Università degli Studi di Trieste
Department of Physics

PhD Thesis

# Covariance matrix estimation for the statistics of galaxy clustering

PhD Student:

**Manuel Colavincenzo**

Supervisor:

**Prof. Pierluigi Monaco**

Co-Supervisors:

**Prof. Stefano Borgani**
**Dott. Emiliano Sefusatti**

# Contents

# List of Figures

# Chapter 1

# Introduction

This Ph.D. thesis is focused on the study of statistical analysis of the large-scale distribution of galaxies. The accurate determination of the cosmological parameters has become one of the key activities in modern cosmology: the understanding of the nature of the dark components of the Universe, dark matter and dark energy, as well as the study of the behavior of gravity at very large scales, is fundamental to improve our knowledge on the history and evolution of the Universe. The structures we observe today result from the growth of initial perturbations due to gravitational instability. The spectrum of these perturbations, and therefore the statistical properties of the galaxy distribution, is defined by the cosmological model. Choosing between different models requires accurate theoretical predictions of the observables and precise modeling of their uncertainties.

Observations of the Cosmic Microwave Background (CMB), (Komatsu et al., 2011; Planck Collaboration et al., 2015), the electromagnetic radiation relic of the early stage of the Universe ($\sim$380'000 years after the Big Bang), in combination with measurements of the expansion of the Universe, e.g. (Freedman et al., 2001) and Large Scale Structures (LSS) probes, e.g. (Allen et al., 2011; Percival et al., 2011; Anderson et al., 2014; Kiessling et al., 2015), have allowed cosmology to enter the era of high precision. Presently, CMB observations alone are able to determine the values of cosmological parameters with per-cent accuracy. The $\Lambda$CDM model has passed this tough observational test, and is now considered as the standard model for cosmology, although it is based on the two unknown ingredients of dark matter and dark energy. The latter manifests itself as a cosmological constant. Deviations of its equation of state (p=$w\rho$) from the simplest $w = -1$, where $w$ is the dark energy parameter, would represents deviation from the $\Lambda$CDM model.

In the last decades a great effort has been devoted to the analysis of the distribution of matter on cosmic scales. To explain the large scale structures in the low redshift ($z < 3$) Universe, we have to account for some degree of clumpiness in the early stage of the Universe evolution, otherwise matter would fail to cluster and the Universe, because of the expansion, would become a homogeneous rarefied gas of elementary particles. CMB observations assure that the primordial fluctuations are small, because we observe a little degree of inhomogeneities with respect the homogeneous background. As an example, in figure (1.1) we show four slices of the Millenium Simulation (Springel et al., 2005) reproducing the distribution of dark matter different epochs: the upper left figure shows the Universe at z=18.3 when it was highly homogeneous; here, we cannot identify structures like clusters of galaxies or filaments, that connect clusters and groups of galaxies. Moving to low redshift, upper right (z=5.7) and bottom left (z=1.4) figures, we note that, because of gravitational instability, the distribution of cosmic structures becomes less homogeneous and cluster start becoming prominent structures forming at intersection of filament. At redshift z=0, bottom right figure, we show the Universe we observe today. All the structures are differentiated, and we can easily identify the halo of galaxy clusters and long filaments.

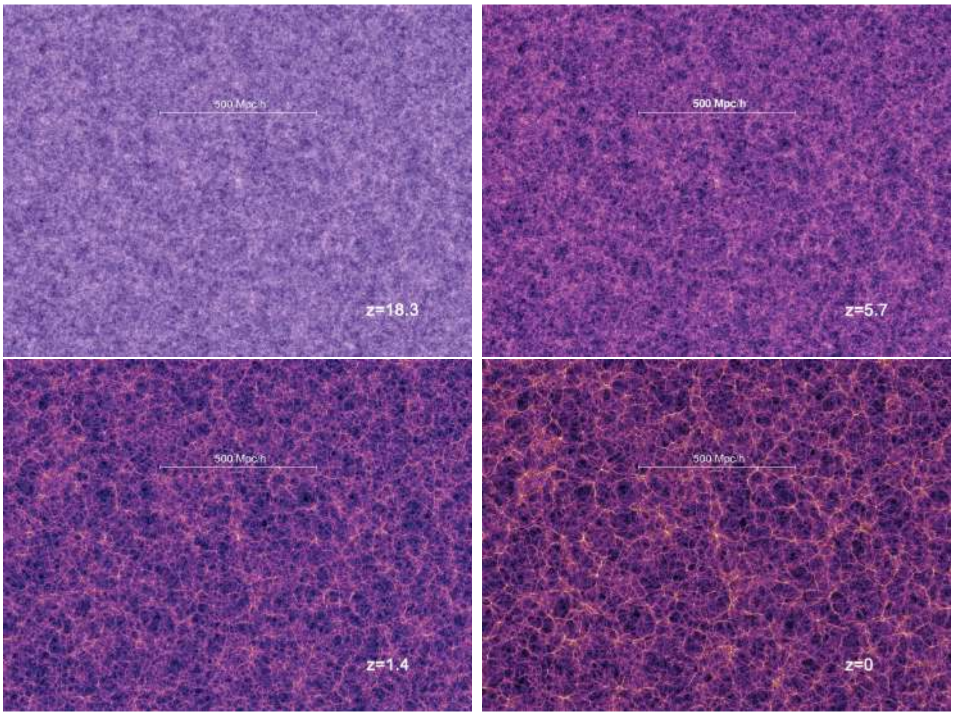


Figure 1.1: Four slices of thickness 15 $h^{-1}$ Mpc, cut from the Millennium Run, showing the dark matter distribution at different redshift, from z=18.3 to z=0. Intensity colors encodes surface density and color encodes local velocity dispersion (Croton et al., 2006).

The analyses that we will discuss in this thesis are relevant for the high precision observations to be obtain with forthcoming and future galaxy surveys like DESI (Levi et al., 2013)), Extended Baryon Oscillation Spectroscopic Survey (eBOSS), Euclid (Laureijs et al., 2011), Wide-Field Infrared Survey Telescope (WFIRST, Green et al., 2012) and the Square Kilometer Array (SKA), that will measure galaxy clustering with sub-percent accuracy.

A wealth of information on cosmological parameters is enclosed in the *two-point correlation function* and in its counterpart in Fourier space, the *Power Spectrum*. The modeling of these two quantities is fundamental in order to describe the clustering strength of different structures on different scales. The ability to extract useful constraints from these statistical measures depends on the accuracy of the modeling of the two-point statistics. As we have already highlighted, the gravitational instability, that is a non-linear process, is responsible for the distribution of the galaxies that we observe; for this reason the evolution of structures cannot be Gaussian, even if the initial conditions are Gaussian. If the evolution of clustering is non-linear, we have to take into account that higher order statistics, *three-point correlation function* or the *bispectrum* in Fourier space, need to be accurately modeled.

In recent years, a great effort has been devoted to provide accurate estimates of the matter power spectrum, using different approaches: perturbation theory (Bernardeau et al., 2002), halo model (Cooray and Sheth, 2002) and simulations (Heitmann et al., 2009, 2010).

The analysis of the clustering statistics, two- and three-point functions, is fundamental as we have already pointed out, but in order to perform any statistical analysis of the large-scale galaxy distribution, it is of primary importance to quantify the errors on these quantities. This information is carried out by the **covariance matrix**. The diagonal of this matrix, the *variance*, represents the squared deviation of the measures from the mean value, while the other elements of the matrix provide information on, if and how, the behavior of a certain quantity we observe on same scale affects the behavior of the same quantity on other scales. If the off-diagonal elements are different from zero, it means that the errors at different scales are correlated. At this point it is important to stress the difference between precision and accuracy: the first one refers to the quantification of the random errors, the second refers to the systematic errors.

Present day large-scale galaxy surveys can observe very large volumes with high precision that has to be paralleled with just as much precision in the determination of the errors. As we will stress in the rest of this thesis, to evaluate the errors, or in general the covariance matrices, we need many realizations of our Universe. The only realization we have access to is the one where we live, so what we can do is to simulate other realizations. In this sense we are assuming that the cosmic density field, that describes the distribution of matter in the Universe, is a random field generated as a random realization out of an ensemble of possible realizations.

Numerical simulations are generally used to generate large sets of realizations of the Universe (Scoccimarro and Sheth, 2002; Takahashi et al., 2009; Sato et al., 2011; Harnois-Déraps and Pen, 2012; Dodelson and Schneider, 2013; Li et al., 2014a; Blot et al., 2015, 2016). As we will describe in chapter 6, the accurate evaluation of the covariance matrix requires a large number of simulations, but the precision requirements in terms of volume and resolutions make unfeasible to proceed with computationally expansive N-body simulations. Various methods to reduce the number of simulations are now used, such as shrinkage (Schäfer and Strimmer, 2005; Pope and Szapudi, 2008) and tapering (Kaufman et al., 2008; Paz and Sánchez, 2015). An alternative to N-body simulations is given by **approximate methods** that allows us to produce a large number of synthetic galaxy catalogs at relatively low cost (a factor of ∼1000 in terms of computing time). These techniques can be divided in two main classes: the methods that use perturbation theory to follow the particle orbits, such as PINOCCHIO (Monaco et al., 2002), PTHalos (Scoccimarro and Sheth, 2002), COLA (Tassev et al., 2013), AugmentedLPT (Kitaura and Heß, 2013), and the methods that start from a non-linear density field and then populate this density field with dark matter halos, the structure hosting galaxies, such as PATCHY (Kitaura et al., 2014), EZmocks (Chuang et al., 2015), HALOGEN (Avila et al., 2015).

All these approaches can take great advantage of analytic modeling of the covariance matrix (Scoccimarro et al., 1999; Sefusatti et al., 2006; de Putter et al., 2012; Takada and Hu, 2013; Grieb et al., 2016; Pearson and Samushia, 2016; Mohammed et al., 2017).

The majority of the above cited papers are devoted to the study of the *matter* density field. From observations we get information on the galaxy density field and we infer that galaxies tend to form in overdense matter regions; for this reason we can say that galaxies are *biased tracers* of the matter density field. Less attention has been paid, so far, to analytic predictions of the *galaxy* power spectrum covariance matrix, since galaxy bias and discreetness effects make its modeling more complex. Finally, in the analysis of real surveys, we should account for the selection function that defines the galaxy sample and its geometry that can introduce spurious correlations between different scales. It is fundamental to include all these features in the modeling of the covariance matrix, in preparation of future galaxy surveys. We do not expect that the analytic predictions can fully replace numerical evaluations of the covariance matrix, but they can help to reduce the number of realizations needed for its evaluation.

In this thesis we study the problem of covariance matrix estimation for clustering. We first study the power spectrum and the bispectrum, quantities that encode a great part of the clustering information. Then, as we have highlighted at the beginning of this introduction, we focus on the computation of the uncertainties on these quantities. To proper quantify the errors we have to include different effects that introduce different complications in the modeling.

In order to check the accuracy of the analytic predictions, we have also looked at the covariance from a numerical point of view; for the covariance analysis a large number of galaxy catalogs are needed to lower the statistical noise that affects in particular the off-diagonal terms of the covariance matrix. For this reason the theory is supported by simulations, that are carried out by resorting to "approximate methods", that allow us a fast production of galaxy catalogs at the expense of a less correct characterization of the non-linear scales. At the same time we take advantage of the massive production of simulations based on fast approximate methods in the regime where the analytic modeling cannot be applied. A comparison of different techniques against exact N-body simulations is required to test the accuracy of the approximations in reproducing the covariance matrix.

In this PhD thesis specific attention is given to the survey that will be carried out by the Euclid[1] satellite, an ESA space telescope equipped with a 1.2 m diameter mirror combined with a high large field-of-view visible imager, a near infrared 3-filter photometer and a slitless spectrograph, aimed at studying the accelerated expansion of the Universe; the imprints of dark energy and gravity will be tracked by using two complementary cosmological probes to capture signatures of the expansion history of the Universe and the growth of cosmic structures: weak gravitational lensing and galaxy clustering. Euclid will cover 15'000 deg$^2$ of extra-galactic sky, with 40 deg$^2$ deep fields. The main photometric redshift range will be between 1 and 3: Euclid will provide a good photometric[2] redshifts for $\sim 2 \times 10^8$ galaxies; in the range 0.9-1.8 it will measure spectroscopic[3] redshifts for $\sim 50 \times 10^6$ galaxies. Euclid will also observe higher redshift, in particular the brightest galaxies at redshift $z > 7$ and the brightest quasars at redshift $z > 8$. The H$\alpha$ emission, that occurs when a hydrogen electron falls from its third to second lowest energy level, will be the main spectral features for the determination of spectroscopic; the limiting flux limit at a wavelength of 1200 nm will be of $\sim 2 \times 10^{-16}$ erg s$^{-1}$ cm$^{-2}$. In this thesis we will focus on the spectroscopic survey.

This thesis is divided in eight chapters. Chapter 2 presents a review of the standard cosmological model, with particular attention to galaxy clustering properties and the large scale structure of the Universe. In the first part of the chapter we highlight the main observational probes of the standard cosmological model including a description of the dark components of our Universe. We describe then, the theory of structure formation reviewing the main statistical properties of the cosmic density field introducing observable

[1] https://www.euclid-ec.org/

[2] The photometric redshift is obtained from the comparison of the magnitudes of galaxies in different bands with the expected magnitude from templates.

[3] The spectroscopic redshift is obtained from the measurements of the shifts in the spectrum lines

quantities, such as the two-point correlation function, the power spectrum and the higher-order statistics and focusing on the evolution of cosmological density perturbations. In the last part of the chapter we focus on the relation between the matter and galaxy density fields.

In chapter 3 we focus on the two-point function clustering estimator, both in configuration and Fourier space introducing quantities that have been used in the rest of the thesis. We then introduce the general estimator for the power spectrum covariance matrix that is of fundamental importance for all the following chapters. We then focus on two of the main probes for the constraints on cosmological parameters, such as Baryonic acoustic oscillation and redshift-space Distortion and on the more recent results obtained from large-scale galaxy surveys.

In chapter 4 we briefly overview the basic principles of N-body simulations; then we focus on the approximate methods that are of primary importance for this thesis. N-body simulations are capable to describe the evolution of the density inhomogeneities from large scales, in the linear regime, to the deep non-linear regime, so that they are the main tools to describe highly non-linear scales. The approximate methods are less accurate in the description of the non-linear scales, but they excel in computation speed allowing to produce a large number of realizations in relative short time.

In chapter 5 we present a study of some of the systematic affecting the clustering of biased tracers, that will have an impact on the analysis of future surveys. We model the systematic using a "mask", that acts like a selection function, changing the observed galaxy number density and introducing spurious correlations between different scales, simulating effects due to, for instance, the zodiacal light and the Milky Way extinction. We provide a theoretical description of our idealized mask model and its effects on the measurements of the power spectrum. We analyze all the corrections to the power spectrum and its covariance due to the presence of this foreground, including its coupling with the cosmological signal. The analysis, and the test of the mask model itself, are carried out using a large set of cosmological galaxy catalogs made with the approximate method PINOCCHIO (Monaco et al., 2002). This work has been published in Colavincenzo et al. (2017).

Within the Euclid collaboration I have been involved in the comparison of the covariance matrix obtained using different approximate methods. In chapter 6 we show the analysis aimed at testing these techniques against N-body simulations; in particular we want to prove that covariance matrices of clustering measures, obtained using the fast methods, are unbiased; we also want to quantify their impact on the errors on cosmological parameters. The test involves the comparison of the most relevant statistical measures of galaxy clustering, namely the halo power spectrum, the two-point correlation function and the bispectrum. After a description of the different simulations used for the comparison, this chapter is mainly focused on the results obtained for the halo power spectrum and for the bispectrum. These two analyses will be published in two separate papers (Blot et al. in preparation for the power spectrum, Colavincenzo et al. in preparation for the bispectrum)

In chapter 7 we focus our attention on the theoretical modeling of the galaxy power spectrum covariance matrix. The aim is to investigate the quasi-linear regime which contains a large fraction of the information needed to constraint cosmological parameters. To reach this goal, we test the theoretical prediction for the covariance matrix of the power spectrum of biased objects, in the particular case of halos, in the simple case of a cubic box with periodic boundary conditions in real space. One of the main purposes of this work is to properly account for shot-noise and galaxy bias. This work will be presented in a paper (Colavincenzo et al. in preparation).

Chapter 8 is dedicated to the bispectrum analysis. To this purpose we use a model for the galaxy bispectrum to constrain the halo bias parameters. Before describing our analysis, we stress the complexity of the bispectrum measurement, pointing out the analytic procedures that can help when the numerical approach becomes unfeasible. Then we describe the statistical analysis we have carried out assuming a specific model for the galaxy bias. In the last part we show the results further stressing the importance of the approximate methods in the covariance analysis. This bispectrum analysis will be presented in a paper (Colavincenzo et al. in preparation).

Finally, in the conclusions, we summarize the results, emphasizing the role of the covariance matrix for the study of the large-scale distribution of galaxies. The analyses we have carried out are all devoted to recover more precise and robust constraints on cosmological parameters, in preparation for the great effort that will be required in the interpretation of the data from Euclid, and other future galaxy surveys. In the final part of the last chapter we will also discuss future developments of the analyses presented in this thesis.

# Chapter 2

# Large-scale structure

The study of the spatial distribution of galaxies is of great interest in cosmology. A great deal of information is enclosed in galaxy clustering and it can be extracted by large galaxy surveys. Learning about the dynamics of gravitation instability and statistical properties of galaxy distribution is at the base of a deeper understanding of the observations. As an example (figure 2.1) we show a portion of the sky observed by the Sloan Digital Sky Survey (SDSS, York et al., 2000). In this figure the Earth is located at the center; the redshift represents the distance between the observer and the galaxies. As we shall see in the next chapters, redshift can be used as distance indicator when the *peculiar velocity* is defined. The observed galaxies are not distributed uniformly and homogeneously, but they tend to stay in long *filaments*. Between these filaments we find large regions with a very small number of galaxies: we call these regions *voids*.

In the first section of this chapter we describe the standard cosmological model, focusing on the pillars of the expanding Universe theory. Then we illustrate the so called *Dark Sector*, by reviewing briefly the dark matter and the dark energy role. Then we will focus on the theory of structure formation that is of main interest for the goals of these thesis and on the relation between the galaxy and matter density fields.

These chapters are based on the textbooks Coles and Lucchin (1995); Dodelson (2003); Liddle (2003); Mo et al. (2010)

## 2.1 Standard cosmological model

The standard cosmological model is based on two main *First Principles*:

- the Universe, on average and on sufficiently large scales, is homogeneous and isotropic;

- the *Copernican principle* states that we are not privileged observers of

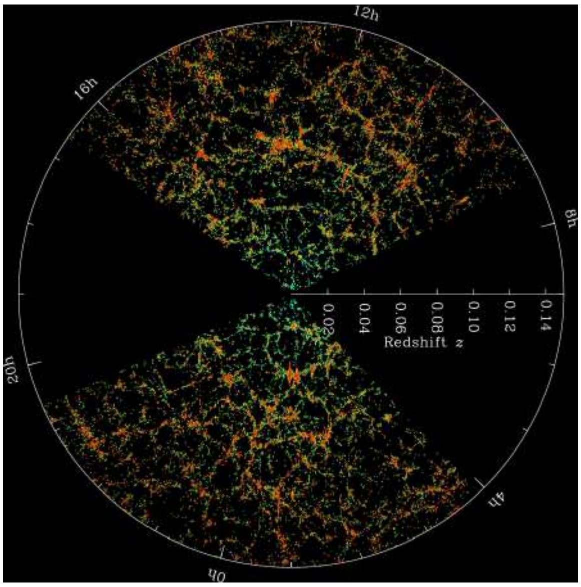


Figure 2.1: Slices through the SDSS 3-dimensional map of the distribution of galaxies. Earth is at the center, and each point represents a galaxy, typically containing about 100 billion stars. The outer circle is at a distance of two billion light years. Both slices contain all galaxies within -1.25 and 1.25 degrees declination. Credit: M. Blanton and the Sloan Digital Sky Survey.

the Universe.

These two principles can be translated in mathematical language using Einstein's theory of General Relativity (GR), according to which the space-time structure of the Universe is determined by the matter distribution in it. In this sense, GR is a *geometrical* theory because the distribution of matter in the Universe determines the geometry of the Universe itself. The most generic metric that describes a homogeneous and isotropic Universe is given by the Robertson-Walker metric:

$$ds^2 = (cdt)^2 - a^2(t)\Big[\frac{dr^2}{1-Kr^2} + r^2(d\theta^2 + \sin\theta^2 d\phi^2)\Big] \ , \tag{2.1}$$

where $ds^2$ is the line element, $c$ is the speed of light, $a(t)$ is the *scale factor*, $K$ is the *curvature parameter* and (r,$\theta$,$\phi$) are the comoving spatial coordinates.

It is worth to analyze all the information contained in the relation 2.1: first of all, it incorporates the Copernican principle, because the curvature parameter $K$ is constant and this corresponds to have no preferred location in the Universe; the geometry is fixed by the value of $K$: values of 1,0,-1 correspond, respectively, to a 3D-sphere (closed Universe), to Euclidean space (flat Universe) and to an hyperbolic space (open Universe). The last piece of information is contained in the scale factor: if $a(t)$ is not constant and it evolves with time, then the Universe is expanding or contracting so that the physical distance between two points changes. The evolution history of the Universe is determined by the study of the evolution of the scale factor with time. To this purpose we can define the Hubble parameter as the rate of change of the scale factor:

$$H(t) \equiv \frac{da/dt}{a} \ . \tag{2.2}$$

The expansion of the Universe was observed for the first time by Edwin Hubble. In his paper (Hubble, 1929) he showed that galaxies are moving away from us and their recession velocity increases proportionally to their distance. When a photon is emitted by a galaxy, it is received by an observer with a frequency different from the emitted one because of the expansion of the Universe. We can quantify this shift by introducing the *redshift*:

$$z = \frac{\nu_e}{\nu_o} - 1 = \frac{1}{a} - 1 \tag{2.3}$$

where $\nu_e$ is the frequency of the emitted photon while $\nu_o$ is the observed ones. $z$ is connected to the recession velocity of the galaxy. These observations translate into the Hubble law:

$$cz = H_0 r \qquad \text{with} \quad z \ll 1 \ , \tag{2.4}$$

with $H_0$ the Hubble constant today given by:

$$H_0 = 100 \ h^{-1}\mathrm{Mpc}^{-1} \ \mathrm{km \ s}^{-1} \ . \tag{2.5}$$

As we have already mentioned, the structures that we observe have grown from an initial perturbation in the mean density field; the growth of the structures induces velocities that deviate from the pure Hubble expansion. We call *peculiar velocity*, $v_{\mathrm{pec}}$, the velocity of a galaxy with respect to the frame comoving with the CMB. We can characterized the galaxy velocity considering a contribution from the Hubble expansion and another one coming from the motion with respect to the expansion:

$$v_r = H_0 r + v_{pec} \ . \tag{2.6}$$

In section (3.2.2) we will describe the effects on clustering due to the presence of the peculiar velocity.

The general equation that relates the space-time geometry with the matter-energy content in the Universe is the Einstein's equation:

$$G_{\mu\nu} - g_{\mu\nu}\Lambda = \frac{8\pi G}{c^4} T_{\mu\nu} \ , \tag{2.7}$$

where $G_{\mu\nu}$ is the Einstein tensor, $G$ is the the Newton's constant, $T_{\mu\nu}$ is the stress-energy tensor, $\Lambda$ the cosmological constant, $g_{\mu\nu}$ is the metric tensor

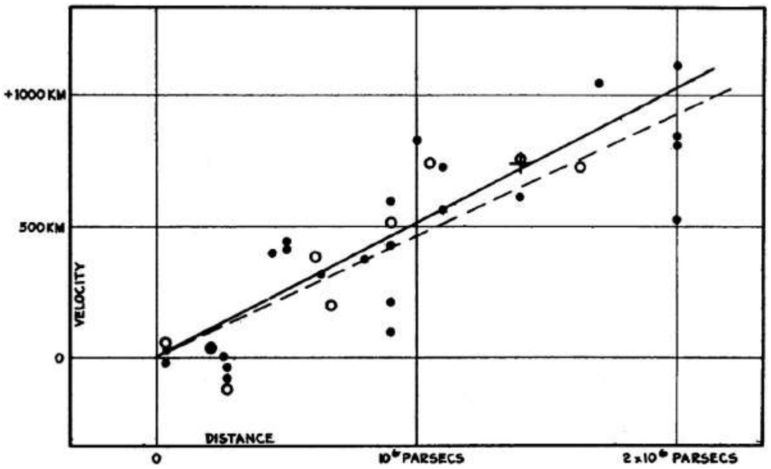


Figure 2.2: Original Hubble diagram (Hubble, 1929). Radial velocity of distant galaxies against distances. The black dots are galaxies corrected for the sun's motion, white circle are not. The solid and the dashed line are, respectively, the best fit of the two groups.

and $c$ the speed of light. According to the standard model, the Universe is filled with a matter component, including the usual Baryonic matter and the non-Baryonic collisionless Dark Matter (see section 2.1.2), a radiation component including photons and neutrinos and a cosmological constant component (see section 2.1.2) with negative pressure. The properties of these constituents are described by a symmetric stress-energy tensor. In the rest frame of the fluids, assuming they are *perfect isotropic fluids*, the stress-energy tensor takes the following form:

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & \mathcal{P} & 0 & 0 \\ 0 & 0 & \mathcal{P} & 0 \\ 0 & 0 & 0 & \mathcal{P} \end{pmatrix} \tag{2.8}$$

with $\mathcal{P}$ the pressure of the fluid and $\rho$ its density. Studying the evolution of $T_{\mu\nu}$ it is possible to derive the evolution of all the components as function of time. The conservation equation for an expanding Universe is:

$$\frac{d\rho}{dt} - 3H(t)\Big(\rho + \frac{\mathcal{P}}{c^2}\Big) = 0 \ , \tag{2.9}$$

where $H(t)$ is defined in eq. 2.2; each of the components we have described above evolve independently following the conservation equation 2.9. Assuming that dark energy is given by the cosmological constant, the equation of state is p $= w\rho c^2$, for each component and the solution is $\rho(a) \propto a^{-3(1+w)}$.

For the matter component $w = 0$ so $\rho_m \propto a^{-3}$, for the radiation component $w = 1/3$ and $\rho_r \propto a^{-4}$ and for the cosmological constant component

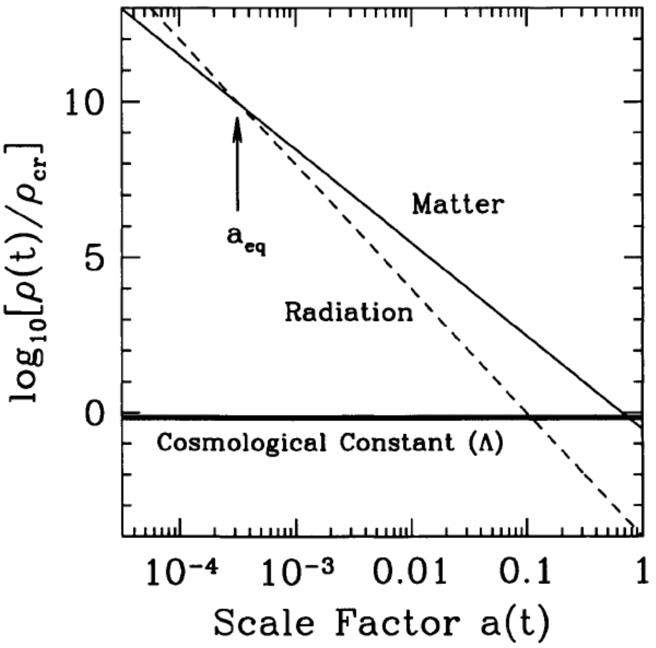


Figure 2.3: Energy densities as function of the scale factor for flat Universe. The solid line shows the non-relativistic matter while the dashed one the radiation. The horizontal solid black line is the dark energy component that remains constant during the evolution. The relative energy per component is normalized with respect to the critical density. Dodelson (2003)

$w = -1$ so that the dark energy density is constant in time, $\rho_\Lambda = \text{const}$. In figure 2.3 we plot the energy density for each component, matter, radiation and cosmological constant as a function of the scale factor a(t). As we can see the Universe undergoes different epochs, each of them characterized by one dominant component: in the early stage the radiation is the most relevant one (small values of the scale factor); then, at a time called *matter-radiation equivalence ($a_{eq}$)* the radiation and the matter components are equally important; after the equivalence, matter becomes dominant. During this time the dark energy component is orders of magnitude smaller than radiation and matter and therefore it is negligible. This is true until $a(t) < 1$; today, at $a(t) = 1$, the Universe has entered in a new phase dominated by the dark energy component.

Taking the 00 component and the trace of the Einstein equations 2.7, we can derive the two Friedmann equations:

$$\frac{\dot{a}^2}{a^2} = \frac{Kc^2}{a^2} + \frac{8\pi G\rho}{3} + \frac{\Lambda c^2}{3} \tag{2.10}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + \frac{3\mathrm{p}}{c^2}) + \frac{\Lambda c^2}{3} \ , \tag{2.11}$$

where the dot stays for the time derivative, $\rho$ is the density, $p$ the pressure, $\Lambda$ the cosmological constant and $K$ the space curvature.

The first Friedmann equation can be arranged to make explicit the dependence on the dimensionless density cosmological parameters:

$$\Omega_m \equiv \frac{\rho_m}{\rho_{cr}}; \quad \Omega_\Lambda \equiv \frac{\rho_\Lambda}{\rho_{cr}}; \quad \Omega_r \equiv \frac{\rho_r}{\rho_{cr}} , \tag{2.12}$$

the curvature parameter is defined as $\Omega_k \equiv 1 - \Omega_m - \Omega_\Lambda - \Omega_r$ . Eq. 2.10 becomes:

$$H^2(t) = H_0^2[\Omega_m (1+z)^3 + \Omega_r (1+z)^4 + \Omega_k (1+z)^2 + \Omega_\Lambda ] . \tag{2.13}$$

In eq. 2.12 we have introduced the critical density $\rho_{cr}$ that is the total energy density for a flat Universe with $\Lambda = 0$:

$$\rho_{cr} \equiv \frac{3H_0^2}{8\pi G} = 2.775 \times 10^{11} \ h^{-1}\,\mathrm{M}_\odot/(\,h^{-1}\,\mathrm{Mpc})^3 . \tag{2.14}$$

In the next section we will describe one of the main observations at the base of the standard cosmological model: the *Cosmic Microwave Background* (CMB).

