouple from each other (in the early universe this happens to neutrinos and photons), but in an approximate treatment this can be neglected. This is called the perfect fluid approximation.

An important class of energy tensor perturbations are adiabatic perturbations, for which the pressure and density perturbations satisfy the following simple relationship

$$\delta p = \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}}\delta\rho\,. \tag{2.18}$$

If the pressure is a function of energy density only, the perturbations are necessarily adiabatic. The adiabatic perturbations are particularly simple in the sense that the state of the cosmic fluid at a point $(t, \vec{x})$ of the perturbed spacetime corresponds to its state at the same point in the background spacetime at some slightly earlier or later time $t + \delta t$ (δt being different at each point $\vec{x}$). Observations from the CMB, for example, are compatible with perturbations being initially purely adiabatic (Planck Collaboration, 2020c) and this also a natural prediction from single-field inflation models (see the later paragraph on summary of inflation).

As is often the case when dealing with time and spatial derivatives, the cosmological perturbations are easier to track in Fourier space. This also allows us to discuss how perturbation of different "size" or scale (ie. different wavelengths) evolve differently. An arbitrary perturbation $h(\eta, \vec{x})$ expanded as plane waves is

$$h(\eta, \vec{x}) = \sum_{\vec{k}} h_{\vec{k}}(\eta) e^{i\vec{k}\cdot\vec{x}}. \tag{2.19}$$

The wave vector $\vec{k}$ is the comoving wave vector ie. it does not scale with the expansion of the universe. The physical wavelength is given by $\lambda = a\frac{2\pi L}{|\vec{k}|}$. Here we have assumed a periodic box of fiducial comoving volume L^3. In the end of the day we can take the limit $L \to \infty$ and replace Fourier sums with integrals. Within the first order perturbation theory the different $\vec{k}$-modes evolve independently and we can write the equations for each mode separately.

We are now in the position of writing down the Einstein equations for the metric perturbations. It turns out that in the absence of anisotropic stress the Bardeen potentials are equal and we have only one degree of freedom Φ in the metric perturbation. Thus, in the presence of scalar perturbations and within the perfect fluid approximation the Einstein equations take in Fourier space the form

$$\left(\frac{k}{\mathcal{H}}\right)^2 \Phi = \frac{3}{2}\left[\delta + 3(1+w)\frac{\mathcal{H}}{k}v\right] \tag{2.20}$$

$$\mathcal{H}^{-1}\Phi' + \Phi = \frac{3}{2}(1+w)\frac{\mathcal{H}}{k}v \tag{2.21}$$

$$\mathcal{H}^{-2} + 3\mathcal{H}^{-1}\Phi' + \left(1 + \frac{2\mathcal{H}'}{\mathcal{H}^2}\Phi\right) = \frac{3}{2}\frac{\delta p}{\bar{\rho}}\,. \tag{2.22}$$

Here I have defined the density contrast $\delta \coloneqq \delta\rho/\bar{\rho}$, which is the most commonly used perturbation quantity. Given a model for the energy components and the initial conditions for the perturbations this set of equations defines their evolution. The equations have two limits.

1. Superhorizon perturbations; perturbations of scales much larger than the so-called Hubble length H^{-1}, for which $k \ll \mathcal{H}$.

2. Subhorizon perturbations; perturbations of scales much smaller than the Hubble length, for which $k \gg \mathcal{H}$.

The horizon size is of the same order as the observable part of the universe so what we observe in practice are the subhorizon perturbations. Until recent times the Hubble length has increased (apart from the very early universe, see below) and larger and larger scales have "entered the horizon" and become observationally accessible. Outside the horizon the perturbations evolve slowly. For these scales a particularly useful quantity is the so-called comoving curvature perturbation. In terms of the metric perturbation Φ (in the comoving Newtonian gauge) it is given by

$$\mathcal{R} = \frac{5+3w}{3+3w}\Phi - \frac{2}{3+3w}H^{-1}\dot{\Phi}\,, \tag{2.23}$$

where w is the equation of state parameter of the background fluid. $\mathcal{R}$ is a gauge invariant quantity and for adiabatic perturbations it stays constant outside the horizon. Because of this property it is useful for describing perturbations until they enter the horizon and start to evolve.

As mentioned, many observations suggest that the matter content in the Universe is dominated by the extremely weakly interacting dark matter (for a review, see eg. Bertone et al., 2005). Even though ordinary baryonic matter can acquire significant pressure when coupled to radiation (prior to formation of the CMB) or when heated, dark matter dominates the formation of structure and thus pressureless matter is a good approximation for many aspects of structure formation. In the case of pressureless subhorizon matter perturbations, Eqs. 2.20–2.22 take the form (now in terms of cosmic time)

$$\dot{\delta} + \frac{ikv}{a} = 0 \tag{2.24}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(av) + ik\Phi = 0 \tag{2.25}$$

$$-k^2\Phi = 4\pi G a^2 \bar{\rho}\delta\,. \tag{2.26}$$

The evolution of sub-horizon matter perturbations in the weak-field limit (guaranteed by the smallness of perturbations) corresponds to a homogeneously expanding fluid under Newtonian gravity, with the expansion parameterized by $a(t)$. In this picture the metric perturbation Φ can be identified with the gravitational potential.

Inserting Eqs. 2.24 and 2.26 into Eq. (2.25) gives the pressureless Jeans equation

$$\ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\delta = 0\,. \tag{2.27}$$

Since this equation does not depend on k it also holds for the real-space perturbation $\delta(\vec{x})$. The equation has two solutions, a growing mode δ_+ and a decaying mode δ_-. The exact behavior depends on the background cosmology via the Hubble parameter $H(t)$. The current understanding is that the Universe was radiation dominated for the first $\sim 50\,000$ years and is currently transitioning to the dark energy dominated era. In the epoch between matter dominates. So, for most of its history the expansion has been dominated by matter. As we will see later, the matter dominated case is also particularly important for the study of the formation of the large-scale structure. The behavior of the density contrast in the matter dominated epoch turn out to be very simple

$$\delta_+ \propto a\,, \qquad \delta_- \propto a^{-3/2}\,. \tag{2.28}$$

More general case, which is becoming more and more important as the matter density dilutes, includes also the dark energy Ω_Λ:

$$\delta_+ \propto \sqrt{\Omega_\mathrm{m} a^{-3} + \Omega_\Lambda} \int^a \frac{x^{3/2}\mathrm{d}x}{[1 + \Omega_\Lambda/\Omega_\mathrm{m} x^3]^{3/2}}\,, \tag{2.29}$$

$$\delta_- \propto \sqrt{\Omega_\mathrm{m} a^{-3} + \Omega_\Lambda}\,, \qquad (\Omega_\mathrm{m} + \Omega_\Lambda = 1)\,. \tag{2.30}$$

After a while the decaying mode dies out and the only evolution in the density field is that the growing mode grows proportional to the background cosmology dependent growth function $\delta(a) \propto D(a)$. The evolution of the growth function is captured by the growth rate

$$f \coloneqq -\frac{\mathrm{d}\ln D(z)}{\mathrm{d}\ln(1+z)}\,. \tag{2.31}$$

In matter dominated and ΛCDM cosmologies the expansion history is completely determined by H_0 and Ω_m and thus $f = f(\Omega_\mathrm{m})$ and we can define the growth index

$$\gamma \coloneqq \frac{\mathrm{d}\ln f}{\mathrm{d}\ln \Omega_\mathrm{m}(z)}\,. \tag{2.32}$$

In ΛCDM

$$f(a) = \frac{1}{1+(\Omega_\Lambda/\Omega_\mathrm{m})a^3}\left[\frac{a\sqrt{\Omega_\mathrm{m}}}{I(a)} - \frac{3}{2}\right], \tag{2.33}$$

$$\Omega_\mathrm{m}(a) = \frac{1}{1+(\Omega_\Lambda/\Omega_\mathrm{m})a^3}\,, \tag{2.34}$$

with $\Omega_\Lambda = 1 - \Omega_\mathrm{m}$ and $I(a)$ equals the right-hand side of Eq. (2.29) (which equals the growth function $D(a)$ up to a normalization). A good approximation for Eq. (2.33) is $f(z) \approx \Omega_\mathrm{m}(z)^{0.55}$ (see eg. Lahav et al., 1991), which implies that $\gamma = \text{constant} \approx 0.55$. This is an important result since it is general to all ΛCDM models and deviations from it would mean the standard model is incorrect. We will see in Chapter 4 that the growth function can be inferred from measurements of the large scale galaxy distribution. Euclid, for example, is expected to determine the growth index to an accuracy better than ± 0.02 (Laureijs et al., 2011).

The perturbations are thought to be produced in the very early universe in the process of inflation, see eg. Liddle and Lyth (2000). Inflation is a period of rapid, almost exponential expansion of the Universe. This expansion is driven by some quantum field[5]. The amplitude of the field, and thus its energy density, fluctuates randomly on microscopic scales. The rapid expansion of space stretches these fluctuations onto macroscopic scales and they "freeze" to become classical perturbations (during the inflation $\mathcal{H}$ grows and scales "exit the horizon"). Eventually the field transfers through not-fully-understood processes its energy to standard model particles that inherit the density perturbations. Since the production of the perturbations is inherently a random process we cannot predict the value of, for example, the density contrast at different points in space, neither can we predict the amplitude of each Fourier component. What we can predict are their statistical properties.

If the Fourier components of a perturbation h follow a Gaussian distribution, which is a very general prediction of inflation, its statistical properties are fully captured by the power spectrum, defined by

$$P_h(k) \coloneqq \langle |h_{\vec{k}}|^2 \rangle. \tag{2.35}$$

Here $\langle\,\rangle$ denotes the ensemble average and the assumption of isotropy shows up in that the power spectrum only depends on the magnitude of $\vec{k}$. Particularly important is the power spectrum of the matter density perturbations $P(k) = \langle |\delta_k|^2 \rangle$, known also as matter power spectrum. Many inflationary models predict a spectrum for the curvature perturbation (2.23) that is close to scale-invariant:

$$P_\mathcal{R}(k) = A_s^2 \left(\frac{k}{k_p}\right)^{n_s - 1}\,, \tag{2.36}$$

where A_s is the amplitude of perturbations at some pivot scale k_p and n_s is called the spectral index. It is close to, but not exactly one which means small deviations from scale invariance (Planck Collaboration, 2020c). As discussed earlier, for adiabatic perturbations this spectrum stays constant outside the horizon. It can be converted into matter power spectrum by converting first into the metric perturbation via Eq. (2.23) and then into density perturbations via equations

[5]There are models for many-field inflation. However, the adiabaticity of primordial perturbations, for example, suggest that one field dominates the process.

Eqs. (2.20)–(2.22). When the equation of state parameter w of the cosmic fluid stays constant, Eq. (2.23) can be solved for Φ to obtain

$$\Phi = -\frac{3+3w}{5+3w}\mathcal{R}\,, \tag{2.37}$$

up to a decaying part. So, outside the horizon also Φ stays constant, except for when w changes, which happens when the Universe transitions between domination of different energy components, for example from radiation to matter domination. The times of these transitions introduce special scales to the power spectrum, most importantly the scale $k_{\rm eq}$ that enters the horizon during matter-radiation equality.

The linear large-scale evolution outlined above is often encoded into so-called transfer function $T(t,k)$. It translates the initial value of the power spectrum into corresponding value at later times

$$\delta_k(t) = \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)T(k,t)\mathcal{R}\,. \tag{2.38}$$

Normalization $\frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)$ is chosen because for scales that enter the horizon during matter domination

$$\delta_k(t) = \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)\mathcal{R}\,, \tag{2.39}$$

and thus $T(t,k) \approx 1$ for many currently observable scales. Accurate computation of the transfer function needs numerical codes or analytic fitting formulae. Examples of the former are CAMB (Lewis et al., 2000) and CLASS (Lesgourgues, 2011) and of the latter the transfer functions by Bardeen et al. (1986) and Eisenstein and Hu (1999). Like shown earlier, after entering the horizon the matter density perturbations evolve according to the growth function $D(a)$.

2.3 Gravitational collapse

In the previous section I outlined the evolution of density perturbations within the first order perturbation theory. However, the structures we can observe, such as galaxies and their clusters, have undergone significant non-linear evolution. Ultimately, this evolution can only be accurately followed by numerical simulations, for a review see eg. Dolag et al. (2008). We can, however, gain some insight and find context for some quantities defined later in Chapter 3 by considering couple of simple collapse models.

As mentioned, matter starts to dominate the expansion of the Universe at the age of some tens of thousands of years. The CMB measurements tell us that when the Universe was 380 000 years old the density perturbations were as small as $\sim 10^{-5}$, so the perturbations are well within the linear regime deep into the matter dominated epoch. On the other hand, for a standard set of ΛCDM parameters, dark energy was a subdominant component until $\sim$4 billion years ago and all the non-linear structures we observe are at least of the order of billion years old. Thus, in the context of the study of the non-linear collapse of structures a reasonable first approximation for the background cosmology is a flat matter dominated FLRW universe. This is also known as the Einstein-de Sitter model after its original proposers. The Einstein-de Sitter model significantly simplifies calculation which allows us to gain valuable insight. In more precise calculations, dark energy should be included.

If we assume a spherically symmetric overdensity embedded in a homogeneous FLRW universe, the evolution of the perturbation can be solved analytically. A standard calculation can be found in textbooks, Kolb and Turner (1990) for example. For illustration, I will review the main points. The aforementioned overdense region follows the expansion law of a closed ($\Omega > 1$) FLRW Universe. This expansion reaches its maximum at some time $t_{\rm max}$, after which the expansion turns into contraction. The expansion law can be expressed with an auxiliary variable called the development angle ψ:

$$\begin{aligned} a(\psi) &= a_{\rm i}\frac{\Omega_{\rm i}}{2(\Omega_{\rm i}-1)}(1-\cos\psi)\,,\\ t(\psi) &= H_{\rm i}^{-1}\frac{\Omega_{\rm i}}{2(\Omega_{\rm i}-1)^{3/2}}(\psi-\sin\psi)\,. \end{aligned} \tag{2.40}$$

Here the subscript i refers to the value of the quantity at some early time when Ω is still very close to unity. The solution of (2.40) traces a cycloid curve in the (t, a) plane. The expansion reaches its maximum at $\psi = \pi$ and the region collapses into a point by $\psi = 2\pi$. We can also solve for the density parameter and the Hubble parameter as a function of the development angle

$$\Omega(\psi) = \frac{2}{1+\cos\psi}\,, \tag{2.41}$$

$$H(\psi) = 2H_{\rm i}\frac{(\Omega_{\rm i}-1)^{3/2}}{\Omega_{\rm i}}\frac{\sin\psi}{(1-\cos\psi)^2}\,. \tag{2.42}$$

On the other hand, the Hubble parameters in the background universe and in the overdensity follow their respective Friedmann equations

$$\begin{aligned}\overline{H}^2 &= \frac{8\pi G}{3}\overline{\rho}\,,\\ H^2 &= \frac{8\pi G}{3}\Omega^{-1}\rho\,.\end{aligned} \tag{2.43}$$

The density within the overdense region can be expressed in terms of the density contrast as $\rho = (1+\delta)\overline{\rho}$, so that

$$1+\delta = \Omega\frac{H^2}{\overline{H}^2}\,. \tag{2.44}$$

We can express the initial density contrast in terms of the initial density parameter within the overdensity by expanding Ω, H and $\overline{H}$ in the early times when $\psi \ll 1$ and $(\Omega-1) \ll 1$

$$\begin{aligned}&\Omega_{\rm i} \approx 1+\frac{1}{4}\psi^2 \quad \text{and} \quad \frac{H_{\rm i}^2}{\overline{H}_{\rm i}^2} \approx 1-\frac{1}{10}\psi^2\\ \Rightarrow \quad &1+\delta_{\rm i} \approx 1+\frac{3}{20}\psi^2 \quad \Rightarrow \quad \delta_{\rm i} \approx \frac{3}{5}(\Omega_{\rm i}-1)\,.\end{aligned} \tag{2.45}$$

In a matter dominated universe the linear perturbations grow proportional to the scale factor a (of the background Universe) and $a \propto t^{2/3}$. Thus, by the time the spherical overdensity has reached its maximum expansion, the linear theory would predict a density contrast of

$$\delta_{\rm max} = \frac{\overline{a}_{\rm max}}{\overline{a}_{\rm i}}\delta_{\rm i} = \left(\frac{t_{\rm max}}{t_{\rm i}}\right)^{2/3}\delta_{\rm i} \approx \left(\frac{3\pi}{4}\right)^{2/3}\frac{\delta_{\rm i}}{\Omega_{\rm i}-1} \approx \frac{3}{5}\left(\frac{3\pi}{4}\right)^{2/3} \approx 1.0624\,. \tag{2.46}$$

In the third step I have approximated that

$$t_{\rm max} \approx \frac{\pi}{2}\overline{H}_{\rm i}^{-1}\frac{1}{(\Omega_{\rm i}-1)^{3/2}} \quad \text{and} \quad t_{\rm i} = \frac{2}{3}\overline{H}_{\rm i}^{-1}\,.$$

The first approximation comes from $\Omega_{\rm i} - 1 \ll 1$ and second one from the matter dominated expansion law and the fact that at the early times the overdensity is small so that $H_{\rm i} \approx \overline{H}_{\rm i}$. After the overdensity starts to contract it takes another $t_{\rm max}$ for the region to have fully collapsed so that the linearly extrapolated density contrast is

$$\delta = 2^{2/3}\delta_{\rm max} \approx 1.6865\,. \tag{2.47}$$

This number is of central importance in the study of the large scale structure of the Universe. The standard practice is to study the linear evolution of the density field up to this value, which is called the critical overdensity, $\delta_{\rm c}$, after which the overdense region is thought have collapsed into a gravitationally bound stable structure. The above treatment can be generalized to the case of ΛCDM Universe, see eg. Nakamura and Suto (1997) and Mo et al. (2010). In this case the expansion laws of the background Universe and the overdensity become more complicated. It turns out, however, that this does not change the value of the critical overdensity significantly. In Mo et al. (2010) they obtain the approximate value of

$$\delta_{\rm c} \approx 1.686\,[\Omega_{\rm m}(2t_{\rm max})]^{0.0055}\,. \tag{2.48}$$

With such a weak dependence on Ω_m in the background Universe, the value $\delta_\mathrm{c} \approx 1.69$ is a good approximation for any reasonable cosmology.