### 2.1.1 The cosmic microwave background

In 1965 Penzias and Wilson discovered what we call the cosmic microwave background.

The CMB is one of the most powerful tools we have to study the early Universe. From the Big Bang to shortly after $\sim$ 380'000 years the Universe was too hot and dense to let the photons travel freely: at that time the Universe was filled with an uniform interacting plasma of photons, electrons and baryons. The photons were absorbed and re-emitted many times and blackbody spectrum set up. Around z$\sim$1100 the Universe expanded and cooled down and photons started to travel freely. What we observe are the last-scattered photons that bring information on that blackbody spectrum. The three main all-sky satellites through to study the CMB are: the NASA experiment Cosmic Background Explorer (COBE, Mather et al., 1994), launched in 1989, the Wilkinson Microwave Anisotropy Probe (WMAP, Bennett et al., 2003) launched in 2001 and the ESA experiment Planck (The Planck Collaboration, 2006) launched in 2009.

COBE showed the perfect blackbody spectrum of the CMB with a temperature $T = 2.728 \pm 0.002$ K (fig. 2.4) and for the first time it has detected temperature anisotropy. WMAP had as main target the detection of the CMB acoustic oscillations on scales smaller then 7°, the angular resolution of COBE.

Planck's main target was to obtain the highest precision, highest resolution and the cleanest CMB maps setting the accuracy to the fundamental

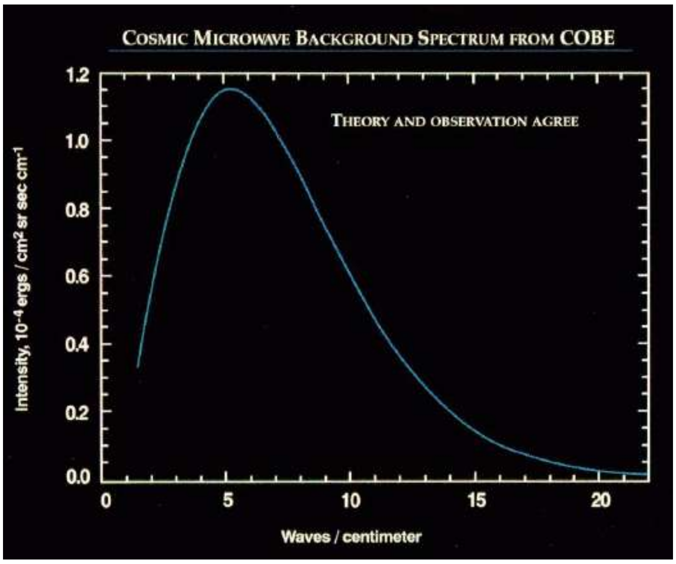


Figure 2.4: Comparison between the theory prediction for a blackbody spectrum and COBE (FIRAS) observations (Mather et al., 1994). The observation points are hidden by the theoretical curve and they have an error smaller then the thickness of the solid line.

astrophysical limit: a resolution 3 times better than that one of WMAP and sensitivity 5 times better.

This anisotropies are due to the initial perturbation in the density field. Observing the CMB we measure the microwave temperature of the sky. The temperature fluctuations are defined as:

$$\frac{\Delta T}{T}(\hat{\mathbf{n}}) \equiv \frac{T(\hat{\mathbf{n}}) - \bar{T}}{\bar{T}} \ , \tag{2.15}$$

where $\hat{\mathbf{n}} = (\theta, \varphi)$ is the position on the sky and $\bar{T}$ the average temperature. It is useful to expand the left- handside of eq. 2.15 in spherical harmonics:

$$\frac{\Delta T}{T}(\hat{\mathbf{n}}) = \sum_{l,m} a_{lm} Y_{lm}(\theta, \varphi) \ , \tag{2.16}$$

where $a_{lm}$ are the harmonic coefficients and $Y_{lm}$ are the Laplace spherical harmonics. As we will see later in section 2.2.1 for the density field, the CMB temperature field can be considered as one realization of a random field. In this way we can quantify the anisotropies introducing the power spectrum of the temperature fluctuations:

$$C_l \equiv \langle |a_{lm}|^2 \rangle^{1/2} \ , \tag{2.17}$$

where $\langle ... \rangle$ is an ensemble average.

Table 2.1: Parameters of the base ΛCDM cosmology computed from the 2015 baseline Planck likelihoods (Planck Collaboration et al., 2016b). $\Omega_b$ is the baryonic density parameter; $\Omega_c$ dark matter density parameter; $n_s$ the spectral index; $H_0$ the Hubble constant; $\Omega_m$ matter density parameter; $\sigma_8$ fluctuation amplitude at $8\,h^{-1}$ Mpc.

| Parameter | Planck [all spectra] |
|---|---|
| $\Omega_b\ h^2$ | 0.02225±0.00016 |
| $\Omega_c h^2$ | 0.1198±0.0015 |
| $n_s$ | 0.9645±0.0049 |
| $H_0$ | 67.27±0.66 |
| $\Omega_m$ | 0.3156±0.0091 |
| $\sigma_8$ | 0.831±0.013 |

In the left panel of figure 2.5 we show the temperature power spectrum fluctuation as a function of the multipole $l$. The amplitude of each peak is strictly connected to the cosmological parameters showed in table 2.1. Studying the CMB spectrum it is possible to put tight constraints on the cosmological model and on the initial conditions for structure formation. Looking at the power spectrum of temperature fluctuations obtained by Planck, figure 2.5, we can see that the ΛCDM model fits perfectly the observational points. The main results of these CMB dedicated surveys are the following:

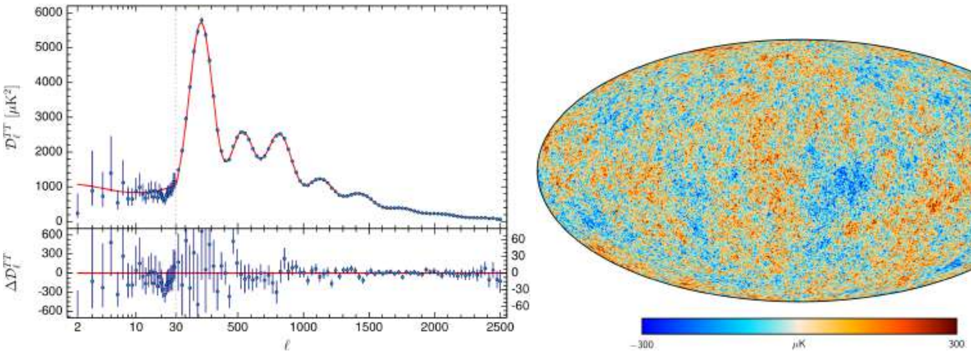


Figure 2.5: On the left the power spectrum of temperature fluctuation (Planck Collaboration et al., 2016c); on the right the CMB map (Planck Collaboration et al., 2016a).

- the Universe is remarkably isotropic on large scales;
- the small ($\sim 10^{-5}$) temperature anisotropies in the CMB spectrum give us information about the initial fluctuation that act as seed for the formation of structures we observe today;

- the CMB is a powerful tool to put constraints on the cosmological parameters: the heights and the positions of the detected peaks depend not only on the density of baryonic matter, but also on the total mean density of the Universe, Hubble constant and other cosmological parameters (table 2.1) .

The fact that the Universe is highly homogeneous and isotropic at large scales acts as a confirmation of the assumption at the base of the standard cosmological model and, at same time, the detection of the small temperature anisotropies are a consequence of the primordial density anisotropies that are responsible for the formation of structures we observe today.

### 2.1.2 The dark sector

The standard ΛCDM model is a successful way to describe the Universe, as we have showed in the previous sections. Even if this model can predict various observations, is based on a Universe composed with 95% of *dark components*. These two components, *dark matter* and *dark energy* are under deep investigation and their nature is still unclear.

Most of the matter in the Universe appears to be in some form which does not emit light and interacts only gravitationally. This disagreement between the mass inferred from observations and the predicted total matter has lead to the search for an additional dark matter component.

Different possibilities were proposed to explain the nature of this elusive component during the years. A historical review of the dark matter nature is not the purpose of these sections, but we will highlight the main observations which prove its existence.

The second dark component, dark energy, was introduced because observations indicate that the Universe is not only expanding, but its expansion is accelerated. The implementation of the acceleration can be achieved in two ways: one consists of modifying the Einstein theory of gravity, the other one is to assume the existence of some cosmological constant of some exotic fluid with negative pressure that is dominant over the other components in the Universe.

### 2.1.3 Dark Matter

The first dark matter evidence can be traced back to Zwicky that in 1933 computed the total mass needed by the Coma Cluster to be stable. He found that the mass was 400 times larger than the luminous mass in stars. This was the first hint that something was missing in the total matter count.

The idea of an additional unknown and undetectable component was shown to be as a real possibility in the 1970s. The papers of Ostriker et al. (1974) and Einasto et al. (1974) showed the necessity to have massive halos of invisible material that permeates and surrounds the Milky Way and nearby galaxies to justify the motion of their satellite galaxies. Using 21 cm and optical measurements of the rotation curves of spiral galaxies (Roberts and Rots, 1973; Rubin et al., 1978; Corbelli and Salucci, 2000) it became clear that without the additional dark matter component, the galaxies rotational velocity should fall-off at large radii. In figure 2.6 we show the rotational velocity of M33 galaxy in function of the distance from the center; the points are the measures by Corbelli and Salucci (2000) while the solid line is the best-fitting model. The other lines stay for all the contributions to the rotational velocity: the dark matter (dot-dashed line), the stellar disc (short-dashed line) and the gas contribution (long-dashed line); as we can see if we neglect the halo term, the luminous matter only cannot reproduce the observations.

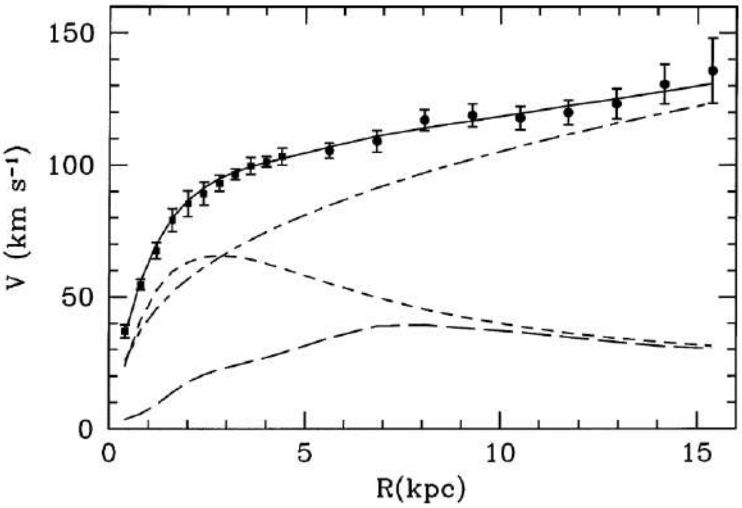


Figure 2.6: M33 rotation curve (points) compared with the best-fitting model (continuous line). Also shown is the dark matter contribution (dot-dashed line), the stellar disc (short-dashed line) and the gas contribution (long-dashed line) (Corbelli and Salucci, 2000).

Between 1980s and 2000s the problem of dark matter shifted from its existence to its nature. From the study on Big Bang Nucleosynthesis (BBN) we know that the observed quantity of baryons cannot explain alone all the matter in the Universe. From Planck we know that $\Omega_m = 0.315$ and only a small fraction is due to baryons, $\Omega_b = 0.045$. Through the observations of galaxy clustering, we know that structures tend to form *hierarchically*: smaller structures form prior to larger ones. Assuming the existence of a a *Cold Dark Matter* (CDM) component makes possible the structures we observe, to form: baryons can fall in the potential wells given by the presence of dark matter.

Moreover, as we have already mentioned, the position and the high of the peaks in the CMB spectrum depends on the total matter density in the Universe. The presence of a CDM component guarantees that the CMB fluctuations are suppressed, as the model predicts, but assures the formation of structures, as explained later. A fluid that allows for these properties is supposed to be non-relativistic at the epoch of decoupling or, alternatively, it should never be coupled with the other components. When baryons decouple from radiation they can fall into the dark matter potential wells. The presence of these potential wells allows the large-scale structures formation we observe today. In figure 2.7 we show a slice of the Universe at different epochs obtained with an N-body simulation using two different cosmological models. It is possible to reproduce the distribution of the galaxy clusters (yellow circles) only assuming the presence of the dark matter component (upper three panels). A universe dominated by baryonic matter only (lower three panels), is characterized by a low level of clustering that is not in agreement with observations. It is worth to summarize the main problems

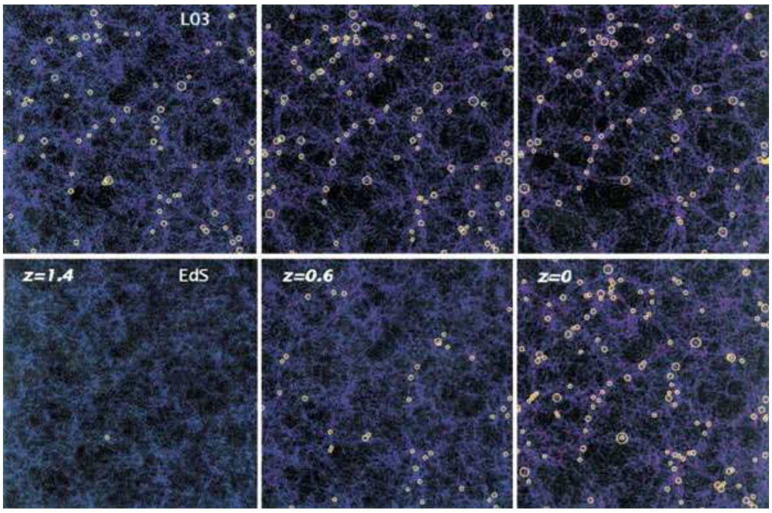


Figure 2.7: Comparison between two N-body simulations (see section 4) for two different cosmological model: the upper panels describe a flat low-density model with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$ and the lower panels show an Einstein-de-Sitter model (EdS) with $\Omega_m = 1$. The yellow circles are the position of galaxy clusters as they would be seen in X-rays observation with a temperature $T > 3$ keV (Borgani and Guzzo, 2001).

solved by the introduction of a non-baryonic weakly interactive dark matter component:

- it explains the value of the matter density parameter and makes it compatible with the value of the baryonic density parameter;
- it explains the small fluctuations observed in the CMB;

- it solves the rotational curve problem of spiral galaxies;
- the observed large-scale structure is possible only thanks to the dark matter potential wells in which the baryon can fall after they decouple from the radiation;
- it explains the acoustic oscillation observed in the galaxy clustering spectrum (see section 3.2.1).

### 2.1.4 Dark Energy

In 1998 the studies of Type Ia supernovae of two independent papers (Riess et al., 1998; Perlmutter et al., 1999) showed that the expansion of the Universe is accelerated. Within the Einstein general relativity theory, this means the existence of an additional component with negative pressure. The nature of this new fluid is not yet clear and we call it *Dark Energy.* If we consider valid the Einstein theory, we can explain the dark energy with a *cosmological constant.* An other possibility is that dark energy is a scalar field with its potential energy larger than its kinetic energy; a model like this is called *quintessence.* Furthermore, modification of the theory of gravity can explain the observational effects we confer to dark energy.

Even if the true nature of dark energy is still under investigation, we have many probes that are in agreement with a Universe dominated by a cosmological constant. After the first discover using Type Ia supernovae, the measurements of the acoustic peaks it the CMB spectrum have allowed to put tight constraints on dark energy: it gave support to the theoretical model of a spatially flat Universe, e.g. $\Omega_m + \Omega_{DE} = 1$. In figure 2.8 we show how the observations of the Type Ia supernovae have improved in terms of observed redshift; in particular the figure shows sixty type Ia supernova from low redshift to high redshift, comparing their distribution with different cosmological model with and without cosmological constant.

Another important probe that we will discuss in section (3.2.1), comes from the observations of the BAO in the galaxy two-point correlation function that encloses information on the expansion history, therefore on the energy content of the Universe.

## 2.2 Theory of structure formation

The Universe today appears populated by large scale structures. The standard cosmological model predicts these structures to be grown, due to gravity, from small initial fluctuations. The same fluctuations can be also observed in the CMB as we have described in section (2.1.1). In this section we will describe the properties of the cosmic density field and the equations that regulate its evolution. In the second part of this section we will focus

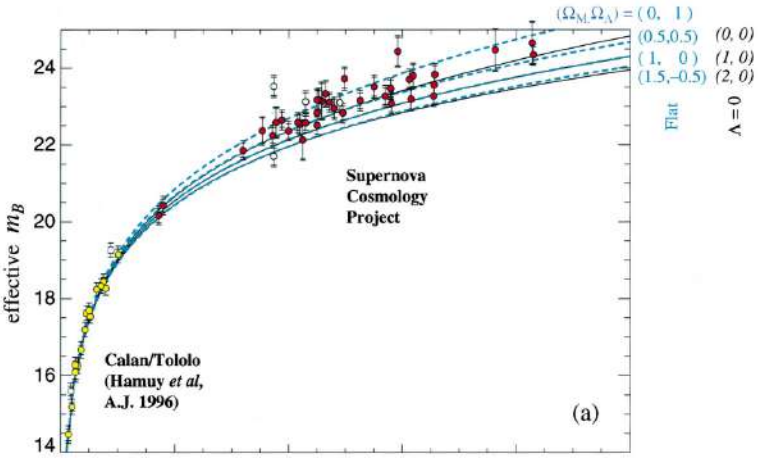


Figure 2.8: 42 high-redshift type Ia supernovae from the Supernova Cosmology Project and 18 low-redshift type Ia supernovae from the Supernova Survey, plotted on a linear redshift scale to display details at high redshift. The solid curves are the theoretical for a range of cosmological models with zero cosmological constant: on top, (1, 0) in middle, and (2, 0) on bottom. The dashed curves are for a range of cosmological models: ($\Omega_m$, $\Omega_\Lambda$)=(0, 1) top, (0.5, 0.5) second from top, (1, 0) third from top, and (1.5, -0.5) on bottom.

on perturbation theory introducing quantity we will use in the rest of the thesis. In all the section we consider the post-recombination Universe and the *Newtonian limit*: the structures are smaller than the horizon size so that the relativistic effects are negligible.

### 2.2.1 Cosmic density field as stochastic field

For the standard cosmological model the Universe is filled with ideal fluids characterized by density $\rho$, pressure $p$ and velocity $\mathbf{v}$. We know that the dark matter component is dominant over matter and radiation, so we will describe its evolution under the action of the gravitational field with potential $\Phi$, considering the contribution of the other components to be negligible.

The density of this fluid can be expressed as a mean density times a small variation given by the *density contrast* $\delta(\mathbf{x}, t)$:

$$\rho(\mathbf{x}, t) = \bar{\rho}(t)(1 + \delta(\mathbf{x}, t)) \ , \tag{2.18}$$

where $\mathbf{x}$ are comoving coordinates ($\mathbf{r} = a(t)\mathbf{x}$ with $\mathbf{r}$ the proper coordinates) and $\mathbf{v}$ is the velocity of the fluid element with respect to the comoving observer at $\mathbf{x}$. The general definition for the density contrast comes directly from eq. 2.18:

$$\delta(\mathbf{x},t) = \frac{\rho(\mathbf{x},t)-\bar{\rho}}{\bar{\rho}} \ . \tag{2.19}$$

We assume that the $\delta(\mathbf{x},t)$ is a *stochastic* field with a probability distribution function given by:

$$P(\delta_1,\delta_2,...,\delta_n)\ d\delta_1 d\delta_2 ... d\delta_n \ ; \tag{2.20}$$

the moments of this distribution are:

$$n = 1 \quad : \quad \langle \delta \rangle = 0 \tag{2.21}$$

$$n = 2 \quad : \quad \langle \delta_1 \delta_2 \rangle = \xi(r) \tag{2.22}$$

$$n > 2 \quad : \quad \langle \delta_1 \delta_2 ... \delta_n \rangle = \int \delta_1^{l_1} \delta_2^{l_2} ... \delta_n^{l_n} P(\delta_1,\delta_2,...,\delta_n)\ d\delta_1 d\delta_2 ... d\delta_n \tag{2.23}$$

We define $\langle ... \rangle$ as *ensemble average*, that is the mean over all realizations of a certain event. The problem, in cosmology, is that we study the Universe that has one unique realization, the Universe that we observe. The *Ergodic* assumption helps to overcome this issue: the **ensemble** average is equivalent to **spatial** average over one realization of a random field if the volume is large enough:

$$\langle f(\mathbf{x}) \rangle \equiv \frac{1}{V} \int_V d^3\mathbf{x}\ f(\mathbf{x}) \ , \tag{2.24}$$

where $f(\mathbf{x})$ is a generic function and V is a generic volume.

From eq. 2.23, the first moment comes directly from the definition of $\delta$; the second moment is the *two-point correlation function* defined as:

$$\xi(x) \equiv \xi(|\mathbf{x}_1 - \mathbf{x}_2|) = \langle \delta_1(\mathbf{x}_1) \delta_2(\mathbf{x}_2) \rangle \ , \tag{2.25}$$

with $x = |\mathbf{x}_1 - \mathbf{x}_2|$. The spatial isotropy and homogeneity assumed by the standard cosmological model on large scale assure that the two-point correlation function does not depend on the particular direction. We can think at the two-point function as the excess probability to find a pair of galaxies separated by a distance $|\mathbf{x}_1 - \mathbf{x}_2|$ with respect to a random distribution.

If $\mathbf{x}_1 = \mathbf{x}_2$, the two-point correlation function reduces to the variance of the density field:

$$\xi(0) = \langle \delta^2(\mathbf{x}) \rangle = \sigma^2. \tag{2.26}$$

All the relations we have described above are valid in configuration space.

In same cases it is useful look to a density field as the superposition of many **modes**. This can be achieved using a Fourier analysis. Assuming that the density field is periodic within some box of side L, then we can write $\delta$ as the sum of all the modes inside the box:

$$\delta(\mathbf{x}) = \sum_k \delta_{\mathbf{k}} \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}} \ ; \tag{2.27}$$

if we let the box become arbitrarily large, then the sum will go over to an integral:

$$\delta(\mathbf{x}) = (2\pi)^3 \int d^3\mathbf{k}\ \delta_{\mathbf{k}}\ \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}}\ . \tag{2.28}$$

The Fourier representation of $\delta(\mathbf{x})$ is obtained inverting eq. 2.28:

$$\delta_{\mathbf{k}} = \int \frac{d^3\mathbf{k}}{(2\pi)^3}\delta(\mathbf{x})\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{x}}\ . \tag{2.29}$$

If $\delta(\mathbf{x})$ is a random field, $\delta_{\mathbf{k}}$ is a random field as well. In general $\delta_k$ is a complex quantity, but we work with real objects, so the fallowing relation is valid:

$$\delta_{\mathbf{k}} = \delta^*_{-\mathbf{k}}\ , \tag{2.30}$$

where $\delta^*$ is the complex conjugate of $\delta$. As we have done in configuration space, we can define the moments of this distribution; the first moment is zero as in configuration space: $\langle\delta_{\mathbf{k}}\rangle = 0$. The second moment, the two-point function in Fourier space is given by:

$$\langle\delta_{\mathbf{k}_1}\delta_{\mathbf{k}_2}\rangle = \delta_D^{(3)}(\mathbf{k}_1 + \mathbf{k}_2)P(k)\ , \tag{2.31}$$

where $\delta_D^{(3)}(\mathbf{k})$ is the 3-dimensional delta of Dirac, defined as

$$\delta_D^{(3)}(\mathbf{k}) = \int \frac{d^3\mathbf{r}}{(2\pi)^3}e^{-i\mathbf{k}\cdot\mathbf{r}}\ . \tag{2.32}$$

The $P(k)$ function we introduce is called *Power Spectrum* and it can be demonstrated to be the Fourier transform of the two-point correlation function:

$$P(k) = \int \frac{d\mathbf{r}^3}{(2\pi)^3}\ \xi(r)e^{-i\mathbf{k}\cdot\mathbf{r}}\ . \tag{2.33}$$

It is important to stress that the two-point correlation function is characterized by the so called *integral constraint*:

$$\int_0^\infty dr\ r^2\xi(r) = 0\ . \tag{2.34}$$

If $\xi(0) = \langle\delta^2(x)\rangle = \sigma^2 > 0$ for the integral constraint this means that the two-point correlation function must pass through zero at same (large) separation $r$.

### 2.2.2 Collisionless fluid

As we have already specified, according to the ΛCDM model the usual baryonic matter is only part of the total matter component of the Universe. The other part is given by the dark matter, a collisionless non-baryonic matter, as we have seen in section 2.1.2. In order to study this kind of non-baryonic fluid we need to use the collisionless Boltzmann's equation:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \dot{\mathbf{x}}\frac{\partial f}{\partial \mathbf{x}} + \dot{\mathbf{v}}\frac{\partial f}{\partial \mathbf{v}} = 0 \ , \tag{2.35}$$

where $f \equiv f(\mathbf{x}, \mathbf{v})$ is the distribution function in the 6D phase-space $(\mathbf{x}, \mathbf{v})$ and $\dot{\mathbf{x}} = -\nabla\Phi$. Eq. 2.35 states the conservation of the distribution of particles in the phase space.

The collisionless Boltzmann equation can also be written in the form of Vlasov equation:

$$\frac{\partial f}{\partial t} + \frac{1}{ma^2}\mathbf{p}\cdot\nabla f - m\nabla\Phi\frac{\partial f}{\partial p} = 0 \ , \tag{2.36}$$

where $\mathbf{p} = ma\mathbf{v}$ with m the mass of the particle. In principle we can have an infinite number of equations corresponding to the velocity moments of $f$. Starting from eq. 2.36 and considering the zeroth order we obtain the mass conservation equation (continuity equation):

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\sum_j \frac{\partial}{\partial x_j}[(1+\delta)\langle v_j\rangle] = 0 \ , \tag{2.37}$$

where the velocity $\mathbf{v}$ is replaced by the mean streaming velocity $\langle\mathbf{v}\rangle$; from the first moment we can derive the Euler equation:

$$\frac{\partial\langle v_i\rangle}{\partial t} + \frac{\dot{a}}{a}\langle v_i\rangle + \frac{1}{a}\sum_j\langle v_j\rangle\frac{\partial\langle v_i\rangle}{\partial x_j} = -\frac{1}{a}\frac{\partial\Phi}{\partial x_i} - \frac{1}{a(1+\delta)}\sum_j\frac{\partial}{\partial x_j}[(1+\delta)\sigma_{ij}^2] \ , \tag{2.38}$$

where $\sigma_{ij}^2 \equiv \langle v_i v_j\rangle - \langle v_i\rangle\langle v_j\rangle$ is the stress tensor. We can continue with the higher moments and obtain the dynamical equations for $\langle v_i v_j\rangle$, $\langle v_i v_j v_l\rangle$ and so forth. In this sense the dynamics is given by an infinite numbers of equations of the velocity moments. We can do some justified assumption to truncate this infinite series of equations: assuming the stress tensor to be small ($\langle v\rangle \ll 1$) we can obtain the equation describing the evolution of perturbations for a pressurless fluid we describe in the next section. For realistic fluid the hypothesis of small stress tensor is not valid and one has to proceed solving the set of Vlasov equations using perturbation theory.

### 2.2.3 Linear solution for a pressureless fluid

We can consider a pressureless fluid in the regime where the perturbations evolve linearly ($\delta \ll 1$) and the displacement are small. In this case we have to work with the linearized equations of motion of the fluid:

$$\frac{\partial\delta}{\partial t} - \nabla\cdot\mathbf{v} = 0 \ , \tag{2.39}$$

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\dot{a}}{a} = -\frac{\nabla \Phi}{a} \ . \tag{2.40}$$

Using these two equations and the Poisson equation

$$\nabla^2 \Phi = 4\pi G \bar{\rho} a^2 \delta \ , \tag{2.41}$$

we can derive the linear evolution of perturbations. In Fourier space we have:

$$\frac{\partial^2 \delta}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial \delta}{\partial t} = 4\pi G \bar{\rho} \delta \ , \tag{2.42}$$

where $\bar{\rho}$ is the mean density of the fluid and $\delta_k$ is the density field in Fourier space defined as:

$$\delta_{\mathbf{k}} = \int \frac{d^3\mathbf{x}}{(2\pi)^3} \ \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} \ . \tag{2.43}$$

Because of the linearization of the equations of motion, we lose some degree of freedom; in particular we lose the vector component of the velocity, the *vorticity*. It is possible to show that the Euler equation 2.40 contains the extra degree of freedom due to the vorticity component. It is possible to neglect it in linear theory because it scales like $a^{-1}$, but it becomes important in then non-linear regime.

The solution of eq. 2.42 has the following form:

$$\delta(x,t) = D(t)\delta_i(x) \ , \tag{2.44}$$

where we define the *linear growth factor* $D(t)$. Substituting eq. 2.44 in eq. 2.42 we can find the two solutions for the growth rate, corresponding to the growing and decaying modes, that we call respectively $D_+$ and $D_-$:

$$D_- \ \propto \ H(z) \tag{2.45}$$

$$D_+ \ \propto \ H(z) \int_z^\infty dz' \ \frac{1+z'}{E^3(z)} \ , \tag{2.46}$$

where $H(t)$ is the Hubble function defined in eq. 2.2. The $E(z)$ function contains the information about cosmology:

$$E(z) = \frac{H(z)}{H_0} = [\Omega_r(1+z)^4 + \Omega_m \ (1+z)^3 + \Omega_k \ (1+z)^2 + \Omega_\Lambda \ ]^{1/2} \ . \tag{2.47}$$

For closed universe with cosmological constant, the solution is given by:

$$D_+ \sim \frac{5}{2} H_0^2 \, \Omega_m \ {}_0 H(a) \int_0^a \frac{da'}{(a'H(a'))^3} \ . \tag{2.48}$$

For a flat Universe ( $\Omega_m \ = 1$) without cosmological constant in the matter era the solutions reduce to:

$$D_- \ \propto \ a \tag{2.49}$$

$$D_+ \ \propto \ a^{-3/2} \ . \tag{2.50}$$

Using the Poisson equation we can study also the evolution of the potential perturbation:

$$\Phi_{\mathbf{k}} \propto D(a)/a \; ; \tag{2.51}$$

For a flat Universe without cosmological constant $D(a) \propto a$ therefore the potential does not evolve; for an open Universe the potential decays because the linear growth rate is suppressed.

### 2.2.4 Non-linear evolution and higher order correlators

In section 2.2 we have showed the evolution of perturbations under the assumption that the density field $\delta$ is small; however perturbations grow in time and when the density field becomes larger then 1 we enter in the non-linear regime. The structures we observe are highly non-linear so the study of this regime cannot be neglected as long as we aim to a precise description of structures evolution.

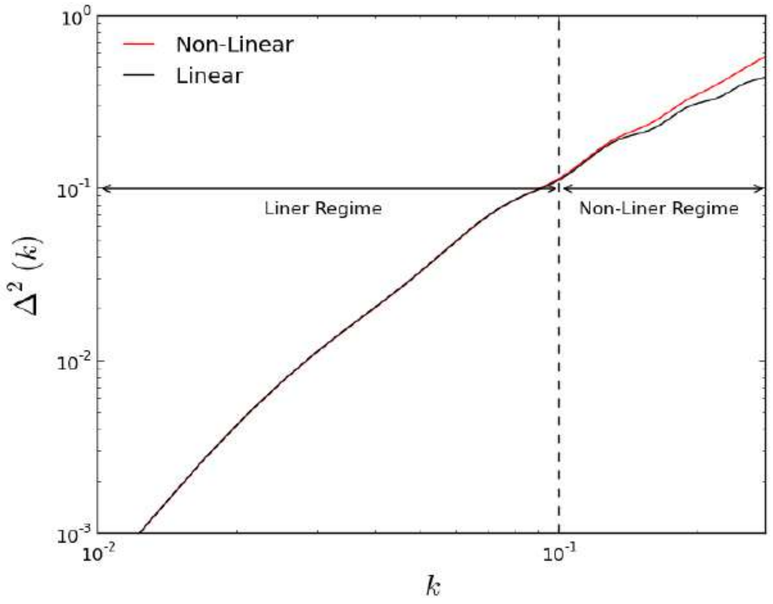


Figure 2.9: Adimensional power spectrum $\Delta(k) = 4\pi k^3 P(k)$. In black the linear power spectrum, in red the non-linear one; the comparison between the two allows to determine a linear regime up to $k \sim 0.1\,h^{-1}\,$Mpc and a non-linear regime for larger wavenumbers (small scales).

In the previous section we have defined the two-point correlation function both in configuration and in the Fourier space, so we can describe the non-linear corrections to these correlators together with the higher order correlators starting from the non-linear solution for the evolution of the density field.

In the Eulerian framework, we can write the expansion of the density field as:

$$\delta_{\mathbf{k}}^{NL} = \delta_{\mathbf{k}}^{(1)} + \delta_{\mathbf{k}}^{(2)} + \delta_{\mathbf{k}}^{(3)} + ... = \delta_{\mathbf{k}}^{L} + \delta_{\mathbf{k}}^{(2)} + \delta_{\mathbf{k}}^{(3)} + ..., \tag{2.52}$$

where the first term is the linear field and all of the other contributions are higher-order corrections. For the quantities we will use in the following chapters of this thesis we are interested in the second order and third order corrections:

$$\delta_{\mathbf{k}}^{(2)} = \int d^3\mathbf{k}_1 d^3\mathbf{k}_2 \; \delta_D(\mathbf{k} - \mathbf{k}_{12}) F_2(\mathbf{k}_1, \mathbf{k}_2) \delta_{\mathbf{k}_1}^{L} \delta_{\mathbf{k}_1}^{L} \tag{2.53}$$

$$\delta_{\mathbf{k}}^{(3)} = \int d^3\mathbf{k}_1 d^3\mathbf{k}_2 d^3\mathbf{k}_3 \; \delta_D(\mathbf{k} - \mathbf{k}_{123}) F_3(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \delta_{\mathbf{k}_1}^{L} \delta_{\mathbf{k}_1}^{L} \delta_{\mathbf{k}_3}^{L} \;, \tag{2.54}$$

where $\delta_D$ is the Dirac delta and $F_2(\mathbf{k}_1, \mathbf{k}_2)$ and $F_3(\mathbf{k}_1, \mathbf{k}_2)$ are the two symmetric kernels, $\mathbf{k}_{12} = \mathbf{k}_1 - \mathbf{k}_2$ and $\mathbf{k}_{123} = \mathbf{k}_1 - \mathbf{k}_2 - \mathbf{k}_3$. The second order kernel is given by:

$$F_2(\mathbf{k}_1, \mathbf{k}_2) = \frac{5}{7} + \frac{1}{2} \frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{k_1 k_2} \Big( \frac{k_1}{k_2} + \frac{k_2}{k_1} \Big) + \frac{2}{7} \Big( \frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{k_1 k_2} \Big)^2 \;, \tag{2.55}$$

while the third order kernel is given by a summation of the non-symmetric kernel $F_3$

$$\begin{aligned} F_3(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = \; & F_2(\mathbf{k}_1, \mathbf{k}_2) \; \Big[ \frac{1}{3} + \frac{1}{3} \frac{\mathbf{k}_1 \cdot (\mathbf{k}_2 + \mathbf{k}_3)}{(\mathbf{k}_2 + \mathbf{k}_3)^2} + \frac{4}{9} \frac{\mathbf{k} \cdot \mathbf{k}_1}{k_1^2} \frac{\mathbf{k} \cdot (\mathbf{k}_2 - \mathbf{k}_1)}{(\mathbf{k}_2 + \mathbf{k}_3)^2} \Big] \\ & - \frac{2}{9} \frac{\mathbf{k} \cdot \mathbf{k}_1}{k_1^2} \frac{\mathbf{k} \cdot (\mathbf{k}_2 + \mathbf{k}_3)}{(\mathbf{k}_2 + \mathbf{k}_3)^2} \frac{\mathbf{k}_3 \cdot (\mathbf{k}_2 + \mathbf{k}_3)}{k_3^2} + \frac{1}{9} \frac{\mathbf{k} \cdot \mathbf{k}_2}{k_2^2} \frac{\mathbf{k} \cdot \mathbf{k}_3}{k_3^2} \;. \end{aligned} \tag{2.56}$$

with all possible permutations of variables. The general expression for the kernels we used here are derived in Goroff et al. (1986) and Jain and Bertschinger (1994).

If we want to obtain the first non-linear contribution to the power spectrum we have to consider the density field up to the third order in eq. 2.54. Considering Gaussian initial conditions we derive the *1-loop* power spectrum:

$$P_{\text{non-linear}}(k) = P_L(k) + P_{22}(k) + P_{13}(k) \;, \tag{2.57}$$

where $P_{22}$ and $P_{13}$ are given by:

$$P_{22}(k) = 2 \int d^3\mathbf{q} \; F_2^2(\mathbf{k}_{-}\mathbf{q}, \mathbf{q}) P_L(|\mathbf{k}_{-}\mathbf{q}|) P_L(\mathbf{q}) \tag{2.58}$$

$$P_{13}(k) = 6 P_L(k) \int d^3\mathbf{q} F_3(\mathbf{k}, \mathbf{q}, \mathbf{k} - \mathbf{q}) P_L(\mathbf{q}) \;. \tag{2.59}$$

We use the name *1-loop* in analogy with particle physics: 1-loop quantities are characterized by one integral.

Moving to higher-order statistics, we can define the three-point function in Fourier space, called *bispectrum*, defined as:

$$\langle \delta_{\mathbf{k}_1} \delta_{\mathbf{k}_2} \delta_{\mathbf{k}_3} \rangle = \delta_D^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B(k_1, k_2, k_3) \tag{2.60}$$

where $\delta_D^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)$ ensures that the three wavenumber must form a closed triangle $\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 = 0$. Using the second order corrections to the density field and assuming Gaussian initial fluctuations, we can write the *tree-level* bispectrum (Verde et al., 1998) as follow:

$$B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = 2F_2(\mathbf{k}_1, \mathbf{k}_2) P_L(k_1) P_L(k_2) + \text{perm} \ , \tag{2.61}$$

where $F_2$ is defined in eq. 2.55. Again we use the term tree-level in analogy with particle physics to designate quantities that do not involve integrals.

For the analyses we will show in the rest of the thesis it is important to give the expression for the the four-point correlation function, the *trispectrum*:

$$\langle \delta_{\mathbf{k}_1} \delta_{\mathbf{k}_2} \delta_{\mathbf{k}_3} \delta_{\mathbf{k}_4} \rangle = \delta_D^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 + \mathbf{k}_4) T(k_1, k_2, k_3, k_4) \ . \tag{2.62}$$

As we have already done for the bispectrum, we can equally write the tree-level trispectrum using the third order corrections to the density field:

$$\begin{aligned} T(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3, \mathbf{k}_4) &= 4F_2(\mathbf{k}_{13}, -\mathbf{k}_2) F_2(\mathbf{k}_{13}, -\mathbf{k}_1) F_2(\mathbf{k}_{13}, -\mathbf{k}_2) P_L(k_1) P_L(k_2) P_L(k_{13}) + 11 \text{ perm.} \\ &+ 6F_3(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) P_L(k_1) P_L(k_2) P_L(k_{13}) + 3 \text{ perm.} \ , \end{aligned} \tag{2.63}$$

where $F_3$ is the symmetric third order kernel.

Here we describe only the tree-level case, because these are the objects that we use in the following chapters.

### 2.2.5 Lagrangian perturbation theory

In the previous section we considered the evolution of perturbations when the density field is small ($\delta \ll 1$). This approach is called *Eulerian Perturbation Theory* (EPT). A complementary approach is *Lagrangian Perturbation Theory* (LPT); in this approach the equations describe the evolution of a fluid element along its trajectory. The benefit of this approach is that it is possible to study the effects that would be non-linear in the Eulerian framework, but still linear in the Lagrangian one.

If the position of the fluid element is $\mathbf{q}$ at time $t = 0$, at time $t_1 > 0$ it will be

$$\mathbf{x}(t) = \mathbf{q} + \psi(\mathbf{q}, t) \ , \tag{2.64}$$

where $\psi$ is the displacement from the initial position. The number of particles enclosed in the volume of the fluid element is conserved:

$$\bar{\rho} \ d^3\mathbf{q} = \bar{\rho}(1 + \delta(\mathbf{x})) \ d^3\mathbf{x} \ . \tag{2.65}$$

From this relation we can see that:

$$1 + \delta(\mathbf{x}) = \left[\det\left|\frac{\partial x_i}{\partial q_j}\right|\right]^{-1} . \tag{2.66}$$

Using the Poisson equation and the total conservation of the number of particles we can derive the evolution equation for the displacement $\boldsymbol{\psi}$:

$$\frac{\partial^2}{\partial t^2}(\nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}) + 2H\frac{\partial}{\partial t}(\nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}) = -4\pi G\bar{\rho}\ \frac{1 - J(\mathbf{q})}{J(\mathbf{q})}\ , \tag{2.67}$$

where

$$J(\mathbf{q}) = 1 + \nabla_{\mathbf{q}} \cdot \boldsymbol{\psi}(\mathbf{q}) + \left[\frac{1}{2}(\nabla_{\mathbf{q}} \cdot \boldsymbol{\psi})^2 + \sum_{i,j} \psi_{i,j}\psi_{i,j}\right] + \mathbf{O}(\psi^3) \tag{2.68}$$

is the Jacobian defined in the righthand side of eq. 2.66.

### 2.2.6 Zeldovich approximation

If we are interested in the linear regime solution, eq. 2.67 becomes:

$$\frac{\partial^2}{\partial t^2}(\nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}) + 2H\frac{\partial}{\partial t}(\nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}) = -4\pi G\bar{\rho}\nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}\ , \tag{2.69}$$

At the first order we do not need to solve eq. 2.69. From the conservation of matter we know that:

$$1 + \delta = \frac{1}{J}\ ; \tag{2.70}$$

For small displacements, at first order $J \simeq 1 - \nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}$, so that we can write a relation between the Lagrangian density contrast and the displacement $\psi$:

$$1 + \delta \simeq 1 - \nabla_{\mathbf{x}} \cdot \boldsymbol{\psi}^{(1)}\ , \tag{2.71}$$

where $\boldsymbol{\psi}^{(1)}$ is the displacement at the first perturbative order.

Equation 2.71 goes under the name of *Zeldovich Approximation* (ZA) for the displacement $\psi$). We know that the density field evolves with time through the linear growth rate, $\delta(t) = D(t)\delta_i$, so the displacement is linked with $D(t)$:

$$\nabla_{\mathbf{q}} \cdot \boldsymbol{\psi}^{(1)} = -D(t)\delta_{\mathbf{q}}\ . \tag{2.72}$$

The ZA describes the trajectory of a fluid element as a straight line, with the distance traveled proportional to $D(t)$, eq. 2.72. We call *shell crossing* when two fluid elements cross their trajectories; when this happens the density diverges.

This can be used to explain the structures gravitational collapse we observe in the Universe. The diagonalizable matrix $T_{ij} = \frac{\partial \Psi_i}{\partial q_j}$ has three eigenvalues: $D\lambda_1$, $D\lambda_2$ and $D\lambda_3$; the density field is:

$$\rho(\mathbf{x}, t) = \rho_0(t)\frac{1}{(1 - D(t)\lambda_1(\mathbf{q}))(1 - D(t)\lambda_2(\mathbf{q}))(1 - D(t)\lambda_3(\mathbf{q}))}\ . \tag{2.73}$$

At the time of shell crossing $D(t) = \lambda^{-1}(\mathbf{q})$ and $\rho \to \infty$. Studying the values of the eigenvalues at the shell crossing it is possible to select the type of collapsed structure:

- $\lambda_1 > \lambda_2 > \lambda_3 > 0$: the collapse happens in one direction given by the eigenvector corresponding to $\lambda_1$; the collapsed structure is two-dimensional. We call it *Pancake*;
- $\lambda_1 = \lambda_2 > \lambda_3$: the collapse happens in two directions given by the eigenvectors of $\lambda_1$ and $\lambda_2$, the collapsed structure is one-dimensional and in this case we call it *Filament*;
- $\lambda_1 = \lambda_2 = \lambda_3$: the collapse happens in three directions; the structure is zero-dimensional and we call it *Knot*.

### 2.2.7 Higher order Lagrangian pertubation theory

The Zeldovich approximation we have described in the previous section is valid for planar geometry. To improve the prediction obtained with the ZA we have to consider higher-order corrections to eq. 2.72. The second order Lagrangian Perturbation Theory (2LPT) can provide remarkable improvements with respect ZA in describing the properties of the density and velocity field. The correction to the ZA displacement is given by:

$$\nabla_{\mathbf{q}} \cdot \mathbf{\Psi}^{(2)} = \frac{1}{2} D_2(t) \sum_{i \neq j} (\Psi^{(1)}_{i,i} \Psi^{(1)}_{j,j} - \Psi^{(1)}_{i,j} \Psi^{(1)}_{j,i}) \ , \tag{2.74}$$

where $\Psi_{i,j} \equiv \partial \Psi_i / \partial \mathbf{q}_j$, $D_2(t)$ is the second order growth factor and $\Psi^{(1)}$ is the ZA displacement. It is convenient to define the Lagrangian potential $\phi^{(1)}$ and $\phi^{(2)}$ so that the position of the fluid element reads:

$$\mathbf{x}(\mathbf{q}) = \mathbf{q} - D_1(t) \nabla_{\mathbf{q}} \phi^{(1)}(\mathbf{q}) + D_2(t) \nabla_{\mathbf{q}} \phi^{(2)}(\mathbf{q}) \ ; \tag{2.75}$$

$\phi^{(1)}$ is associated with the first perturbative order, ZA, while $\phi^{(2)}$ is the first correction to the ZA. 2LPT requires the determination of the potentials, $\phi^{(1)}$ and $\phi^{(2)}$, this means we have to solve two Poisson equations:

$$\begin{aligned} \nabla^2_{\mathbf{q}} \phi^{(1)}(\mathbf{q}) &= \delta_{\mathbf{q}} & (2.76) \\ \nabla^2_{\mathbf{q}} \phi^{(2)}(\mathbf{q}) &= \sum_{i>j} [\phi^{(1)}_{,ii}(\mathbf{q}) \phi^{(1)}_{,jj}(\mathbf{q}) - (\phi^{(1)}_{,ij}(\mathbf{q}))^2] \ , & (2.77) \end{aligned}$$

where $\phi_{,ii} \equiv \partial^2 \phi / \partial \mathbf{q}_i \partial \mathbf{q}_j$.

The main reason why 2LPT works better than ZA is that 2LPT describes the correction to the ZA displacement due to gravitational field effects; 2LPT is only the first correction to the firs order ZA. It is possible to go to the third order in the displacement field $\Psi$ (3LPT); 3LPT can improve the agreement with exact N-body simulations in same cases, but it becomes more costly due to the increasing number of Poisson equations needed to solve. More details on the 3LPT calculation are given in appendix A of Catelan (1995), Monaco (1997), while details on the comparison of 2LPT and 3LPT are given in Munari et al. (2017).

### 2.2.8 From LPT to EPT

In the previous section we derived the relation between the displacement and the density field under the Zeldovich approximation, 2LPT and 3LPT. Assuming $\mathbf{\Psi}$ to be small at the scales of interest, we can derive the expression for the density field in Zeldovich approximation in Fourier space:

$$\delta^{ZA}_{\mathbf{k}} = D(t)\delta^{L}_{\mathbf{k}} + \frac{1}{2}\int d^3\mathbf{k}_1 d^3\mathbf{k}_2\ \delta_D(\mathbf{k}-\mathbf{k}_{12}) F_2^{ZA}(\mathbf{k}_1,\mathbf{k}_2)\delta^L_{\mathbf{q}_1}\delta^L_{\mathbf{q}_2}\ , \qquad (2.78)$$

where $F_2^{ZA}(\mathbf{k}_1,\mathbf{k}_2)$ is defined as

$$F_2^{ZA}(\mathbf{k}_1,\mathbf{k}_2) = 1 + \frac{\mathbf{k}_1\cdot\mathbf{k}_2}{k_1 k_2}\Big(\frac{k_1}{k_2}+\frac{k_2}{k_1}\Big) + \Big(\frac{\mathbf{k}_1\cdot\mathbf{k}_2}{k_1 k_2}\Big)^2 \qquad (2.79)$$

Eq. 2.78 is obtained thanks to the fact that for small displacements we can write:

$$\Psi^{(1)} = -iD(t)\frac{\mathbf{k}}{k^2}\delta_{\mathbf{k}}\ . \qquad (2.80)$$

In figure 2.10 we show a schematic representation of the information we can derive moving from LPT to EPT. The solid black arrows means that from the Zeldovich approximation in EPT, we can derive the full linear theory in EPT and in the same way, from the 2LPT displacement analysis we can derive the full linear theory and a full prediction for the non-linear kernel $F_2$, eq. 2.55. At the same time, with the dashed black arrow, we show that from the ZA we can derive a an approximated prediction for the non-linear kernel

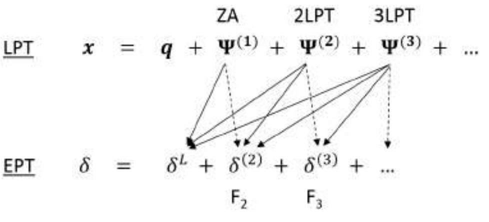

ZA
2LPT
3LPT
LPT
$x = q + \Psi^{(1)} + \Psi^{(2)} + \Psi^{(3)} + ...$
EPT
$\delta = \delta^L + \delta^{(2)} + \delta^{(3)} + ...$
$F_2$
$F_3$

Figure 2.10: Schematic representation of the difference between LPT and EPT. The first line represents the evolution of the fluid element described by its position, $\mathbf{q}$, and its displacements from the initial position $\mathbf{\Psi}$, while the second line the evolution of perturbation described by the density field $\delta$ and its non-linear correction. The arrows mean that we can move from LPT to EPT and we can recover the information obtained with an EPT analysis.

$F_2$, eq. 2.79, and from the 2LPT we can obtain an approximated prediction for the non-linear kernel $F_3$. We can proceed in this way looking at higher order corrections and moving from LPT to EPT.

## 2.3 Galaxy correlations

In this section we address the problem of probing the cosmic density field, with a particular attention to the large-scale structures analysis. What we have access from observations are galaxies, that tend to form in those regions of the Universe where the matter density is higher than other places. Moreover the evolution of galaxies is non-linear, as we have described in section 2.2.4, and is also non-local (Chan and Scoccimarro, 2012). For these reasons the relation between the galaxy and matter density fields is not trivial, because galaxies are biased with respect mass distribution, and this bias is given by different contributions.

In the following sections we describe the relation between the galaxy and matter fields, starting from the most simple case of linear bias, and then proceeding with the inclusion of the complication coming from non-linearity and non-locality.

### 2.3.1 Sampling

We define the cosmic density field as a continuous field so that for any position and at any time we can define $\delta(\mathbf{x}, t)$ (eq. 2.19). In our case, $\delta_g(\mathbf{x}, t)$ is the galaxy density field. For numerous applications we cannot use directly the continuous field, but we have to work with set of points that accurate represents the continuous density field. The transition from the continuous field to the set of points is obtained, usually, through a *Poissonian sampling* of the density field. In this case the entire space is divided into sub-volumes such that in each sub-volume the number of particles is distributed according to Poisson distribution with mean equals to the mean density of the cell. The mean number of particles in each sub-volumes is:

$$\langle N(\mathbf{x}) \rangle_P = [1 + \delta_g(\mathbf{x})] \bar{n} \Delta V \ , \tag{2.81}$$

where $N(\mathbf{x})$ is the number of particles in each sub-volume, $\Delta V$ the volume of each sub-volume and $\langle ... \rangle_P$ is the average over the Poisson distribution. We are interested in the probability of finding a pair of particles from a random choice of a pair of sub-volumes separated by a comoving distance equal to $\mathbf{x}$:

$$\langle N(\mathbf{x}_i) N(\mathbf{x}_i + \mathbf{x}) \rangle_{\mathbf{x}_i} = (\bar{n} \Delta V)^2 [1 + \xi(x)] \ ; \tag{2.82}$$

this probability is proportional to the two-point correlation function defined in eq. 2.25, that as we have already highlighted describes the excess probability, with respect to a Poisson distribution, to find a pair of particles in a pair of randomly chosen sub-volumes with comoving separation $x$.

### 2.3.2 Bias

The simplified way to relate the matter and the density fields is to consider the two fields to be proportional to each other:

$$\delta_g(\mathbf{x}) = b\delta(\mathbf{x}) \; , \tag{2.83}$$

where $b$ is for now a proportionality constant. Eq. 2.83 implies that also the two-point statistics is biased: $\xi_g(x) = b^2\xi(x)$ and the same is valid also for the power spectrum. This way to model the galaxy density field we observe is simplified; if we want to include also the non-Gaussian properties that characterized the galaxy field we have to consider a more general relation. We can assume, for example, a generic **local** relation between the galaxy density field and the matter density field:

$$\delta_g(\mathbf{x}) = \frac{n_g(\mathbf{x}) - \bar{n}_g}{\bar{n}_g} = F[\rho_m(\mathbf{x})] \; , \tag{2.84}$$

where $F$ is used to describe a generic functional relation between the fields. Considering the galaxy density field smoothed on some scales and small on large scales, we can write it as Taylor expansion of the matter density field:

$$\delta_g(\mathbf{x}) = b_0 + b_1\delta(\mathbf{x}) + \frac{1}{2}b_2\delta^2(\mathbf{x}) + \mathcal{O}(\delta^3(\mathbf{x})) \; . \tag{2.85}$$

where we introduce the galaxy bias parameters $b_0, b_1, b_2, ...$ and $\delta(\mathbf{x})$ without the subscript is the matter density field. Truncating the expansion at the second order in the density field and imposing $\langle\delta_g(\mathbf{x})\rangle = 0$, we obtain:

$$\delta_g(\mathbf{x}) = b_1\delta(\mathbf{x}) + \frac{1}{2}b_2\Big(\delta^2(\mathbf{x}) - <\delta^2(\mathbf{x})>\Big) \; ; \tag{2.86}$$

we call $b_1$ linear bias and $b_2$ quadratic non-linear bias. It is important to stress that apart from the linear bias we have an infinite number of non-linear bias parameters equals to the perturbative order we decide to keep in.

The linear bias, $b_1$, can be determined from simulations using the two-point statistics:

$$\xi_g(r) = b_1^2\langle\delta(\mathbf{x}_1)\delta(\mathbf{x}_2)\rangle = b_1^2\xi(r) \tag{2.87}$$

$$P_g(k) = b_1^2 P(k) \tag{2.88}$$

where $\xi_g(r)$ is the two-point correlation function of the galaxy and $\xi(r)$ the one for the matter; the same is valid in Fourier space.

In real observations the linear bias is degenerate with the amplitude of the matter fluctuations, so we need at least an other quantity to its determination. To this address it is used the galaxy bispectrum that is linked to the matter bispectrum and power spectrum through the linear bias and the quadratic bias:

$$B_g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = b_1^3 B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) + b_1^2 b_2 \Big[P(k_1)P(k_2) + 2 \text{ perm.}\Big] \ , \qquad (2.89)$$

where $B_g$ is the galaxy bispectrum and $B$ the matter one; $P$ is the linear matter power spectrum. For the determination of the bias parameters it is useful to introduce the *reduced bispectrum*:

$$Q_g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \equiv \frac{B_g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)}{P_g(k_1)P_g(k_2) + 2 \text{ perm.}} = \frac{1}{b_1} Q(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) + \frac{b_2}{b_1^2} \ , \qquad (2.90)$$

because this quantity is highly sensitive to the values of the bias parameters: looking at eq. 2.90 the first term $1/b_1$ represents a change in the shape-dependence of the reduced bispectrum while $b_2/b_1^2$ is an additive constant. If we know at least two triangular configurations, then we can fit eq. 2.90 for each configuration and look at the variation of the amplitude and shape; this makes it possible to fix $b_1$ and $b_2$ using only the bispectrum.

The assumption of locality was put in question for the first time by Chan et al. (2012) and Baldauf et al. (2012): they fitted the bispectrum from simulation finding the value of $b_1$ to be different from that one obtained by the power spectrum from the relation 2.88. In order to understand if this difference is due to the locality assumption, Chan et al. (2012) modified eq. 2.85 in the following way:

$$\delta_g(\mathbf{x}) = b_1\delta(\mathbf{x}) + \frac{1}{2} b_2 \delta^2(\mathbf{x}) + \gamma_2 \mathcal{G}_2 \ , \qquad (2.91)$$

where $\gamma_2$ is the *non-local* galaxy bias parameter and $\mathcal{G}_2$ is given by:

$$\mathcal{G}_2 = (\nabla_{ij}\phi)^2 - (\nabla^2\phi)^2 \ , \qquad (2.92)$$

where $\phi$ is the gravitational potential. $\nabla\phi$ is directly related to the displacement field in Zeldovich approximation; it follows that $\mathcal{G}_2$ is an invariant of $\nabla_{ij}\phi$ which measure the variations in displacements that affect the clustering. We can write $\mathcal{G}_2$ so that it matches the value obtained from simulations (Chan et al., 2012):

$$\mathcal{G}_2 = \int d^3\mathbf{q}_1 \int d^3\mathbf{q}_2 \ \delta_D(\mathbf{q} - \mathbf{q}_{12})[\cos\theta_{12} - 1]\delta_{\mathbf{q}_1}\delta_{\mathbf{q}_2} \ , \qquad (2.93)$$

with $\cos\theta_{12} \equiv \hat{q}_1 \cdot \hat{q}_2$. Through this model, that includes also the non-locality of the gravitation collapse, they were capable to recover the value of $b_1$ obtained from the power spectrum.

# Chapter 3

# Probes of the galaxy distribution

In this chapter we will introduce the main estimator (in all the text they are quantities characterized by an hat) for the two-point statistic of clustering both in configuration and Fourier space. In the last section we will describe two of the main probes to constrain the cosmological parameters, the Baryonic Acoustic Oscillation (BAO) and the Redshift Space Distortion (RSD) and a brief overview of the state of the art of cosmological constraints coming from the past and future galaxy surveys is provided.

## 3.1 Measurements of galaxy clustering

Assuming a relation between mass and galaxy density, we can study the spatial clustering of galaxies to infer the mass distribution and to constrain the cosmological parameters.

When we observe a galaxy samples, we do not measure the real positions of the galaxies, but their positions in redshift space that are affected by their peculiar velocities. In the this section we will examine the distortion that modifies the true galaxy distribution in redshift space. A second factor we have to take into account, in real observations, is the presence of some *selection function* $W(\mathbf{x})$, which provides the probability that a galaxy, satisfying some criterion, is included in the sample. Any survey is characterized by a selection function that will change according to the observational target.

First of all we describe some of the estimators for the two-point statistic of clustering in configuration and Fourier space. In particular we consider the two-point correlation function and the power spectrum for the cases of periodic box without selection function; we describe only the power spectrum in real space in the case we include a selection function and the redshift-power spectrum to include the effect of anisotropy in the distribution of galaxy. Then we focus on the evaluation of the covariance matrix of the two-point statistics that is a crucial point of this thesis.

### 3.1.1 Two point clustering estimators

The two-point correlation function, as we have mentioned in previous section (eq. 2.82), is one of the main quantity to study the clustering of galaxies. Although we have already provided the statistical definition of the two-point correlation function, when considering a generic redshift survey the two-point function is given by counting the pairs of given distance in the sample, with a certain selection function, with respect to the galaxies in a random sample with the same selection function of the survey sample:

$$\hat{\xi}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} - 1 \tag{3.1}$$

where $DD(r)\Delta r$ is the number of observed pairs with separation in the range $r \pm \Delta r/2$, $RR(r)\Delta r$ is the expected number of pairs in a random sample and $DR(r)$ is the number of cross-pair between the real and random samples in the same range. Eq. 3.1 is the *Landy and Szalay* estimator (Landy et al., 1998) and it is built to minimize the variance of the two-point correlation function. The number of objects in the random catalog is expected to be larger than the size of the observed sample, because we want that the shot-noise is as low as possible with respect to the observed sample.