The bound concentrations of dark matter resulting from the gravitational collapse are called halos. To get an estimate for the true, non-linear density of the halos (in contrast to the linearly interpolated value of 1.69) a common assumption is that a collapsing structure virializes at half of its radius of maximum expansion. Within the spherical collapse model in an Einstein-de Sitter universe this leads to density of factor of $\Delta_\mathrm{c} = 18\pi^2 \approx 178$ larger than the mean density at the time of the collapse. This factor is roughly 200 so that the size of a halo is often defined by the radius within which the density is 200 times the mean density at the time the halo formed. If dark energy is included, the overdensity of the collapsed halo will depend on Ω_m. This dependence can be approximated by

$$\Delta_\mathrm{c}(z) = 18\pi^2 + 82[\Omega_\mathrm{m}(z) - 1] - 39[\Omega_\mathrm{m}(z) - 1]^2 \tag{2.49}$$

(Bryan and Norman, 1998). However, in practice $\Delta_\mathrm{c} = 200$ is still often used to define halos.

The treatment above holds for both homogeneous spherical regions and spherically symmetric concentric shells. In the latter case the each overdense region is a spherical shell, thin enough to be approximated homogeneous. In this case the density profile needs to drop towards larger radii so that the inner spheres would collapse first and the outer shells would not cross them prematurely. However, spherical symmetry of either kind is an oversimplification. A more general assumption is that the linear overdensities will evolve into ellipsoids that will undergo the non-linear collapse, see eg. Bond and Myers (1996). In this case the region will first collapse along the shortest axis of the ellipsoid and form flat disk, followed by the collapse along the two shorter axes.

The evolution and collapse of an ellipsoidal overdensity can be studied within the so-called Zel'dovich approximation, presented in Zel'dovich (1970). In this approach the evolution of the density field is traced by mass elements of an initial density field (the density field at some time deep in the linear regime), whose locations will in time be displaced along straight trajectories proportionally to the gradient of the gravitational potential Φ. In an Einstein-de Sitter universe the displacement of a mass element as a function of the scale factor is is given by

$$\Delta\vec{x} = -\frac{D(a)}{4\pi G\bar{\rho}a^3}\nabla\Phi_\mathrm{i}(\vec{x}_\mathrm{i})\,, \tag{2.50}$$

where $\Phi_\mathrm{i}(\vec{x}_\mathrm{i})$ is the initial gravitational potential at the initial position of the mass element and growth function is normalized so that initially $D(a_\mathrm{i}) = 1$. Turns out that the elliptical collapse is determined by the eigenvalues of what is known as the deformation tensor, $\partial_j\partial_k\left(\Phi_\mathrm{i}/4\pi G\bar{\rho}a^3\right)$. In terms of its three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ the density contrast is given by

$$1 + \delta = \frac{1}{[1 - \lambda_1 D(a)]\,[1 - \lambda_2 D(a)]\,[1 - \lambda_3 D(a)]}\,. \tag{2.51}$$

In the linear regime when $\lambda_1 D(a) \ll 1$, $\delta = D(a)(\lambda_1 + \lambda_2 + \lambda_3) = D(a)\delta_\mathrm{i}$, where δ_i is the initial density contrast[6], ie. the evolution matches the linear prediction. The novelty of the approach is postulating that the approximation holds even when $\lambda_1 D(a) \sim 1$ as the region collapses.

The initial overdensity and the eigenvalues of the deformation tensor enter the critical density of the elliptic collapse in the form of the following two parameters

$$e \coloneqq \frac{\lambda_1 - \lambda_3}{2\delta\mathrm{i}}\,, \tag{2.52}$$

$$p \coloneqq \frac{\lambda_1 + \lambda_3 - 2\lambda_2}{2\delta\mathrm{i}}\,. \tag{2.53}$$

$e \geq 0$ and measures ellipticity in the (λ_1, λ_3) plane and p the corresponding oblateness ($0 \leq p \leq e$) or prolateness ($0 \geq p \geq -e$). By fitting to numerical solutions of the full ellipsoidal evolution, Sheth et al. (2001) show that in an Einstein-de Sitter universe a good estimate for the linearly extrapolated overdensity corresponding to the collapse along the longest principal axis can be obtained by solving

$$\frac{\delta_e}{\delta_\mathrm{c}} = 1 + \beta\left[5(e^2 \pm p^2)\frac{\delta_e^2}{\delta_\mathrm{c}^2}\right]^\gamma \tag{2.54}$$

[6] $\lambda_1 + \lambda_2 + \lambda_3 = \mathrm{Tr}\left[\partial_j\partial_k\left(\Phi_\mathrm{i}/4\pi G\bar{\rho}a^3\right)\right] = \nabla^2\left(\Phi_\mathrm{i}/4\pi G\bar{\rho}a^3\right) = \delta_\mathrm{i}$

for δ_e. Here δ_c is again the critical overdensity for spherical collapse and the plus (minus) sign is used for negative (positive) values of p. A good agreement with the numerical results is found with values $\beta \approx 0.47, \gamma \approx 0.615$. Eq. (2.54) shows that the critical overdensity for ellipsoidal collapse, when defined by the collapse of the longest axis, is larger than for spherical collapse meaning that assumption of ellipsoidal collapse will postpone the formation of structures compared to the spherical case.

Chapter 3

Halo statistics

The previous chapter dealt with a continuous density field. What we mostly observe, however, are compact objects, such as galaxies and galaxy clusters, distribution of which is thought to trace the underlying matter distribution. There is strong evidence that these objects reside in halos of dark matter[1]. These halos are collapsed, gravitationally bound objects with relatively well-defined mass and spatial size. In the previous chapter I outlined a simple model for formation of such an object. It turned out that the mass within a region will undergo a gravitational collapse if the linearly extrapolated density contrast reaches a value greater than $\delta_{\rm c} \approx 1.69$ within that region.

The study of formation and evolution of the dark matter halos is based on the idea that halos that form at a given time correspond to the peaks of the density field that reach $\delta_{\rm c}$ at that time. It is therefore important to understand the statistics of such peaks. In Sec 2.2 I outlined that in the linear regime and during the matter dominated epoch the matter density contrast grows proportional to the linear growth rate, $\delta(\vec{x},t) \propto D(t)$. Within this approximation the density field at any time is just a scaled version of the field $\delta(\vec{x},t_{\rm i})$ at some early time $t_{\rm i}$. Thus, a lot can be learned by inspecting the properties of this initial density field, denoted by $\delta(\vec{x})$ from now on.

Mathematically speaking, the density field can fluctuate on arbitrarily small distance scales. We, however, are interested in fluctuations of some finite extent that will form structures of observable size. This is achieved by considering a smoothed density field

$$\delta(\vec{x};R) \coloneqq \int \delta(\vec{x}')W(\vec{x}+\vec{x}';R){\rm d}^3x\,, \tag{3.1}$$

where the density field is convolved with a filter function $W(\vec{x};R)$ of radius R. The filter function defines the averaging of spatial scales below the scale R. There is some freedom in choosing the filter but a common choice is a top-hat filter defined by

$$W(\vec{x};R) = 1/V(R) \quad \text{for} \quad |\vec{x}| \le R\,, \tag{3.2}$$

$$W(\vec{x};R) = 0 \quad \text{for} \quad |\vec{x}| > R\,. \tag{3.3}$$

Normalization $V(R)$ corresponds to the volume of the spatial region the window function averages field values over and is obtained by integrating the window over the whole space, so that $V(R) = (4\pi/3)R^3$. Alternatively, we can use the average mass within the filter radius as the variable for the filtering scale: $M \coloneqq \overline{\rho}V(R) = \overline{\rho}\frac{4\pi}{3}R^3$. Another useful filter, especially in more theoretical considerations, is the k-space top-hat filter:

$$\widetilde{W}_k(\vec{k};R) = 1 \quad \text{for} \quad |\vec{k}| \le 1/R \tag{3.4}$$

$$\widetilde{W}_k(\vec{k};R) = 0 \quad \text{for} \quad |\vec{k}| > 1/R\,, \tag{3.5}$$

where $\widetilde{W}_k$ denotes the Fourier transform of W_k. In coordinate space this becomes

$$W_k(\vec{x};R) = \frac{1}{2\pi^2R^3}\frac{\sin y - y\cos y}{y^3}\,, \quad y \coloneqq |\vec{x}|/R\,. \tag{3.6}$$

[1] The kinematics of stars (galaxies in case of clusters) and gas require significantly larger concentrated mass than the luminous matter in these objects. In the case of clusters the excess matter also shows up in gravitational lensing.

The integral of W_k over the whole space diverges, but formally its volume can be defined by $W_k(0;R)V_k(R) = 1 \Rightarrow V_k(R) = 6\pi^2 R^3$.

3.1 Statistics of density peaks

Since collapsed objects correspond to density peaks that reach a certain height, an estimate for number density and spatial distribution of such peaks is useful information. Assuming the density field is Gaussian – an assumption supported by both the theory of inflation and observations on linear scales – these statistics can be obtained rigorously for the peaks of the smoothed density field rather straightforwardly. For a detailed calculation see Bardeen et al. (1986). Here I will just quote the most important results.

The peak height can be parameterized relative to the expected variance of the linear density field below some scale R

$$\sigma^2(R) \coloneqq \langle \delta(\vec{x}, R)^2 \rangle = \frac{1}{2\pi^2} \int P(k) \widetilde{W}^2(kR) k^2 \mathrm{d}k \,, \tag{3.7}$$

where $P(k)$ is the linear matter power spectrum and $\widetilde{W}(kR)$ is the Fourier transform of a top-hat filter. Since $\sigma^2(R)$ is the variance of the smoothed density field, $\sigma(R)$ is its root mean square (RMS). Same way as a filter corresponds to a mass, $\sigma(R)$ corresponds to density fluctuations at a mass scale $M = \overline{\rho}\frac{4\pi}{3}R^3$. The peak height is then defined as $\nu \coloneqq \delta(\vec{x};R)/\sigma(R)$ ie. it measures how the perturbation compares to typical fluctuations of the density field at the same scale. Below we will also need the so-called spectral moments of the density field

$$\sigma_\ell^2(R) \coloneqq \frac{1}{2\pi^2} \int k^{2\ell} P(k) \widetilde{W}^2(kR) k^2 \mathrm{d}k \,. \tag{3.8}$$

Here ℓ takes values $\ell = 0, 1, 2, \ldots$ and $\ell = 0$ corresponds to the RMS fluctuations defined in Eq. (3.7).

If the density field is Gaussian, so is the smoothed density field. This allows deriving the number density of peaks by inspecting the expected number of extremal values of the field under constraints on the height of the peak and that the extremum is a local maximum. The comoving number density of peaks in the height range $\nu + \mathrm{d}\nu$ turns out to be given by the following, rather complicated expression

$$n_\mathrm{p}(\nu)\mathrm{d}\nu = \frac{1}{(2\pi)^2 R_*^3} e^{-\nu^2/2} G(\gamma, \gamma\nu)\mathrm{d}\nu \,. \tag{3.9}$$

Here

$$R_* \coloneqq \sqrt{3}\frac{\sigma_1(R)}{\sigma_2(R)} \,, \quad \gamma \coloneqq \frac{\sigma_1^2(R)}{\sigma_2(R)\sigma_0(R)} \,,$$

and

$$G(\gamma, y) \coloneqq \frac{1}{\sqrt{2\pi(1-\gamma^2)}} \int_0^\infty \exp\left[-\frac{(x-y)^2}{2(1-\gamma^2)}\right] f(x)\mathrm{d}x \,,$$

with

$$f(x) \coloneqq \frac{x^3 - 3x}{2} \left\{ \mathrm{erf}\left[\left(\frac{5}{2}\right)^{1/2} x\right] + \mathrm{erf}\left[\left(\frac{5}{2}\right)^{1/2} \frac{x}{2}\right] \right\} + \left(\frac{2}{5\pi}\right)^{1/2} \left[\left(\frac{31x^2}{4} + \frac{8}{5}\right) e^{5x^2/8} + \left(\frac{x^2}{2} - \frac{8}{5}\right) e^{-5x^2/2}\right] \,,$$

where erf is the error function

$$\mathrm{erf}(x) \coloneqq \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} \mathrm{d}y \,. \tag{3.10}$$

To find out how the peak number density modulates spatially we can study the density field smoothed at two scales, R_b and R_s (b refers to "background" and s to "smoothed") such that

$R_\mathrm{b} > R_\mathrm{s}$. The number density of peaks of height $\delta_\mathrm{s}/\sigma_\mathrm{s} = \nu_\mathrm{s}$ given that the background density field has the value $\delta_\mathrm{b}/\sigma_\mathrm{b} = \nu_\mathrm{b}$ in the same region can be written as

$$n_\mathrm{p}(\nu_\mathrm{s}|\nu_\mathrm{b})\mathrm{d}\nu_\mathrm{s} = \frac{n_\mathrm{s}(\nu_\mathrm{s},\nu_\mathrm{b})\mathrm{d}\nu_\mathrm{s}\mathrm{d}\nu_\mathrm{b}}{P(\nu_\mathrm{b})\mathrm{d}\nu_\mathrm{b}}\,. \tag{3.11}$$

Here $P(\nu_\mathrm{b})\mathrm{d}\nu_\mathrm{b}$ is the probability that the background field has amplitude $\nu_\mathrm{b} + \mathrm{d}\nu_\mathrm{b}$ and is simply Gaussian for the underlying Gaussian density field. $n_\mathrm{p}(\nu_\mathrm{s},\nu_\mathrm{b})$ is the number density of peaks at locations where the background field has amplitude $\nu_\mathrm{b} + \mathrm{d}\nu_\mathrm{b}$. It can be derived the same way as equation (3.9) for the number density of peaks but with the extra restricting condition $\delta_\mathrm{b}/\sigma_\mathrm{b} = \nu_\mathrm{b}$. In the limit where we only take into account large-scale fluctuations of the background field $R_\mathrm{b} \gg R_\mathrm{s}$ the conditional peak number density can be written as

$$n_\mathrm{p}(\nu_\mathrm{s}|\nu_\mathrm{b})\mathrm{d}\nu_\mathrm{s} = \frac{1}{(2\pi)^2 R_*^3} e^{-\nu_\mathrm{s}^2/2} G(\gamma_\mathrm{s}', \gamma_\mathrm{s}'\nu_\mathrm{s})\mathrm{d}\nu_\mathrm{s}\,, \tag{3.12}$$

where

$$\gamma_\mathrm{s}' \coloneqq \frac{\gamma_\mathrm{s}}{\sqrt{1-\epsilon^2}}\,, \quad \nu_\mathrm{s}' \coloneqq \frac{\nu_\mathrm{s} - \epsilon\nu_\mathrm{b}}{\sqrt{1-\epsilon^2}}\,, \quad \epsilon \coloneqq \langle \nu_\mathrm{s}\nu_\mathrm{b}\rangle\,.$$

It can be shown (see eg. Bower, 1991) that the covariance of the small-scale and the background fields $\epsilon = \sigma_\mathrm{b}/\sigma_\mathrm{s}$ if they are defined via the k-space top-hat filter. For real-space top-hat filter this holds approximately. If the spectral index of of the power spectrum of the density perturbations is larger than -3, $\sigma(R)$ decreases with increasing R and thus in the limit $R_\mathrm{b} \gg R_\mathrm{s}$ $\epsilon \to 0$ and

$$n_\mathrm{p}(\nu_\mathrm{s}|\nu_\mathrm{b}) \approx n_\mathrm{p}(\nu_\mathrm{s}'), \quad \nu_\mathrm{s}' \coloneqq (\delta_\mathrm{s} - \delta_\mathrm{b})/\sigma_\mathrm{s}\,. \tag{3.13}$$

So in this limit the effect of background density field is to shift the peak height from $\delta_\mathrm{s}/\sigma_\mathrm{s}$ to $(\delta_\mathrm{s} - \delta_\mathrm{b})/\sigma_\mathrm{s}$. To see what kind of overdensity of peaks this produces one can study its ratio to the unconditional number density of peaks

$$\delta_\mathrm{p}(\nu_\mathrm{s}|\nu_\mathrm{b}) \coloneqq \frac{n(\nu_\mathrm{s}|\nu_\mathrm{b})}{n(\nu_\mathrm{s})} - 1\,. \tag{3.14}$$

If we assume that $\delta_\mathrm{b} \ll \delta_\mathrm{s}$ (which is true when the peaks are high enough and background smoothing scale large enough), which implies that $\epsilon\nu_\mathrm{b}/\nu_\mathrm{s} \ll 1$, we can obtain

$$\delta_\mathrm{p}(\nu_\mathrm{s}|\nu_\mathrm{b}) = \frac{\nu_\mathrm{s}^2 - g_1}{\delta_\mathrm{s}}\delta_\mathrm{b}\,, \quad \text{with} \quad g_1 \coloneqq \left.\frac{\partial \ln G(\gamma_s, y)}{\partial y}\right|_{y=\gamma_\mathrm{s}\nu_\mathrm{s}}\,. \tag{3.15}$$

The above number densities were derived for the initial density field. All the quantities involved, namely ν, γ, R_*, are different ratios of δ and σ_l that, in the linear regime, scale identically in time. Because of this the results hold for arbitrary time. What is not taken into account, however, is the dynamic evolution of the volume of the overdensity δ_b (it will contract relative to the global expansion due to the overdensity). Initially, when the density fluctuations were small, the mass of the matter within the region was $M = V_\mathrm{L}\bar{\rho}$. The initial volume V_L is called the Lagrangian volume. Since matter is conserved, as the density contrast grows the same mass is at a later time contained in a volume determined by $M = V_\mathrm{E}\bar{\rho}(1 + \delta_\mathrm{b})$. This evolved volume is called the Eulerian volume. Thus, the evolution of the volume results in an enhancement factor $V_L/V_E = 1 + \delta_\mathrm{b}$ for the ratio of conditional and unconditional number densities and to first order in δ_b we finally obtain

$$\delta_\mathrm{p} = \left(1 + \frac{\nu_\mathrm{s}^2 - g_1}{\delta_\mathrm{s}}\right)\delta_\mathrm{b} =: b_\mathrm{p}(\delta_\mathrm{s}; R_\mathrm{s})\delta_\mathrm{b}\,. \tag{3.16}$$

Here I have defined the bias factor $b_\mathrm{p}(\delta_\mathrm{s}; R_\mathrm{s})$.