As we have done with the two-point correlation function, we can consider an estimator for the counterpart in Fourier space, the power spectrum, in the simple case of periodic cubic box of side L; using the reality condition 2.30 we have:

$$\hat{P}_{tot}(k) = \frac{k_f^3}{N_k} \sum_{\mathbf{q} \in k} \delta_{\mathbf{q}} \delta_{-\mathbf{q}} \ , \tag{3.2}$$

where $k_f = 2\pi/L$, $N_k$ are all the modes within each shell, the sum runs over all the modes within the shell and the density contrast is defined as:

$$\delta_{\mathbf{k}} = \frac{1}{k_f^3} \frac{1}{N_k} \sum_i e^{i\mathbf{k}\cdot\mathbf{x}_i} \ , \quad \text{for } k \neq 0 \tag{3.3}$$

where the sum runs over all the galaxies in the box.

In this case the total power spectrum, including the shot-noise contribution, is given by:

$$\hat{P}_{tot} = \langle \hat{P}(k) \rangle = \hat{P}(k) + \hat{P}_{\rm SN} \ , \tag{3.4}$$

where $\hat{P}_{\rm SN}$ is the *shot-noise* estimator, that in the more simple case is given by the Poissonian noise, 1/N, with N the total number of galaxy in the box. In order to recover the cosmological power spectrum we have to subtract the shot-noise contribution.

As we have already mentioned in the initial part of this section the real surveys are characterized by a selection function.

One possibility is to enclose the survey into a box, (Baumgart and Fry, 1991), and define an auxiliary field (Feldman et al., 1994) given by:

$$F(\mathbf{x}) = w_k(\mathbf{x})\theta(\mathbf{x})[n(\mathbf{x}) - \alpha n_r(\mathbf{x})] \ , \tag{3.5}$$

where $n(\mathbf{x}) = \bar{n}(1+\delta(\mathbf{x}))$ is the observed density, $n_r(\mathbf{x})$ is the density from a random catalog with no correlation; as we have already stressed, the density of the random catalog is higher than the observed one, because we want that the shot-noise in the random catalogs is as small as possible compared to the observed sample; so, we define $\alpha$ as the ratio $\bar{n}/\bar{n}_r$; $\theta(\mathbf{x})$ is the window function with value 1 or 0 for points inside or outside the survey; $w_k(\mathbf{x})$ is a weight defined as:

$$w_k(\mathbf{x}) = \frac{1}{1 + (2\pi)^3\bar{n}(\mathbf{x})P(k)} \ . \tag{3.6}$$

$w_k(\mathbf{x})$ is obtained imposing that the fractional variance of the power spectrum $\sigma_P^2(k)/P^2(k)$ is minimized. As it it clear from eq. 3.6, the definition of the weight requires a preliminary estimation of the power spectrum we want to measure. Usually we can assume $P(k) \equiv P(k_*) \equiv P_*$, with $k_*$ the scale of interest for our analysis; for simplicity $\bar{n}$ is assumed to be constant. The Fourier transform of eq. 3.5 is:

$$F_{\mathbf{k}} = k_f^3 \sum_{\mathbf{q}} W(\mathbf{q} - \mathbf{k})\delta_{\mathbf{q}} \ , \tag{3.7}$$

where $W(\mathbf{k}) = w_k\theta_{\mathbf{k}}$ and $k_f$ is the fundamental frequency of the box given by $2\pi/L$ with $L$ the dimension of the box. The Fourier transform gives the convolution between the weighted window function and the galaxy density field. The power spectrum of this auxiliary field is then:

$$P_F(k) = k_f^6 \sum_{\mathbf{p}} |W(\mathbf{k} - \mathbf{p})|^2 P_{tot}(p) \ , \tag{3.8}$$

where $P_{tot}$ is the power spectrum of the box including the shot-noise contribution. For a top-hat window function if $k \gg 1/L$ we can approximate the auxiliary power spectrum as:

$$P_F(k) \simeq P(k)k_f^6 \sum_{\mathbf{p}} |W(\mathbf{k} - \mathbf{p})|^2 + k_f^3 \sum_{\mathbf{p}} |W(\mathbf{k} - \mathbf{p})|^2 \frac{1}{N} \ . \tag{3.9}$$

The estimator for the power spectrum is given by subtracting the shot-noise contribution and dividing by the selection function. As we can see from the equation above, the presence of the window function modifies the power spectrum, so that the quantity we have to estimate is both the convolution between the power spectrum and the window function itself, and the shot-noise term.

In all the relations we have described, we did not consider the redshift dependence. In the introduction of this section, we pointed out that cosmological surveys are carried out in redshift space and that peculiar velocities distort the spatial distribution of cosmological objects (see section 3.2.2). This distortion makes the galaxy distribution anisotropic and we have to include this effect in the power spectrum estimator. In the following we describe the estimator for the redshift power spectrum.

The general expression for the power spectrum in redshift space can be given in terms of Legendre polynomials (Taylor and Hamilton, 1996):

$$P(\mathbf{k}, z) = P(k, \mu, z) = \sum_{l=0,2,4,...} P_l(k, z)\mathcal{L}_l(\mu)(2l + 1) \ , \tag{3.10}$$

where $\mathcal{L}_l$ are the Legendre polynomials, $\mu$ is the the directional cosine between the line of sight direction and $\mathbf{k}$. For $l = 0$ we define the monopole $P_0(k, z)$ as the angular averaged power spectrum; for $l = 2$ we have the quadrupole $P_2(k, z)$ quantifying the leading anisotropy in the power spectrum due to the redshift-space distortion. The quadrupole is an important quantity that can be used to improve the constraints on the bias and cosmological parameters (Yamamoto et al., 2005).

### 3.1.2 Covariance matrix of clustering

The modeling of the errors on the two-point correlation function is fundamental as well the determination of the two-point function itself. The covariance matrix of LS estimator for a periodic box has been derived in Peebles (1973) Hamilton (1993) and Bernstein (1994) taking into account the non-Gaussian and discreetness effects. A simplified formula can be obtained, in term of the power spectrum, in the Gaussian limit where non-Gaussian and discreteness contributions can be neglected:

$$C(\hat{\xi}_i, \hat{\xi}_j) = \frac{(2\pi)^5}{V} \int dk \ k^2 P^2(k) J_{1/2}(kr_i) J_{1/2}(kr_j) \ , \tag{3.11}$$

where $J_{1/2}$ is a Bessel function.

As already pointed out by Bernardeau et al. (2002), the issue of cosmic error computation is recurrent in cosmological surveys, so that various techniques have been proposed in the literature. We can rely on two main methods, called **bootstrap** (Barrow et al., 1984) and **jackknife** (Tukey, 1958). The first one consists in building an ensemble of sub-samples with the same number of galaxies of the initial sample. The galaxies in each sub-sample are chosen picking randomly from the initial sample with replacement, in the sense the sub-sample can include the same galaxy more than once, while others may not be included at all. In the second case the initial sample is divided in a set of disjoint sub-samples each containing $N/N_{jackk}$ galaxies, where N is the number of galaxies in the initial sample and $N_{jackk}$ that one of the specific sub-sample. After the ensemble of sub-samples is obtained we proceed with the evaluation of the covariance matrix.

As for the two-point statistics we can move in Fourier space and consider a general estimator for the power spectrum covariance matrix in a periodic box (Hamilton, 1997; Scoccimarro et al., 1999):

$$C(\hat{P}_{tot,i}, \hat{P}_{tot,j}) = \frac{2}{N_{k_i}} \hat{P}^2_{tot}(k_i)\delta_{ij} + k_f^3 \tilde{T}_{tot}(k_i, k_j) \ , \tag{3.12}$$

where $N(k_i)$ is the number of modes inside a k-shell of width $\Delta k_i = 4\pi k_i^2 dk_i$, $\hat{P}_{tot}(k)$ is the power spectrum including the shot-noise contribution given by eq. 3.4, $k_f$ is the fundamental frequency and $\tilde{T}_{tot}(k_i, k_j)$ is the average of the trispectrum $T(\mathbf{k}_i, -\mathbf{k}_i, \mathbf{k}_j, -\mathbf{k}_j)$ over the angle $\theta$ between the vectors $k_i$ and $k_j$, including all the shot-noise contributions:

$$\tilde{T}_{tot}(k_i, k_j) \equiv \frac{1}{2} \int_{-1}^{1} d\cos\theta T(\mathbf{k}_i, -\mathbf{k}_i, \mathbf{k}_j, -\mathbf{k}_j) \ . \tag{3.13}$$

For next future surveys, as we will stress in the following chapters, we require a very large number of independent realizations of the Universe, so we have to proceed with numerical simulations. In section 4 we highlight the main properties of a full N-body simulation and we observe that this kind of numerical approach is not a real possibility when we have to produce thousands of galaxy catalogs. Using analytic approximations (see section 4.2) is possible to reduce the computing time required by a full N-body simulations in exchange of accuracy. As we will also show in chapter 6 these procedures appears to be a valid alternative when a large number of synthetic realizations is required. It is worth also to mention that the requirement in terms of number of simulations needed for future surveys has lead to the development of particular techniques to evaluate the covariance matrices of clustering using a reduce number of simulations, but at same time preserving the accuracy we can reach using the brute force approach. We describe some of these methods in section 6.1.

## 3.2 Cosmological constraints from clustering

In this section we briefly review the Baryonic acoustic features impress in the power spectrum of CMB temperature fluctuation as in the two-point statistics of clustering and the redshift distortions induced by the peculiar velocity of galaxies. These two observations can be used to but tight constraints on the cosmological parameters. In the last section we will discuss the state of the art in the determination of the cosmological parameters, looking at the present and near past galaxy survey achievements.

### 3.2.1 Baryonic acoustic oscillation

The BAO is a signature imprinted by a series of sound waves that propagated in the hot plasma of tightly coupled photons and baryons in the early Universe (Eisenstein and Hu, 1998). In section 2.1.1 we have already described the main features of the CMB, stressing the fundamental importance of the oscillations in the temperature power spectrum. The same oscillations are present in galaxy clustering. Using perturbation theory it is possible to compute the comoving distance, $r_s$, that the sound waves can travel from the Big Bang to the epoch of recombination, when photons decouple from matter:

$$r_s = \int_0^{t_{rec}} \frac{c_s(t)}{a(t)}\, dt = \int_{z_{rec}}^{\infty} \frac{c_s(z)}{H(z)}\, dz \ , \tag{3.14}$$

where $c_s$ is the wave sound speed.

If we look at configuration space we can consider an overdensity of photons and baryons propagating from a primordial overdensity peak of all species (dark matter, baryons, neutrinos, and photons). The wave propagates during the radiation-dominated epoch and slows down during the matter dominated epoch. At time of recombination photons decouple from baryons and they start to travel free, causing the weave to stall. A spherical shell of baryons overdensity of radius equal to the distance traveled by the wave until photons decoupling, eq. 3.14, take shape. During the time between the radiation and the matter epoch, the dark matter overdensity does not evolve because it was already decoupled from the matter-radiation fluid at time of primordial overdensity. After recombination the gravitational instability increases because the mutual attraction between baryons and dark matter; this interaction lets the perturbations to grow. In the final stage the total matter overdensity is located at the center of a spherical shell of 150 Mpc radius (Eisenstein et al., 2007) (figure 3.1). The comoving size, $r_{||}$ and $r_{\perp}$ of a generic object or feature, at fixed redshift, is related to the observed size, $\Delta z$ and $\Delta\theta$, by the Hubble parameter $H(z)$ and by the angular diameter $D_A(z)$:

$$\begin{aligned} r_{||} &= \frac{c\Delta z}{H(z)} && \text{along line of sight} && (3.15)\\ r_{\perp} &= (1+z)D_A(z)\Delta\theta && \text{transverse direction .} && (3.16) \end{aligned}$$

From eqs. 3.15 and 3.16 it is clear that the measurement of the observed size gives access to $r_{||}H(z)$ and to $r_{\perp}/D_A(z)$. If the physical scales $(r_{||}, r_{\perp})$ of the feature we are studying are known, we can have an estimation of $H(z)$ and $D_A(z)$. BAO can be used to measure the matter and baryon density allowing us to have an estimate of the characteristic scale, $s$, of the BAO in matter or galaxy clustering. In configuration space the BAO appears as a bump the the two-point correlation function, meaning that we are observing an excess in

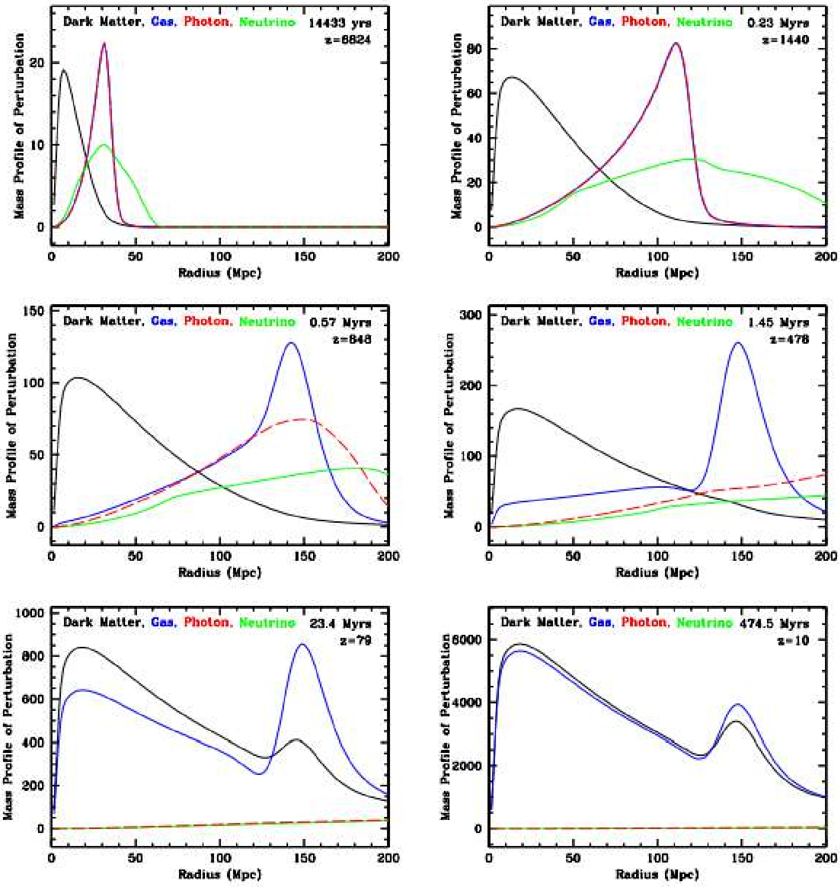


Figure 3.1: Linear-theory response to an initially overdensity at the origin. In each plot it is shown the evolution of perturbation for the four species: dark matter (black), baryons(blue), photons (red) and neutrinos (green), at different redshifts. (Eisenstein et al., 2007)

the number of pairs at that particular scale $s$; in Fourier space it appears as a series of harmonic oscillations with peaks and troughs of their amplitudes located at multiples of $k = \pi/s$. In figure 3.2 we show both the two-point correlation function and the power spectrum from SDSS-BOSS compared with the best-fit models, with particular attention to the BAO (lower plots). Operatively we look at the distribution of galaxies finding the characteristic scale where there is an excess clustering due to the BAO. When this scale is measured in comoving coordinate, knowing the correspondent physical scale we can determine independently the parameters $H(z)$ and $D_A(z)$.

The Hubble and the angular diameter parameters are strictly related to the dark energy density parameters and the dark energy equation of state parameter $w$, so BAO features can be used as powerful probe to constraints

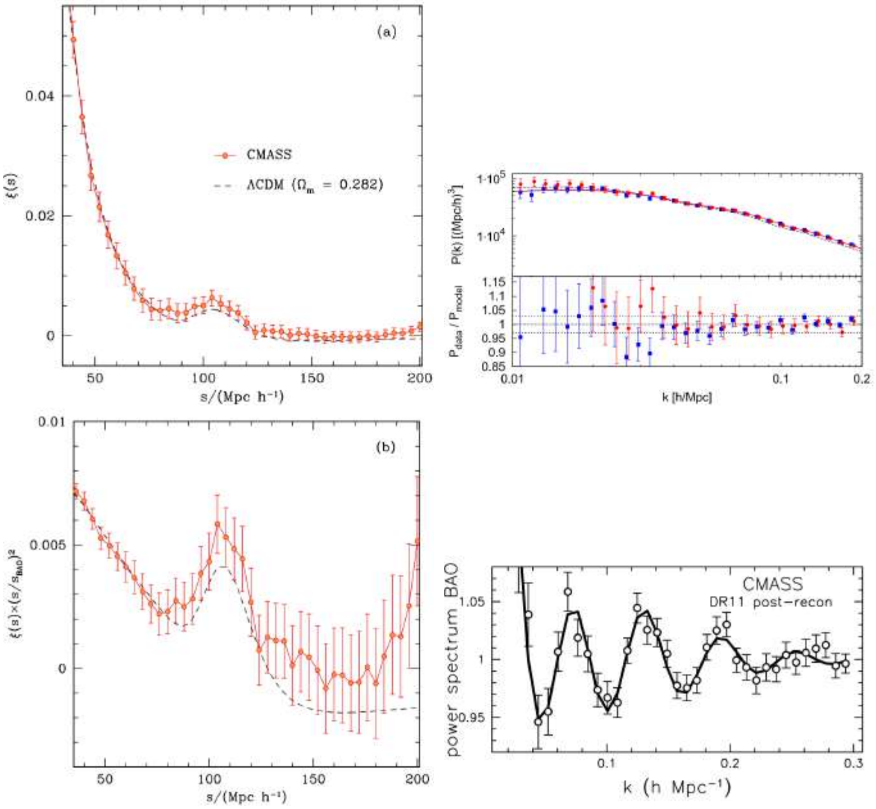


Figure 3.2: In the top left panel the spherically averaged redshift-space two-point correlation function of the full CMASS sample with error bars obtained from a set of 600 mocks catalogs (Manera et al., 2013). The dashed line corresponds to the best-fitting CDM model obtained by combining the information from the shape of the correlation function and CMB measurements (Sánchez et al., 2017). In the bottom left panel the same as the upper left panel, but rescaled by $(s/s_{BAO})^2$, where $s_{BAO} = 107.2\,h^{-1}\,\mathrm{Mpc}$ (Sánchez et al., 2017). In the top right panel we show the power spectra for the North galactic cap (blue dots) and for the South galactic cap (red dots) (Gil-Marín et al., 2015). The red and blue solid lines correspond to the best-fit model; the bottom sub-panels show the ratio between the power spectrum measurements and the best-fit models; in the bottom right panel panel the power spectrum of CMASS DR11 galaxies, divided by a smooth, no-BAO power spectrum. Solid curve is the best-fit model (Anderson et al., 2014).

the dark energy parameter.

### 3.2.2 Redshift-space distortion

The measurements of cosmological distances are very complicated, being affected by large systematic. On the other hand, redshift is easier to obtain, but it is not a direct measure of galaxy distance because of the peculiar velocity of the galaxy itself. If the redshift is used as a distance indicator we observe a distortion of the true galaxy distribution. If we consider a mass shell with small overdensity ($\delta \ll 1$), its expansion will be decelerated, but its peculiar velocity is not enough large to compensate the Hubble expansion. In redshift space we will observe the mass shell squashed along the line of sight. When the density starts to increase, in the quasi linear regime ($\delta \sim 1$), the mass shell is just turning around, its peculiar in-fall velocity is equal to the Hubble velocity. In redshift-space, an observer at large distance sees the shell totally collapsed. A mass shell that has already turned around appears flattened along the line-of-sight, if its in-fall velocity is less than twice the Hubble ones. At smaller radii we enter in the non-linear regime. The mass distribution has collapsed and the peculiar in-fall velocities are larger than the Hubble velocity because of the random contributions due to the small-scales; in this case we observe the so called *Finger-of-God*, because the distribution is elongated along the line of sight. It is clear that the Hubble expansion and the peculiar velocity change the observed clustering on different scales.

The relation between the redshift-space position and the real distance is:

$$s = r + v_r \ , \tag{3.17}$$

where s is the distance of an object in redshift space, r is the distance in real space and $v_r$ is the radial component of the peculiar velocity, that is responsible for the change in the galaxy density field seen in redshift space.

The computation of the redshift-space distortion requires the definition of the density field in redshift space $\delta(\mathbf{s})$. Using the conservation of the total number of galaxies, it is possible to derive the relation between $\delta(\mathbf{r})$ and $\delta(\mathbf{s})$:

$$1 + \delta(\mathbf{s}) = \frac{r^2}{(r + v_r)^2}\Big(1 + \frac{\partial v_r}{\partial r}\Big)^{-1}\Big(1 + \delta(\mathbf{r})\Big) \ . \tag{3.18}$$

Eq. 3.18 is a general expression valid for any perturbative order. If we reduce to the case in which $\delta \ll 1$ and we use the plane-parallel approximation (Kaiser, 1987) we have, in Fourier space:

$$\delta_{\mathbf{k}}^{(s)} = (1 + \beta\mu_{\mathbf{k}}^2)\delta_{\mathbf{k}} \ , \tag{3.19}$$

where

$$\beta \equiv \frac{f(\,\Omega_m\,)}{b} \equiv -\frac{1}{b}\frac{\mathrm{dln}D(z)}{\mathrm{dln}(1+z)} \tag{3.20}$$

and $\mu_{\mathbf{k}} \equiv k_z/k$. Using eq. 2.31 the power spectrum of the density field in redshift space is:

$$P^{(s)}(k,\mu) = (1+\beta\mu_{\mathbf{k}}^2)^2 P(k) \ . \tag{3.21}$$

When the linear approximation is not valid, eq. 3.21 should be corrected to include the non-linear corrections needed to describe the finger-of-god. The *dispersion model* by Peacock and Dodds (1994) and (Hamilton, 1998) considers the power spectrum in redshift space as the linear one, times the contribution of the pair-wise velocity. The redshift space distortion can be used to put constraints on the cosmological model, because the amplitude of the power spectrum in redshift space depends on $f(\Omega_m)$ through $\beta$. We cannot measure directly the value of $\beta$, so we need to find a way to disentangle its value from the power spectrum. The usual procedure is to take an harmonic expansion of eq. 3.21. The monopole and the quadrupole, in linear approximation, are:

$$P_0^{(s)}(k) \equiv \Big(1+\frac{2}{3}\beta+\frac{1}{5}\beta^2\Big)P(k) \tag{3.22}$$

$$P_2^{(s)}(k) \equiv \Big(\frac{4}{3}\beta+\frac{4}{7}\beta^2\Big)P(k) \ ; \tag{3.23}$$

the ratio $P_2/P_0$ depends only on $\beta$, so measuring $P_0$ and $P_2$ using eq. 3.21 we can fix the $\beta$ value. $\beta$ is given by $f(\Omega_m)/b$, where $\Omega_m$ is the matter density parameter and $b$ is the linear bias. Putting constraints on $\beta$ using the ratio $P_2/P_0$ allow us to constraint the matter density parameter and consequently testing the standard cosmological model. Measurements from 2dFGRS, (Peacock et al., 2001), show a value of $\beta = 0.43$ that implies $\Omega_m \sim 0.3$, value that is compatible with the ΛCDM model. This result is obtained assuming some b-value for the 2dFGRS galaxies.

### 3.2.3 State of the art

During the last decade, an increasing number of surveys were used to extract cosmological information from large scales structures: Six degrees Field Galaxy Survey (SdFGs, Beutler et al., 2011), Sloan Digital Sky Survey I-II (SDSS, Eisenstein et al., 2005), WiggleZ Dark Energy Survey (WZDES, Blake et al., 2011; Kazin et al., 2014), VIMOS Public Extragalactic Redshift Survey(VIPERS, Guzzo et al., 2014), Baryon Oscillations Spectroscopy Survey(BOSS, Anderson et al., 2014; Cuesta et al., 2016; Ross et al., 2017; Sánchez et al., 2017). Future large scales and high redshift experiments, such as (DESI, Schlegel et al., 2011; Levi et al., 2013)), Extended Baryon Oscillation Spectroscopic Survey (eBOSS), Large Synoptic Survey Telescope (LSST, LSST Science Collaboration et al., 2009), Euclid (Laureijs et al., 2011), Wide-Field Infrared Survey Telescope (WFIRST, Green et al., 2012) and the Square Kilometer Array (SKA, Schilizzi et al., 2008), will give high precision observations that, together with the previous results, will allow to improve our constraining power. Putting together all the observational

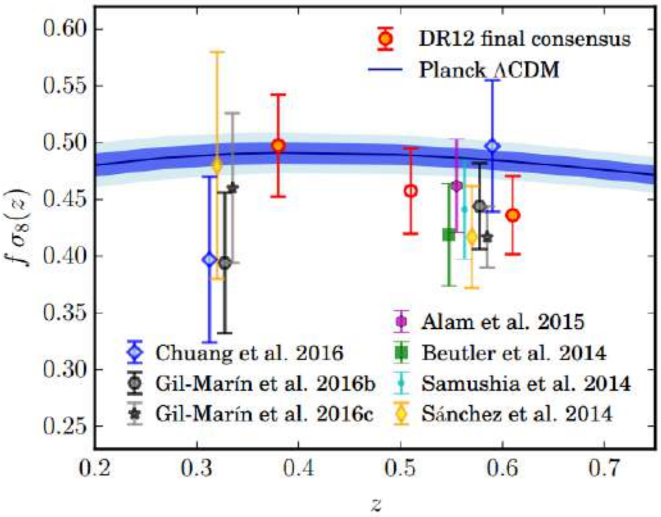


Figure 3.3: Comparison $f\sigma_8$ measurements across previous BOSS measurements in DR11 and DR12 samples. The blue shaded region is the conditional constraint of $f\sigma_8$ assuming Planck $\Lambda$ background cosmology.

results coming from the large scales surveys that have gathered data between 2014 and late 2016, we can have an idea of both the high precision clustering measurements, and of the advancing in modeling the clustering properties. In figure 3.3 (Alam et al., 2017) it is showed a comparison of the $f\sigma_8$ values from BOSS measurements from 2014 to 2016 with the predictions from Planck ΛCDM model, while in figure 3.4 they show the same comparison, but using measurements coming from other surveys (2dfGRS, 6dFGS, GAMA, WiggleZ, VIPERS). These comparisons also act as a validity test of GR on large scales. Clustering measurements can be used in combination with CMB especially to constrain the cosmological parameter related to the dark energy in figure 3.5 we show the $\Omega_k$-$w$ and the $w_0$-$w_a$ constraints. Near future surveys will allow to reduce the uncertainty on these constraints and start to put real boundaries on the right cosmological model. Looking at figure 3.5 we notice that including in the analysis, data from Planck and BOSS together with those one from Joint Lightcurve Analysis (JLE) SNe (Betoule et al., 2014), the constraints can be very tight and they seem to be in agreement with a flat ΛCDM model. In figure 3.6 we show the Euclid forecasts for the dark energy cosmological parameters from the Euclid Definition Study Report (Laureijs et al., 2011). In combination with Planck results, Euclid can improve upon current constraints by over a factor of 100. These constraints will allow each of the broad classes of dark energy models to be tested.

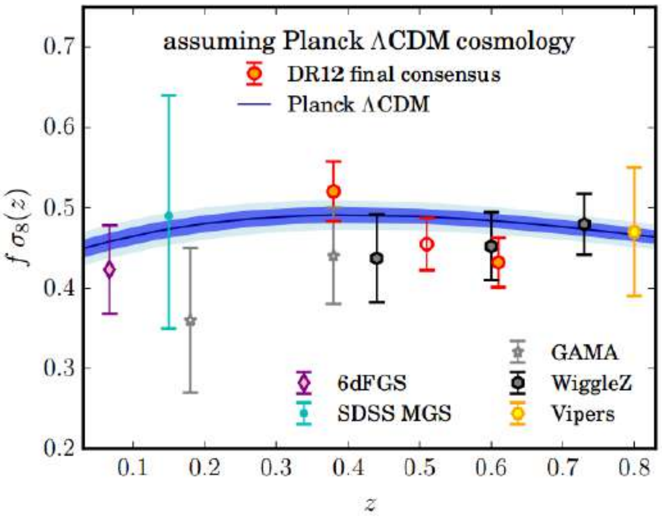


Figure 3.4: Results from (Alam et al., 2017) compared with the measurements of the 2dfGRS (Percival et al., 2004) and 6dFGS (Beutler et al., 2012), the GAMA (Blake et al., 2012), the WiggleZ (Blake et al., 2013), the VVDS (Guzzo et al., 2008), and the VIPERS (de la Torre et al., 2013) surveys, as well as the measurements from the SDSS-I and -II main galaxy sample (Howlett et al., 2015) and the SDSS-II LRG sample (Oka et al., 2014). The blue shaded line is the same of figure 3.3.

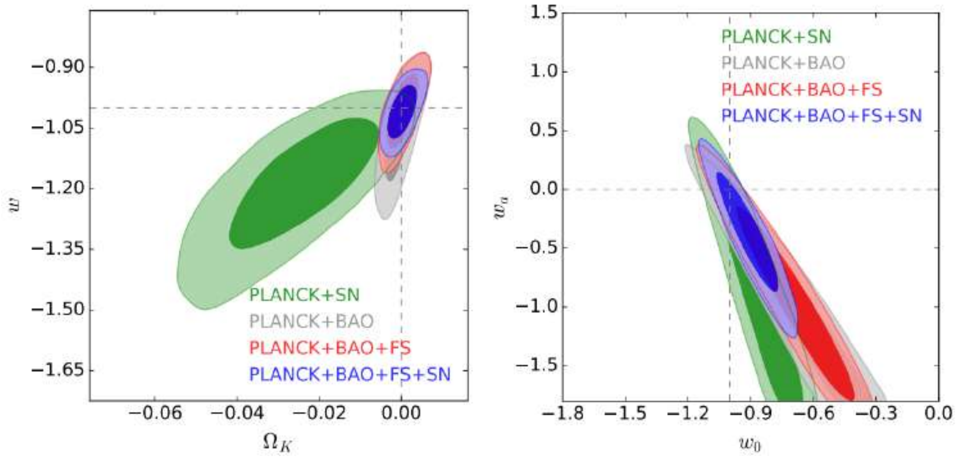


Figure 3.5: Parameter constraints for the owCDM (left), model which considers a variation in the spatial curvature keeping constant the equation of state for dark energy, and $w_0 w_a$CDM (right), model which allows a time-evolving equation of state. The plots show the comparison from BAO and BAO+FS to those with Type Ia SNe (Alam et al., 2017).