Eq. (3.16) is an important fact about the Gaussian density fields; the overdensity of density peaks is enhanced compared to the large-scale mass overdensity and the enhancement is larger for higher peaks. The current understanding is that the collapsed structures we observe today, that is, galaxies and galaxy clusters, are formed at the locations of these very same peaks and that more massive objects correspond to higher peaks. Thus, we expect their distribution to be a biased tracers of the underlying mass distribution and the bias grows with the mass of the objects.

3.2 Mass function

What we would like to have are predictions for the number density and clustering properties of halos as a function of their observational properties, most importantly masses. Ultimately, these predictions will lead to models for the clustering bias of galaxy clusters, which are necessary for the cosmological analysis presented in Paper III, for example. Unfortunately, the treatment in the previous section does not allow these predictions due to the so-called cloud-in-cloud problem. One could think that a density peaks of sufficient height can be interpret as collapsed halos of known mass $M \propto \overline{\rho} R^3$. However, a mass element associated with a peak $\delta_1 = \delta(\vec{x}; R_1)$ can also be associated with a peak $\delta_2 = \delta(\vec{x}; R_2)$, where $R_2 > R_1$. The situation is that when $\delta_1 > \delta_2$, the overdensity will reach the critical value for collapse first at scale R_1 and collapse into a halo of mass M_1 and later merge to be part of a halo of mass M_2, and should thus be excluded from the number density of mass M_1 halos. If $\delta_1 < \delta_2$ the mass element will collapse directly into a halo of mass M_2 and in this case should be excluded from the M_1 halos as well.

What is needed is a way to partition the initial linear density field into disjoint regions, each of which will form a single collapsed object at later times. Statistics of such a partitioning and consequently the number density of halos of a given mass at a given time can be estimated using the so-called excursion set formalism, due to Bond et al. (1991). They formulate the problem in terms of the density field extrapolated linearly to the present day, namely $\delta_0(\vec{x})$. This way $\delta(\vec{x}, t) = D(t)\delta_0(\vec{x})$ and $D(t)$ grows monotonically to its present time value of unity. The question now becomes at what times do regions of the static field $\delta_0(\vec{x})$ reach a time varying threshold $\delta_\mathrm{c}(t) = \delta_\mathrm{c}/D(t)$.

In this approach a convenient mass variable is $S \coloneqq \sigma^2(M) \coloneqq \left\langle \delta_0^2(\vec{x}, M) \right\rangle$, which, if the spectral index of the matter power spectrum $n_\mathrm{s} > -3$, is a monotonically decreasing function of M. For a fixed location $\vec{x}$, the value of the smoothed density field $\delta_0(\vec{x}; M)$ varies randomly as the filtering scale is varied (which corresponds to varying mass and thus S) and forms a trajectory in the $(S, \delta_0(\vec{x}; S))$ plane. The idea of the excursion set formalism is to postulate that the fraction of mass contained in collapsed halos of mass $M > M_1$ at time t_1 is identified with the fraction of all such trajectories that cross the threshold $\delta_\mathrm{c}(t_1)$ first time at value S_1. See Fig. 3.1 for visual representation of the situation.

The task is then to calculate the fraction of such trajectories. This can be done analytically in the case when the density field is smoothed with a k-space top-hat filter function, defined in Eq. (3.4). In this case the smoothed density field is

$$\delta_\mathrm{s}(\vec{x}; R) = \int \mathrm{d}^3 k \widetilde{W}_k(\vec{k}R) \delta_{\vec{k},0} e^{i\vec{k}\cdot\vec{x}} = \int_{|\vec{k}|<k_\mathrm{max}} \mathrm{d}^3 k \delta_{\vec{k},0} e^{i\vec{k}\cdot\vec{x}}\,, \tag{3.17}$$

where $k_\mathrm{max} = 1/R$, $\delta_{\vec{k},0}$ is the Fourier transform of $\delta_0(\vec{x})$. The reason for using a k-space top-hat filter is that when k_max is increased to $k_\mathrm{max} + \Delta k$ (which corresponds to an increment in the filter radius and mass) the corresponding change $\Delta\delta_\mathrm{s}$ is a Gaussian random variable with variance

$$\sigma^2(k_\mathrm{max} + \Delta k) - \sigma^2(k_\mathrm{max})\,, \tag{3.18}$$

where

$$\sigma^2(k_\mathrm{max}) = \frac{1}{2\pi^2} \int_{k<k_\mathrm{max}} P(k) k^2 \mathrm{d}k\,. \tag{3.19}$$

What is important is that the distribution of $\Delta\delta_\mathrm{s}$ is independent of $\delta_\mathrm{s}(\vec{x}; k_\mathrm{max})$ which means that the steps taken in the (S, δ_s) trajectory are uncorrelated. Thus the trajectories are those of a Markovian random walk and their distribution can be calculated analytically. The probability for a "particle" (ie. a trajectory) that executes a Markovian random walk to be located between δ_s and $\delta_\mathrm{s} + \Delta\delta_\mathrm{s}$ at S is determined by the diffusion equation

$$\frac{\partial \Pi}{\partial S} = \frac{1}{2} \frac{\partial^2 \Pi}{\partial^2 \delta_\mathrm{s}^2}\,, \tag{3.20}$$

where S acts as a time variable and δ_s as spatial displacement. We would like to solve the equation with the boundary condition that the particle starting from $(0, 0)$ reaches (S, δ_s) without reaching

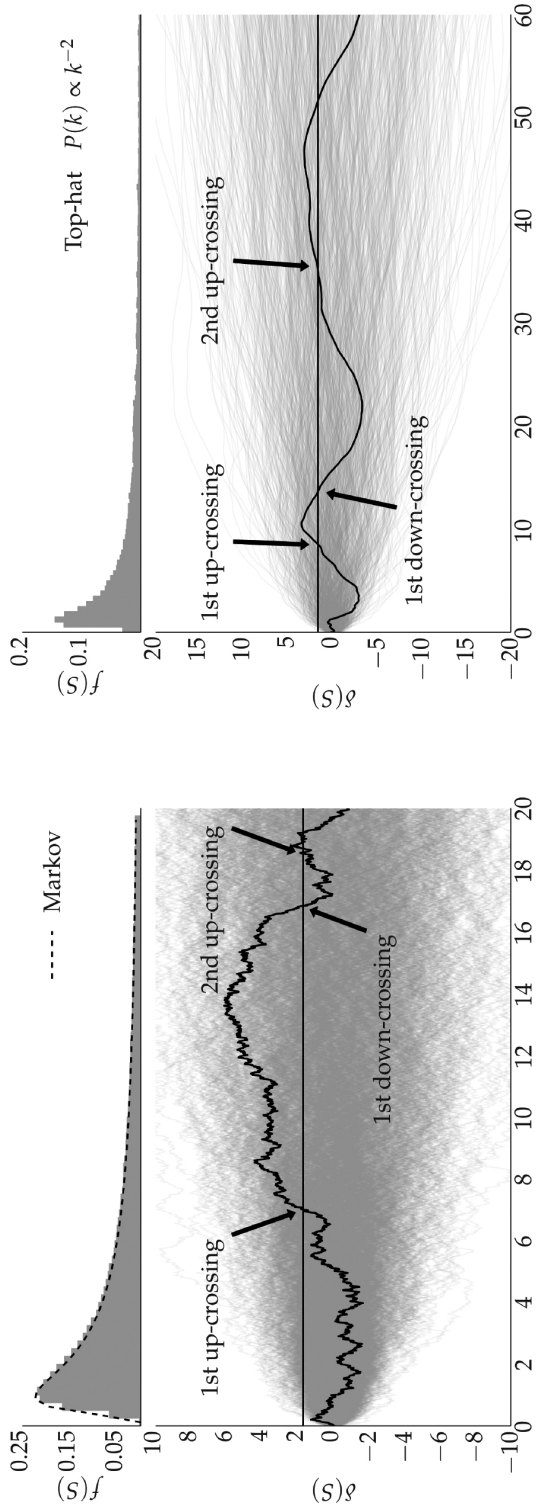

Figure 3.1: Each lower panel shows an ensemble of trajectories in the $(S, \delta_0(\vec{x}, S))$ plane. The left panels show the case corresponding to a density field filtered with the k-space top-hat filter and the right panel corresponding to the real-space top-hat filter. In each panel, one of the trajectories is highlighted to show where it up-crosses the critical barrier for the first time. The plot also shows the first down-crossing and the second up-crossing. The first up-crossing defines the mass-scale the collapsing region is associated with. Upper panels show the first up-crossing distribution estimated using 106 trajectories. The dashed line in the case of the k-space top-hat filter is an analytical prediction. Figure from Nikakhtar et al. (2018).

δ_c before that. This problem is known as diffusion equation with an absorbing barrier and the solution was found by Chandrasekhar (1943)

$$\Pi(\delta_s, S) = \frac{1}{\sqrt{2\pi S}} \left\{ \exp\left(-\frac{\delta_s^2}{2S}\right) - \exp\left[-\frac{(2\delta_c - \delta_s)^2}{2S}\right] \right\} . \tag{3.21}$$

With this solution we can calculate the fraction of trajectories that reach critical density for the first time at $S > S_1$:

$$F(> S_1) = \int_{-\infty}^{\delta_c} \Pi(\delta_s, S_1) \mathrm{d}\delta_s \tag{3.22}$$

By the original ansatz this gives the fraction of mass in halos of $M < M_1$, namely $F(M < M_1)$. If all the mass in halos of some mass[2] $F(M > M_1) = 1 - F(M < M_1)$ and the halo mass function becomes

$$\begin{aligned} n(M,t)\mathrm{d}M &= -\frac{\overline{\rho}}{M}\frac{\partial F(> M)}{\partial M}\mathrm{d}M \\ &= \frac{\overline{\rho}}{M}\frac{\partial F(> S)}{\partial S}\frac{\mathrm{d}S}{\mathrm{d}M}\mathrm{d}M \\ &= \frac{\overline{\rho}}{M}\frac{1}{\sqrt{2\pi}}\frac{\delta_c}{S^{3/2}}\exp\left(-\frac{\delta_c^2}{2S}\right)\left|\frac{\mathrm{d}S}{\mathrm{d}M}\right|\mathrm{d}M . \end{aligned} \tag{3.23}$$

This is often expressed in terms of the variable $\nu \coloneqq \delta_c(t)/\sigma(M)$

$$n(M,t)\mathrm{d}M = \frac{\overline{\rho}}{M^2} f(\nu) \left|\frac{\mathrm{d}\ln\nu}{\mathrm{d}\ln M}\right| \mathrm{d}M , \tag{3.24}$$

where

$$f(\nu) = \sqrt{\frac{2}{\pi}}\nu \exp\left(-\frac{\nu^2}{2}\right) \tag{3.25}$$

is the so-called multiplicity function. The variable ν, sometimes called the significance, quantifies how rare fluctuations of amplitude $\delta_c(t)$ are given the RMS fluctuations of the density field at the scale corresponding to mass M.

In a more general case when the step-size and its variance can depend on the location δ_s (which is the case for filters other than the k-space top-hat filter) Eq. (3.20) will contain extra terms

$$\frac{\partial \Pi}{\partial S} = -\frac{\partial(\mu\Pi)}{\partial \delta_s^2} + \frac{1}{2}\frac{\partial^2(\Sigma^2\Pi)}{\partial^2\delta_s^2} , \tag{3.26}$$

where μ and Σ^2 are the drift and variance parameters, respectively, defined by

$$\mu \coloneqq \lim_{\Delta S\to\infty}\frac{\langle \Delta\delta_s|\delta_s\rangle}{\Delta S} , \quad \text{and} \quad \Sigma^2 \coloneqq \lim_{\Delta S\to\infty}\frac{\langle (\Delta\delta_s)^2|\delta_s\rangle}{\Delta S} .$$

This more general case cannot be solved analytically. One can, however, simulate an ensemble of trajectories subject to Eq. (3.26) and count the fraction of those that cross the barrier $\delta_c(t)$ at a given time and obtain an estimate of the multiplicity function. This is exactly what was done in the upper panels of Fig 3.1.

The excursion set formalism only works in a statistical sense. It predicts the distribution of masses of halos born at a given time but it cannot be used to predict the mass of the halo a particular mass element ends up in. This can be seen eg. by the following example. A spherical region of radius R has mass M that corresponds to the critical density at a time t. If one looks at a point $\vec{x}$ inside the spherical region at distance r from its center the overdensity around this will be below the critical density for all the radii larger than $R - r$. Thus, the mass element at $\vec{x}$ will,

[2]Within the excursion set formalism this corresponds to $\sigma^2(M) \to \infty$ as $M \to 0$, which means that every trajectory reaches δ_c since the "time" goes to infinity. This happens if the linear power spectrum has $n_s > -3$. In reality, even CDM particles have finite velocities and the effective n_s may be < -3 as $k \to \infty$. $\sigma^2(M)$ then approaches a finite value for $M \to 0$ and some mass will not be contained in collapsed objects.

according to the excursion set interpretation, be part of a halo of mass scale $M(R-r)^3/R^3$, which is smaller than M. So for a particular mass element the first upcrossing mass scale would only be the lower limit for the mass of the halo it ends up to.

Everything above was based on the universal critical density δ_c of the spherical collapse model. However, as mentioned in Sec. 2.2 the collapse will more generally be ellipsoidal. For a Gaussian density field, it is possible to derive a distribution function for the ellipticity and prolateness of Eq. (2.54), given the smoothed density field δ_s, as was done by Doroshkevich (1970)

$$\mathcal{P}(e,p|\delta_s) = \frac{1125}{\sqrt{10\pi}} e(e^2-p^2)\left(\frac{\delta_s}{\sigma}\right)^5 \exp\left[-\frac{5\delta_s^2}{2\sigma^2}(3e^2+p^2)\right] . \tag{3.27}$$

e and p are two new random fields that change stochastically as the filter mass is varied. For each value of e and p the critical overdensity is given by Eq. (2.54). It is in principle possible to derive distribution of $(\Delta\delta_s, \Delta e, \Delta p)$ given (δ_s, e, p), execute a set of random walks in (S, δ_s, e, p) and compute the fraction of trajectories that cross the barrier $\delta_c(e,p)$ at a given mass scale. A simpler, yet effective treatments starts with the observation that the density of Eq. (3.27) has its maximum at $p=0$ and $e=(\sigma/\delta_s)/\sqrt{5}$. So, for the most probable ellipsoid that reaches the critical density $p=0$ and $e=(\sigma/\delta_e)/\sqrt{5}$. Inserting this into Eq. (2.54) we obtain

$$\delta_e(t) = \delta_c(t)\left\{1+\beta\left[\frac{\sigma^2}{\delta_c(t)}\right]\right\} . \tag{3.28}$$

Here, again, $\beta \approx 0.47$, $\gamma \approx 0.615$. Even though this critical density only holds for the peak of the ellipticity distribution, it can be used to approximate the average dependence of δ_e on σ. Since the dependence of the critical density on $S=\sigma^2$ is now known, the multiplicity function needed for the mass function can be computed using the excursion set formalism in the presence of a "moving barrier" (ie. a barrier that depends on S), shape of which is known. Sheth et al. (2001) do this and find numerically a multiplicity function that can be approximated by

$$\begin{aligned} f(\nu) &= A\left(1+\frac{1}{\nu^{2q}}\right)\sqrt{\frac{2}{\pi}}\nu\exp\left(-\frac{\nu^2}{2}\right) \\ &= A\left(1+\frac{1}{\nu^{2q}}\right) f_{sc}(\nu) , \end{aligned} \tag{3.29}$$

where $f_{sc}(\nu)$ is the multiplicity function of Eq. 3.25, $q \approx 0.3$ and $A \approx 0.3222$. The value for A is fixed by the requirement that the integral of $f(\nu)$ over ν gives unity so that all the mass in the Universe is contained in halos of some mass, however small.

The multiplicity function (3.29) actually fits the simulated halo distribution best when allowing a scaling of the peak height by $\nu \to \nu' = \sqrt{a}\nu$, where they find the best agreement with $a=0.707$. This additional parameter is not completely ad hoc. Its value is to some extent determined by the halo finding algorithm used to analyze the simulations (and thus, in essence, by the definition of a halo within a simulation). The friends-of-friends method, for example, includes a free parameter, the linking length[3] l. Increasing the linking length will lead to larger number of massive halos and vice versa. The value is often set to $l \approx 0.2$ times the mean inter-particle separation in order to obtain halos with mass densities that match the predictions of the spherical collapse model. Since the computation of multiplicity function (3.29) assumes elliptical collapse, the exact optimal value for l is no longer obvious. The role of the parameter a can be seen as a way for the multiplicity function to be adjusted to a specific value of l.