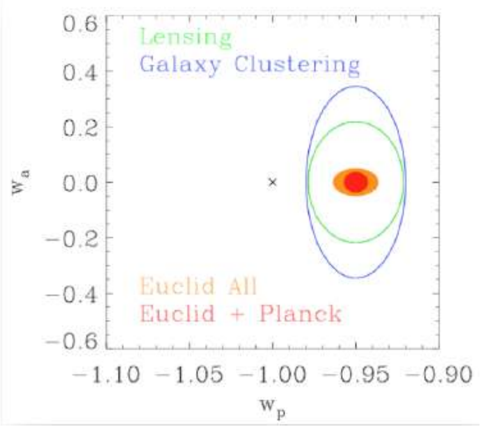


Figure 3.6: The expected constraints from Euclid in the dynamical dark energy parameter space. We show lensing only (green), galaxy clustering only (blue), all the Euclid probes (lensing+galaxy clustering+clusters+ISW; orange) and all Euclid with Planck CMB constraints (red) (Laureijs et al., 2011).

# Chapter 4

# Numerical methods for cosmological simulations

The possibility to have access to a large number of cores, that speed up the number of computations, and the potentiality of large memory storages has given the possibility to study the evolution of complex dynamical systems that in general have complex analytic solutions.

We can study the Universe using powerful numerical techniques; usually we consider the mass distribution in the Universe as described by particles or sampled on a grid. Knowing the equation of motions, it is possible to obtain, numerically, the evolution of each particles.

In this chapter, first we describe briefly what is an “exact” N-body simulations, underlining the main techniques and then we discuss of a various number of *approximate methods* that we will use in the next chapters of this thesis.

As we will also show in chapter 6, the approximated methods allow us to produce large sets of galaxy catalogs that are needed to evaluate the covariance matrix. Future large galaxy survey will require a very large number of realizations ($\sim 10'000$) that cannot be obtained with full N-body simulations.

## 4.1 N-body simulations

Cosmological N-body codes are used to calculate the non-linear growth of structures in the universe by following the trajectories of N particles interacting between each others through gravity. An N-body code is characterized by two main conditions: first some *initial condition* must be set in a way to represent the prediction of linear theory; second the evolution of structures described by the code has to be free of distortion due to the small number of particles and the finite size of the simulation volume.

One particle in a N-body simulation represents a large number of dark matter particles; this means that the interaction between two mass particles mimics the interaction between two fluid elements. This assumption leads to the introduction of a force *softening scale*: the fluid elements should feel much less force than two mass particles. The scale at which the force has to be softened is comparable with the average inter-particle separation.

We have already pointed out that the mass particles of an N-body simulation interact only because of gravity. To run a simulation it is needed to evaluate the force, between the particles, and then to solve the equation of motions. In cosmology we have to take in account that the Universe is expanding with time, so the equations of motion are given by the classical Newtonian equations where the expansion is controlled by the scale factor $a(t)$.

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{v}_i \tag{4.1}$$

$$\frac{d\mathbf{v}_i}{dt} = -2H(t)\mathbf{v}_i - \frac{1}{a^2}\nabla_{\mathbf{x}}\Phi_i \tag{4.2}$$

$$\nabla^2_{\mathbf{x}}\Phi = 4\pi G a^2[\rho(\mathbf{x},t) - \bar{\rho}(t)] \; , \tag{4.3}$$

where $H(t)$ is the Hubble parameter, $\Phi$ the gravitational potential due to the density perturbations and we use comoving coordinates for convenience.

The calculation of the force between particles is needed to integrate the equation of motion forward in time and it is the most consuming task for a N-body simulation. For this reason many techniques have been developed focus on this aspect (Bertschinger, 1998). For the purpose of this thesis we will not go into the details of this topic, but we will focus our attention on the approximated methods that are mainly used in the following chapters and for all the analysis we have carried out.

## 4.2 Approximate methods

As we have stressed in the introduction, future surveys will be capable to observe billions of galaxies; at this level, the constraints on the cosmological parameters will be highly influenced by systematic, as we will show in chapter 5. Moreover an accurate evaluation of the covariance matrices of clustering is required as we will show in chapter 6. Taking under control these two quantities is one of the first requirement for the data analyses. One possible way to address these issues is to simulate large number of galaxy catalogs with the same features of the specific survey. A large number of simulated catalogs is required to lower the noise in the estimate of the covariance of the clustering measurements, so that it could be used in the likelihood estimation and then in the cosmological parameters determination.

N-body simulations can provide realistic galaxy catalogs, but the number of realizations needed to estimate the covariance with the precision required by next future surveys is too large; for this reason, a program based only on

N-body simulations is unfeasible. A valid alternative is to take advantage of specific approximations to obtain the large scales density field and the Dark Matter (DM) halo distribution. We describe methods based on two type of approaches: those ones (PINOCCHIO, PTHalos, COLA, AugmentedLPT) that take advantage of LPT to follow the particles evolution from the beginning to the formation of the halo, and those ones (PATCHY, EZmocks, HALOGEN) that use sophisticated bias model to populate the density field with halos, calibrated on some big simulation.

Using these approximated methods allows one to produce a large number of synthetic galaxy catalogs reducing the computing time of a factor $\sim$ 1000. The price to pay is a loss in accuracy, particularly at small scales (k$\sim$0.5 $h^{-1}$ Mpc) compared with N-body simulations. The following sections are based on the review on approximate methods by Monaco (2016). In the rest of the thesis a large part of the analyses will be done using the PINOCCHIO, but in chapter (6) we will also show a comparison between some of the methods we describe in the this sections.

### PINOCCHIO

The PINOCCHIO (PINpointing Orbit Crossing Collapsed HIerarchical Objects) algorithm (Monaco et al., 2002, 2013) is based on (i) generating a linear density field on a grid, as usually done for the initial conditions of an N-body simulation; (ii) estimating the time at which each grid point (or particle) collapses, using a combination of ellipsoidal collapse model and excursion set theory (Sheth et al., 2001); (iii) grouping together collapsed particles into DM halos with an algorithm that mimics their hierarchical assembly. Displacements of particles (and DM halos) from their initial positions are computed using LPT. Starting from the linear density field (i), previously smoothed, the collapse time for each particle is computed. A halo is set up using these collapsed particle with an algorithm that mimic the hierarchical clustering. After the halo is formed it has to increase its mass and this is done using single particle, that ends up at some point in that halo, and by merging with smaller halos. The history of a halo is given by its tree of merging, saved time-step after time step and output at the end. The accuracy of the algorithm is given by its precision to assign the right halo mass and to put the halos in the right final position. Using LPT it is possible to have different levels of precision: with 3LPT the halo power spectrum is accurate within 10% at $k \sim 0.5$ in redshift space (Munari et al., 2017).

PINOCCHIO has been used to build galaxy catalogs for the VIPERS survey (de la Torre et al., 2013) and is being used for the Euclid preparatory science.

### Augmented LPT

An improvement of the standard LPT technique, the Augmented LPT (ALPT), is proposed by (Kitaura and Heß, 2013). This method is based on splitting the displacement field into a long- and a short-range component. 2LPT is used to compute the long-range component, while for the short-range component they use the solution coming from the spherical collapse approximation. Using these two different approximations for the two different components they can improve the a standard LPT based methods when compared with a N-body simulations, but an additional free parameters is required to define the transition between the long- and short-range regimes.

### PTHALOS

This second scheme, Perturbation Theory Halos (PTHalos), (Scoccimarro and Sheth, 2002) is based on the generation of a density field with 2LPT on which a FoF (Friend of Friend halo finder) algorithm is run. Contrary to PINOCCHIO, PTHalos uses a spherical collapse model. After the halos are defined, their mass is rescaled to reproduce a specific mass function. In this last step, the information of the halo mass is completely lost because each halo is forced to be on the specific mass function, but still the code is highly predictive on halo bias and clustering. This techniques has been used to produce galaxy catalogs for the BOSS survey.

### Particle-Mesh based schemes

One of the main difference between an N-body simulation and a PINOCCHIO or PTHalos simulation is that the two approximated methods do not try to resolve the inner structure of halos, while a large amount of time of an N-body simulation is dedicated to solve the complex non-linear orbits of halo particles. One way to speed-up an N-body simulation is to use a particle-mesh scheme with few time step: after the distribution of particles has been generated, dark matter halos are extracted using an halo-finder code. The recovery of the halos is strongly connected with the size of the mesh used to compute the density. Usually the size of the mesh is 1/2 or 1/3 of the inter-particle distance and for this reason the memory requirements are very high. A simulation that follows this schema is in any case “approximated” compared with an usual N-body simulation because it is does not describe accurately the force at particle-particle level, as a proper N-body simulation does.

Two main example of Particle Mesh based methods, optimized to reproduce the linear growth rate with few numbers of time-step are: the COmoving Lagrangian Acceleration (COLA) algorithm by Tassev et al. (2013) and the FastPM scheme by Feng et al. (2016).

The basic idea of COLA is to decouple the large and small scales to accurate describe both scales, but reducing the computing time with respect a full N-body simulation: it solves the large scales using LPT and the small scales using an N-body simulation. To split the two regime, the equations of motion have to be recast by going in a frame comoving with LPT observers. COLA has been used in the production of galaxy catalogs for the WiggleZ survey (Drinkwater et al., 2010) and the main galaxy sample of SDSS.

In FastPM, assuming that the velocity evolves following the ZA, the kick (update the position of a particle) and drift (update the velocity of a particle) factor are redefined so that the linear growth factor is forced to be recovered with few time-steps. In their paper Feng et al. (2016), show that they can achieve an improvement of the 2LPT solution using 5 time-steps.

Different PM codes have been used to produce catalogs for the BOSS survey and for Euclid.

### PATCHY

The PATCHY (Perturbation Theory Catalog generator of Halo and galaxY distribution) algorithm by Kitaura et al. (2014) uses the matter density field from Augmented LPT, described above, to built an halo distribution that fit a given simulation at the 3-point level. The idea is to built a stochastic bias model using a set of theoretically justified free parameters. The halo density field is a power-law of the matter density field and depends on parameters that describe a threshold, an exponential cutoff and the normalization; then the sampling of the halo density field is done with a dispersion larger than Poisson so that the stochasticity of bias is reproduced. The mass of the halo is assigned to reproduce the halo-dependent bias (Zhao et al., 2015). All the free parameters that determine the halo density field are fixed using a large exact N-body simulation.

PATCHY has been run to produce a large set of catalogs, 12'288, for the BOSS survey.

### EZmocks

EZmocks (Effective Zeldovich approximation mock catalog) by (Chuang et al., 2015) generates (i) a dark matter density field on a grid predicted using ZA reproducing the clustering properties with a bias model (ii) then the probability distribution function of halos in the BigMultiDark simulation (Klypin et al., 2016) is mapped in the ZA density field (iii) the amplitude of power spectrum and bispectrum is fitted with a specific density threshold and (iv) the shape of the power spectrum is modified, changing the slope of the initial power spectrum with a scale-dependent function; (v) the BAOs are fitted by enhancing the amplitude of the oscillations in the initial power spectrum; finally (vi) the velocity field is computed within ZA.

### Halogen

This is the more recent approximated method by Avila et al. (2015). The density field is generated using LPT on a grid and resampled on twice the inter-particle distance; the halo distribution is given so that it samples an analytic mass function. The sites for haloes are identified using random particles from the 2LPT snapshot and the cells containing the halos follow a probability proportional the power of the density field ($P \propto \rho^{\alpha}$), without overlapping particles. The mass conservation is guaranteed lowering the cell mass by the halo mass. The halo velocity is fixed taking in to account the velocity dispersion within a cell from a reference simulation. The parameter $\alpha$ is obtained in each mass bin using a standard $\chi^2$ minimization technique.

## 4.3 Discussion

The methods we have described above are characterized by different properties and they allow us to obtain different levels of accuracy. The methods based on Lagrangian theory (PINOCCHIO, PTHalos, COLA), can be very predictive because they find simulated halos at the object-by-object level. Between these codes, PM-based codes are quick N-body solvers, but the memory requirements are very high because they are also required to solve small halos. To the contrary PINOCCHIO does not require accuracy below the inter-particle distance, so its memory request is lower than COLA, but its accuracy in placing halos is limited by LPT.

Methods like PATCHY, EZmocks and HALOGEN are faster than the LPT based methods, but they need a reference full N-body simulation for the calibration of the parameters and this step requires many evaluation of the clustering statistics. These codes can be used to produce a large number of realizations of the Universe, but taking in mind that their productivity is lowered with respect other methods we have described above, because for each cosmology they need to calibrate the parameters on a large simulation and they do not predict the halo mass function or halo merger histories, that are given for example by PINOCCHIO.

As we have already highlighted at the beginning of the previous section, these approximate methods use a different approach to reach the same goal, that is the fast and cheaper production of realizations of the Universe, compared with a full N-body simulation. Having access to a large number of catalogs in a relatively small amount of time allow us to evaluate the clustering covariance matrices. In chapter 6 we describe some of these techniques, studying how well they can reproduce the results of a full N-body simulation and if they can be used for cosmological purpose, such as the determination of cosmological parameters.

# Chapter 5

# Uncertainty in the visibility mask of a survey and its effect on the clustering of biased tracers

## 5.1 General introduction to the foreground problem

Next large-volumes galaxy surveys will allow us to investigate very large galaxy sample, so that with very high statistics, the error budget will be dominated by systematics. On the largest scales, at or beyond the BAO scale, it will become of fundamental importance to keep under control the effect of foregrounds, due both to the zodiacal light and to the Milky Way (galactic extinction and stellar contamination above all). These will act by modulating the survey depth on the sky. A similar modulation will be due to instrumental or survey features, like 0-point offset of photometric calibration (see for instance the calibration of BOSS photometry Padmanabhan et al., 2008; Ross et al., 2011); in the following, for simplicity we will refer to foreground removal as the process that we are addressing, but our approach is equally valid for these systematics. Careful characterization of foregrounds will make it possible to subtract them. However, the residuals from this subtraction will, in most cases, be highly correlated on the sky, thus mimicking large-scale structure. So, the error on foreground subtraction must be properly propagated to correctly assess the error on parameter estimation.

A great effort has been devoted to understanding the effect of Galactic foreground and foreground removal for the 21-cm line emission in the reionization epoch (e.g. Santos et al., 2005; Jelić et al., 2008), 21-cm intensity mapping survey at low redshift (e.g. Wolz et al., 2014). For the LSS an example of foreground analysis is given by the BOSS survey: they analyzed the potential systematic effects on the galaxy observed density (Ross et al., 2012) finding that the major contributions come from stellar density and Galactic extinction.

In this work we focus on the issue of how the uncertainty in the removal of foregrounds (or other similar systematics, as said above) propagates to the measurement of clustering at the two-point level and to its covariance. The number of galaxies in an observed sample is given, on average, by the integral of the galaxy luminosity function from a luminosity threshold, determined by the survey flux limit, to infinity. In realistic cases, the flux limit is modulated on the sky by foregrounds, and these are typically correlated on large angular scales. A visibility mask will quantify this effect in order to remove it, but this removal will be done with some uncertainty. This will result in a modulation of the luminosity threshold, that will propagate to the number density of observed objects, creating fake large-scale structure.

To address this issue, we have used the approximate method PINOCCHIO (Munari et al., 2017) to run 10,000 simulations of a box of 1.5 $h^{-1}$ Gpc. We will consider DM halo "mock" catalogs at redshift $z = 1$, where it is possible to have observational access to large volumes and the lower level of non linearities allows approximate methods to be more accurate. We use DM halos in place of galaxies as biased tracers of the density field, and their mass in place of galaxy luminosity. This simplification of the procedure is acceptable in this context, as long as clustering on very large scales is considered and halo mass is simply used to implement the effect of a varying "luminosity" threshold. As a note, a one-to-one correspondence between luminosity and mass is equivalent to a simplified halo occupation distribution (HOD) model, as we will comment in section 5.3.

To derive analytic predictions of the clustering of these mock catalogs, we build a toy model based on the following assumptions. (i) A mass-independent bias scheme is implemented. DM halos and galaxies share the property of having a mass- or luminosity-dependent bias, but this greatly complicates the analytic approach. We implement mass-independent bias by shuffling halo masses among the objects, as explained in section 5.3. A coincise analysis of the mass-dependent bias case will be outlined in section 2.3.2. (ii) We consider an idealized geometry for the mask. In a plane-parallel approximation, the plane of the sky is identified with the $x - y$ plane. This is tiled with squares of physical length $l$, and for each tile the residual of foreground subtraction is quantified by drawing a random number from a Gaussian distribution. No correlation among tiles is considered, so the length $l$ is to be interpreted as the projection (in a flat sky approximation), at the observation redshift, of the angular correlation length of the residual foreground.

Using measurements of the power spectrum on the 10,000 mock catalogs (with and without imposing a mask), and comparing them with analytic predictions, we will show that it is possible to fully quantify the impact of the visibility mask on the power spectrum of biased tracers. This can be written as the sum of a pure cosmological term, a pure mask term, and a term involving their convolution. The same computation for the covariance matrix of the power spectrum is much more complicated, because the convolution term gives rise to a long list of mixed terms that are not easy to compute analytically, even in this idealized setting. This leads to the conclusion that the covariance matrix of the power spectrum cannot be simply written as the sum of a cosmology term and a mask term. The result we discuss in this chapter have been published in Colavincenzo et al. (2017).

## 5.2 Power spectrum of biased tracers in the presence of foregrounds

In this section we derive some simple analytical expressions describing the corrections to the power spectrum (and its covariance) of a galaxy sample defined by a given, nominal luminosity threshold $L_0$ when some foregrounds induce local variations $\delta L$ to the effective threshold that depend on the position on the sky.

### 5.2.1 Luminosity function and galaxy number density

Let us consider a flux-limited sample of galaxies with luminosity function $\bar{\Phi}(L)$, and let $\Phi(\mathbf{x},\, L)dL$ be the galaxy number density at the position $\mathbf{x}$ with luminosity between $L$ and $L+dL$ so that $\bar{\Phi}(L) \equiv \langle \Phi(\mathbf{x},\, L)\rangle$, with $\langle \dots \rangle$ (and the bar) denoting averages over a very large volume. In general, $\Phi(\mathbf{x},\, L)$ cannot be factorized into the product of a luminosity-dependent ($\bar{\Phi}(L)$) and a position-dependent function; if this were the case, the amplitude of clustering would be independent of luminosity. This means that $\Phi(\mathbf{x},\, L)$ encodes the information of luminosity-dependent bias.

If our determination of a galaxy luminosity is not affected by foregrounds, the galaxy number density of a sample of galaxy characterised by the lower luminosity threshold $L_0$ will be given by

$$n(\mathbf{x}; L_0) = \int_{L_0}^{\infty} dL\, \Phi(\mathbf{x}, L)\,, \tag{5.1}$$

while its mean value will be

$$\bar{n}(L_0) = \int_{L_0}^{\infty} dL\, \bar{\Phi}(L)\,, \tag{5.2}$$

We can characterize spatial fluctuations in the number density of galaxies of luminosity $L$ by means of the galaxy overdensity $\delta_\Phi(\mathbf{x},\, L)$ defined by the relation

$$\Phi(\mathbf{x},\ L) = \bar{\Phi}(L)\,[1 + \delta_\Phi(\mathbf{x},\, L)]\,. \tag{5.3}$$

For a sample of galaxies with luminosity threshold $L_0$ we define instead the overdensity $\delta(\mathbf{x};\, L_0)$ by means of the relation

$$n(\mathbf{x};\, L_0) = \bar{n}(L_0)\,[1 + \delta(\mathbf{x};\, L_0)]\,. \tag{5.4}$$

It follows that the two overdensities $\delta_\Phi(\mathbf{x},\, L)$ and $\delta(\mathbf{x};\, L_0)$ are related by

$$\bar{n}(L_0)\,\delta(\mathbf{x};\, L_0) = \int_{L_0}^{\infty} dL\,\bar{\Phi}(L)\,\delta_\Phi(\mathbf{x},\, L)\,. \tag{5.5}$$

In an observed sample, our measurement of the galaxy luminosity $L$ will be influenced by foregrounds. The two most obvious cases are galaxy extinction, that will decrease the observed flux, and zodiacal light, that will increase the sky noise; contamination by field stars or survey features (e.g., seeing conditions from the ground or solar aspect ratio from space), or survey features like modulations of 0-point of photometric calibration are other examples. For a fixed observed flux limit, the true limiting magnitude, and then the true density, will be modulated by these foregrounds, or by any *residual* of a foreground removal procedure. We expect, with some generality, such residuals to be highly correlated on the sky and we model here, in a very simple way, how this correlation affects the measurement of the galaxy power spectrum and its covariance.

We are interested in studying how a modulation of the intrinsic flux limit propagates to the observed galaxy density and its correlation functions. To this aim, we assume that the effect of residual foregrounds consists in changing locally the luminosity threshold $L_0$ by a quantity $\delta L(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector that defines the position on the sky. We further assume that such perturbations to $L_0$ are small, i.e. $\delta L/L_0 \ll 1$, and that, as residuals, they have vanishing spatial mean, that is $\langle \delta L(\boldsymbol{\theta})\rangle = 0$, The *observed* galaxy number density of a sample with nominal threshold $L_0$ will then be written as

$$\begin{aligned} n_{\rm obs}(\mathbf{x};\, L_0) &= \int_{L_0+\delta L(\boldsymbol{\theta})}^{\infty} dL\,\Phi(\mathbf{x},\, L)\,, \\ &= n(\mathbf{x};\, L_0) + \delta n(\mathbf{x};\, L_0)\,, \end{aligned} \tag{5.6}$$

where the second contribution on the r.h.s., defined as

$$\delta n(\mathbf{x};\, L_0) = \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\,\Phi(\mathbf{x},\, L)\,, \tag{5.7}$$

represents the correction due to the mask foreground residuals. We notice that:

$$\begin{aligned}\delta n(\mathbf{x};\, L_0) &= \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L)\, [1+\delta_\Phi(\mathbf{x},\, L)] \\ &= \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L) \\ &+ \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L)\, \delta_\Phi(\mathbf{x},\, L)\,. \end{aligned} \tag{5.8}$$

In the last equation, the first contribution on the r.h.s., describes the effect of the fluctuations in luminosity $\delta L(\boldsymbol{\theta})$ on the mean density and has therefore only an angular dependence. The second contribution, instead, accounts for the effect of the fluctuations in the threshold on the density perturbation and it is therefore expected to be subdominant (although not necessarily negligible).

It is important to stress that, even if $\langle \delta L(\boldsymbol{\theta}) \rangle = 0$, we cannot expect that the ensemble average of the correction vanishes, i.e. $\langle \delta n(\mathbf{x};\, L_0) \rangle = 0$, because of the nonlinear dependence on $\delta L(\boldsymbol{\theta})$. In particular, the mean of the second contribution of equation (5.8) vanishes due to the fact that density perturbations at high redshift are expected to be uncorrelated to any foreground residual, and that $\langle \delta_\Phi \rangle = 0$ by definition. The first term can be Taylor-expanded:

$$\int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L) \simeq \bar{\Phi}(L_0)\, \delta L + \frac{1}{2}\frac{d\bar{\Phi}}{dL}(L_0)(\delta L)^2 + \ldots \tag{5.9}$$

It is clear that $\langle \delta n(\mathbf{x};\, L_0) \rangle$ will be non-zero at the second-order in $\delta L(\boldsymbol{\theta})$. For these reasons, in the definition of the *observed* galaxy overdensity $\delta_{\rm obs}(\mathbf{x};\, L_0)$ given by the usual expression

$$n_{\rm obs}(\mathbf{x};\, L_0) \equiv \bar{n}_{\rm obs}(L_0) \left[1 + \delta_{\rm obs}(\mathbf{x};\, L_0)\right], \tag{5.10}$$

the mean value $\bar{n}_{\rm obs}(L_0)$ does not equal to the true mean density $\bar{n}(L_0)$.

### 5.2.2 The case of luminosity-independent bias

A possible analytical description of galaxy perturbations in the presence of residual foregrounds consists in Taylor-expanding the observed number density $n_{\rm obs}(L_0)$ in the threshold perturbations $\delta L(\boldsymbol{\theta})$. Instead, in what follows we will make the assumption that the quantity $\Phi(\mathbf{x},\, L)$ can be factorized as the product of a luminosity-dependent and a position-dependent function. In terms of equation (5.3):

$$\Phi(\mathbf{x},\, L) \simeq \bar{\Phi}(L)\, [1+\delta_\Phi(\mathbf{x})]\,. \tag{5.11}$$

This factorisation is clearly unphysical, as it amounts to neglecting any dependence of bias on luminosity, but it allows a great simplification of the calculations.

For instance, we have

$$n(\mathbf{x};\, L_0) = \bar{n}(L_0)\,[1+\delta(\mathbf{x})]\,, \tag{5.12}$$

with $\delta(\mathbf{x}) = \delta_\Phi(\mathbf{x})$. In particular, the corrections due to foregrounds become

$$\begin{aligned} \delta n(\mathbf{x};\, L_0) =& \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L) \\ &+ \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L)\, \delta(\mathbf{x}) \\ =&[1+\delta(\mathbf{x})] \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L)\,. \end{aligned} \tag{5.13}$$

Introducing now the notation

$$\delta n_{\text{mask}}(\boldsymbol{\theta};\, L_0) = \int_{L_0+\delta L(\boldsymbol{\theta})}^{L_0} dL\, \bar{\Phi}(L) \tag{5.14}$$

for the perturbations in the number density exclusively due to foregrounds (and therefore only dependent on the angle $\boldsymbol{\theta}$) we can write the *observed* galaxy number density, dropping the explicit dependence on $L_0$, as

$$n_{\text{obs}}(\mathbf{x}) = [1+\delta(\mathbf{x})]\,[\bar{n} + \delta n_{\text{mask}}(\boldsymbol{\theta})]\ . \tag{5.15}$$

The observed density contrast $\delta_{\text{obs}}(\mathbf{x})$, accounting for both cosmological perturbations and foregrounds effect is defined as

$$n_{\text{obs}}(\mathbf{x}) \equiv \bar{n}_{\text{obs}}\,[1+\delta_{\text{obs}}(\mathbf{x})]\,. \tag{5.16}$$

Noting that $\bar{n}_{\text{obs}} = \bar{n} + \langle \delta n_{\text{mask}}(\boldsymbol{\theta})\rangle$ we have then

$$\bar{n}_{\text{obs}}\,[1+\delta_{\text{obs}}(\mathbf{x})] = \bar{n}_{\text{obs}}\,[1+\delta(\mathbf{x})]\,[1-\delta_{\text{mask}}(\boldsymbol{\theta})]\,, \tag{5.17}$$

where we introduced the density contrast

$$\delta_{\text{mask}}(\boldsymbol{\theta}) \equiv \frac{\delta n_{\text{mask}}(\boldsymbol{\theta}) - \langle \delta n_{\text{mask}}(\boldsymbol{\theta})\rangle}{\bar{n}_{\text{obs}}}\,. \tag{5.18}$$

It is important to stress that $\delta_{\text{mask}}$ can be seen both as a function of the sky coordinate $\boldsymbol{\theta}$ and as a function of space coordinate $\mathbf{x}$, subject to the constraint of being constant along the lines of sight. Finally we can express the observed galaxy density contrast $\delta_{\text{obs}}(\mathbf{x})$ in terms of the actual density contrast $\delta(\mathbf{x})$ and the mask-induced relative density corrections $\delta_{\rm mask}(\mathbf{x})$ (expressed as a function of $\mathbf{x}$) as

$$\delta_{\rm obs}(\mathbf{x}) = \delta(\mathbf{x}) - \delta_{\rm mask}(\mathbf{x}) - \delta_{\rm mask}(\mathbf{x})\, \delta(\mathbf{x})\,. \tag{5.19}$$

Adopting the following convention for the Fourier Transform

$$\delta(\mathbf{k}) = \frac{1}{(2\pi)^3}\int d^3\mathbf{x}\; e^{i\mathbf{k}\cdot\mathbf{x}}\delta(\mathbf{x})\ , \tag{5.20}$$

the perturbed density contrast, eq. (5.19), in Fourier space will read:

$$\begin{aligned}\delta_{\rm obs}(\mathbf{k}) =& \delta(\mathbf{k}) - \delta_{\rm mask}\,(\mathbf{k}) - \int d^3\mathbf{q}\,\delta(\mathbf{q})\,\delta_{\rm mask}(\mathbf{k}-\mathbf{q}) \\ =& \delta(\mathbf{k}) - \delta_{\rm mask}(\mathbf{k}) - \delta_{\rm conv}(\mathbf{k}) \end{aligned} \tag{5.21}$$

Here we introduced $\delta_{\rm conv} \equiv \delta \otimes \delta_{\rm mask}$ as the convolution of $\delta(\mathbf{x})$ and $\delta_{\rm mask}(\mathbf{x})$.

It is important to stress that, in our simplified model we can write the observed density contrast $\delta_{obs}$, eq. (5.19), as a function of the independent quantities $\delta$ and $\delta_{mask}$.