3.3 Clustering bias

By combining Eq. (3.24) with a suitable multiplicity function, such as that of Eq. (3.29), we now have an estimate for the number density of halos within a given mass range. This allows us to compute how the number density varies with the large-scale fluctuations of the density field, similarly to what was done in Sec. 3.1. Ultimately this will yield an estimate of the bias factor

[3] Any particle that finds another particle within a distance l is linked to it to form a group.

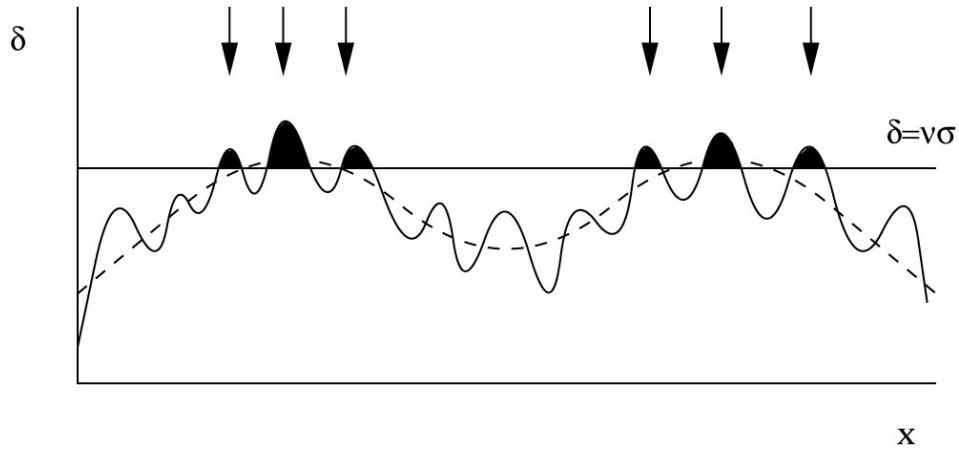

Figure 3.2: Schematic representation of the peak-background split. The horizontal line is the critical density for the collapse and the dashed curve shows the slowly-varying background field. The shaded regions pointed by the arrows are the peaks of the density field that are expected to be sites of halo formation. Figure from Peacock (1999).

for a halo population of a given mass. It can be done using the so-called peak-background split approximation, see, for example, Schmidt et al. (2013) for a detailed account. In this approach the density perturbations smoothed at the scale R_s, namely $\delta(\vec{x}; R_\mathrm{s})$, are though to consist of a large-scale part $\delta_\mathrm{b} \coloneqq \delta(\vec{x}; R_\mathrm{b})$ and a small-scale part $\delta_\mathrm{s} \coloneqq \delta(\vec{x}; R_\mathrm{s}) - \delta_\mathrm{b}$, where $R_\mathrm{s} \ll R_\mathrm{b}$. Fig. 3.2 is a schematic representation of the division. The small-scale part is responsible for the collapse of the halos and the the effect of the large-scale density fluctuations is only to shift the background density within the region of interest by a constant offset δ_b, which will alter the collapse time of halos. The scale of this division makes no difference as long as it is larger than the halo scale and smaller than the scale of the correlations we are interested in.

To see how the background field modulates the halos formation consider the following. In general relativity the evolution of a spherical perturbation is independent of the external universe (Birkhoffs's theorem). This means that a region with physical density of ρ_c will collapse at the same proper time in a universe with background density $\bar{\rho}$ as in a universe with background density $\bar{\rho}' = (1+\delta_\mathrm{b})\bar{\rho}$. In this perturbed background the absolute critical density perturbation is given by $\rho_c - \bar{\rho}' = (1+\delta_\mathrm{c})\bar{\rho} - (1+\delta_\mathrm{b})\bar{\rho} = (\delta_\mathrm{c} - \delta_\mathrm{b})\bar{\rho}$. On the other hand, the significance of these fluctuations can be expressed in terms of physical densities as

$$\nu' = \frac{\rho_c - \bar{\rho}'}{\delta\rho_\mathrm{RMS}}\,, \tag{3.30}$$

where $\delta\rho_\mathrm{RMS}$ is the RMS fluctuation of the physical density field at the scale R_s. This is not affected by a constant offset, so it has the same value $\bar{\rho}\sigma(R_\mathrm{s})$ as in the unperturbed background. The modified significance now becomes

$$\begin{aligned} \nu' &= \frac{(\delta_\mathrm{c} - \delta_\mathrm{b})\bar{\rho}}{\bar{\rho}\sigma(R_\mathrm{s})} \\ &= \frac{\delta_c - \delta_\mathrm{b}}{\sigma(R_\mathrm{s})}\,. \end{aligned} \tag{3.31}$$

Similarly to what was found in Sec. 3.1, the effect of a large-scale fluctuation on the local number density of halos is thus to shift the critical density by $-\delta_\mathrm{b}$. Since $|\delta_\mathrm{b}| \ll 1$, we can Taylor

expand the multiplicity function

$$\begin{aligned} f(\nu') &\approx f(\nu) - \frac{\mathrm{d}f}{\mathrm{d}\delta_\mathrm{c}}\delta_\mathrm{b} \\ &= f(\nu) - \frac{1}{\sigma(R_\mathrm{s})}\frac{\mathrm{d}f}{\mathrm{d}\nu}\delta_\mathrm{b} \end{aligned} \tag{3.32}$$

Using Eq. (3.24) we can then see that the ratio of perturbed and unperturbed halo number densities is given by

$$\frac{n'(M,t,\delta_\mathrm{b})}{n(M,t)} = 1 - \frac{1}{\sigma(M)f(\nu)}\frac{\mathrm{d}f}{\mathrm{d}\nu}\delta_\mathrm{b} \tag{3.33}$$

Assuming the multiplicity function of Eq. (3.29), this yields

$$\frac{n'(M,t,\delta_\mathrm{b})}{n(M,t)} = 1 + \left(\nu^2 - 1 + \frac{2q}{1+\nu^{2q}}\right)\frac{\delta_\mathrm{b}}{\delta_c(t)}\,, \tag{3.34}$$

Like in the case of Eq. (3.15), this does not take into account the dynamical evolution of the volume of the overdense region. The evolution is again given by the ratio of Lagrangian and Eulerian volumes. Including this factor of $(1+\delta_\mathrm{b})$, we have up to first order in δ_b for the excess number density of halos

$$\delta_\mathrm{h}(M,t,\delta_\mathrm{b}) = \frac{n'(M,t,\delta_\mathrm{b})}{n(M,t)} - 1 = \left(1 + \frac{\nu^2-1}{\delta_\mathrm{c}(t)} + \frac{2q/\delta_\mathrm{c}(t)}{1+\nu^{2q}}\right)\delta_\mathrm{b}\,. \tag{3.35}$$

From this expression we can read the bias factor $b(M,t) \coloneqq \delta_\mathrm{h}(M,t,\delta_\mathrm{b})/\delta_\mathrm{b}$. It is often more convenient to define the bias factor as a function of redshift. This is done by using redshift as the time variable for the critical density and in the matter dominated epoch the scaling is simply $\delta_\mathrm{c}(z) = \delta_\mathrm{c}/(1+z)$. Again, the expression (3.35) can be augmented with the scaling of the peak height to $\tilde{\nu} = \sqrt{a}\nu$. The advantage of the above derivation is that the bias factor is now formulated in terms of the same parameters as the mass function.

Eq. (3.35) was originally derived by Sheth and Tormen (1999) as a fix for the poor agreement of the spherical collapse multiplicity function of Eq. (3.25) with simulations. Still, after the introduction of this bias relation, the quest for an even more accurate mass function and bias-to-mass relation has been an active field of research, see eg. Jenkins et al. (2001), Warren et al. (2006), and Tinker et al. (2010). One of the more recent improvements regarding the agreement with simulations is that of Bhattacharya et al. (2011). They introduced an additional parameter p to the multiplicity function of Eq. (3.29) to obtain

$$f(\nu) = A\sqrt{\frac{2}{\pi}}\left(1 + \frac{1}{(\sqrt{a}\nu)^{2q}}\right)(\sqrt{a}\nu)^p \exp\left(-\frac{a\nu^2}{2}\right)\,. \tag{3.36}$$

Correspondingly, the parameter p also enters the mass-to-bias relation

$$b(M) = 1 + \frac{a\nu^2 - p}{\delta_\mathrm{c}} + \frac{2p/\delta_\mathrm{c}}{1+(\sqrt{a}\nu)^{2q}}\,. \tag{3.37}$$

This new parameter is purely ad hoc and is not motivated by the excursion set formalism or any other physical principle. At redshift $z = 0$ they find the best-fit values presented in the first row of Table 3.1. This model was later revised by Comparat et al. (2017) using the MultiDark simulation suite[4], see Prada et al. (2012) and Klypin et al. (2016). They report the constraints presented in the second and the third row of Table 3.1. The reason for two sets of results is that one can go two ways about fixing the mass-to-bias relation of Eq. (3.37); either by fitting the parameters by comparing the mass function from Eq. (3.36) to simulations, and assuming it also accurately describes the bias, or by directly measuring the large-scale bias as a function of halos mass. Comparat et al. (2017) find that the former calibration does not accurately reproduce the clustering observed in their simulation suite and by a direct measurement of the clustering bias obtain the values presented on the third row of Table 3.1.

[4] `https://www.cosmosim.org/`

A	a	p	q	Calibration
0.333	0.788	0.807	1.795	B11 HMF
0.280 ± 0.002	0.903 ± 0.007	0.640 ± 0.026	1.695 ± 0.038	C17 HMF
free	0.740 ± 0.008	0.61 ± 0.02	1.64 ± 0.03	C17 clustering

Table 3.1: Best-fit values for the parameters of the mass function from Bhattacharya et al. (2011) (B11) and Comparat et al. (2017) (C17). Shown for C17 are the parameters inferred from the halo mass function (HMF) and the halo clustering.

For a fixed calibration the bias relation in Eq. (3.37) depends on the cosmology via the growth function and the RMS mass fluctuations defining the significance $\nu = \delta_c/[D(t)\sigma(M)]$. Thus, the agreement between the predicted and measured clustering biases of halo populations of different masses can be used as a test of a cosmological model. In Paper III we use the bias relation of Comparat et al. (2017) to carry out along these lines a proof-of-concept type of analysis to a galaxy cluster sample. For details, see Sec. 4.4.

Chapter 4

Two-point correlation functions

In Sec. 2.2 I defined the power spectrum of the density perturbations. This is a quantity predicted naturally from theory. However, a quantity that is more closely connected to observations is the two-point correlation function (2PCF)

$$\xi(\vec{r}) \coloneqq \langle \delta(\vec{x})\delta(\vec{x}+\vec{r}) \rangle \,. \tag{4.1}$$

It quantifies how the density perturbation at one point depends on perturbations at nearby regions ie. measures strength of structures on different length scales. The 2PCF is closely connected to the power spectrum; they form a Fourier transform pair:

$$\xi(\vec{r}) = \frac{1}{(2\pi)^3} \int \mathrm{d}^3 k P(\vec{k}) e^{i\vec{k}\cdot\vec{r}} \,, \tag{4.2}$$

$$P(\vec{k}) = \int \mathrm{d}^3 r \xi(\vec{r}) e^{-i\vec{k}\cdot\vec{r}} \,, \tag{4.3}$$

and in theoretical considerations they can be used interchangeably.

4.1 Discrete objects

In practice we cannot observe the continuous density field but discrete objects, such as galaxies and their clusters. Correspondingly, the density field $\rho(\vec{r})$ will not be the matter density but the number density of such objects. In this case the 2PCF is defined as the excess probability of finding an object at separation $\vec{r}$ from another, already observed object

$$\mathrm{d}P = \langle \rho \rangle \left[1 + \xi(\vec{r})\right] \mathrm{d}V. \tag{4.4}$$

Here $\langle \rho \rangle$ is the mean number density and $\mathrm{d}V$ is a volume element at separation $\vec{r}$. This definition agrees with the one in Eq. (4.1) when $\delta(\vec{x})$ are interpreted as number density fluctuations. Generally, the separation vector $\vec{r}$ does not need to be a three-dimensional spatial separation. One can study correlations as a function of angular separation or separations along one or more fixed directions. The latter case is discussed in more detail below. The most suitable form of the 2PCF depends on the survey details such as angular and redshift coverage and which cosmological parameters one wants to constrain by the measurement.

As discussed in Chapter 3, the perturbations in galaxy or cluster number density, denoted by δ_i, are thought to be proportional, but not equal to those in the matter density, namely δ_m, so that $\delta_i = b_i \delta_\mathrm{m}$. b_i is the bias factor and it is different for objects of different types, such as different galaxy populations or clusters of different masses. In Sec.3.3 I outlined how the bias of dark matter halos of different masses can be estimated. For galaxy clusters the situation is rather simple, they are thought to be in a one-to-one relation with virialized dark matter halos (eg. Moscardini et al., 2000). Since the 2PCF is essentially the square of the density contrast the proportionality constant for a biased 2PCF is the square of the linear bias, $\xi_\mathrm{i}(\vec{r}) = b_i^2 \xi_\mathrm{m}(\vec{r})$. For example Marulli et al. (2012) have used this simple model to successfully explain clustering of X-ray selected galaxy clusters.

Formation of individual galaxies is a more complex process and they cluster on smaller, non-linear scales. Treating their clustering bias as a simple proportionality constant is consequently only an approximation. In principle it could depend on, for example, redshift and scale. Also, the relation between the matter and number density fluctuations does not need to be exactly linear.

The number density fluctuations can be seen as a combination of two random processes. The first process is the one that generates the matter density fluctuations in the early Universe and the second one populates the density field with point sources. The 2PCF of galaxies or clusters thus contains two types of variance; that of the density field, known as the cosmic variance, and that of discrete sampling of the density field, known as shot noise or sampling variance. Statistical uncertainty in a correlation function measurement is a combination of these two and in general cannot be calculated analytically. It can be estimated from data itself using resampling techniques, such as bootstrapping or jackknifing. However, for an accurate estimate one needs to simulate an ensemble of galaxy or cluster distributions that contain the cosmic and sampling variance and follow the characteristics of the actual survey and compute the 2PCF variance from this mock data set. Some aspects of the error analysis using mock catalogs are discussed in Sec. 4.3.

In principle the 2PCF could be defined as a function of two positions, $\vec{x}, \vec{x'}$ but the assumption of statistical homogeneity forces the 2PCF to only depend on the separation of the positions, $\vec{r} = \vec{x} - \vec{x'}$. Statistical isotropy would in theory further imply that the 2PCF only depends on the magnitude of the separation, $|\vec{r}|$. In an observational setup the isotropy is, however, broken. The radial positions of galaxies or clusters cannot be measured directly but are instead inferred from their redshifts. These redshifts consist of a cosmological redshift, which is determined by their comoving distance, and a redshift caused by their peculiar velocities. The latter will distort the inferred locations along the line-of-sight direction. These distortions come in two flavors. On small scales, where galaxies move essentially randomly within their hosting halos, the structures will be elongated along the line-of-sight due to galaxies being randomly red(blue)shifted to appear further away (closer). On larger scales the galaxies will stream coherently into (away from) over(under)densities. This coherent motion will lead to over(under)dense region to appear contracted (elongated) along the line-of-sight direction. See Fig. 4.1 for a diagrammatic representation. The random motion is more important on scales $\lesssim 1h^{-1}$Mpc and the coherent flow on scales $\gtrsim 1h^{-1}$Mpc (Peacock et al., 2001).

A solution to this problem is to measure correlations along and perpendicular to the line-of-sight directions and integrate over the line-of-sight direction. The idea is that the distortions average out when they are integrated over a large enough region. In addition to redshift space distortions caused by the peculiar velocities, the integration also smooths out random redshift measurement errors (caused by, for example, lack of spectroscopic measurements). One obtains this way a projected correlation function

$$w_p(r_p) = 2\int_0^{\pi_{\rm max}} \xi(r_p, r_\pi) \mathrm{d}r_\pi \,, \tag{4.5}$$

where r_p and r_π are separations along the directions perpendicular and parallel to the line of sight direction, respectively. This definition assumes that the spatial region under study is small compared to its distance from us (ie. it covers a small angle). In this case the redshift space distortions of all the galaxies or clusters can be thought to occur along a single line-of-sight direction. This is called the plane-parallel approximation. The integration limit $\pi_{\rm max}$ should be chosen large enough so that the correlations will be negligible above that scale and increasing it would only increase noise. Since the underlying real-space correlation function is assumed to be homogeneous and isotropic the projected correlation function should match that of a prediction from $\xi(r_p, r_\pi) = \xi\left(\sqrt{r_p^2 + r_\pi^2}\right)$, which gives

$$w_p^{\rm theory}(r_p) = 2\int_{r_p}^{\pi_{\rm max}} \xi(r) \frac{r\mathrm{d}r}{\sqrt{r^2 - r_p^2}} \,. \tag{4.6}$$

In addition to trying to get rid of the redshift space distortions it is possible to exploit them. The peculiar velocities are caused by the inhomogeneities of the cosmic density field, so on large scales we can use linear perturbation theory Eqs. (2.24)–(2.26) to predict statistics of the velocity

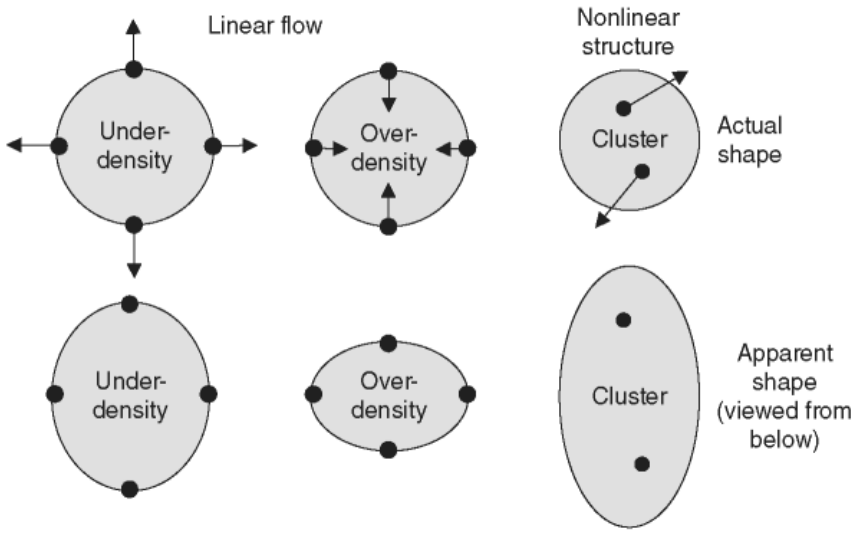

Figure 4.1: Diagrammatic representation of redshift space distortions. Top row shows the actual galaxy positions and bottom row distorted positions. Figure from `https://ebrary.net/77703/sociology/redshift_space_distortions`.

field and thus those of the redshift space distortions. Seminal theoretical progress within this approach was made by Kaiser (1987). The anisotropy of the correlation function can be quantified via its multipole moments

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1}\xi(s,\mu)L_\ell(\mu)\mathrm{d}\mu\,, \tag{4.7}$$

where $\mu \coloneqq \cos\theta$ is the cosine of the angle between the separation vector of two objects and the line-of-sight direction, s is the magnitude of the separation vector and $L_\ell(\mu)$ is the ℓth Legendre polynomial. The odd multipoles ($\ell = 1, 3, \ldots$) vanish since the corresponding Legendre polynomials are odd and $\xi(s,\mu)$ is even in μ (since the 2PCF is defined in terms of pairs of objects and a pair has no distinction between separations of $\vec{r}$ and $-\vec{r}$) so that the product integral vanishes. To have a simple prediction for the even correlation function multipoles we need the following assumptions.

- Density and velocity perturbations and their gradients are small (linear perturbations).
- The objects of interest are far away from us so that the Hubble recession velocity dominates over peculiar velocities.
- Redshifts cover a narrow range $\Delta z \ll 1$. This corresponds to measuring correlation function in redshift bins within each of which the Hubble parameter is approximately constant.
- The plane parallel approximation is valid.

Within these assumptions it turns out that only the first three even multipoles $\ell = 0, 2, 4$ are expected to be non-zero and an estimate for them can be obtained in terms of the parameter

$\beta \coloneqq f/b$, where f is the growth rate defined in Sec. 2.2 and b is the clustering bias, as follows

$$\xi_0(s) = \left(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right)\xi(s)\,, \tag{4.8}$$

$$\xi_2(s) = \left(\frac{4}{3}\beta + \frac{4}{7}\beta^2\right)\xi(s) + \left(-4\beta - \frac{12}{7}\beta^2\right)\frac{J_3(s)}{s^3}\,, \tag{4.9}$$

$$\xi_4(s) = \frac{8}{35}\beta^2\xi(s) + \frac{12}{7}\beta^2\frac{J_3(s)}{s^3} - 4\beta^2\frac{J_5(s)}{s^5}\,. \tag{4.10}$$

Here $\xi(s)$ is the 2PCF without the redshift space distortions and

$$J_\ell(s) \coloneqq \int_0^s \xi(r) r^{\ell-1} \mathrm{d}r\,. \tag{4.11}$$

Eqs. (4.8)–(4.10) show that if we know the clustering bias b we can measure the evolution of the growth rate f by measuring the the correlation function multipoles at various redshifts.

4.2 2PCF estimators and the split method

In practice the 2PCF of a galaxy or cluster distribution is estimated from a catalog of sources. The catalog is a list of object positions in real or redshift space and possibly weights. As discussed in Sec. 4.1, the 2PCF of a point distribution is defined as excess probability of finding pairs of objects separated by $\vec{r}$. An estimator for the 2PCF is thus based on counting numbers of pairs of objects within separation bins. These counts are denoted by $\mathrm{DD}(\vec{r})$ ("data-data" pairs). In the simplest case a pair is included if the absolute value of the separation vector between the points falls within $|\vec{r}| \pm \frac{1}{2}\Delta r$. This should then be compared to the case when the points are not clustered (but are generated through a Poisson process). For this purpose one needs to generate a set of randomly distributed points that has the same angular and redshift coverage and contains the same selection effects as the actual catalog. This is called the random catalog and the pair counts within this catalog are denoted by $\mathrm{RR}(\vec{r})$ ("random-random" pairs).

With these ingredients the simplest estimator for the 2PCF is

$$1 + \xi(\vec{r}) = \frac{N_\mathrm{R}(N_\mathrm{R}-1)}{N_\mathrm{D}(N_\mathrm{D}-1)}\frac{\mathrm{DD}(\vec{r})}{\mathrm{RR}(\vec{r})}. \tag{4.12}$$

Here N_D and N_R are the number of points in the data and random catalogs, respectively. Normalizing the pair counts by the total number of pairs within each catalog, that is $\frac{1}{2}N_i(N_i-1)$, would not be necessary if both the catalogs had the same number of points. However, one usually uses a random catalog that is larger than the data catalog to decrease the sample variance related to the finite number of random points. Dividing by RR counts can be seen as a form of Monte Carlo integration and the accuracy increases with the number of random points. Thus, it is usually more convenient to express everything in terms of the normalized pair counts, which I will denote by lower case letters, so that $\mathrm{dd}(\vec{r}) \coloneqq 2\mathrm{DD}(\vec{r})/\left[N_\mathrm{D}(N_\mathrm{D}-1)\right]$ and so on.

In addition to DD and RR counts one can also count DR ("data-random") pairs, that is, pairs between the data and the random catalog. For normalization purposes, number of such pairs is $N_\mathrm{D}N_\mathrm{R}$. Including the DR counts is useful for minimizing edge effects[1]. The simplest way to take advantage of them is that of Davis and Peebles (1983)

$$\xi(\vec{r}) = \frac{\mathrm{dd}(\vec{r})}{\mathrm{dr}(\vec{r})} - 1\,. \tag{4.13}$$

A smaller variance in the 2PCF estimate can be achieved with the following estimator, proposed by Hamilton (1993)

$$\xi(\vec{r}) = \frac{\mathrm{dd}(\vec{r})\,\mathrm{rr}(\vec{r})}{\mathrm{dr}(\vec{r})^2} - 1. \tag{4.14}$$

[1]Points near the edges of the survey region have fever pairs than those further away from the edges. DR counts take this into account by measuring the number of random points around an average galaxy or cluster instead of an average location, as is done by RR counts.

However, both of these two estimators are biased, meaning that their expectation value is not the true 2PCF. This problem is fixed by the currently most widely used estimator, proposed by Landy and Szalay (1993)

$$\xi(r) = \frac{\mathrm{dd}(\vec{r}) - 2\mathrm{dr}(\vec{r}) + \mathrm{rr}(r)}{\mathrm{rr}(\vec{r})}\,. \tag{4.15}$$

I will call this the LS (Landy-Szalay) estimator. At the limit of infinitely large random catalog it is unbiased and also provides the minimum variance when $\xi \ll 1$. In this limit the estimator by Hamilton has approximately same variance as the LS estimator but is usually not favored due to the bias[2]. In addition, when comparing a a number of 2PCF estimators using simulations of galaxy clusters, Kerscher et al. (2000) show that the LS estimator is significantly less sensitive to the size of the random catalog than the Hamilton estimator and stands out as the recommended estimator for practical applications.

The estimate given by Eq. (4.15) will be more accurate the more pairs of each type we have. Also the computational cost (ie. how long will it take to compute the estimate with given number of CPU cores) of the LS estimator is determined by how many pairs need to be counted. To get a rough estimate for the computational cost consider the following. The total number of pairs within a catalog is $\frac{1}{2}N_i(N_i-1) \approx \frac{1}{2}N_i^2$ (or N_iN_j between two catalogs) but one usually is interested in the 2PCF up to some scale $r_{\max}$ that is smaller than the largest separations within the survey. Thus, we only need to count a fraction f of all the pairs. Often we are interested in a range of scales for which $|\xi(r)| \ll 1$ ie. the clustering is weak and the fraction of data pairs within $r < r_{\max}$ does not differ significantly from a random distribution. Thus, we have the same f for all the pair counts and the computational cost is proportional to $\frac{1}{2}f(N_\mathrm{D}^2+2N_\mathrm{R}N_\mathrm{D}+N_\mathrm{R}^2)$. As mentioned, one typically uses $N_\mathrm{R} \gg N_\mathrm{D}$ to minimize the error in the Monte Carlo integration. We parameterize the number of random points by $M \coloneqq N_\mathrm{R}/N_\mathrm{D}$. This way the computational cost is $\frac{1}{2}fN_\mathrm{D}^2(1+2M+M^2)$. Note that this estimate does not take into account the various code overheads, such as reading of the data and random catalogs. It is valid when the pair counting dominates the computation time.

The baseline value for the Euclid survey, for example, is $M = 50$, so that ~ 25 times more time is spent on counting the random-random than the data-random pairs. On the other hand, the variance of the pair counts, that make up the variance of the 2PCF estimate, is roughly inversely proportional to the number of pairs in each bin (variance of a Poisson process). The smaller number of DR pairs contribute more to the uncertainty of the 2PCF estimate but the computational cost is dominated by the larger number of RR pairs. This suggests that it might be possible to achieve significant savings in computational time with insignificant loss of accuracy by excluding some of the RR pairs. In Paper I we study in detail how the bias and covariance of the LS 2PCF estimate depend on the properties of the random catalog and show both analytically and numerically that this is indeed the case.

The method we present was used independently in 2PCF work before Paper I (for example Zehavi et al., 2011) and it was already hinted at by the original Landy and Szalay paper. However, it had not been studied in detail in the literature. The idea is that one produces, instead of one large random catalog corresponding to $M \gg 1$, a set of M_S (we call this the split factor) catalogs with $(M/M_\mathrm{S})N_\mathrm{D}$ points. This can be thought of as splitting the large random catalog into M_S sub-catalogs, hence the name. Each small random catalog should be statistically equivalent to the full random catalog. In principle the split factor could be any number $1 \le M_\mathrm{S} \le M$ but we show that $M_\mathrm{S} = M$ is the ideal value for achieving the best accuracy for a fixed computational cost so I will take it to be fixed to this value for the rest of the thesis. So, in the optimal split method the random catalog is split into M sub-catalogs of size N_D. The DR and RR pairs are then counted within each separate random catalog but not between and hence

$$\mathrm{DR} = \sum_{i=1}^{M} \mathrm{DR}_i\,, \quad \mathrm{RR} = \sum_{i=1}^{M} \mathrm{RR}_i\,.$$

Here DR_i and RR_i are the pairs between the data catalog and the ith random sub-catalog and within the ith random sub-catalog, respectively. The DR counts are equal to those of the LS

[2]In the limit of $\xi \to 0$ the bias is proportional to $1/N_\mathrm{D}$ so that it is more important for smaller galaxy/cluster catalogs.

estimator but the RR counts are significantly lower. The number of RR pairs to count in the LS estimator is proportional to $M^2 N_\mathrm{D}^2$ but in the split method $M N_\mathrm{D}^2$, since we have M random catalogs of size N_D, so that M times fewer random pairs need to be counted.

All the split pair counts can be expressed in terms of M and N_D. The number of DD pairs is still $N_\mathrm{D}(N_\mathrm{D}-1)/2$. The number of DR pairs is $M N_\mathrm{D}^2$ and the number RR pairs is $M N_\mathrm{D}(N_\mathrm{D}-1)$. We can redefine

$$\mathrm{dr}(\vec{r}) \coloneqq \frac{1}{M}\sum_i \frac{\mathrm{DR}_i(\vec{r})}{N_\mathrm{D}^2} = \frac{1}{M}\sum_i \mathrm{dr}_i(\vec{r}) \tag{4.16}$$

$$\mathrm{rr}(\vec{r}) \coloneqq \frac{1}{M}\sum_i \frac{\mathrm{RR}_i(\vec{r})}{N_\mathrm{D}(N_\mathrm{D}-1)} = \frac{1}{M}\sum_i \mathrm{rr}_i(\vec{r}) \tag{4.17}$$

and in terms of these redefined pair counts the split LS estimator can be written as

$$\xi(\vec{r}) = \frac{\mathrm{dd}(\vec{r}) - 2M^{-1}\sum_i \mathrm{dr}_i(\vec{r})}{M^{-1}\sum_i \mathrm{rr}_i(\vec{r})} + 1\,. \tag{4.18}$$

In the paper we show using simulated mock catalogs that for $M = 50$ the use of the split method reduces the computation time of a 2PCF estimate by a factor of more than ten without noticeable increase in the estimator bias or variance. The advantages of the split method are summarized in more detail by Fig. 4.2. This figure shows the mean variance of 2PCF estimates over scales of 80–120h^{-1}Mpc[3] as a function of the computational cost of the corresponding estimate. The figure includes comparison to a method dubbed "dilution", which was proposed as an alternative to the split method. In the dilution method a single subset of the full random catalog is used when counting the RR pairs. "dilution = 0.5", for example, simply means using half of the random points when counting the RR pairs. As with the split method, the subset should be statistically equivalent to the full random (ie. it only differs by the point density). In the case of the DR counts the full random catalog is used. The other comparison case is simply using the standard LS estimator with a smaller random catalogs. As shown by the figure, and Paper I in more detail, the split method provides the smallest variance at a given computation time and allows a significant saving in computational cost with a negligible increase in variance.

4.3 2PCF covariance and the linear construct method

To be able to make conclusions based on estimated 2PCFs it is necessary to assess their accuracy. A typical goal for a galaxy survey is to determine cosmological parameters based on comparing measured and predicted 2PCF, or alternatively the corresponding power spectra. To to be able to estimate confidence limits for the parameter estimates we need a likelihood function for the parameters. Usually this is done by developing a model for the 2PCF as a function of the cosmological parameters and then computing a prediction at different points in the parameter space. The parameter space scanning is often implement using Markov Chain Monte Carlo (MCMC) sampling methods, a concrete cosmological example being Lewis and Bridle (2002). In this framework the parameter likelihood is determined by the difference of the measured and predicted 2PCF values at each point in the parameter space and the probability distribution the 2PCF measurement realizations are thought to be drawn from.

In galaxy clustering analyses this distribution is commonly approximated by simple a Gaussian function (Cole et al., 2005; Okumura et al., 2008, for example). This is mathematically convenient, and can be motivated by, for example, the central limit theorem (see eg. Chapter 16 of Peacock, 1999). However, Schneider and Hartlap (2009) show analytically by using the non-negativity of the matter power spectrum that this can only be the case approximately. Schneider and Hartlap (2009), Keitel and Schneider (2011), and Wilking and Schneider (2013) have studied ways to transform correlation functions in order to obtain more Gaussian-distributed quantities but the results are not applicable to real data. Regarding power spectrum estimates, Wang et al. (2019) have studied the non-Gaussianity of their probability distributions and devised a transformation

[3]These scales were selected to include the important BAO scale.

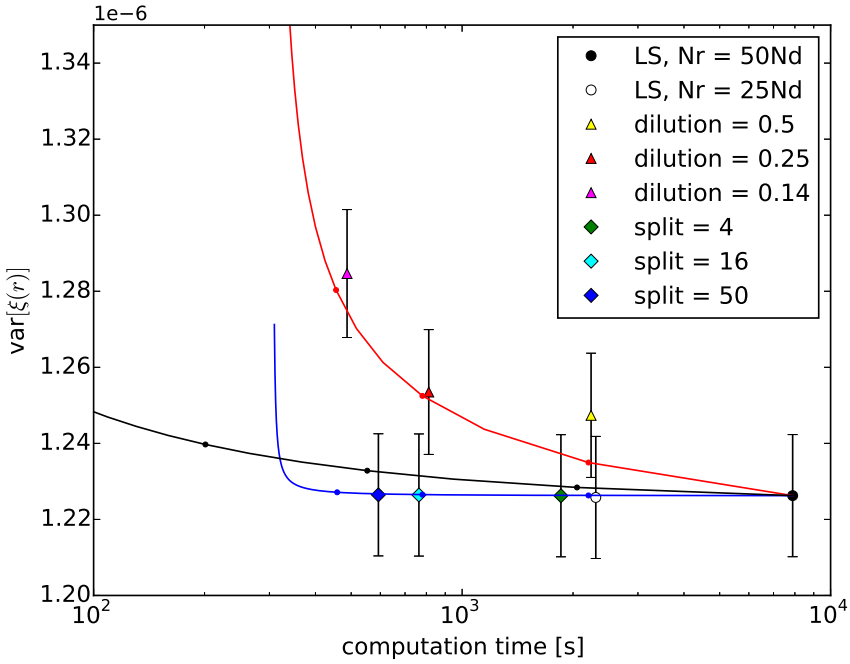

Figure 4.2: The measured variance (mean variance over the range $r = 80$–$120h^{-1}$Mpc) as function of computational cost (mean computation time) for the different 2PCF estimators (markers with error bars). The solid lines (blue for the split method, red for dilution, and black for standard LS with $N_\mathrm{R} < 50N_\mathrm{D}$) are our theoretical predictions for the increase in variance and computation time ratio when compared to the standard LS, $N_\mathrm{R} = 50N_\mathrm{D}$ case, and the small dots on the curves correspond to the measured cases (except for LS they are, from right to left, $N_\mathrm{R} = 25N_\mathrm{D}$, $12.5N_\mathrm{D}$, and $(50/7)N_\mathrm{D}$; only the first of which was measured). Figure from Paper I.

scheme to "Gaussianize" measured spectra. Nevertheless, in practice Gaussianity might well be sufficient approximation for certain 2PCF estimates, Gaussianized or not. The distribution of the estimates in a specific setup can be studied with simulated mock catalogs. Indeed, in the case of the the simulations we use to validate the methods in Paper III, for example, the Gaussian approximation seems to be sufficient for describing the distribution of the measured 2PCF values.