### 5.2.3 Power spectrum

The real-space two-point correlation function for the observed overdensity $\delta_{\rm obs}(\mathbf{k})$ can be simply expanded as

$$\begin{aligned}\langle\delta_{\rm obs}(\mathbf{k}_1)\,\delta_{\rm obs}(\mathbf{k}_2)\rangle \ =\ & \langle\delta(\mathbf{k}_1)\,\delta(\mathbf{k}_2)\rangle + \langle\delta_{\rm mask}(\mathbf{k}_1)\delta_{\rm mask}(\mathbf{k}_2)\rangle + \\ & \langle\delta_{\rm conv}(\mathbf{k}_1)\,\delta_{\rm conv}(\mathbf{k}_2)\rangle\,, \end{aligned} \tag{5.22}$$

since $\langle\delta\,\delta_{\rm mask}\rangle = \langle\delta\,\delta_{\rm conv}\rangle = \langle\delta_{\rm mask}\,\delta_{\rm conv}\rangle = 0$ (they involve averages of either $\delta$ or $\delta_{\rm mask}$). The total, observed power spectrum $P_{\rm obs}(\mathbf{k})$ will therefore be given by:

$$P_{\rm obs}(\mathbf{k})\ =\ P_{\rm cosmo}(k)\ +\ P_{\rm mask}(\mathbf{k})\ +\ P_{\rm conv}(\mathbf{k})\,, \tag{5.23}$$

where $P_{\rm conv}$ is the convolution of the cosmological and mask power spectra:

$$P_{\rm conv}(\mathbf{k}) = \int d^3\mathbf{q}\,P_{\rm mask}(\mathbf{q})\,P_{\rm cosmo}(|\mathbf{k}-\mathbf{q}|)\ . \tag{5.24}$$

This term is of great importance because it couples the cosmological signal with the noise coming from the mask. Moreover, the integral generates scale mixing, thus transferring power among different scales. The procedure presented above is analogous to that of computing the effect of the variance of the window function of a survey on the cosmological power spectrum (Takada and Hu, 2013).

Introducing a simple estimator for the power spectrum such as

$$\hat{P}(k_i) = \frac{1}{N_{k_i}}\sum_{\mathbf{q}\in k_i}\delta_{\mathbf{q}}\delta_{-\mathbf{q}} \tag{5.25}$$

where $\mathbf{q} \in k_i$ denotes a sum over all modes for which $k = |\mathbf{k}|$ is in the $i$-th bin of size twice the fundamental frequency, $k_f = 2\pi/L$, of the box, we can define the power spectrum covariance matrix as

$$C_{ij} \equiv \mathrm{cov}[\hat{P}(k_i),\, \hat{P}(k_j)] = \langle \delta\hat{P}(k_i)\delta\hat{P}(k_j)\rangle \tag{5.26}$$

where $\delta\hat{P}(k_i) = \hat{P}(k_i) - \langle \hat{P}(k_i)\rangle$ is the deviation of the $\hat{P}(k)$, measured in a given realization, from its ensemble average. It is easy to see that the covariance of the observed power spectrum $P_{\mathrm{obs}}$, eq. (5.23), can then be written as

$$\begin{aligned}
C_{ij}^{\mathrm{obs}} &\equiv \mathrm{cov}(\hat{P}_{\mathrm{obs}}(k_i),\, \hat{P}_{\mathrm{obs}}(k_j)) \\
&= \mathrm{cov}[\hat{P}_{\mathrm{cosmo}}(k_i),\, \hat{P}_{\mathrm{cosmo}}(k_j)] + \mathrm{cov}[\hat{P}_{\mathrm{mask}}(k_i),\, \hat{P}_{\mathrm{mask}}(k_j)] + C_{ij}^{mixed} \\
&= C_{ij}^{cosm} + C_{ij}^{mask} + C_{ij}^{mixed}\,,
\end{aligned} \tag{5.27}$$

i.e., as a sum of the covariance of the cosmological power spectrum, the covariance of the mask power spectrum $P_{\mathrm{mask}}(\mathbf{k})$ plus a mixed term accounting for several contributions that can be written as a function of higher order correlation functions of the density field and of the mask:

$$\begin{aligned}
C_{ij}^{mixed} = \; & \langle \hat{P}_{\mathrm{conv}}(k_i)\hat{P}_{\mathrm{conv}}(k_j)\rangle - \langle \hat{P}_{\mathrm{conv}}(k_i)\rangle\langle \hat{P}_{\mathrm{conv}}(k_j)\rangle + \\
& \langle \hat{P}_{\mathrm{cosmo}}(k_i)\hat{P}_{\mathrm{conv}}(k_j)\rangle - \langle \hat{P}_{\mathrm{cosmo}}(k_i)\rangle\langle \hat{P}_{\mathrm{conv}}(k_j)\rangle + \\
& \langle \hat{P}_{\mathrm{mask}}(k_i)\hat{P}_{\mathrm{conv}}(k_j)\rangle - \langle \hat{P}_{\mathrm{mask}}(k_i)\rangle\langle \hat{P}_{\mathrm{conv}}(k_j)\rangle + \\
& \langle \hat{P}_{\mathrm{cosmo}}(k_i)\hat{G}(k_j)\rangle + \langle \hat{P}_{\mathrm{mask}}(k_i)\hat{G}(k_j)\rangle + \\
& \langle \hat{P}_{\mathrm{conv}}(k_i)\hat{G}(k_j)\rangle + \langle \hat{G}(k_i)\hat{G}(k_j)\rangle
\end{aligned} \tag{5.28}$$

where $\hat{G} = 2\delta_{\mathbf{q}}\delta_{\mathrm{mask},\mathbf{q}} - \delta_{\mathbf{q}}\delta_{\mathrm{conv},\mathbf{q}} + \delta_{\mathrm{mask},\mathbf{q}}\delta_{\mathrm{conv},\mathbf{q}}$, with $\mathbf{q} \in k_i$. Eq. (5.28) shows all the terms that come from the coupling of the cosmological signal with the mask. The mixed terms

$$\begin{aligned}
C_{ij}^{obs} \supset \quad & \langle \hat{P}_{\mathrm{cosmo}}(k_i)\hat{G}(k_j)\rangle + \langle \hat{P}_{\mathrm{mask}}(k_i)\hat{G}(k_j)\rangle + \\
& \langle \hat{P}_{\mathrm{conv}}(k_i)\hat{G}(k_j)\rangle + \langle \hat{G}(k_i)\hat{G}(k_j)\rangle
\end{aligned} \tag{5.29}$$

are written in implicit form. We recall that

$$\hat{G} = \delta_{\mathbf{q}}\delta_{\mathrm{mask},\mathbf{q}} - \delta_{\mathbf{q}}\delta_{\mathrm{conv},\mathbf{q}} + \delta_{\mathrm{mask},\mathbf{q}}\delta_{\mathrm{conv},\mathbf{q}}\ , \tag{5.30}$$

Inserting eq. (5.30) into eq. (5.29) we end up with four mix terms that we call $C_{\mathrm{mix}}^{i}$, with $i = 1...4$. Let's start with the first contribution that come from the first line of eq. (5.29):

$$\begin{aligned}
C_{\mathrm{mix}}^{1} = \; & \frac{1}{N_{k_i}N_{k_j}} \sum_{\mathbf{q}\in k_i}\sum_{\mathbf{p}\in k_j} \langle \delta_{\mathbf{p}}\delta_{-\mathbf{p}}\delta_{\mathrm{mask},\mathbf{q}}\delta_{\mathrm{conv},-\mathbf{q}} - \delta_{\mathbf{p}}\delta_{-\mathbf{p}}\delta_{\mathbf{q}}\delta_{\mathrm{conv},-\mathbf{q}} - \\
& \delta_{\mathbf{p}}\delta_{-\mathbf{p}}\delta_{\mathbf{q}}\delta_{\mathrm{mask},-\mathbf{q}}\rangle = \\
& \frac{1}{N_{k_i}N_{k_j}} \sum_{\mathbf{q}\in k_i}\sum_{\mathbf{p}\in k_j} \int d^3s\ \langle \delta_{\mathbf{p}}\delta_{-\mathbf{p}}\delta_{-\mathbf{q}}\rangle\langle \delta_{\mathrm{mask},\mathbf{q}}\delta_{\mathrm{mask},-\mathbf{q}}\rangle + \mathrm{cc}
\end{aligned} \tag{5.31}$$

for all $\mathbf{q}$ different from zero, with cc for complex conjugate. For the same reason $C^2_{\mathrm{mix}} = 0 = C^3_{\mathrm{mix}}$. The only non zero contribution is:

$$\begin{aligned} C^4_{\mathrm{mix}} &= \frac{1}{N_{k_i} N_{k_j}} \sum_{\mathbf{q}\in k_i} \sum_{\mathbf{q}\in k_j} [\langle \delta_{\mathrm{mask},\mathbf{q}} \delta_{\mathrm{mask},\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{q}} \delta_{\mathrm{conv},-\mathbf{p}} - \\ & \delta_{\mathrm{mask},\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{q}} \delta_{\mathrm{conv},-\mathbf{p}} - \delta_{\mathrm{mask},\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{q}} \delta_{\mathrm{mask},-\mathbf{p}} - \\ & \delta_{\mathbf{q}} \delta_{\mathrm{mask},\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{q}} \delta_{\mathrm{conv},-\mathbf{p}} + \delta_{\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{q}} \delta_{\mathrm{conv},-\mathbf{p}} + \\ & \delta_{\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{mask},-\mathbf{p}} \delta_{\mathrm{conv},\mathbf{q}} - \delta_{-\mathbf{q}} \delta_{\mathrm{mask},\mathbf{q}} \delta_{\mathrm{mask},\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{p}} + \\ & \delta_{\mathbf{q}} \delta_{\mathrm{mask},-\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{conv},-\mathbf{p}} + \delta_{\mathbf{q}} \delta_{\mathrm{mask},-\mathbf{q}} \delta_{\mathbf{p}} \delta_{\mathrm{mask},-\mathbf{p}} \rangle] = \\ & \frac{1}{N_{k_i} N_{k_j}} \sum_{\mathbf{q}\in k_i} \sum_{\mathbf{p}\in k_j} \Big\{ \int d^3 s_1 d^3 s_2 \ \langle \delta_{-\mathbf{q}} \delta_{-\mathbf{p}} \rangle \langle \delta_{\mathrm{mask},\mathbf{q}} \delta_{\mathrm{mask},\mathbf{p}} \delta_{\mathrm{mask},-\mathbf{q}+\mathbf{s}_1} \delta_{\mathrm{mask},-\mathbf{p}+\mathbf{s}_2} \rangle + \\ & \qquad \int d^3 s_1 d^3 s_2 \ \langle \delta_{\mathbf{q}} \delta_{\mathbf{p}} \delta_{-\mathbf{q}} \delta_{-\mathbf{p}} \rangle \langle \delta_{\mathrm{mask},-\mathbf{q}+\mathbf{s}_1} \delta_{\mathrm{mask},-\mathbf{p}+\mathbf{s}_2} \rangle + \\ & \qquad \langle \delta_{\mathbf{q}} \delta_{\mathbf{p}} \rangle \langle \delta_{\mathrm{mask},-\mathbf{q}} \delta_{\mathrm{mask},-\mathbf{p}} \rangle \Big\} \end{aligned} \tag{5.32}$$

As we can see from the previous relations, the mixed terms are convolutions of high order correlators of both cosmological and mask fields.

It is possible to further expand the expressions, considering that all the 4-point correlators can be written in the form:

$$\langle \delta_1 \delta_2 \delta_3 \delta_4 \rangle = \langle \delta_1 \delta_2 \delta_3 \delta_4 \rangle_{\mathrm{connected}} + \langle \delta_1 \delta_2 \rangle \langle \delta_3 \delta_4 \rangle + \mathrm{perm.}\ . \tag{5.33}$$

We expect that the cosmological connected part are equal to zero, but we cannot assume that the same is valid for the mask connected part.

There will be a scale at which perturbations $\delta$ and $\delta_{mask}$ are of the same order of magnitude, in which case there is no obvious reason why mixing terms should be small. A full analytical computation of the additional covariance contribution in $C^{mixed}_{ij}$ is discouraging even in the context of our simple model. Instead, to quantify the various terms we will resort to a numerical assessment taking advantage of the large number of DM halo catalogs produced for this project, to which we will add the effect of a mask as explained in the next section.

In section (5.2) we made the simplifying assumption that the quantity $\Phi(\mathbf{x}; L)$ can be factorized into luminosity-dependent and position-dependent functions (equation 5.11). This condition simplifies the calculations, but it is clearly an approximation sinces it implies luminosity-independent bias. It is not the aim of this paper to study this other case, but it is worth to show what happens when we drop our approximation.

Let us consider the number density of halos with mass greater than the nominal threshold $M_0$ to be given by

$$n(\mathbf{x}, M_0) = \int_{M_0}^{\infty} dM \ \Phi(\mathbf{x}; M) \ , \tag{5.34}$$

then the observed number density, after the mask perturbation is

$$n^{obs}(\mathbf{x};M_0)=\int_{M_0\,[1+A(\boldsymbol{\theta})]}^{\infty}dM\;\Phi(\mathbf{x};M)\,. \tag{5.35}$$

For small perturbations to the mass threshold, i.e. $\sigma_A \lesssim 1$, we can Taylor-expand $n^{obs}$ with respect to $A(\boldsymbol{\theta})$ to get

$$\begin{aligned} n^{obs}(\mathbf{x};M_0) &= n(\mathbf{x},M_0)+\frac{\partial n^{obs}}{\partial A}\Big|_{A=0}A \\ &+\frac{1}{2}\frac{\partial^2 n^{obs}}{\partial A^2}\Big|_{A=0}A^2+\mathcal{O}(A^3)\ . \end{aligned} \tag{5.36}$$

that we can formally express as

$$\begin{aligned} n^{obs}(\mathbf{x},M_0) &= n(\mathbf{x},M_0)+n^{(1)}(\mathbf{x},M_0)A(\boldsymbol{\theta}) \\ &+n^{(2)}(\mathbf{x},M_0)A^2(\boldsymbol{\theta})+\mathcal{O}(A^3)\ , \end{aligned} \tag{5.37}$$

where

$$n^{(1)}(\mathbf{x},M_0) \equiv \frac{\partial n}{\partial(\ln M_0)} \tag{5.38}$$

$$n^{(2)}(\mathbf{x},M_0) \equiv \frac{1}{2}\frac{\partial^2 n}{\partial(\ln M_0)^2}\ . \tag{5.39}$$

We can now expand in $A$ the halo density contrast, defined as

$$\delta_h^{obs}(\mathbf{x},M_0)\equiv\frac{n^{obs}(\mathbf{x},M_0)}{\bar{n}^{obs}(\mathbf{x},M_0)}-1\ , \tag{5.40}$$

and obtain

$$\begin{aligned} \delta_h^{obs} &= \delta_h(1-\frac{1}{2}C_2\sigma_A^2)+C_1A+(C_1\delta_h+\tilde{\epsilon})A \\ &+\frac{1}{2}C_2(A^2-\sigma_A^2)+\frac{1}{2}(C_2\delta_h+2C_1\tilde{\epsilon}+\tilde{\eta})A^2 \\ &+\mathcal{O}(A^3)\ . \end{aligned} \tag{5.41}$$

where $\delta_h$ is the cosmological halo density contrast and where we defined

$$C_1(M_0) \equiv \frac{\partial\ln\bar{n}}{\partial\ln M_0} \tag{5.42}$$

$$C_2(M_0) \equiv \frac{M_0^2}{\bar{n}}\frac{\partial^2\bar{n}}{\partial M_0^2} \tag{5.43}$$

$$\tilde{\epsilon} \equiv M_0\frac{\partial\delta_h}{\partial M_0} \tag{5.44}$$

$$\tilde{\eta} \equiv M_0^2\frac{\partial^2\delta_h}{\partial M_0^2}\ . \tag{5.45}$$

It is easy to see that if eq. (5.11) holds, then $\tilde{\eta} = \tilde{\epsilon} = 0$ and the density contrast reduces to the form of eq. (5.19). From equation 5.41, one can show that the two-point correlation function in Fourier space is given by

$$\begin{aligned}\langle \delta^{obs}_{\mathbf{k}_1} \delta^{obs}_{\mathbf{k}_2} \rangle &= \delta_D(\mathbf{k}_{12}) P_{\delta\delta}(k_1)(1 - C_2\sigma^2_A) + \langle A_{\mathbf{k}_1} A_{\mathbf{k}_2} \rangle \\ &+ \int d^3q [P_{\delta\delta}(q) + 2C_1 P_{\delta\tilde{\epsilon}}(q) + C_1^2 P_{\tilde{\epsilon}\tilde{\epsilon}}(q)] \\ &\times \langle A_{\mathbf{k}_1 - \mathbf{q}} A_{\mathbf{k}_1 + \mathbf{q}} \rangle \\ &+ \frac{1}{2}[C_2 P_{\delta\delta}(k_1) + 2C_1 P_{\delta\tilde{\epsilon}}(k_1) + C_1^3 P_{\tilde{\epsilon}\tilde{\eta}}(k_1)] \\ &\times \int d^3q \langle A_{\mathbf{q}} A_{\mathbf{k}_{12} - \mathbf{q}} \rangle \\ &+ (\mathbf{k}_1 \leftrightarrow \mathbf{k}_1) + \mathcal{O}(A^3) \ . \end{aligned} \tag{5.46}$$

Since we are interested in a comparison with eq. (5.23), using eq. (5.21) with $\delta^{obs}$ from eq. (5.41) we obtain the monopole of the observed power spectrum to be given by

$$\begin{aligned} P_0^{obs} &= P_{\delta\delta}(1 - C_2\sigma^2_A) + P_{AA,0}(k) \\ &+ \int d^3q [P_{\delta\delta}(q) + 2C_1 P_{\delta\tilde{\epsilon}}(q) + C_1^2 P_{\tilde{\epsilon}\tilde{\epsilon}}(q)] P_{AA,0}(|\mathbf{q} - \mathbf{k}|) \\ &+ [C_2 P_{\delta\delta}(k_1) + 2C_1 P_{\delta\tilde{\epsilon}}(k_1) + C_1^3 P_{\tilde{\epsilon}\tilde{\eta}}(k_1)]\sigma^2_A \\ &+ \mathcal{O}(A^3) \ , \end{aligned} \tag{5.47}$$

where

$$P_{AA,0}(k) = \frac{k_f^3}{V_p(k)} \int_k d^3\mathbf{q} \langle |A_{\mathbf{q}}|^2 \rangle \tag{5.48}$$

is the monopole of the mask power spectrum. The main difference with (5.23), is given by the presence of the power spectra $P_{\delta\tilde{\epsilon}}$, $P_{\delta\tilde{\eta}}$ and $P_{\tilde{\eta}\tilde{\eta}}$, all dependent on derivatives of the density contrast $\delta_h$. Considering a local bias expansion as $\delta_h = \sum_n b_n(M)\delta^n_m$ these additional terms could be expressed in terms of derivatives of the halo bias functions $b_n(M)$.

A modulation of the mass threshold implies a different halo selection, and then a different bias. Because the relation between bias and threshold halo mass is not linear, we expect that the bias of halos subject to a mask $A(\boldsymbol{\theta})$ will be different from the bias of unmasked halos, even when $\langle A(\boldsymbol{\theta}) \rangle = 0$. As a consequence, we expect the masked and unmasked catalogs not to match at small scales, as they do in Figure 5.3.

We will not go into the computation details of the full covariance matrix, because we can directly use mocks to see the effect of introducing the halo bias.

For the following section, in which we show the results of our analysis, we consider the esier case of luminosity-independent bias.

## 5.3 Simulated catalogs

### 5.3.1 Cosmological catalogs

As mentioned in the Introduction, our choice is to use DM halos in place of galaxies as biased tracers. Moreover, we will use the mass $M$ of the DM halo in place of the galaxy luminosity $L$. In particular, the nominal mass threshold, i.e. the minimal mass defining the halo sample (in absence of foregrounds) will be denoted as $M_0$, corresponding to the $L_0$ of the previous section. This is equivalent to applying a minimal HOD model (Cooray and Sheth, 2002) with one galaxy per halo and a linear relation between halo mass and luminosity. This is known to be unrealistic, but we considered this approximation proper for the idealised case presented in this paper.

The simulated catalogs we used for all the measurements are DM halo catalogs obtained with the approximate method PINOCCHIO (Monaco et al., 2002, 2013, see (Munari et al., 2017) for a review of approximate methods). We use the latest version of the code, V4.1, presented in Munari et al. (2017), where displacements were computed with LPT up to the third order, resulting in a sizable improvement of the predicted power spectrum: the wavenumber at which the prediction of $P(k)$ drops by 10 per cent, with respect to an N-body simulation run on the same initial conditions, increases from $k = 0.1\ h^{-1}\,\mathrm{Mpc}$ to $\sim 0.3 - 0.5\ h^{-1}\,\mathrm{Mpc}$ at redshift 0 or 1. This lack of accuracy is not relevant for the present analysis, that is mostly focused on large scales.

We generated 10,000 realizations of a cubic 1500 $h^{-1}\,\mathrm{Mpc}$ box, sampled with $1000^3$ particles. This is, to out knowledge, the largest set of catalogs of DM halo catalogs ever presented. The cosmological parameters are $\Omega_m = 0.285$, $\Omega_\Lambda = 0.715$, $\Omega_b = 0.044$, $h = 0.695$ and $\sigma_8 = 0.285$. We used outputs at $z = 1$, where, as mentioned in the introduction, it is possible to have observational access to large scales and PINOCCHIO is more accurate. The particle mass is $M_\mathrm{p} = 2.67 \times 10^{11}\ h^{-1} M_\odot$.

We will consider a halo sample defined by a mass threshold $M_0 = 50 \times M_\mathrm{p}$. With this choice we have approximately $500,000$ halos in each catalog, corresponding to a number density of $1.5 \times 10^{-4}\ h^3\,\mathrm{Mpc}^{-3}$.

### 5.3.2 Implementation of the mask toy model

One side of the simulation box will be serving as the field-of-view in a distant-observer approximation. For simplicity we model patches in the sky characterised by a constant, uniform foreground residual as square tiles covering the box side mentioned above. An "effective" threshold for halo detection will then be defined, as a correction for the nominal one $M_0$, for the whole volume (along the line-of-sight) behind a given tile. Fig. 5.1 provides a pictorial representation of our toy model.

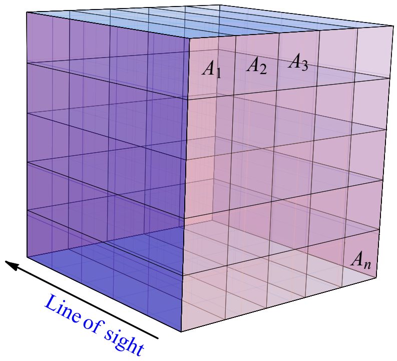


Figure 5.1: Every simulation box containing the halo catalogs is assumed to represent a cosmological volume in the distant-observer approximation. Patches of equal foreground error residual are modelled as square tiles covering the field-of-view, corresponding to one side of the box.

For each halo catalog, we produce a different foreground mask consisting of a correction to the mass threshold $M_0$ for each tile across the field-of-view. We will describe the relative variation of threshold as the two-dimensional quantity

$$A(\boldsymbol{\theta}) \equiv \frac{\delta M(\boldsymbol{\theta})}{M_0}\,. \tag{5.49}$$

Since $A(\boldsymbol{\theta})$ represents the effect of a residual foreground, we will assume $\langle A(\boldsymbol{\theta})\rangle = 0$, the bracket representing ensemble averages.

We divide the sky plane into square tiles of length $l$, within which $A$ is kept constant, so we can write:

$$A(\boldsymbol{\theta}) = \sum_{i=1}^{N_t} A_i \Theta_i(\boldsymbol{\theta}) \tag{5.50}$$

where the function $\Theta_i(\boldsymbol{\theta}) = 1$ if the angular position $\boldsymbol{\theta}$ falls inside the $i$-th tile and zero otherwise, and $N_t$ is the total number of tiles. The coefficients $A_i$ are assigned as independent random numbers, drawn from a Gaussian distribution with standard deviation $\sigma_A$, so $A$-values in nearby tiles are uncorrelated. The length $l$ therefore represents the physical correlation length induced by the foreground residual; it will correspond to the projection, at the observation redshift, of an angular correlation scale.

The production of masked halo samples proceeds as follows. First, the DM halo masses provided by PINOCCHIO are modified so as to be continuous. Indeed, the discreteness due to the particle mass can be of the same order of the correction to the mass threshold $\delta M$, leading to spurious effects in the number density that would affect the covariance matrix. This procedure is applied to all halos with more than 30 particles (we altogether ignore smaller groups); this is smaller than the 50 particles mass cut mentioned above, because the mask modulation will decrease the mass cut in ~half of the sky tiles. Calling $\alpha$ the logarithmic slope of the DM halo mass function around $M_0$ (and computing $\alpha$ from the avaraged mass function of the 10,000 mocks), the halo mass $M$ of a halo made of $N$ particles ($M_{\rm old} = NM_{\rm p}$) is modified as follows:

$$M_{\rm new} = M_{\rm old} \left\{ 1 + r \left[ \left( \frac{N+1}{N} \right)^{\alpha} - 1 \right] \right\}^{1/\alpha} \tag{5.51}$$

where $r$ is a random number between zero and one. Second, in order to remove the mass dependence of halo bias, that invalidates equation (5.11), in each mock catalog halo masses are randomly shuffled among all the halos, thus preserving the halo mass function. In this way, imposing a mass cut is equivalent to a sparse sampling, and halos with different mass cuts will have a similar clustering amplitude. Finally, the catalog is selected by applying the position-dependent mass cut $M_0 + \delta M(\boldsymbol{\theta}) = M_0 \, [1 + A(\boldsymbol{\theta})]$.

### 5.3.3 Analytical predictions

In terms of the adimensional field $A(\boldsymbol{\theta})$, and adopting now the halo mass as proxy for the galaxy luminosity, we can rewrite eq. (5.18)

$$\begin{aligned} \delta_{\rm mask}(\mathbf{x}) &= \frac{1}{\bar{n}_{obs}} \int_{M_0\,[1+A(\boldsymbol{\theta})]}^{M_0} dM \, \bar{\Phi}(M) - \frac{\langle \delta n_{mask} \rangle}{\bar{n}_{obs}} \\ &= -\frac{M_0 \, \bar{\Phi}(M_0)}{\bar{n}_{obs}} \, A(\boldsymbol{\theta}) + \mathcal{O}(A^2) - \frac{\langle \delta n_{mask} \rangle}{\bar{n}_{obs}} \, , \end{aligned} \tag{5.52}$$

showing that the field $A(\boldsymbol{\theta})$ represents, modulo a $-$ sign, the overdensity due to the mask, $\delta_{\rm mask}$, up to a multiplicative constant.

From the definiton of $A(\boldsymbol{\theta})$, eq. (5.50), it is simple to derive explicitly its Fourier transform

$$A_{\mathbf{k}} = \frac{L_{\rm box} l^2}{(2\pi)^3} \sum_i A_i e^{ik_x x_i} e^{ik_y y_i} j_0 \left( \frac{k_x l}{2} \right) j_0 \left( \frac{k_y l}{2} \right) \delta^K_{k_z,0} \tag{5.53}$$

where $j_0(x)$ is the zeroth-order Bessel function and $\delta^K$, the Kronecker symbol. The power spectrum of $A(\boldsymbol{\theta})$ is given by

$$P_A(k_x, k_y, k_z) = \frac{L_{\rm box} l^2}{(2\pi)^3} k_f \sigma_A^2 j_0^2 \left( \frac{k_x l}{2} \right) j_0^2 \left( \frac{k_y l}{2} \right) \delta_D(k_z) \tag{5.54}$$

where we took the continuum limit by replacing $\delta^K(\mathbf{k})/k_f^3 \to \delta_D(\mathbf{k})$ for $V \to \infty$. This term scales as $P_A \propto l^2 \sigma_A^2$, so that at $k < 2\pi/l$ it will grow not only, as expected, with the variance of the residuals but also with the correlation length $l$. Finally, it is equally simple to write down the convolution of the power spectrum $P_A(\mathbf{k})$ with a power spectrum $P(k)$ as

$$\begin{aligned} P_{\rm conv,A}(\mathbf{k}) &= \int d^3q\, P_A(\mathbf{q})\, P_{\rm cosmo}(|\mathbf{k}-\mathbf{q}|) \\ &= \frac{\sigma_A^2 l^2 L_{\rm box} k_f}{(2\pi)^2} \int_{k_f}^{k_{max}} dq_x\, dq_y\, j_0^2\left(\frac{q_x l}{2}\right) j_0^2\left(\frac{q_y l}{2}\right) \\ &\quad \times P_{\rm cosmo}\left(\sqrt{(k_x-q_x)^2+(k_y-q_y)^2+k_z^2}\right). \qquad (5.55) \end{aligned}$$

This term, as the previous one, scales as $\propto l^2\sigma_A^2$. The integral in equation (5.55) can be computed numerically, once $P_{\rm cosmo}(k)$ is given. The theoretical prediction for the mask power spectrum makes it possible to analytically compute the convolution term of eq. (5.24), at least at linear order in $A(\boldsymbol{\theta})$.

$P_A(k)$ and $P_{\rm conv,A}(k)$ represent, up to a multiplicative factor, analytical predictions, respectively, for $P_{\rm mask}(k)$ and $P_{\rm conv}(k)$, since, as we will see, corrections due to higher order terms in $A(\boldsymbol{\theta})$ are small. Of particular interest is the relation between the variance of the mask-induced overdensity, $\sigma^2_{\rm mask} \equiv \langle\delta^2_{\rm mask}\rangle$, and the variance of the relative error on the mass threshold, $\sigma_A^2$. To first order in $A$ we have

$$\sigma^2_{\rm mask} \simeq \frac{M_0^2\bar{\Phi}^2(M_0)}{\bar{n}^2(M_0)}\sigma_A^2\,, \qquad (5.56)$$

where $\bar{\Phi}(M_0)$ represents now the halo mass function and $\bar{n}(M_0)$ the number density of objects above the threshold.

### 5.3.4 Power spectrum estimators

The power spectrum of the halo catalogs was measured using the estimator of Sefusatti et al. (2016), that provides a sophisticated procedure to minimize the impact of aliasing coming from the estimate of density of a set of particles on a $600^3$ grid points. All the $k$-bins are multiples of the fundamental frequency of the box, $k_f = 2\pi/L = 0.041\ h^{-1}\,$Mpc, while the Nyquist frequency is $k_{Nyq} = N_g k_f/2 = 1.256\ h^{-1}\,$Mpc, where $N_g = 600$ is the grid size. The shot noise contribution has always been subtracted. Given that the cosmological density field is isotropic in our case, we present here results for the monopole of the power spectrum; clearly the mask will induce non-zero multipoles, that will contaminate the redshift space distorsion signal; we do not address this point in this paper.

The density field for the estimation of the mask power spectrum, $P_{\rm mask}(k)$, was obtained directly from the two-dimensional field $A(\boldsymbol{\theta})$ as follows: $\delta_{\rm mask}$ is assumed to be equal to the constant value $A_i$ along the whole i-th tile (fig. 5.1), and the so-defined density field is Fourier-transformed without

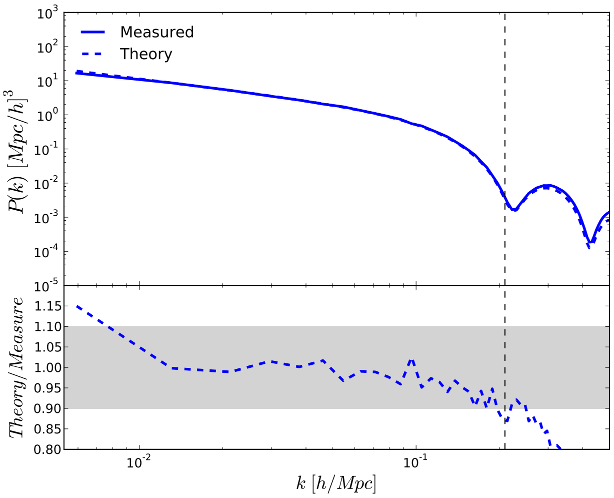


Figure 5.2: Comparison between measured (continuous line) and analytic (dashed line) monopole of the mask power spectrum $P_{\rm mask}(k)$. The black dashed vertical line corresponds to $k = 2\pi/l$. The lower panel gives the residuals.

involving a density estimate on a set of points. As a consequence, the estimation of $P_{\rm mask}(k)$ is not affected by shot noise. As a consistency test, we show in figure (5.2) the monopole of the mask power spectrum, computed analytically from eq. (5.54) and numerically from 10,000 realisations of the mask $A(\boldsymbol{\theta})$. The two results are remarkably consistent at large scales, while at small scales the numerical determination shows some overestimate with respect to the analytic one; this is likely due to sampling effects, but such small differences in the range where the term drops are not a concern for what follows.

We expect the mask to affect large scales because of its own geometry: since points behind a given tile are subject to the same effective threshold $L_0 + \delta L$, they will present some level of induced correlation, even when their separation along the line-of-sight is very large.

## 5.4 Results

The variance of $\delta_{\rm mask}$, $\sigma^2_{\rm mask} \equiv \langle \delta^2_{\rm mask} \rangle$, gives the magnitude of the effect of the mask on the observed density, and we are interested in the scale range where it is comparable to the variance of the density perturbations $\delta$ (of the same order of the tile length and of the BAO scale). In fact, the limit $\sigma^2_{\rm mask} \ll 1$ corresponds to a very good knowledge of the foregrounds, and therefore to a negligible effect of possible residuals, while the opposite limit of large $\sigma^2_{\rm mask}$ should describe a situation of poor knowledge of foregrounds that we are expected to avoid.

We will consider the two values $\sigma_A = 0.05$ and 0.2 (or errors of $\sim 0.05$ and $\sim 0.20$ magnitudes), corresponding respectively to $\sigma_{mask} = 0.07$ and 0.28. The first value may be a good order of magnitude for a residual foreground (this point will be addressed later), while the second value is very pessimistic and is used to emphasize the effects of foreground removal. We will also assume two different values for the size of the tiles, $l = 30$ and 100 $h^{-1}$ Mpc. At $z = 1$ and for the cosmological parameters given above, these comoving scales subtend angles of 0.74 and 2.5 degrees. We also tested the case $\sigma_A = 0.01$; the effect of the mask (for this toy case) is entirely negligible for both the power spectrum and its covariance, so we will not show this case.

### 5.4.1 Power spectrum

Figure 5.3 shows the power spectra for all the four considered cases, with $l = 30$ and 100 $h^{-1}$Mpc (top and bottom panels) and $\sigma_A = 0.05$ and 0.2 (left and right panels). Each panel in the figure is composed by two plots. The upper one shows the monopole of the power spectrum. The coloured lines give the contributions to the power spectrum of eq. (5.23) (denoted by a black line): the blue line is the cosmological power spectrum, $P_{\rm cosmo}(k)$, the red line represents the pure mask contribution, $P_{\rm mask}(k)$, while the magenta line is the convolution term $P_{\rm conv}$ (eq. (5.24)). Solid lines are obtained by measuring the 10,000 mocks as explained in Section 5.3.4; the convolution term is computed by difference:

$$P_{\rm conv} = P_{\rm obs} - P_{\rm cosmo} - P_{\rm mask} \ . \tag{5.57}$$

The red and magenta dashed lines represent the theoretical prediction for the power spectrum of the mask, eq. (5.54), and for the cross-term, $P_{\rm conv}$. The latter is obtained from $P_A(k)$ with the multiplicative factor from eq. (5.52). In the lower plot of each panel we report the ratio between the components and the observed power spectrum, to highlight the relative size of each contribution. In this case we only report the quantities measured from mocks. Vertical lines in all plots mark $k = 2\pi/l$, where the Fourier transform of the mask starts to oscillate. This is equal to $\sim 0.2$ $h$/Mpc for $l = 30$ $h^{-1}$ Mpc, and to $\sim 0.06$ $h$/Mpc for $l = 100$ $h^{-1}$ Mpc.

As anticipated in Figure 5.2, the contribution of the mask power spectrum (the red line in the plots) is important at large scales, $k < 2\pi/l$. Its relevance depends on $l\sigma_A$: for $l = 30$ $h^{-1}$ Mpc it is found to level at about 1 per cent for $\sigma_A = 5\%$ and in excess of 10 per cent for $\sigma_A = 20\%$, while for $l = 100$ $h^{-1}$ Mpc its importance gets increasingly large at large scales even for $\sigma_A = 5\%$, while it dominates at $k < 0.03$ $h$/Mpc for the higher variance case. At $k > 2\pi/l$ the mask is typically negligible, even though the first peaks still get above the 1 per cent level in the high variance case. It is useful to recall

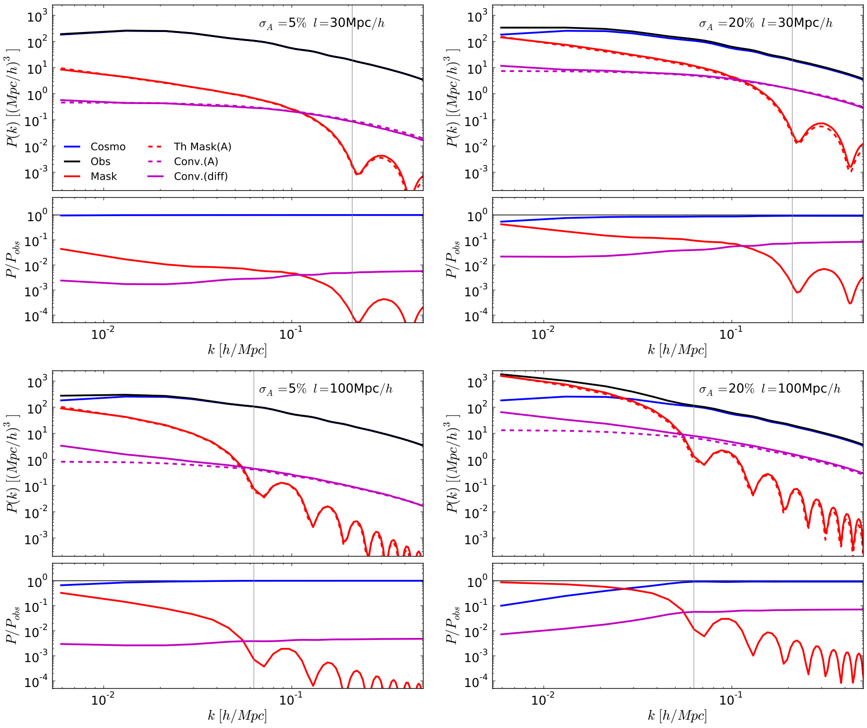


Figure 5.3: Averaged power spectra of mock catalogs. Top panels: tile size $= 30 \times 30$ Mpc$^2/h^2$. Bottom panels: tile size $= 100 \times 100$ Mpc$^2/h^2$. Left panels: $\sigma_A = 5\%$, right panels: $\sigma_A = 20\%$. In all panels the solid lines denote respectively the monopole of total power spectrum (black), cosmological power spectrum (blue), mask power spectrum (red) and convolution term (magenta). These are all measured from catalogs, the last being determined by difference. The dashed red and magenta lines are the theoretical predictions for the mask power spectrum (eq. (5.54)) and the convolution term (eq. (5.55)). The vertical thin lines mark $k = 2\pi/l$. The lower plots show the ratio of the various components with respect to the total power spectrum; in this case we only show the measurements from the mocks.

that these oscillations are simply the result of our toy model, and therefore they do not necessarily have a real physical meaning.

The contribution of the cross-term does not fall down at small scales, but remains at a fraction of the cosmological power spectrum. At small scales, this fraction is always below the 1 per cent level for $\sigma_A = 5\%$, but is found to be $\sim 10$ per cent in the higher variance case; notably, this fraction scales with $\sigma_A$ but not with $l$. The relatively larger contribution of $P_{\rm conv}$ with respect to $P_{\rm mask}$ for $k > 2\pi/l$ is the result of the transfer of large-scale power operated by the convolution and does not show the rather artificial oscillations of the latter. At large scales the convolution term is always overtaken by $P_{\rm mask}$, so it never becomes dominant, its relative weight ranging from tenths of per cent to few per cent. Here we notice a small discrepancy between the theoretical prediction for $P_{\rm conv}$ and the measured one at large scales that could be due to the small difference between the theoretical prediction for the mask power spectrum and the measured one at large scales (fig. 5.2).

The agreement of analytic and measured contributions allows us to be confident in the control of the total power spectrum. In the analysis of the power spectrum covariance we will only use the quantities measured from the 10,000 mock catalogs.

As a concluding remark, the mask power spectrum $P_{\rm mask}(k)$ can easily be important at large scales even when foreground removal is controlled to within a few per cent. The reason lies in the scaling with $(l\sigma_A)^2$: a highly correlated foreground will anyway give a significant contribution to large scales. Conversely, the convolution term gives a roughly constant relative contribution to the power spectrum, that is typically negligible if the uncertainty in the foreground removal is controlled at the few per cent level, but can become important in more pessimistic cases. Because the mask creates power on large scales, within this toy model one could conclude that the BAO scale should be safe at the per cent level if good control, to the few per cent level, is achieved on foreground removal. We will get back to this point in the Conclusions.

### 5.4.2 Covariance

Figure 5.4 shows the variance of the measured power spectra (all the different components), divided by the measured power spectrum squared, $\Delta P^2/P_{\rm obs}^2$. For a purely cosmological Gaussian field it would correspond simply to the inverse of the number of available $k$-modes $1/N_{\mathbf{k}}$; in fact, we checked that our cosmological term is very similar to the Gaussian prediction. Here, the magenta curve represents the mixed mask-cosmology contribution; we recall that $C_{ij}^{\rm mixed}$ is not simply the covariance of the convolution $P_{\rm conv}$, but includes a variety of different combinations of cosmological and error residuals perturbations, in addition to shot-noise. The mixed term is obtained as the

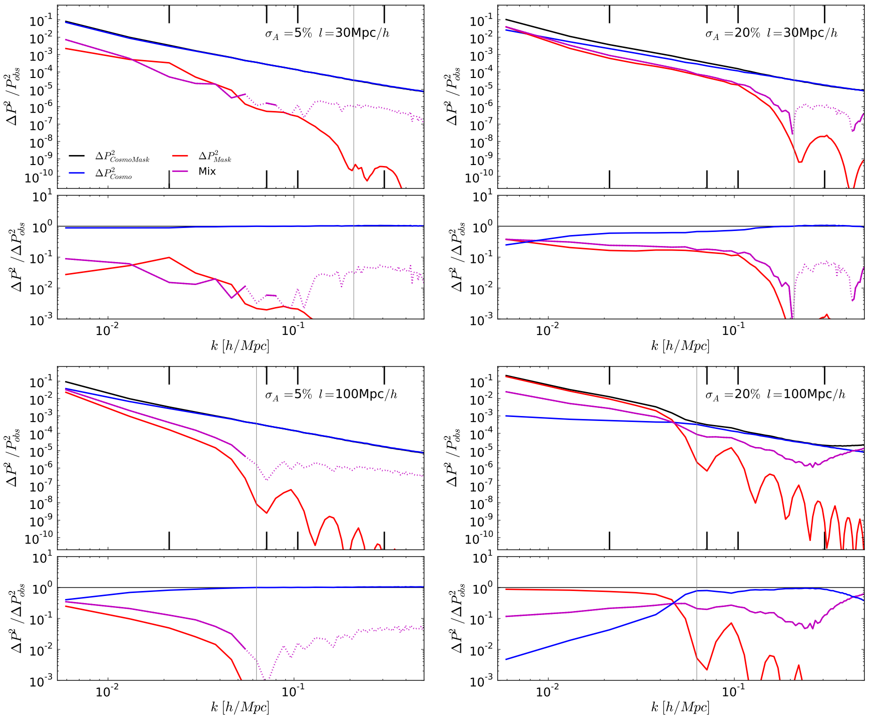


Figure 5.4: Averaged diagonal components of the covariance matrix of mock catalogs. Top panels: tile size = 30 × 30 $h^{-1}$ Mpc. Bottom panels: tile size = 100 × 100 $h^{-1}$ Mpc. Left panels: $\sigma_A = 5\%$, right panels: $\sigma_A = 20\%$. In all panels the solid lines denote respectively the variance of the total power spectrum (black), cosmological power spectrum (blue), mask power spectrum (red) and all mixed terms (magenta). These are determined by difference, dotted lines denote negative values. The vertical thin lines mark $k = 2\pi/l$. The lower plots show the ratio of the various components with respect to the total power spectrum. In all figures, the vertical ticks denote the wavenumbers used to show the off-diagonal terms in Figure 5.5.

difference:

$$\Delta P^2_{\rm mixed} = \Delta P^2_{\rm obs} - \Delta P^2_{\rm cosmo} - \Delta P^2_{\rm mask} \, , \tag{5.58}$$

since all components on the r.h.s. can be measured independently. The dotted parts of these curves denote the place where the mixed term gets negative and then oscillates around zero. In this region the subtraction of shot noise induces an uncertainty that is larger than the signal seeked for (the measurement of the mask power spectrum is not affected by shot noise, so we can detect a much lower signal). The lower plots show the contribution of each component with respect to the total one, measured on masked mock catalogs.

Like the power spectrum case, $k = 2\pi/l$ marks the scale above which the pure mask term is important. Comparing the contribution of different components to the total power spectrum variance we notice that the pure mask component $\Delta P^2_{\rm mask}$ and the mixed one $\Delta P^2_{\rm mixed}$ present a similar scale-dependence and comparable amplitudes, at least in the large-scale range where the mixed terms can be measured. Starting from the configuration with $\sigma_A = 0.05$ and $l = 30 \ h^{-1}\,$Mpc, the pure mask and mixed terms contribute to the total variance by a few per cent. Mixed terms are so small in the first BAO region that they can hardly be measured even with this statistics. So the contribution of mask terms is modest and limited to large scales. Things become pretty different when $\sigma_A$ and $l$ are increased. The high variance case gives contributions of mask and mixed terms well in excess of 10 per cent at the BAO scale, that become dominant at the largest scales sampled by the boxes. In the large tile size case the mixed terms get to the 10 per cent level even with the modest mask variance of 5%, while in the high variance, large tile size case the covariance is completely dominated by the mask term on large scales, while the mixed term remains above the 10 per cent level but still larger than the cosmic variance. This is the only case where the mixed terms give a measurable contribution at scales smaller ($k$ higher) than $2\pi/l$; like the convolution term in the power spectrum, they give a relevant and non-oscillating small-scale contribution.

This analysis only shows the diagonal of the covariance matrix. Off-diagonal terms of the covariance matrix are of great importance, because they get mixed with the diagonal term during the matrix inversion necessary to determine the precision matrix that enters the likelyhood function. We now consider how uncertainties in foreground subtraction affect the covariance between different wavenumbers by studying the cross-correlation coefficients defined as:

$$r_{ij} = \frac{C_{ij}}{\sqrt{C^{obs}_{ii} C^{obs}_{jj}}} \, . \tag{5.59}$$

where $C_{ij}$ is defined by eq. (5.26). These are shows in figure 5.5 for some relevant values of $k_j$ (0.02, 0.07, 0.1 and 0.31 $h$/Mpc), as a function of $k_i$. These $k$-values span the range from very large to non-linear scales, and are marked in Figure 5.4 as vertical ticks.

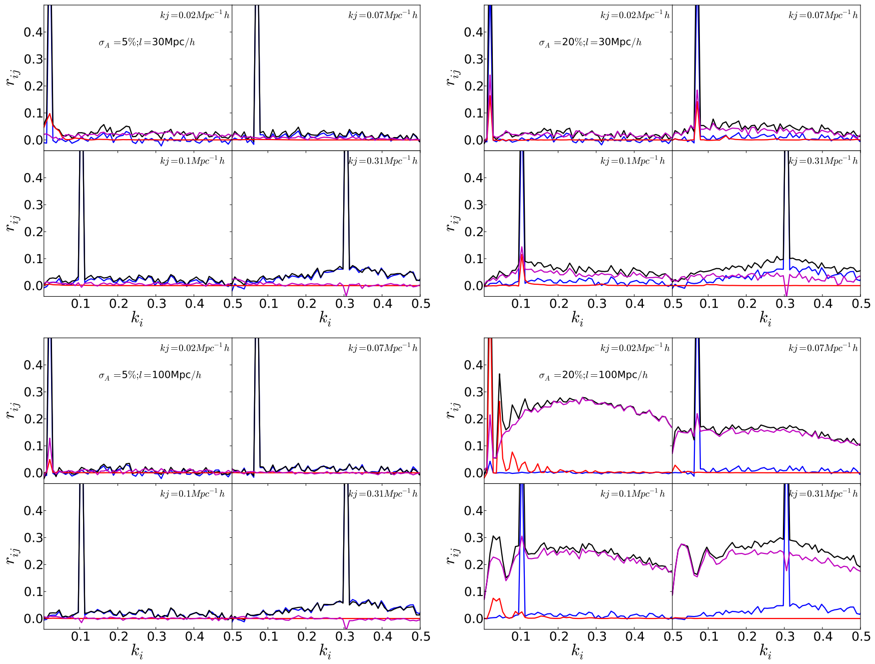


Figure 5.5: Correlation coefficient: Top panels: tile size = 30 × 30 $h^{-1}$ Mpc. Bottom panels: tile size = 100 × 100 $h^{-1}$ Mpc. The color code is the same of the previous figures.

The figure shows how contributions to the normalized covariance due to mask and mixed terms are negligible in the first configuration with $\sigma_A = 5\%$ and $l = 30\,h^{-1}$ Mpc. Off-diagonal terms are small in all cases, the cosmological one being appreciable at the highest $k_j$. However, in the smallest $k_j$ bin the mixed terms give a roughly constant contribution of few per cent. Increasing the scale $l$, we do not notice a larger impact of mask or mixed terms as we did for the diagonal; it seems that off-diagonal terms become significant at $k \ll 2\pi/l$, a regime that is not yet reached at $k = 0.02$ $h$/Mpc in this case. But when $\sigma_A = 0.2$, off-diagonal terms become very significant, amounting to 5 per cent for $l = 30\,h^{-1}$ Mpc and are in excess of 20 per cent for $l = 100\,h^{-1}$ Mpc, independently of scale. In this latter (rather extreme) case the structure of the power spectrum covariance matrix is strongly modified; clearly, an inversion of this matrix without proper account of off-diagonal terms would lead to large errors in parameter determination.

## 5.5 Discussion

In this chapter we have addressed the problem of how the uncertainty in the removal of foregrounds, in an observational survey of biased tracers like galaxies, propagates to the measurement of the cosmological power spectrum and its covariance matrix. For this first investigation we have decided to use a simplified setting, so as to be able to formulate analytic solutions for the two-point statistics. We have used DM halos as biased tracers, and their mass as a proxy of galaxy luminosity, as in a simplified HOD where each halo is populated by a single galaxy. To this aim, we have produced a very large set of 10,000 realizations of 1.5 $h^{-1}$ Gpc boxes, and extracted DM halos from these volumes at redshift $z = 1$ using the PINOCCHIO approximate method. This is, to our knowledge, the largest set of cosmological catalogs of DM halos. We have neglected luminosity- (mass-)dependent bias by randomly shuffling masses among DM halos in each catalog, so as to preserve their mass function. As for the foreground, we have constructed a simple toy model where, in a plane-parallel approximation, the $x - y$ plane of the box is tiled in squares of side $l$, that are characterised by a Gaussian residual foreground of variance $\sigma_A^2$ that propagates to the density through a modulation of the mass limit $M_0$; residuals in different tiles are uncorrelated, so $l$ should be interpreted as the projection, at the observation redshift, of an angular correlation scale of the foreground. The chosen values of 30 and 100 $h^{-1}$ Mpc are subtended, at $z = 1$, by angles of 0.74 and 2.5 degrees respectively.

The main conclusions of our analysis can be summarized as follows:

(1) The residuals of foreground subtraction ("mask") enter the power spectrum of masked catalogs as two terms, the power spectrum of the mask and its convolution with the cosmological power spectrum. This is similar to what happens when a survey geometry is applied to a cosmological volume.

(2) The mask term is significant at $k < 2\pi/l$, while the convolution term is usually smaller in this scale range, but can still be significant at smaller scales due to its scale mixing. Mask and convolution terms scale as $l^2\sigma_A^2$, implying that large correlation lengths of the mask residuals may have a significant effect on large scales even when the foreground removal is controlled to within a few per cent.

(3) Analytic estimations of mask and mixed power spectrum terms give results consistent with those measured from mocks, giving confidence on the level of control of the various terms.

(4) The power spectrum covariance matrix contains not only the cosmological and mask contributions, but also several, additional terms due the coupling of the convolution term with both mask and cosmology. The sum of all these terms can only be determined by difference of measurements of masked mocks, cosmological mocks and pure mask.

(5) Mask and mixed terms are found to have similar effect on the power

spectrum covariance matrix. A 5% accuracy on foreground removal guarantees a modest impact of these terms, with the exception of $k \ll 2\pi/l$ modes, where they can significantly contribute to the diagonal. In this case mixed terms give a roughly constant contribution to non-diagonal elements of the covariance matrix. The higher variance case of $\sigma_A = 20\%$ shows a dramatic impact on the structure of the covariance matrix.

(6) As a consequence of the relevance of mixed terms, a simple modeling of the covariance matrix as the sum of a pure cosmological term and a cosmology-independent term due to the mask appears to be an oversimplification, as the mixed terms couple cosmology and mask.

As long as BAO is the main target of an observational project, the results presented in this paper point to the conclusion that a $\sim 5\%$ error in foreground removal should guarantee a modest impact of the mask on parameter estimation. Indeed, due to the $l^2\sigma_A^2$ scaling, signals with smaller correlation scales will have little impact, while a large correlation length $l$ will mostly impact on larger scales (we limited our analysis to correlation lengths that are subtended, at $z = 1$, by relatively small angles because of the constraints on the box size). These errors can be compared to estimated errors in foreground removal or photometric calibration. Clearly, the case $\sigma_A = 20\%$ is pessimistic, and has been shown only to illustrate the effect of the mask. Photometric calibration can be controlled to the millimag level (Padmanabhan et al., 2008), so its induced errors will likely be negligible. Conversely, Galaxy extinction is know to the few per cent level (Schlegel et al., 1998; Peek and Graves, 2010; Berry et al., 2012) and zodiacal light can have a similar uncertainty far from the ecliptic (Planck Collaboration et al., 2014). $\sigma_A = 5\%$ can then be considered a realistic order of magnitude for the largest contribution to the visibility mask uncertainty. However, this conclusion is based on a very idealized setting, so it should be taken only as an indication, before tests with much more realistic mocks and masks are performed. On the one hand, this toy model is mixing modes on the whole box length; in a realistic survey a redshift bin would span a smaller comoving distance on the line of sight, and this would reduce the impact of mixed terms. On the other hand, a more complex mask like galactic extinction, having power on a range of scales, may easily have a stronger impact than our toy model; luminosity-dependent bias would also add to the covariance in a way that needs to be addressed.

To reduce the impact of a foreground to a desired level, one can of course work to improve the modeling of the foreground and of its correlated residuals. But another way to reduce this impact is to work on the estimator of the two- point statistics, with the aim of minimizing the impact of the residuals. This has been done, in preparation to the DESI survey by Schlegel et al. (2011), Levi et al. (2013)), by Burden et al. (2016) for the two-point correlation function, and by Pinol et al. (2016) for the power spectrum. In the first paper the authors modify the estimator of the correlation function to remove the angular mode contaminated by the incompleteness due to fiber assignment; in the second paper they investigate different methods to define the survey mean density, in particulare taking into accunt the fiber assignment coverage.

Two conclusions from the tests we have presented are robust. Firstly, the impact of foreground removal is of dramatic importance to properly sample the large scales beyond the BAO. This is expected: foregrounds, especially the zodiacal light, are correlated on large angular scales, that are projected to very large scales where the clustering signal is weak. But the scaling with $l\sigma_A$ shows that mode coupling gives a large weight to large-scale correlations, making the control of residual errors of great importance. It is convenient to recall that measurement of non-Gaussianity with error on the $f_{NL}$ smaller than unity, or effects of scale-dependent growth related to modified gravity, should be revealed at scales beyond the power spectrum peak; therefore the effect of foreground residuals are crucial for these measurements. Secondly, a poor control of foregrounds can lead to great changes in the covariance matrix. In particular, a significant presence of non-diagonal terms has deep consequences in the the ability to invert the covariance matrix and produce correct estimations of cosmological parameters and their errorbar. Control of foregrounds to the few per cent level is confirmed to be of paramount importance for large-scale structure.

# Chapter 6

# Covariance matrix comparison

As we have already showed in the previous chapter, because of the high statistical precision expected from next future surveys, the error budget will be dominated by systematics. The accurate modeling of the data is fundamental to constrain the cosmological parameters, but for precision cosmology, it is also important to revise the assumptions made when the observations are compared to theory.

If we call $\mathbf{D}$ the measurements derived from the observations, i.e. the galaxy power spectrum, and $\mathbf{T}(\boldsymbol{\theta})$ the model that depends on set of parameters $\theta$ that we want to determine, then the probability that the data $\mathbf{D}$ correspond to a realization of the model $\mathbf{T}(\boldsymbol{\theta})$ is given by the likelihood function:

$$\mathcal{L}(\mathbf{D}|\boldsymbol{\theta},\mathbf{C}) \propto |\mathbf{C}|^{1/2}\exp\Big[-\frac{1}{2}\sum_{i,j}(D_i - T_i(\theta))C^{-1}_{ij}(D_j - T_j(\theta))\Big] \ ; \qquad (6.1)$$

this expression is valid when the data follow a multi-variate Gaussian distribution with mean $\langle\mathbf{D}\rangle$ and covariance matrix $\mathbf{C}$. Eq. 6.1 comes directly from the *Bayes theorem*:

$$P(T|D) = \frac{P(T)P(D|T)}{P(D)} \ , \qquad (6.2)$$

where $P(T|D)$ is called **posterior**, $P(D|T)$ is the probability of the data given the model (**likelihood**) and $P(T)$ is called **prior**; if we set $P(D) = 1$ and we ignore the prior then we can find the likelihood, eq. 6.1, maximizing $P(T|D)$.

From eq. 6.1 it is clear that an accurate evaluation of the inverse of the covariance matrix, also called *precision matrix*, is needed to compute the likelihood function. Usually, for clustering measurements, the precision matrix is obtained using an ensemble of simulated galaxy catalogs. The main aspect to consider is that we have available a finite number of realizations, so that the precision matrix is affected by some degree of noise (Dodelson and

Schneider, 2013). The level of this noise depends on the number of synthetic catalogs used for the estimation of the precision matrix. It is important to control the noise because it will propagate into the parameter uncertainties (Taylor et al., 2013). As pointed out in Monaco (2016) a large number of galaxy catalogs will be required to correctly assess the covariance matrix of future large-volume galaxy surveys.

The N-body simulations are the usual tools for the production of synthetic galaxy catalogs. As we have described in section 4, with an N-body approach we can describe the interaction of N particles under the effect of gravity, following their evolution also in the deep non-linear regime. The typical simulation volume that will be required by future surveys is of the order of $\sim 4$ Gp $h^{-1}$; resolving the halos that host the faintest galaxies in a survey, that dominate the number and then the measurements on which cosmological parameter estimation is based, requires an halo resolution to be of the order of $\sim 10^{11} M_{\odot}$, the number of particles trajectories to follow is of the order of $\sim 16,000^3$, as pointed out in Monaco (2016). These numbers are for a single big simulation, but the evaluation of the covariance matrices of clustering requires a large number of simulated catalogs to lower the statistical noise and to highlight the cosmological information. Just as an example, for the SDSS-BOSS bispectrum analysis, they use 2048 simulated galaxy catalogs using a k-bin large $6k_f$; for a bin large $3k_f$ there would been needed of $> 10,000$ independent realizations. Moreover to investigate the different cosmological models, we need to run this large simulation characterizing it, each time, with different values for the cosmological parameters. It is not feasible, in terms of computing time and memory, to use full N-body simulations to make thousands realizations for each cosmological model.

In this chapter we describe the possible alternatives to N-body simulation approach. We focus, in particular, on various techniques to reduce the number of galaxy catalogs needed to control the noise in the covariance matrix and on the approximate methods that, using different approximations, allow us to have large number of simulations reducing the computing time with respect to an N-body simulation.

## 6.1 Estimation of the covariance matrix with a reduced number of simulated catalogs

As we have already anticipated in the introduction to this chapter, there are various techniques that allow us to reduce the number of simulated galaxy catalogs that we need to estimate the covariance matrix, but allowing us, at same time, to take under control the noise in the precision matrix.