The logarithm of the Gaussian likelihood is given by

$$\ln \mathcal{L} = -(\xi - \xi_{\text{theory}})^{\text{T}} C^{-1} (\xi - \xi_{\text{theory}}) + \text{constant}\,, \tag{4.19}$$

where the vector ξ contains the estimated 2PCF values, the vector ξ_{theory} contains the theoretically predicted values for the 2PCF (and is a function of the cosmological parameters) and the matrix C^{-1} is the inverse of the covariance matrix of the estimate ξ. The covariance matrix C fully specifies the statistical properties of a Gaussian distributed random vector. Even in the case non-Gaussian distributions the covariance matrix servers a characterization of uncertainty. Its diagonal elements correspond to the variance of the 2PCF estimate at each data point (separation bin, multipole etc.) and the off-diagonal elements quantify how the errors at different points are correlated. The covariance matrix is defined as

$$C(\vec{r}_1, \vec{r}_2) \coloneqq \Big\langle \big[\xi(\vec{r}_1) - \big\langle \xi(\vec{r}_1) \big\rangle\big] \big[\xi(\vec{r}_2) - \big\langle \xi(\vec{r}_2) \big\rangle\big] \Big\rangle\,. \tag{4.20}$$

Here $\vec{r_1}$ and $\vec{r_2}$ denote arbitrary separation bins. We thus need an estimate for this covariance matrix of a 2PCF estimate. The simplest way to obtain such an estimate is to compute the sample covariance, where one simply replaces the ensemble averages with means over N 2PCF realizations drawn from the same distribution as the measured 2PCF

$$\hat{C}(\vec{r}_1, \vec{r}_2) \coloneqq \frac{1}{N-1} \sum_i \big(\xi_i(\vec{r_1}) - \overline{\xi}(\vec{r}_1)\big) \big(\xi_i(\vec{r}_2) - \overline{\xi}(\vec{r}_2)\big)\,, \tag{4.21}$$

where

$$\overline{\xi}(\vec{r}) \coloneqq \frac{1}{N} \sum_i \xi_i(\vec{r}) \tag{4.22}$$

is the sample mean. The denominator in (4.21) is $N-1$ instead of N to yield an unbiased estimate. This is basically because the sample mean is estimated from the same set of 2PCF as the covariance and thus removes one degree of freedom in the averaging process.

The drawback of the sample covariance is that one needs to compute the 2PCF from a large number of mock object catalogs. The variance of the covariance estimate scales roughly as $1/N$ so to get the error of the estimate down to a 10% level accuracy one needs of the order of 100 independent catalog realizations and for a 1% level 10 000 realizations. These mock catalogs should be statistically as similar to the actual galaxy or cluster sample as possible. How to efficiently produce such a set of simulations is a research field of its own and beyond the scope of this thesis. In Paper I we use a set of N-body simulations presented in Grieb et al. (2016) and Lippich et al. (2019) called Minerva. In Paper II we use a set of dark matter halo simulations produced with a more approximate algorithm called PINOCCHIO[4] (PINpointing Orbit Crossing Collapsed HIerarchical Objects) that emulates N-body simulations, see Monaco et al. (2002) and Munari et al. (2017).

Fig. 4.3 shows an example of 2PCF multipoles measured from a set of 5 000 PINOCCHIO mocks. The three 2PCF multipoles $\ell = 0, 2, 4$ are binned in 100 r-bins over scales of 0–200h^{-1}Mpc. The plot shows the mean over the 5 000 realizations and a single realization with the 1σ error envelope (which corresponds to the diagonal of the covariance matrix). The BAO feature at $\sim 100h^{-1}$Mpc can be seen clearly in the monopole. Fig. 4.4 shows an example of a covariance matrix computed from the same set of 2PCF realizations. The covariance matrix elements are plotted as normalized correlation coefficients

$$\rho_{ij} \coloneqq \frac{\hat{C}(r_i, r_j)}{\sqrt{\hat{C}(r_i, r_i)\hat{C}(r_j, r_j)}}\,, \tag{4.23}$$

so that $\rho_{ij} = 1$ for $i = j$ and $-1 \leq \rho_{ij} \leq 1$ for $i \neq j$. The different blocks represent different multipoles and their cross-correlations. The plot shows clear correlations between errors in neighboring r-bins and also between different multipoles.

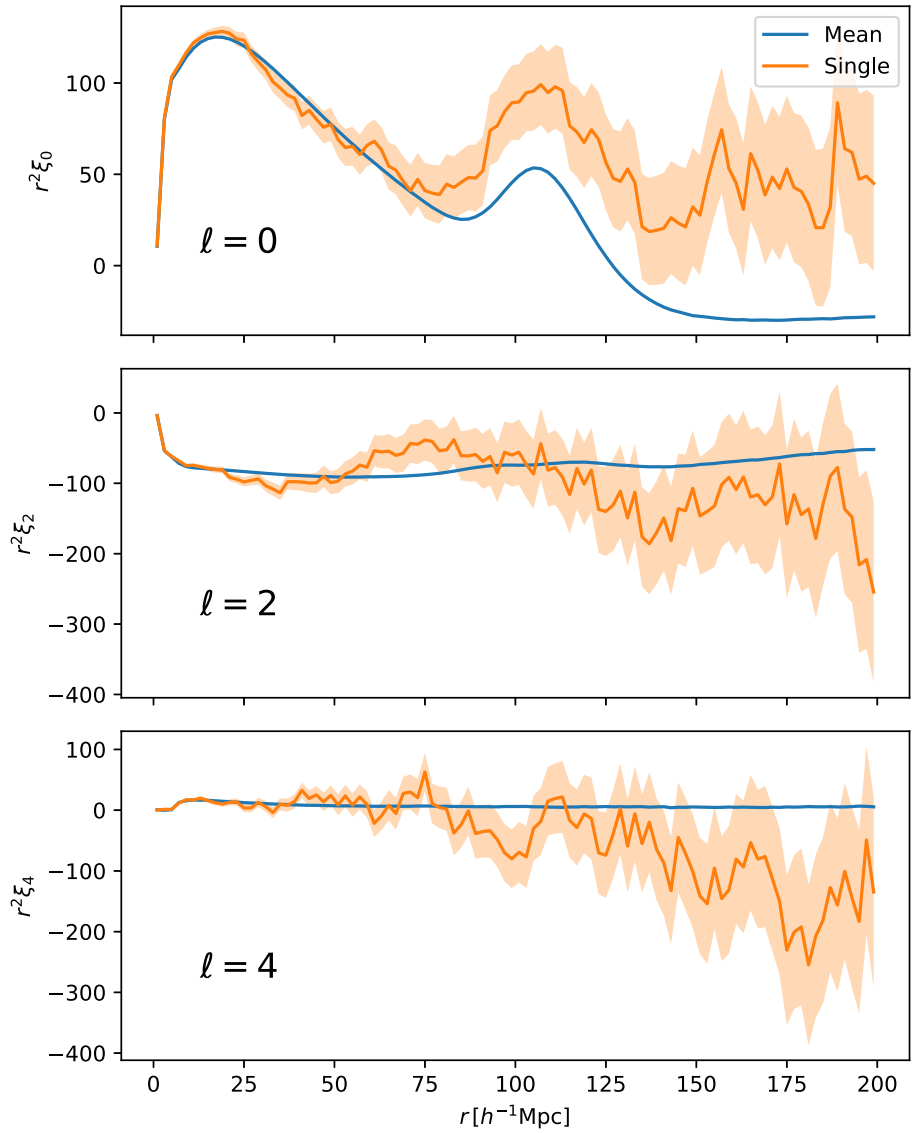

Figure 4.3: Correlation function multipoles measured from 5 000 PINOCCHIO mock catalogs. From top to bottom the panels correspond to monopole, quadrupole and hexadecapole. The blue line is the mean of the sample and the orange line a single realization. The shaded area around the single realization curve is the $1\,\sigma$ error envelope, computed as the standard deviation of the available realizations. Figure from Paper II.

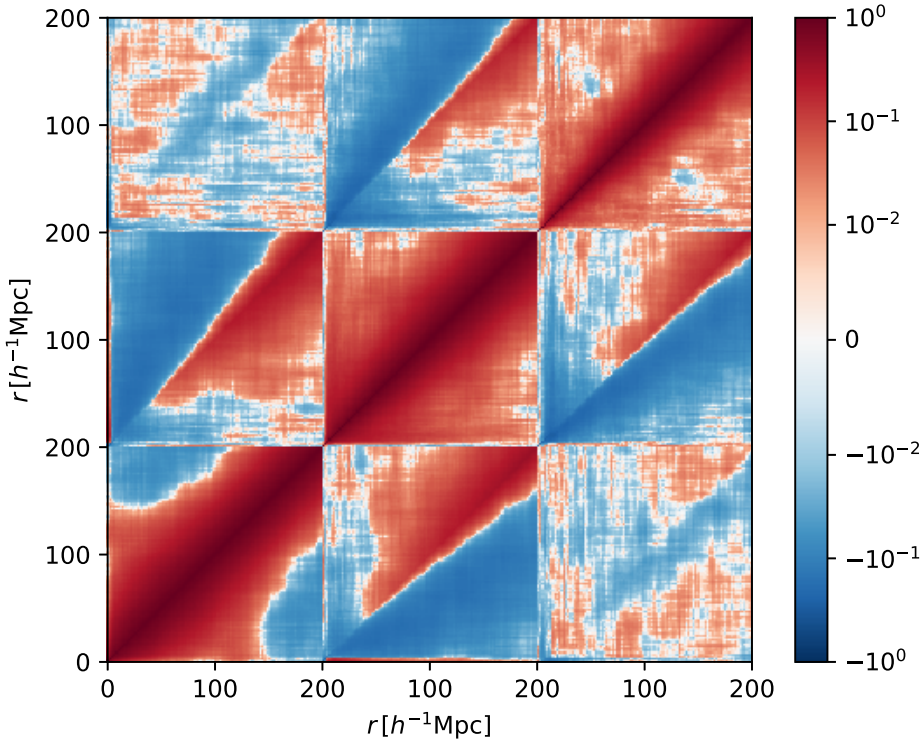

Figure 4.4: Covariance matrix of the correlation function multipoles, measured from the PINOCCHIO mocks. The blocks from left to right and from the bottom to the top row correspond to monopole, quadrupole and hexadecapole respectively. The elements are normalized by the diagonal. Figure adjusted from Paper II.

In Paper II we develop a faster method for computing the covariance of a 2PCF estimate obtained via the split method. To calculate a prediction for the covariance one needs to inspect the fluctuations of the 2PCF estimate around its expected value. This is done by defining the relative fluctuations of different pair counts as

$$\alpha(\vec{r}) \coloneqq \frac{\mathrm{dd}(\vec{r})}{\langle \mathrm{dd}(\vec{r})\rangle} - 1\,, \tag{4.24}$$

$$\beta_i(\vec{r}) \coloneqq \frac{\mathrm{dr}_i(\vec{r})}{\langle \mathrm{dr}_i(\vec{r})\rangle} - 1\,, \tag{4.25}$$

$$\gamma_i(\vec{r}) \coloneqq \frac{\mathrm{rr}_i(\vec{r})}{\langle \mathrm{rr}_i(\vec{r})\rangle} - 1\,. \tag{4.26}$$

In terms of these pair count fluctuations Eq. (4.18) becomes

$$\xi(\vec{r}) = \frac{\langle \mathrm{dd}(\vec{r})\rangle[1+\alpha(\vec{r})]}{\langle \mathrm{rr}(\vec{r})\rangle[1+M^{-1}\sum_i \gamma_i(\vec{r})]} - 2\frac{\langle \mathrm{dr}(\vec{r})\rangle[1+M^{-1}\sum_i \beta_i(\vec{r})]}{\langle \mathrm{rr}(\vec{r})\rangle[1+M^{-1}\sum_i \gamma_i(\vec{r})]} + 1 \tag{4.27}$$

This can be then inserted into the definition of the covariance matrix (4.20). After a rather tedious calculation one obtains expression for the covariance in terms of pair count covariances ($\langle\alpha(\vec{r}_1)\alpha(\vec{r}_2)\rangle$, $\langle\alpha(\vec{r}_1)\beta(\vec{r}_2)\rangle$ etc.). The main point is, however, that the expression is of the form

$$C(\vec{r}_1,\vec{r}_2) = A(\vec{r}_1,\vec{r}_2) + M^{-1}B(\vec{r}_1,\vec{r}_2)\,, \tag{4.28}$$

ie. it has a term that depends on M^{-1} and a term that is independent of the size of the full random catalog. Thus, if one computes the covariance using two sets of small random catalogs (using $M = 1$ and $M = 2$) it is possible to solve for matrices A and B and construct the covariance for an arbitrary M (usually significantly larger than 2). We call this the linear construct (LC) method. An additional saving can be achieved by computing two sets of 2PCF estimates, both with independent random catalogs of size $M = 1$, and constructing the $M = 2$ case by co-adding the DR and RR pairs of the two individual sets. Note that the LC covariance assumes nothing about the distribution of the 2PCF estimates (Gaussianity, for example) but is based purely on the definitions of the covariance and the split LS estimator.

As discussed, the computational cost of the LS estimator is roughly proportional to $N_\mathrm{D}^2(1 + 2M + M^2)$ and in the case of split LS estimator the M^2 term is replaced by M. The LC covariance estimate corresponds to case of split LS with $M = 2$. If we take as a reference the case of split LS with $M = 50$ the speedup is a factor of $(1 + 2 \cdot 50 + 50)/(1 + 2 \cdot 2 + 2) \approx 21.6$. In the paper we show that despite the fact that the LC covariance has the same expectation value as the sample covariance, it has somewhat larger variance[5]. We derive expressions for the covariances of both the sample (for split LS estimator) and LC covariances with which one can estimate the error of the covariance estimate based on the covariance itself. In our simulations it turned out that the extra scatter in the LC covariance means that one needs 1.2–1.8 times more mock catalogs to reach the same accuracy than with the sample covariance. This property is summarized by Fig. 4.5. The plot shows how the sample covariance and the LC covariance converge towards a reference covariance matrix as we increase the number of mock catalogs used to estimate them. The reference covariance was computed from 5 000 independent mock catalogs. The left panel of the plot shows that for a given number of realizations (ie. mock catalogs) the sample covariance is closer to the reference covariance. The right panel, however, shows that for a fixed computational cost the LC covariance is significantly closer to the reference covariance.

As noted, the LC method requires more 2PCF realizations to reach the same precision as the sample covariance. In addition, the handling of large number of small catalogs makes the code overheads more important compared to larger catalogs, so that the computation time is not as strongly dominated by the pair counting. Due to these effects we observe in practise a speedup factor of $\sim$ 5–10, in contrast to the theoretical prediction of $\sim$ 20. This is still a significant improvement and can be further increased by code optimization. The preceding estimate is still,

[4] `https://adlibitum.oats.inaf.it/monaco/pinocchio.html`

[5] Elements of sample-based covariance matrices are themselves random variables and their values are thus subject to some variance.

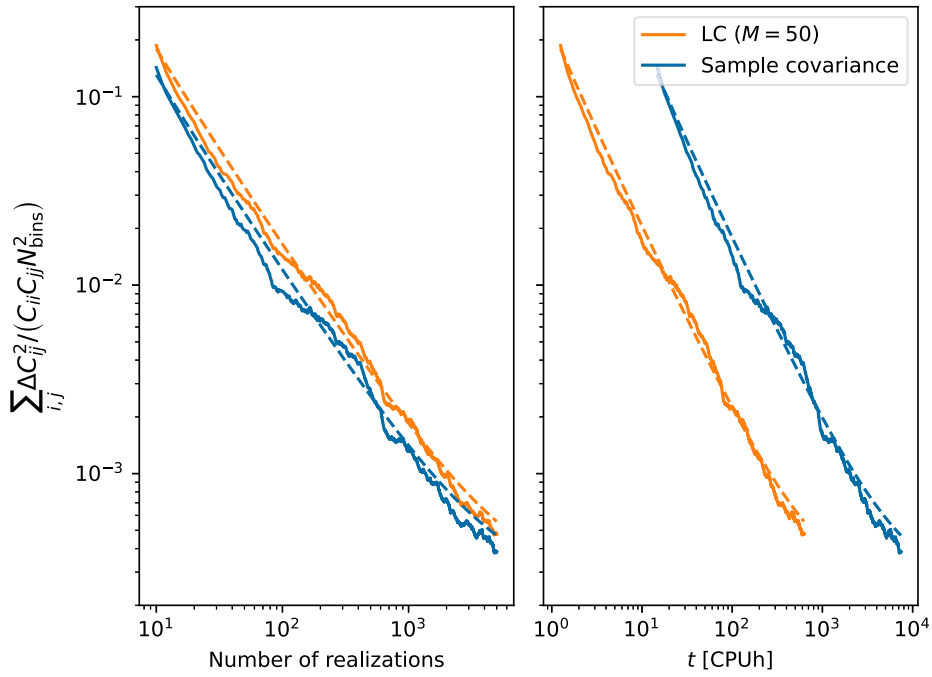

Figure 4.5: Convergence of the covariance of the correlation function multipoles, with respect to the number of realizations (left) and CPU time (right). The y-axis is the mean squared difference between the elements of the LC (orange) or sample (blue) covariance and a reference covariance matrix, normalized by the diagonal elements of the reference matrix to give different scales equal weight. The dashed lines show our theoretical predictions. Figure from Paper II.

in a sense, optimistic. It does not account for the cost of producing the larger number of mock catalogs. This cost strongly depends on the requirements of a particular 2PCF measurement, such as desired mass resolution, simulation volume, whether an N-body simulation or baryonic effects are needed and so on. In the case of the PINOCCHIO mocks that we use to validate the LC method the cost of a single mock catalog is approximately 6 CPUh. The data of Fig. 4.5 shows that in our test roughly 1.4 times more realizations were needed for the LC covariance to match the precision of the sample covariance. If the desired precision for the sample covariance is achieved with 5 000 mocks, LC requires 7 000 mocks, which costs approximately 12 000 extra CPUh. This is actually more than the cost of the 5 000 2PCF realizations needed for the sample covariance. However, a set of simulated mock catalogs is generally useful for various kinds of analyses (even across different surveys), whereas a covariance matrix estimate can only be used along a specific 2PCF measurement. Also, to mimic a narrow redshift bin we used only one fifth of the volume of each PINOCCHIO mock. Should we have used the whole volume the cost of the covariance matrix estimate would have risen significantly, whereas the cost of the mocks would have stayed the same. All in all, the costs of the simulated catalogs and the covariance matrix estimate are not directly comparable but the cost of the simulations should still be kept in mind when judging the performance benefits.

4.4 An example of a 2PCF analysis

In Paper III we analyze an X-ray selected cluster sample called CODEX (COnstrain Dark Energy with X-ray, Finoguenov et al., 2020). The CODEX catalog is compiled by combining data from

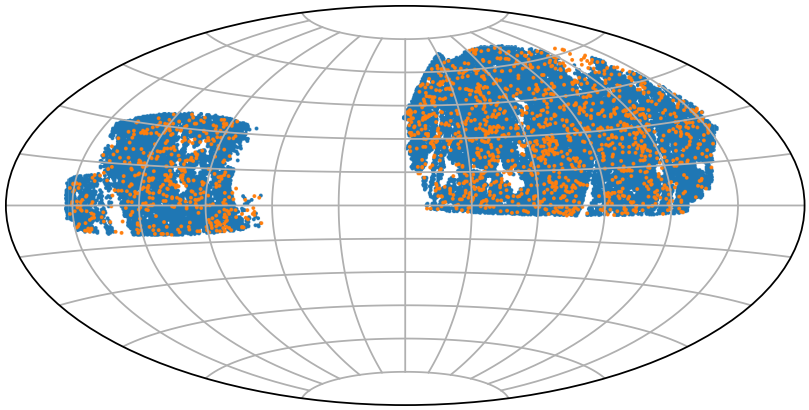

Figure 4.6: Sky footprint of the CODEX catalog in the equatorial coordinate system. Orange points are the clusters used in our analysis and blue points are the corresponding random points. Figure from Paper III.

the ROSAT all-sky X-ray survey and the SDSS galaxy survey on the northern sky. In the ROSAT survey galaxy clusters are identified as compact sources of high X-ray luminosity (caused by the hot intracluster plasma) and in SDSS as concentrations of optically observed galaxies. The catalog has 10 382 clusters in the redshift range $0.05 < z < 0.68$ with sky coverage of roughly 10 000 square degrees. The sky footprint of the catalog is shown in Fig. 4.6. To guarantee the purity of our sample we select a subset of the CODEX catalog based on the cluster redshifts and richnesses[6]. We limit ourselves in the redshift range $0.1 < z < 0.5$. Below these redshifts the cluster finding algorithm (the red-sequence matched-filter Probabilistic Percolation, redMaPPer, Rykoff et al., 2014) starts to suffer from projection effects and above them the richness estimates based on the SDSS imaging become uncertain. We employ a fixed lower limit of 25 and a redshift-dependent limit of $22(z/0.15)^{0.8}$ for cluster richness. Clusters with lower richnesses are prone to on one hand being missed by the cluster finder algorithm and on the other hand being false identifications. For further details, see Finoguenov et al. (2020) and Klein et al. (2019). After these cuts we are left with 1892 clusters for our clustering analysis.

For our analysis the most important cluster properties included in the CODEX catalog are X-ray luminosities and richnesses. Both of these are known to correlate with the total mass of the cluster (including the dark matter halo), see Capasso et al. (2020) and Kiiveri et al. (2021) for the scaling relation used in our analysis. With these proxies we can estimate the mass of each cluster. The goal of the paper is to study how, based on these masses, predictions for the clustering bias agree with the actually measured clustering signal. We use the projected 2PCF as our clustering statistics. Examples of our measurements are shown in Fig. 4.7. Plotted here is a set of projected 2PCF we measured from the CODEX sample within different redshift ranges. These are compared to predictions from linear dark matter power spectra scaled by the square of the bias factor predicted from the cluster masses. We predict the bias factor for each cluster based on its estimated mass using Eq. (3.37) at the redshift of the cluster. The bias of the whole cluster population is computed as a weighted mean

$$\bar{b} = \sum_i b(M_i, z_i) g_i(z_i) \,. \tag{4.29}$$

Here the sum runs over all the clusters in the sample and the factor $g(z_i) \coloneqq D(z_i)/D(0)$ scales each individual bias factor to be relative to the present day matter density instead of the density at the redshift of the cluster. We compute the prediction for the projected 2PCF of the dark matter

[6]Richness of a galaxy cluster is the sum of its member candidates weighted by the probability for each of them to be member of the cluster.

distribution by first Fourier transforming a prediction for the linear perturbation theory power spectrum and then using Eq. (4.6) to predict the projected 2PCF from the underlying isotropic 2PCF. As we can see, the shape and the amplitude of the measured correlation functions are mostly compatible with the theoretical predictions within the estimated error bars.

As mentioned in Sec 3.3, the agreement between the measured and predicted clustering bias can be used as a cosmological test. As a simple example we study the likelihood of parameters $(\Omega_{\rm m}, \sigma_8)$ in the light of our 2PCF measurements. $\Omega_{\rm m}$ is again the present day matter density parameter and σ_8 is a parameter used, in large scale structure surveys in particular, to parameterize the amplitude of the linear power spectrum. It is defined in terms of $\sigma(R)$ as

$$\sigma_8 \coloneqq \sigma(8h^{-1}{\rm Mpc})\,, \tag{4.30}$$

where the scale $8h^{-1}$Mpc roughly corresponds to the spatial scale of galaxy clusters. All the other parameters were kept fixed. This is of course an oversimplification but not a terrible one since the cluster mass function (and consequently their clustering bias) has the strongest dependence on the two parameters varied through the ΛCDM growth function and the amplitude of the density fluctuations at the cluster scales.

We assume the Gaussian likelihood of Eq. (4.19). In our case the vector ξ contains the measured values of the projected 2PCF. The theoretical prediction is $\xi_{\rm theory} = b^2 w_p^{\rm theory}$, where b is the mass-based bias prediction and $w_p^{\rm theory}$ is the projected correlation function of the linear perturbation theory from Eq. (4.6). Both of the factors depend on the cosmological parameters. In principle also the measured 2PCF depends on the cosmology since the measured angles and redshifts need to be converted into comoving distances. The changes introduced by varying the cosmology are negligible compared to the statistical uncertainty of the 2PCF estimate but for completeness we model the effect following Marulli et al. (2012). Distances perpendicular and parallel to line-of-sight, r_p and r_π respectively, are related in two different cosmologies, labeled by 1, 2, by

$$r_{p,1} = \frac{d_{{\rm A},1}(z)}{d_{{\rm A},2}(z)} r_{p,2}\,, \quad r_{\pi,1} = \frac{H_2(z)}{H_1(z)} r_{\pi,2}\,, \tag{4.31}$$

where $d_{\rm A}(z)$ is the angular diameter distance and $H(z)$ the Hubble parameter at redshift z. So, to compare theoretical predictions at varying cosmologies with the measurement at the fiducial cosmology at scales (r_p, π) we evaluate them at scales $[d_{\rm A}(z)/d_{{\rm A},f}(z)]r_{\rm p}$ and $[H(z)_f/H(z)]r_\pi$, where f refers to fiducial cosmology and we take z to be the mean redshift of the sample.

The constraints we obtain for $(\Omega_{\rm m}, \sigma_8)$ from the clustering analysis are summarized in Fig. 4.8. Constraints on σ_8 in particular are rather loose but compatible with values obtained from other measurements (see below). The clustering-based constraints can be further combined with those obtained by comparing observed and predicted mass functions. In the case of X-ray selected clusters one usually uses the X-ray luminosity function (XLF), which is a proxy for the mass function. In Fig. 4.8 we show constraints from the XLF measured from the CODEX clusters (from Finoguenov et al., 2020) and from combining the XLF and 2PCF measurements into a joint likelihood. The individual likelihoods are most degenerate along different directions so the combination has significantly more constraining power. The joint contours are, however, somewhat optimistic since they are just the product of the constituent likelihoods, that are in reality correlated to some extent.

The marginalized parameter constraints we obtain from the combined likelihood are $\Omega_{\rm m} = 0.27^{+0.01}_{-0.02}$ and $\sigma_8 = 0.79^{+0.02}_{-0.02}$. These can be compared with, for example, results from the CMB measurements of the Planck and WMAP satellites. These are $\Omega_{\rm m} = 0.3147 \pm 0.0074$, $\sigma_8 = 0.8101 \pm 0.0061$ (Planck Collaboration, 2020b) and $\Omega_{\rm m} = 0.279 \pm 0.025$, $\sigma_8 = 0.821 \pm 0.023$ (Hinshaw et al., 2013), respectively. There is a slight tension between the value of $\Omega_{\rm m}$ inferred from our cluster measurements with the one obtained from the Planck measurements but as it is shown in Paper III this can be explained by systematic uncertainties in our analysis.

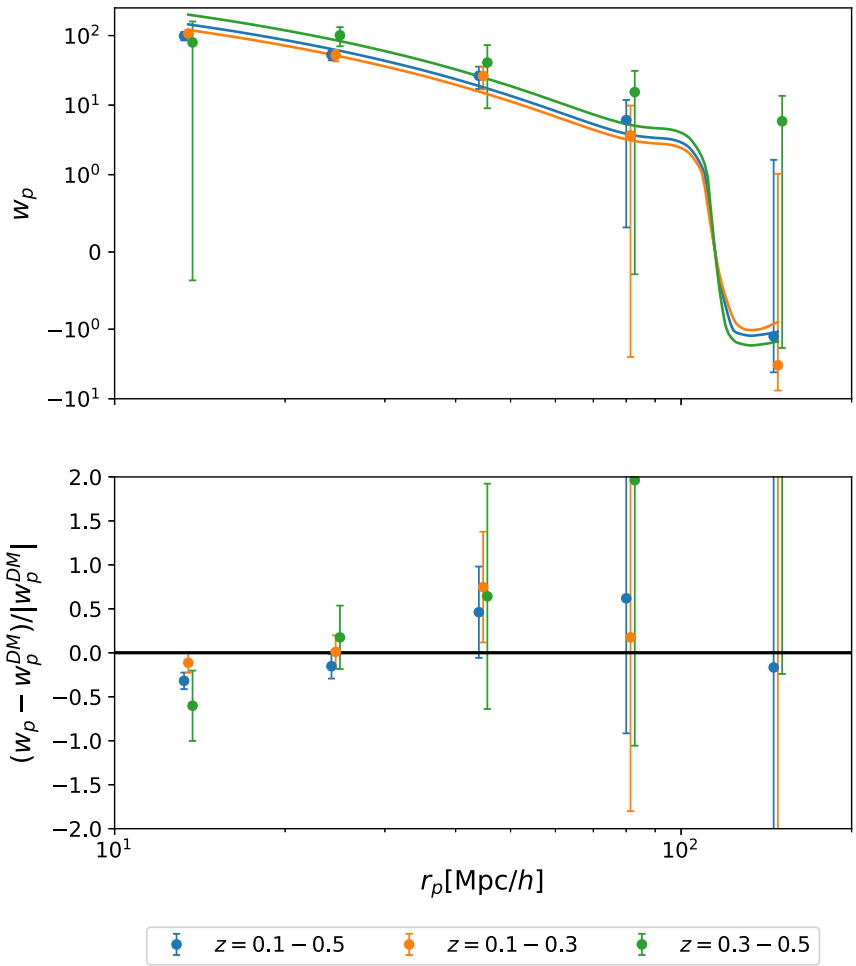

Figure 4.7: Top panel: projected 2PCF in three redshift ranges $0.1 < z < 0.5$, $0.1 < z < 0.3$ and $0.3 < z < 0.5$. The data points are 2PCF estimate from the CODEX clusters. The solid curves of the corresponding colors are predicted dark matter 2PCF scaled by the mass-based bias estimate computed within each cluster sub-sample. Bottom panel: relative difference between the 2PCF estimate and the corresponding prediction. The data points have been shifted horizontally for clarity. Figure from Paper III.

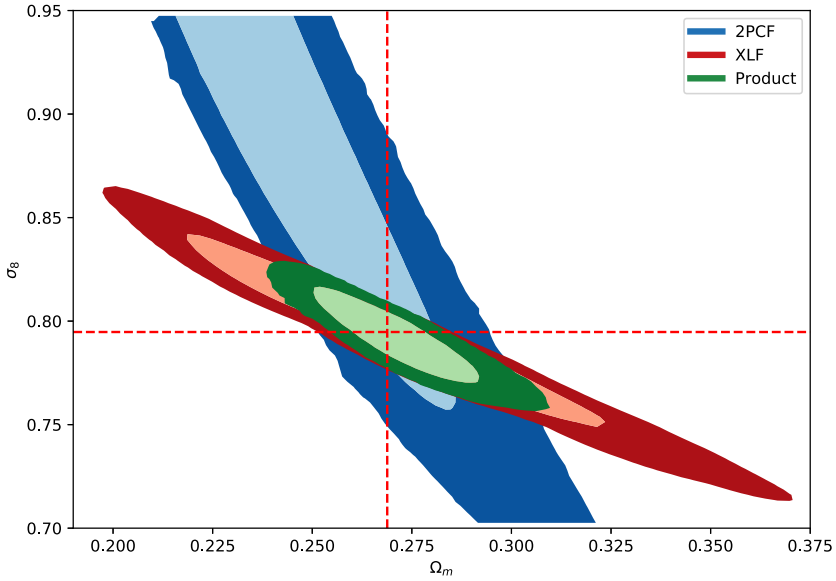

Figure 4.8: Likelihood contours of ($\Omega_{\rm m}$, σ_8) from the cluster two-point correlation function (blue), the X-ray luminosity function (red) and the joint distribution (green). The light and dark contours are the 68 % and 95 % confidence regions, respectfully, and the red dashed lines show the best fit values for the joint likelihood. Figure from Paper III.

Chapter 5

Conclusions

Large galaxy surveys, such as the forthcoming Euclid survey, are and will be an important cosmological probe. Combination of theoretical developments, astronomical observations and large numerical simulations has resulted in a well-developed framework for understanding galaxy clustering and allows it to be exploited efficiently for cosmological purposes. The spatial distribution of galaxies and their clusters contains a wealth of information about the initial conditions and evolution history of the Universe. This information is encoded in statistical quantities, such as power spectra and two-point correlation functions of the large scale galaxy distribution. This thesis, including the accompanying papers, focused on some aspects of the two-point correlation function analysis and its connection to the theory of cosmology.

Two-point correlation function estimators rely on counting pairs of objects. This operation scales as the square of the number of objects. Thus, the computational cost of estimating the two-point correlation functions grows rapidly as the next generation galaxy surveys produce larger and larger galaxy catalogs. The situation becomes even more challenging when one considers the covariance matrix of the two-point correlation function estimate. Currently, the most robust way to estimate the covariance is to measure the sample covariance from a large set of (possibly of the order of 10 000) mock galaxy catalogs. This analysis step is going to take up a significant portion of, for example, Euclids computational budget. Doing it as efficiently as possible is thus far from irrelevant and could save resources for other computationally intensive tasks.

Historically, the two-point correlation function estimators were designed to optimize the bias and variance of the estimate but not necessarily its computational cost. We aimed at filling in this gap by studying systematically how the size of the random catalog affects the variance of a two-point correlation estimate computed using the so-called Landy-Szalay estimator. We show that in a setup resembling the Euclid spectroscopic survey the computational cost of a single two-point correlation function estimate can be reduced by a factor of more than ten without significant loss of accuracy. This is achieved by omitting part of the random-random pairs in a way that preserves the desirable properties of the Landy-Szalay estimator. We call the new method the split method.

We further proceed to study the effect of the size of the random catalog on the covariance matrix of the two-point correlation function estimate produced by the split Landy-Szalay estimator. It turns out that the mock two-point correlation function estimates used to measure the sample covariance can be estimated with a smaller number of random points than the actual two-point correlation function estimate without biasing the covariance matrix estimate. This can be done by combining two sets of mock correlation function estimates using random catalogs of different sizes. Since the new estimate is a linear combination of two covariances we dub the method the Linear Construct. We show both theoretically and using simulations that for a given number of mock catalogs the Linear Construct covariance has somewhat larger variance than the sample covariance. To reach the same accuracy with the Linear Construct method 1.2–1.8 times more mock catalogs are needed. The computational saving per mock catalog is however so large that in a simulated setup mimicking a redshift bin within a galaxy survey the overall cost is reduced by a factor of ~ 5–10. The computational saving of combining the split Landy-Szalay estimator and the Linear Construct covariance is thus of the order of 100 and could be decisive on whether computing the covariance matrix is feasible or not.

The computational efficiency within an analysis pipeline dealing with large amounts of data is a necessity but it is not the point of cosmological surveys. The goal is to test theoretical models and constrain their parameters. This is what the two-point correlation functions and their covariance matrices will be ultimately used for. I presented an example of such an analysis using an X-ray selected sample of galaxy clusters. Compared to galaxies, galaxy clusters have two important advantages. First, they cluster at length scales larger than galaxies and are thus less prone to effects of non-linear evolution. Second, masses of their hosting halos can be estimated observationally and these masses can be in turn used to predict how their clustering is enhanced compared to the underlying matter distribution. This way the clustering bias is no longer a free parameter but a function of cosmology and thus potentially offers extra constraining power. We show that the results from mass-based bias predictions are compatible with the actually observed clustering bias. We also show at the proof-of-concept level that comparing the measured and predicted two-point correlation functions, combined with the mass-based bias prediction, can be used to constrain the parameters Ω_m and σ_8. The constraints we obtain are compatible with the ones obtained from the Cosmic Microwave Background, although rather loose due to the small sample size of 1892 clusters. Euclid, for example, is expected to find of the order of 60 000 galaxy clusters in the redshift range of $z = 0.2$–2.0, making also this kind of analysis more precise.

References

Abazajian, K. et al. (Oct. 2003). "The First Data Release of the Sloan Digital Sky Survey". In: *The Astrophysical Journal* 126.4, pp. 2081–2086.

Abazajian, K. et al. (July 2004). "The Second Data Release of the Sloan Digital Sky Survey". In: *The Astrophysical Journal* 128.1, pp. 502–512.

Abazajian, K. et al. (Oct. 2016). "CMB-S4 Science Book, First Edition". In: *arXiv e-prints*, arXiv:1610.02743.

Akeson, R. et al. (Feb. 2019). "The Wide Field Infrared Survey Telescope: 100 Hubbles for the 2020s". In: *arXiv e-prints*, arXiv:1902.05569.

Aubourg, É. et al. (Dec. 2015). "Cosmological implications of baryon acoustic oscillation measurements". In: *Physical Review D* 92.12, p. 123516.

Bardeen, J. M. et al. (May 1986). "The Statistics of Peaks of Gaussian Random Fields". In: *The Astrophysical Journal* 304, p. 15.

Bennett, C. L. et al. (Oct. 2013). "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results". In: *Astrophysical Journal Supplement* 208.2, p. 20.

Bertone, G. et al. (Jan. 2005). "Particle dark matter: evidence, candidates and constraints". In: *Physics Reports* 405.5-6, pp. 279–390.

Bhattacharya, S. et al. (May 2011). "Mass Function Predictions Beyond ΛCDM". In: *The Astrophysical Journal* 732.2, p. 122.

Bond, J. R. and S. T. Myers (Mar. 1996). "The Peak-Patch Picture of Cosmic Catalogs. I. Algorithms". In: *Astrophysical Journal Supplement* 103, p. 1.

Bond, J. R. et al. (Oct. 1991). "Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations". In: *The Astrophysical Journal* 379, p. 440.

Bower, R. G. (Jan. 1991). "The evolution of groups of galaxies in the Press-Schechter formalism". In: *Monthly Notices of the Royal Astronomical Society* 248, pp. 332–352.

Bryan, G. L. and M. L. Norman (Mar. 1998). "Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons". In: *The Astrophysical Journal* 495.1, pp. 80–99.

Capasso, R. et al. (May 2020). "Mass calibration of the CODEX cluster sample using SPIDERS spectroscopy - II. The X-ray luminosity-mass relation". In: *Monthly Notices of the Royal Astronomical Society* 494.2, pp. 2736–2746.

Chandrasekhar, S. (Jan. 1943). "Stochastic Problems in Physics and Astronomy". In: *Rev. Mod. Phys.* 15 (1), pp. 1–89.

Cole, S. et al. (Sept. 2005). "The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications". In: *Monthly Notices of the Royal Astronomical Society* 362.2, pp. 505–534.

Comparat, J. et al. (Aug. 2017). "Accurate mass and velocity functions of dark matter haloes". In: *Monthly Notices of the Royal Astronomical Society* 469.4, pp. 4157–4174.

Costille, A. et al. (July 2018). "The EUCLID NISP grisms flight models performance". In: *Space Telescopes and Instrumentation 2018: Optical, Infrared, and Millimeter Wave*. Ed. by M. Lystrup et al. Vol. 10698. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 106982B.

Cropper, M. et al. (July 2016). "VIS: the visible imager for Euclid". In: *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*. Ed. by H. A. MacEwen et al. Vol. 9904. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 99040Q.

Davis, M. and P. J. E. Peebles (Apr. 1983). "A survey of galaxy redshifts. V. The two-point position and velocity correlations." In: *The Astrophysical Journal* 267, pp. 465–482.

Delabrouille, J. et al. (Apr. 2018). "Exploring cosmic origins with CORE: Survey requirements and mission design". In: *Journal of Cosmology and Astroparticle Physics* 2018.04, pp. 014–014.

DES Collaboration (July 2022). "Dark Energy Survey Year 3 Results: Constraints on extensions to ΛCDM with weak lensing and galaxy clustering". In: *arXiv e-prints*, arXiv:2207.05766.

DESI Collaboration (Oct. 2016). "The DESI Experiment Part I: Science,Targeting, and Survey Design". In: *arXiv e-prints*, arXiv:1611.00036.

Dolag, K. et al. (Feb. 2008). "Simulation Techniques for Cosmological Simulations". In: *Space Science Reviews* 134.1-4, pp. 229–268.

Doroshkevich, A. G. (Oct. 1970). "Spatial structure of perturbations and origin of galactic rotation in fluctuation theory". In: *Astrophysics* 6.4, pp. 320–330.

Eisenstein, D. J. and W. Hu (Jan. 1999). "Power Spectra for Cold Dark Matter and Its Variants". In: *The Astrophysical Journal* 511.1, pp. 5–15.

Finoguenov, A. et al. (June 2020). "CODEX clusters. Survey, catalog, and cosmology of the X-ray luminosity function". In: *Astronomy & Astrophysics* 638, A114.

Grieb, J. N. et al. (Apr. 2016). "Gaussian covariance matrices for anisotropic galaxy clustering measurements". In: *Monthly Notices of the Royal Astronomical Society* 457.2, pp. 1577–1592.

Hamilton, A. J. S. (Nov. 1993). "Toward Better Ways to Measure the Galaxy Correlation Function". In: *The Astrophysical Journal* 417, p. 19.

Hazumi, M. et al. (Dec. 2020). "LiteBIRD satellite: JAXA's new strategic L-class mission for all-sky surveys of cosmic microwave background polarization". In: *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Vol. 11443. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 114432F.

Heymans, C. et al. (Nov. 2012). "CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey". In: *Monthly Notices of the Royal Astronomical Society* 427.1, pp. 146–166.

Hildebrandt, H. et al. (Feb. 2017). "KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing". In: *Monthly Notices of the Royal Astronomical Society* 465.2, pp. 1454–1498.

Hinshaw, G. et al. (Oct. 2013). "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results". In: *Astrophysical Journal Supplement* 208.2, p. 19.

Hoekstra, H. et al. (Aug. 2006). "First Cosmic Shear Results from the Canada-France-Hawaii Telescope Wide Synoptic Legacy Survey". In: *The Astrophysical Journal* 647.1, pp. 116–127.

Hoekstra, H. et al. (Jan. 2002). "Lensing Results from the Red-sequence Cluster Survey". In: *A New Era in Cosmology*. Ed. by N. Metcalfe and T. Shanks. Vol. 283. Astronomical Society of the Pacific Conference Series, p. 169.

Ivezić, Ž. et al. (Mar. 2019). "LSST: From Science Drivers to Reference Design and Anticipated Data Products". In: *The Astrophysical Journal* 873.2, p. 111.

Jarvis, M. et al. (June 2006). "Dark Energy Constraints from the CTIO Lensing Survey". In: *The Astrophysical Journal* 644.1, pp. 71–79.

Jenkins, A. et al. (Feb. 2001). "The mass function of dark matter haloes". In: *Monthly Notices of the Royal Astronomical Society* 321.2, pp. 372–384.

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

Keihänen, E. et al. (Nov. 2019). "Estimating the galaxy two-point correlation function using a split random catalog". In: *Astronomy & Astrophysics* 631, A73.

Keihänen, E. et al. (Oct. 2022). "Euclid: Fast two-point correlation function covariance through linear construction". In: *Astronomy & Astrophysics* 666, A129.

Keitel, D. and P. Schneider (Oct. 2011). "Constrained probability distributions of correlation functions". In: *Astronomy & Astrophysics* 534, A76.

Kerscher, M. et al. (May 2000). "A Comparison of Estimators for the Two-Point Correlation Function". In: *The Astrophysical Journal* 535.1, pp. L13–L16.

Kiiveri, K. et al. (Mar. 2021). "CODEX weak lensing mass catalogue and implications on the mass-richness relation". In: *Monthly Notices of the Royal Astronomical Society* 502.1, pp. 1494–1526.

Klein, M. et al. (Sept. 2019). "A new RASS galaxy cluster catalogue with low contamination extending to z $\sim$ 1 in the DES overlap region". In: *Monthly Notices of the Royal Astronomical Society* 488.1, pp. 739–769.

Klypin, A. et al. (Apr. 2016). "MultiDark simulations: the story of dark matter halo concentrations and density profiles". In: *Monthly Notices of the Royal Astronomical Society* 457.4, pp. 4340–4359.

Kolb, E. W. and M. S. Turner (1990). *The early universe*. Vol. 69.

Komatsu, E. et al. (Feb. 2011). "Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation". In: *Astrophysical Journal Supplement* 192.2, p. 18.

Lahav, O. et al. (July 1991). "Dynamical effects of the cosmological constant." In: *Monthly Notices of the Royal Astronomical Society* 251, pp. 128–136.

Landy, S. D. and A. S. Szalay (July 1993). "Bias and Variance of Angular Correlation Functions". In: *The Astrophysical Journal* 412, p. 64.

Laureijs, R. et al. (Oct. 2011). "Euclid Definition Study Report". In: *arXiv e-prints*, arXiv:1110.3193.

Lesgourgues, J. (Apr. 2011). "The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview". In: *arXiv e-prints*, arXiv:1104.2932.

Lewis, A. and S. Bridle (Nov. 2002). "Cosmological parameters from CMB and other data: A Monte Carlo approach". In: *Physical Review D* 66.10, p. 103511.

Lewis, A. et al. (Aug. 2000). "Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models". In: *The Astrophysical Journal* 538.2, pp. 473–476.

Liddle, A. R. and D. H. Lyth (2000). *Cosmological Inflation and Large-Scale Structure*.

Lindholm, V. et al. (Feb. 2021). "Clustering of CODEX clusters". In: *Astronomy & Astrophysics* 646, A8.

Lippich, M. et al. (Jan. 2019). "Comparing approximate methods for mock catalogues and covariance matrices - I. Correlation function". In: *Monthly Notices of the Royal Astronomical Society* 482.2, pp. 1786–1806.

Marulli, F. et al. (Nov. 2012). "Cosmology with clustering anisotropies: disentangling dynamic and geometric distortions in galaxy redshift surveys". In: *Monthly Notices of the Royal Astronomical Society* 426.3, pp. 2566–2580.

Mo, H. J. et al. (2010). *Galaxy Formation and Evolution*.

Monaco, P. et al. (2002). "The PINOCCHIO algorithm: pinpointing orbit-crossing collapsed hierarchical objects in a linear density fiel". In: *Monthly Notices of the Royal Astronomical Society* 331, p. 587.

Moscardini, L. et al. (Aug. 2000). "Predicting the clustering of X-ray selected galaxy clusters in flux-limited surveys". In: *Monthly Notices of the Royal Astronomical Society* 316.2, pp. 283–298.

Mukhanov, V. F. et al. (June 1992). "Theory of cosmological perturbations". In: *Physics Reports* 215.5-6, pp. 203–333.

Munari, E. et al. (2017). "Improving fast generation of halo catalogues with higher order Lagrangian perturbation theory". In: *Monthly Notices of the Royal Astronomical Society* 465, p. 4658.

Nakamura, T. T. and Y. Suto (Jan. 1997). "Strong Gravitational Lensing and Velocity Function as Tools to Probe Cosmological Parameters — Current Constraints and Future Predictions —". In: *Progress of Theoretical Physics* 97, p. 49.

Netterfield, C. B. et al. (June 2002). "A Measurement by BOOMERANG of Multiple Peaks in the Angular Power Spectrum of the Cosmic Microwave Background". In: *The Astrophysical Journal* 571.2, pp. 604–614.

Nikakhtar, F. et al. (Aug. 2018). "The Excursion set approach: Stratonovich approximation and Cholesky decomposition". In: *Monthly Notices of the Royal Astronomical Society* 478.4, pp. 5296–5300.

Okumura, T. et al. (Apr. 2008). "Large-Scale Anisotropic Correlation Function of SDSS Luminous Red Galaxies". In: *The Astrophysical Journal* 676.2, pp. 889–898.

Peacock, J. A. (1999). *Cosmological Physics*.

Peacock, J. A. et al. (Mar. 2001). "A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey". In: *Nature* 410.6825, pp. 169–173.

Penzias, A. A. and R. W. Wilson (July 1965). "A Measurement of Excess Antenna Temperature at 4080 Mc/s." In: *The Astrophysical Journal* 142, pp. 419–421.

Percival, W. J. et al. (Nov. 2001). "The 2dF Galaxy Redshift Survey: the power spectrum and the matter content of the Universe". In: *Monthly Notices of the Royal Astronomical Society* 327.4, pp. 1297–1306.

Planck Collaboration (Sept. 2020a). "Planck 2018 results. I. Overview and the cosmological legacy of Planck". In: *Astronomy & Astrophysics* 641, A1.

— (Sept. 2020b). "Planck 2018 results. VI. Cosmological parameters". In: *Astronomy & Astrophysics* 641, A6.

— (Sept. 2020c). "Planck 2018 results. X. Constraints on inflation". In: *Astronomy & Astrophysics* 641, A10.

Prada, F. et al. (July 2012). "Halo concentrations in the standard Λ cold dark matter cosmology". In: *Monthly Notices of the Royal Astronomical Society* 423.4, pp. 3018–3030.

Rykoff, E. S. et al. (Apr. 2014). "redMaPPer. I. Algorithm and SDSS DR8 Catalog". In: *The Astrophysical Journal* 785.2, p. 104.

Schmidt, F. et al. (July 2013). "Peak-background split, renormalization, and galaxy clustering". In: *Physical Review D* 88.2, p. 023515.

Schneider, P. and J. Hartlap (Sept. 2009). "Constrained correlation functions". In: *Astronomy & Astrophysics* 504.3, pp. 705–717.

Semboloni, E. et al. (June 2006). "Cosmic shear analysis with CFHTLS deep data". In: *Astronomy & Astrophysics* 452.1, pp. 51–61.

Sheth, R. K. and G. Tormen (Sept. 1999). "Large-scale bias and the peak background split". In: *Monthly Notices of the Royal Astronomical Society* 308.1, pp. 119–126.

Sheth, R. K. et al. (May 2001). "Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes". In: *Monthly Notices of the Royal Astronomical Society* 323.1, pp. 1–12.

Smoot, G. F. et al. (Sept. 1992). "Structure in the COBE Differential Microwave Radiometer First-Year Maps". In: *The Astrophysical Journal* 396, p. L1.

Spergel, D. N. et al. (Sept. 2003). "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters". In: *Astrophysical Journal Supplement* 148.1, pp. 175–194.

Spergel, D. N. et al. (June 2007). "Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Implications for Cosmology". In: *Astrophysical Journal Supplement* 170.2, pp. 377–408.

Suzuki, N. et al. (Feb. 2012). "The Hubble Space Telescope Cluster Supernova Survey. V. Improving the Dark-energy Constraints above $z > 1$ and Building an Early-type-hosted Supernova Sample". In: *The Astrophysical Journal* 746.1, p. 85.

Tegmark, M. et al. (May 2004). "Cosmological parameters from SDSS and WMAP". In: *Physical Review D* 69.10, p. 103501.

Tinker, J. L. et al. (Dec. 2010). "The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests". In: *The Astrophysical Journal* 724.2, pp. 878–886.

Van Waerbeke, L. et al. (June 2000). "Detection of correlated galaxy ellipticities from CFHT data: first evidence for gravitational lensing by large-scale structures". In: *Astronomy & Astrophysics* 358, pp. 30–44.

Voges, W. et al. (Sept. 1999). "The ROSAT all-sky survey bright source catalogue". In: *Astronomy & Astrophysics* 349, pp. 389–405.

Wang, M. S. et al. (June 2019). "Cosmological inference from galaxy-clustering power spectrum: Gaussianization and covariance decomposition". In: *Monthly Notices of the Royal Astronomical Society* 486.1, pp. 951–965.

Warren, M. S. et al. (Aug. 2006). "Precision Determination of the Mass Function of Dark Matter Halos". In: *The Astrophysical Journal* 646.2, pp. 881–885.

Wilking, P. and P. Schneider (Aug. 2013). "A quasi-Gaussian approximation for the probability distribution of correlation functions". In: *Astronomy & Astrophysics* 556, A70.

York, D. G. et al. (Sept. 2000). "The Sloan Digital Sky Survey: Technical Summary". In: *The Astrophysical Journal* 120.3, pp. 1579–1587.

Zehavi, I. et al. (July 2011). "Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity". In: *The Astrophysical Journal* 736.1, p. 59.

Zel'dovich, Y. B. (Mar. 1970). "Gravitational instability: An approximate theory for large density perturbations." In: *Astronomy & Astrophysics* 5, pp. 84–89.

Zhao, C. et al. (Apr. 2022). "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: cosmological implications from multitracer BAO analysis with galaxies and voids". In: *Monthly Notices of the Royal Astronomical Society* 511.4, pp. 5492–5524.