In Pearson and Samushia (2016) the authors estimate the power spectrum covariance matrix using a theoretically justified parametric model calibrated on simulated galaxy catalogs. The general expression for the power spectrum covariance matrix for a cubic volume is given by Scoccimarro et al. (1999):

$$C_{ij} = \frac{(2\pi)^3}{V(k_i)}\Big(\frac{P_i^2 + n^{-1}}{2\pi k_i^2 \delta k}\delta_{ij} + \bar{T}(k_i, k_j)\Big) \ , \tag{6.3}$$

where $V$ is the volume, $n^{-1}$ is the inverse of the galaxy number density, $\delta_{ij}$ is the Kronecker delta function, $\delta k$ is the size of the bin in Fourier space and $\bar{T}(k_i, k_j)$ is the bin averaged trispectrum. The term proportional to the power spectrum, $P_i^2$, represents the Gaussian part of the covariance matrix while the non-linear evolution of structures induces non-Gaussian corrections given by the trispectrum. The evaluation of eq. 6.3 becomes highly complex when we consider non-trivial survey windows that introduce additional terms. In this case, the computation of the power spectrum covariance matrix is made using an ensemble of simulated galaxy catalogs.

The alternative proposed by Pearson and Samushia (2016) is to use a relatively simple, few-parameter function to approximate the true covariance matrix, that includes the non-linear effects due to a survey window. The main advantage of the fitting function approach is that the covariance matrix elements converge to their true values much faster than the sample variance.

They have evaluated the covariance matrix using 600 galaxy catalogs from BOSS DR11 (Manera et al., 2013) and they have compared it with the covariance obtained with the fitting formula; they use $\sim 100$ catalogs to fit the free parameters. What they found is that the fitting function generated covariance matrices, calibrated with few number of galaxy catalogs, were statistically indistinguishable from the sample covariance matrix generated with 600 catalogs.

The main result of this paper consists in the possibility of having a fitting formula for the power spectrum covariance matrix, that reproduces the true covariance matrix using only a very small number of galaxy catalogs. Future survey will require to modify the functional form of the fitting formula in order to include the power spectrum measurements on much smaller scales, but once this functional form is found the method will work in the same way of the specific case they analyze in the paper.

An alternative to the fitting procedure described above, is given by the work of Pope and Szapudi (2008): they apply the idea of *shrinkage estimation* by Schäfer and Strimmer (2005) to the determination of matter power spectrum covariance matrix with a limited number of realizations. The aim of this paper is to reduce the total noise in the covariance, while preserving as much information as possible on the real covariance from simulations. The *shrinkage* estimation consists in combining a theoretical model with empirical estimate. Given $n$ sets of data, let $\mathbf{x}$ to be the measure for each of them; if $x_i^{(k)}$ is the $k^{th}$ observation of the $i^{th}$ element of $\mathbf{x}$, then the unbiased empirical covariance matrix of the data vector is given by:

$$S_{ij} = \widehat{Cov}(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j) \ , \tag{6.4}$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)}$ is the empirical mean. The shrinkage procedure requires to have also a target covariance matrix, $T_{ij}$ that is based on some model with no or few free parameters. This covariance has a smaller variance with respect to the empirical covariance, but may be biased. The final expression for the covariance matrix is given by a linear combination of the empirical and the target covariances:

$$\mathbf{C} = \hat{\lambda}^{*}\mathbf{T} + (1 - \hat{\lambda}^{*})\mathbf{S} \ , \tag{6.5}$$

where $\hat{\lambda}^{*}$ is called *shrinkage intensity.* The optimal estimator for this parameter, found by Schäfer and Strimmer (2005), is given by:

$$\hat{\lambda}^{*} = \frac{\sum_{i,j}\widehat{Var}(S_{ij}) - \widehat{Cov}(T_{ij}, S_{ij})}{\sum_{ij}(T_{ij} - S_{ij})^2} \ , \tag{6.6}$$

where $\widehat{Cov}(T_{ij}, S_{ij})$ is the estimate of the covariance of the elements of the two covariance matrices. The value of the shrinkage intensity determines the final covariance matrix $\mathbf{C}$: if the estimation of $\hat{\lambda}^{*}$ is larger than one, then the condition $\hat{\lambda}^{*} = 1$ is imposed and in this case, from eq. 6.5, the covariance matrix is the target one; if the estimate $\hat{\lambda}^{*}$ is less than zero, then $\hat{\lambda}^{*}$ is forced to be zero and this corresponds to say that only the empirical covariance matrix is used. The main conclusion of Pope and Szapudi (2008) is that the shrinkage technique can provide a covariance matrix with a precision comparable the empirical one, using a limited number of simulations or jackknife resampling (see 3.1).

The last technique we want to describe is based on the idea of Kaufman et al. (2008) used to speed-up the calculation of maximum likelihood estimation. In Paz and Sánchez (2015) the authors use the *tapering* technique as tool to minimize the impact of the noise in the precision matrix. The basic idea of tapering is that, in many cases, the correlation between distant data pairs is negligible and little information is lost by treating these points as being independent (Kaufman et al., 2008). With this assumption it is possible to use fast numerical methods to evaluate the likelihood function. When applied to cosmological covariance matrix we have to consider that the off-diagonal elements of the matrix can be non-negligible. The idea, in this case, is to first build a new covariance matrix where the off-diagonal terms are reduced. To this address they introduce a specific function, called *tapering matrix* $\mathbf{T}$, defined in the following way:

$$T_{ij} = K(||r_i - r_j||) \ , \tag{6.7}$$

where $r_i$ is the i-th measurement location of the data space and K is a positive compact-support function that sets to zero the off-diagonal entries of the covariance matrix. The new covariance matrix characterized by reduced off-diagonal terms is given by:

$$\mathbf{C}^T = \hat{\mathbf{C}} \circ \mathbf{T} \ , \tag{6.8}$$

where $\hat{\mathbf{C}}$ is the estimate covariance matrix, $\mathbf{C}^T$ is the tapering covariance matrix. The precision matrix is not given simply by the inversion of eq. 6.8, because this is biased by the tapering function itself; moreover we want the precision matrix to include the physical information coming from the off-diagonal part of the covariance matrix, but excluding the noise. An estimation of the expression for the tapering precision matrix comes from the matricial product of the inverse of the tapering covariance matrix and the tapering matrix:

$$\mathbf{\Psi}^t = \Big(1 - \frac{N_b + 1}{N_s - 1}\Big)\Big(\hat{\mathbf{C}} \circ \mathbf{T}\Big)^{-\mathbf{1}} \circ \mathbf{T} \ , \tag{6.9}$$

where $\mathbf{\Psi}$ is the precision matrix and the prefactor provide an unbiased inverse of the covariance matrix.

In the conclusion of the paper they stress that the covariance tapering method leads to smaller errors than the standard technique, without introducing any systematic bias in the estimated parameters. Moreover they show that the optimal tapering parameter depends only on the structure of the underlying covariance matrix and it is insensitive to the bin size or to the number of synthetic samples used in the estimation of the precision matrix. The last result concerns the possibility of using this technique to estimate the covariance matrix reducing the number of realizations: using 600 simulations from SDSS-DR9 BOSS clustering measurements (Anderson et al., 2012), they obtain an uncertainty on BAO shift equivalent to the one obtained with 2'300 simulations, but using the standard covariance technique.

## 6.2 Approximate methods

In the previous section we have described some possible alternatives to estimate the covariance matrix with a reduced number of realizations using also hybrid techniques to connect theoretical model and simulations. An alternative approach is to use simulated galaxy catalogs made with approximate methods, that allow us to have a large number of realizations to precise estimate the covariance matrix and reduce the statistical noise; these techniques can introduce systematic errors that, if not spotted, propagete in the accuracy of cosmological parameters constraints. Within this project we aim to estimate the weight of these errors on the covariance matrix, prceeding with two main tests: first, we compare the covariance matrices obtained with approximate methods with that one obtained using full reference N-body simulation runs with GADGET (Springel, 2005), called Minerva; second, we compare the cosmological parameters constraints obtained with the N-body covariance matrix and with the others covariance matrices. This project is carried out within the Euclid Collaboration Galaxy Clustering Working Group.

The techniques we test are the following: Pinocchio, COLA, Halogen, PeakPatch (all these codes are described in section 4.2). Because we want to compare how the covariance matrices of clustering differ from what we can obtain from full N-body simulations, we produce 300 realizations of the same Universe with each of the code we have listed above and with the full N-body simulation Minerva. To this account we use the same initial conditions of the N-body simulation in order to reproduce the same (large-scale) scatter.

The main simulation properties are the following: periodic cubic box of 1500 $h^{-1}\,\mathrm{Mpc}$ sampled with $1000^3$ particles. The cosmological parameters are $\Omega_m = 0.285$, $\Omega_\Lambda = 0.715$, $\Omega_b = 0.044$, $h = 0.695$ and $\sigma_8 = 0.285$. We use output at $z = 1$. The mass particle is $M_\mathrm{p} = 2.67 \times 10^{11}\ h^{-1} M_\odot$. We choose to focus at this particular redshift, because Euclid will, in particular, observe galaxy in the redshift range 0.9-1.9.

The measures of the clustering statistics are carried out considering two different halo samples. The selection of the halos for the two cases is made with the two following criteria: in the first case, we look at the halo masses so that the sample is obtained considering all the halos above a certain mass threshold. The first bin contains halos with mass larger than $1.06 \times 10^{13} M_\odot$, while the second bin halos with mass larger than $2.66 \times 10^{13} M_\odot$. These mass thresholds correspond, respectively, to a number of particles per halo of 40 (first bin) and 100 (second bin). We call this first selection “Hard” mass cut (see table 6.1). The second selection, that we call abundance matching, consists in fixing the number of halos in each sample by matching the halo mass functions (see table 6.2). In this second case we match the the number of halos in each bin, so we make sure that the differences between the different codes in the halo abundace is within 1%.

In the next two sections we describe the results for the comparison of the power spectrum (Blot et al. in prep.) and bispectrum (Colavincenzo et al. in prep). For both the two statistics, we show the average value over 300 realizations, studying in particular the differences with respect to the values obtained using N-body simulations. For the analysis of the covariance matrices we proceed with two separate steps: first, we compare the diagonal of the covariance matrices highlighting the main differences with the average; then we look at the off-diagonal elements of the matrix by means of the cross-correlation coefficients. It is worth to stress that with this number of realizations we can control the variance, but we expect the cross-correlation coefficients to be very noisy; in this case we compare the noise obtained from the measures on the approximate methods with the noise obtained from the measures on full N-body simulations.

Table 6.1: "Hard" Mass Cut

| Bin | N | M($10^{13}$) |
|---|---|---|
| 1 | 40 | 1.06755 |
| 2 | 100 | 2.66887 |

Table 6.2: Abundance Matched Cut

| | M($10^{13}$) | |
|---|---|---|
| Code | bin 1 | bin 2 |
| Minerva | 1.12092 | 2.66887 |
| COLA | 1.08566 | 2.76556 |
| Pinocchio | 1.04371 | 2.62971 |
| Halogen | 1.12092 | 2.66887 |
| PeakPatch | n.a. | 2.35500 |

| | N | |
|---|---|---|
| Code | bin 1 | bin 2 |
| Minerva | 719004 | 183638 |
| COLA | 716401 | 182070 |
| Pinocchio | 724788 | 184896 |
| Halogen | 721409 | 182016 |
| PeakPatch | n.a. | 183561 |

### 6.2.1 Results: power spectrum

In this section we show and comment the results for the real and redshift-space power spectrum. The study of the two-point statistic is important becasue the Gaussian part of the covariance matrix depends directly from it. Evaluating the power spectrum on the samples obtained with approximate methods, we expect to well reproduce the large-scales and to over- or under-estimate the smaller scales.

All the measuraments for the power spectrum are corrected for the shot-noise contribution, that we consider as Poissonian. This is anyway true only when we consider the power spectrum from n-body simulations, in the other cases we expect it to be sub-Poissonian. In the figures for the average power spectrum we plot the Poissonian shot-noise prediction with a horizontal gray line. The error bars are obtained from the the diagonal of the covariance matrix after dividing for the total number of realizations (300).

In figure 6.1 we show the real power spectrum (without RSD) for the two different cases, "hard" mass cut, upper plots, and abundance matching, bottom plots. In the upper panels we show the average power spectra obtained from the 300 realizations, while in the lower panels we show the relative differences of each of the fast methods with respect the N-body simulation.

For the low mass bins we plot the power spectra from PINOCCHIO, COLA, Halogen and Minerva, while for the high mass bins we include also PeakPatch, that does not have halos with masses in the first bin. In all the

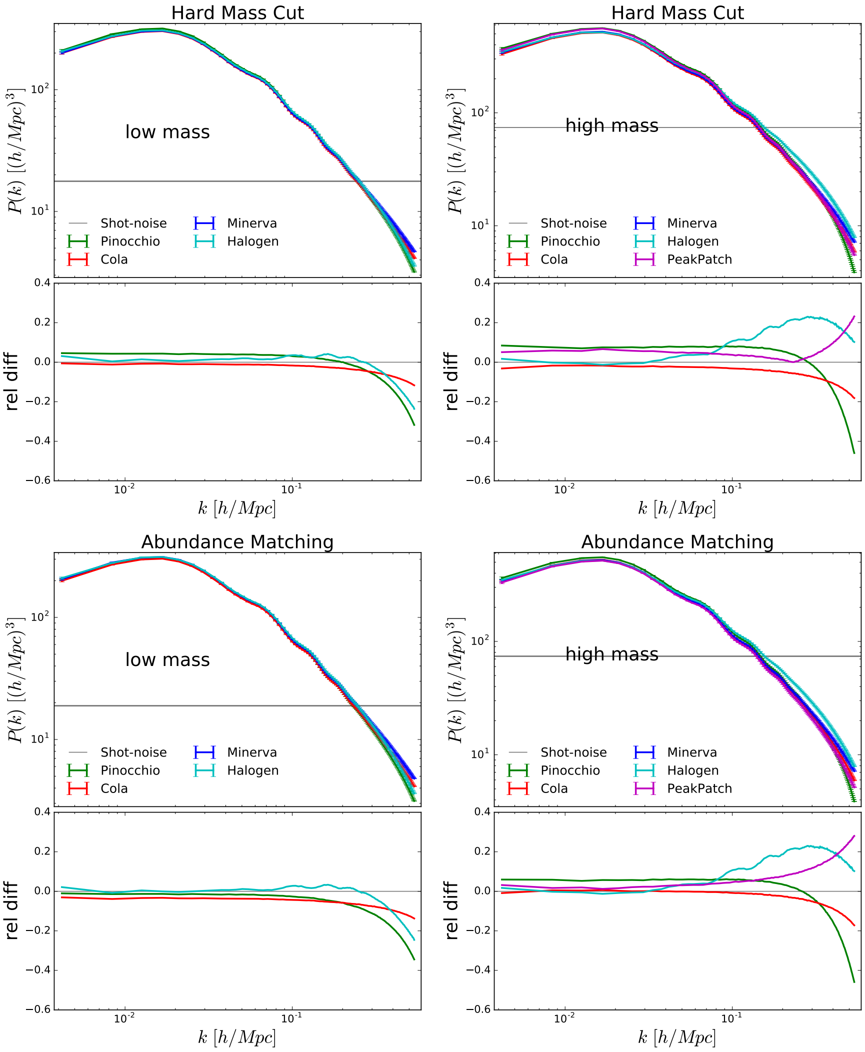


Figure 6.1: Average real space power spectrum on 300 realizations. On the left low mass bins, on the right high mass bins. In the lower panels the relative difference of the approximated methods with respect to Minerva.

cases we can see that for the large scales, small values of k, all the approximate methods well reproduce the simulation with an error below 5%. The agreement remains constant up to 0.2 $h\,\mathrm{Mpc}^{-1}$. For smaller scales the approximate methods start to fail in reproduce the N-body simulations: COLA is the most precise on these scales with a difference of $\sim$ 10% at $k \sim 0.5\,h\,\mathrm{Mpc}^{-1}$; PINOCCHIO works slightly worst with a difference of $\sim$ 10% at $k \sim 0.35\,h\,\mathrm{Mpc}^{-1}$ and larger then 20% at $k \sim 0.5\,h\,\mathrm{Mpc}^{-1}$; Halogen behaves like PINOCCHIO for the low mass “hard” cut, while for the low mass abundance cut is more similar to COLA, instead for both the two high mass bins it overestimates the simulation by 20% starting from $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$; PeakPatch is comparable to PINOCCHIO up to $k \sim 0.2\,h\,\mathrm{Mpc}^{-1}$, but then the relative difference with respect to Minerva increases in the opposite direction of PINOCCHIO. We expect to observe this behaviour at small scales looking at the power spectrum obtained from approximate methods; the small scale structrues evolution is not well reproduced like a full N-body simulation that follows the particle trajectories also in the non-linear regime.

In figure 6.2 we show the average redshift-space monopole power spectrum. As we have done for the real space we plot the two samples with the two mass bins. The large scales, up to $k \sim 0.2\,h\,\mathrm{Mpc}^{-1}$, are well reproduced by all the methods, even if PINOCCHIO and PeakPatch seem to work slightly worse in the case of the high masses for the "hard" cut sample, but still the differences are smaller than 5%. On smaller scales, $k > 0.2\,h\,\mathrm{Mpc}^{-1}$, all the codes, besides Halogen, show a better agreement with simulation when compared with the real space; in all the cases the differences are almost the same of those ones observed on large scales. Halogen is the only one to show the same behavior in both real and redshift space in the case of the monopole.

In figure 6.3 we plot the average quadrupole of the power spectrum. All the codes look comparable on large scales reproducing the N-body simulation up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$. Smaller scales are not correctly reproduced, but for all the codes the quadrupole looks larger than Minerva for scales smaller than $k > 0.3\,h\,\mathrm{Mpc}^{-1}$. Halogen, differently from the other codes, underestimates the simulation by 10% up to $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$. On smaller scales the difference is greater than 50% on scales $k > 0.2\,h\,\mathrm{Mpc}^{-1}$.

In figure 6.4 we plot the average hexadecapole of the power spectrum. Contrary to the power spectrum in real space and the other multiple orders, the hexadecapole shows a high level of noise on large scales, while it oscillates around zero for scale $k > 0.02\,h\,\mathrm{Mpc}^{-1}$. Even if the signal enclosed in the hexadecapole is hidden by noise on large scale, we can see that all the codes show the same correlated noise as the N-body simulation. This is still an important information because we want to test the accuracy of the approximate methods to reproduce the same features of the N-body simulation even if this is noise. We will stress this point at the end of this chapter, in the discussion section.

As we have already noticed, apart from the monopole power spectrum, the real power spectrum as well as the quadrupole and the hexadecapole have a worst behavior on smaller scales ($k > 0.1\,h\,\mathrm{Mpc}^{-1}$) with respect the N-body simulation. This is almost independent from the mass bin or from the selection sample ("hard" or abundance mass cut). We are not surprised by this behavior because the approximations, that characterize the fast methods, are characterized by a less accurate description of the small scales with respect to full N-body simulations. This is not a critical issue because the goal of this project is to use and to test the approximate

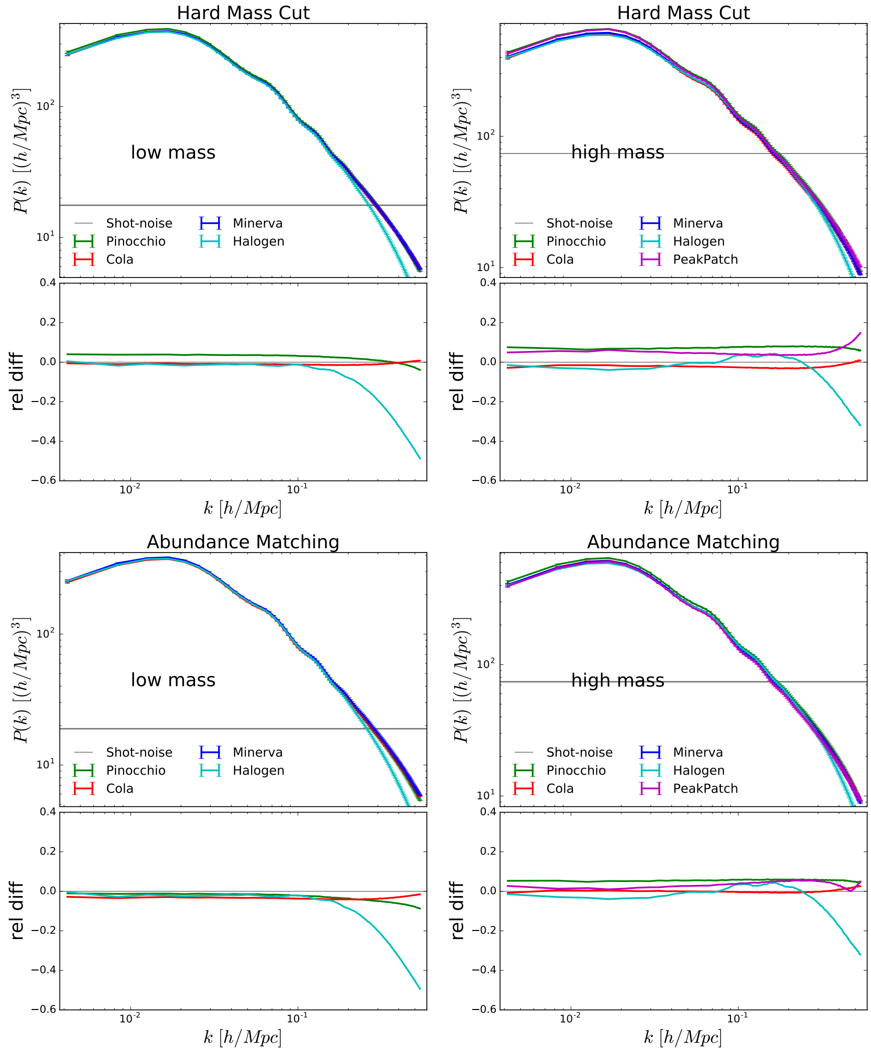


Figure 6.2: As figure 6.1 but in redshift space: monopole.

methods to evaluate the **covariance matrices**, and we expect to reduce the differences we observe having a large number of realizations.

The next figures show the power spectrum variance, i.e. the diagonal of the covariance matrix and the relative difference with respect to Minerva. As we have done for the power spectrum, we show the two different mass cuts and the two mass bins. In figure 6.5 we plot the real power spectrum variance. Looking at the low mass bin of the “hard" mass cut, the N-body variance is reproduced by all the codes with a difference lower than 10%. The best method appears to be COLA with a few per-cent differences on all the scales, while PINOCCHIO and Halogen seem to slightly overestimate the variance. For the high mass bin COLA remains the best code in representing

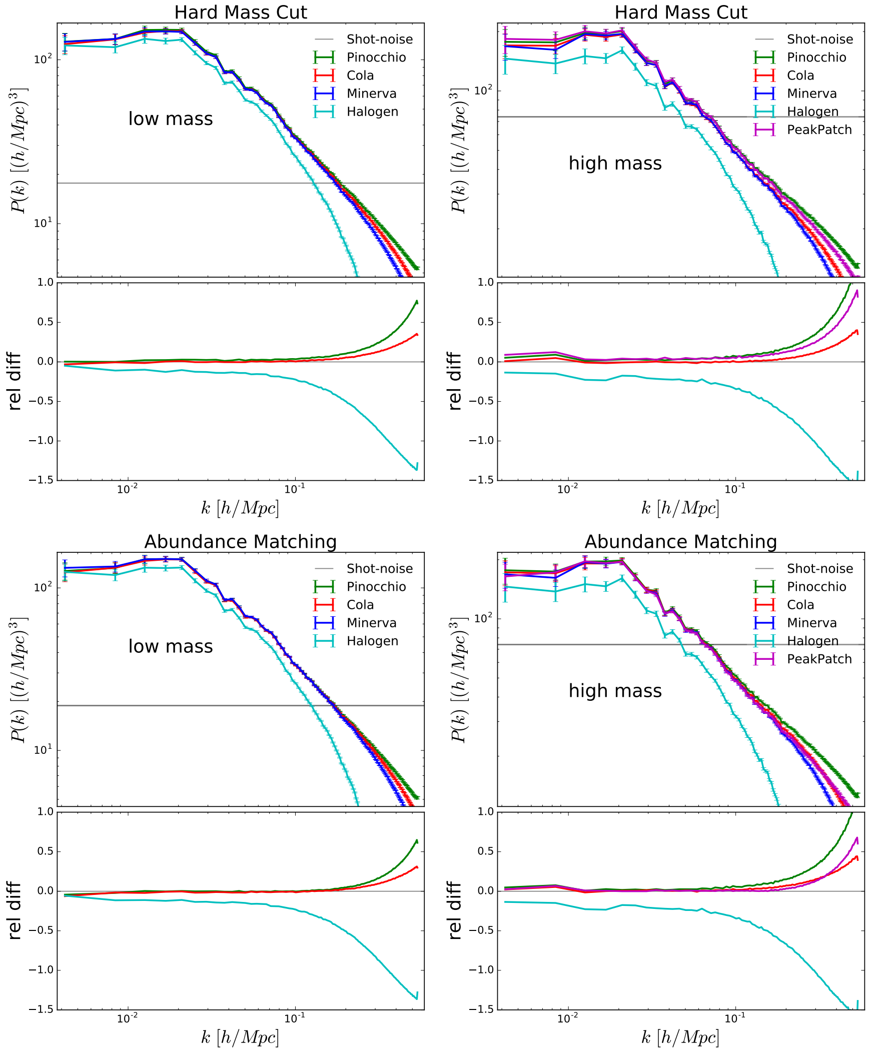
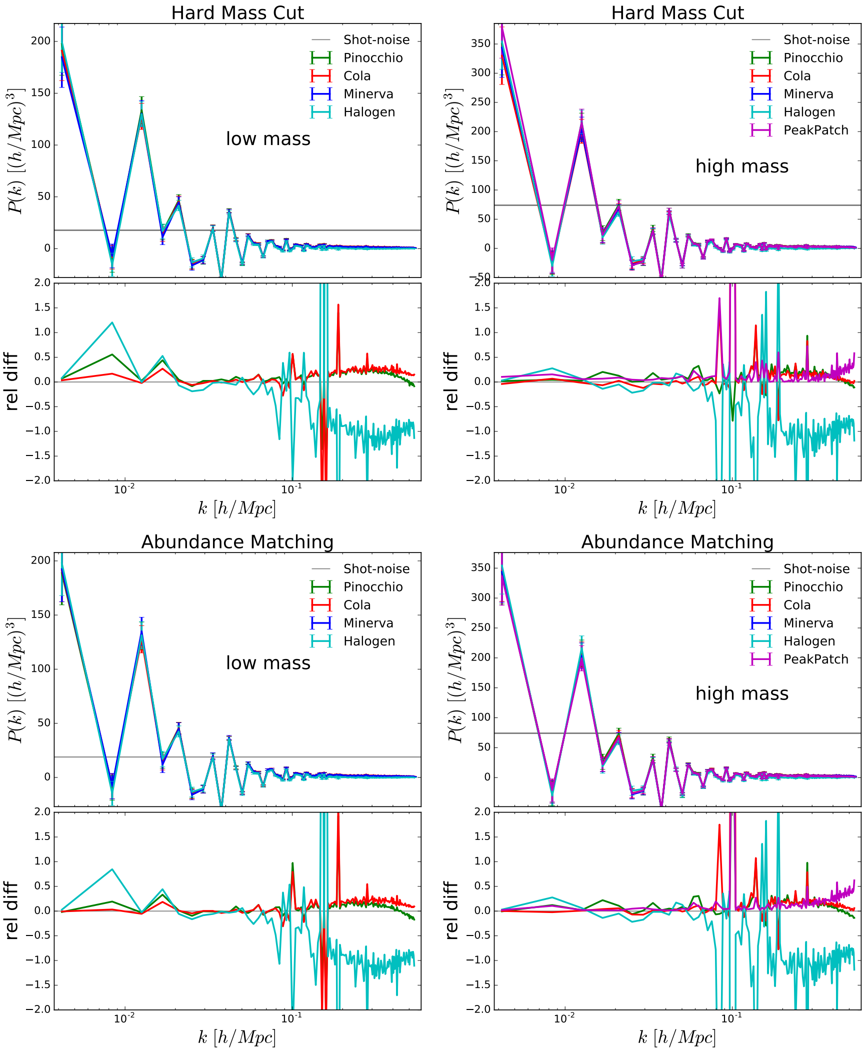


Figure 6.3: As figure 6.1 but in redshift space: quadrupole.

the power spectrum variance, while the other ones overestimate it by 20% an all the scales. When we look at the abundance matching case, the power spectrum variances from the low mass bin appears to reproduce the N-body simulation better when compared with the low mass bin of the “hard" mass cut case. All the codes differ by less than 10%. We can say the same for the high mass bin apart from Halogen that remains similar to the other case.

What it is important to stress, that we will remarks at the end of the chapter, is that contrary to the average power spectrum, the difference between the variance obtained with the different approximate methods and Minerva remains constant even on small scales, while the correspondent power spectrum drops down or goes up so that the difference with simu-



Hard Mass Cut

Hard Mass Cut

Abundance Matching

Abundance Matching

Figure 6.4: As figure 6.1 but in redshift space: hexadecapole.

lation becomes larger than 20%.

In figure 6.6 the variance of the monopole power spectrum is quite similar to the one we have showed in figure 6.5. The Halogen power spectrum variance seems to behave better compared with the real case, in the low mass bins for both the two sample.

The variance of the quadrupole and of the hexadecapole, figures (6.7 and 6.8) look comparable in terms of precision to reproduce the N-body variance. As for the real and monopole case the variances evaluated for the abundance matching samples appears to reproduce the simulation variance with an accuracy higher than the variance obtained from the "hard" mass samples. For the low mass bin Halogen variance drops down for scale $k >$

$0.1\,h\,\mathrm{Mpc}^{-1}$, while for the high mass bin all the codes reproduce better the simulation with differences below 10%.

The variance we just described is only a part of the covariance matrix. As we have stressed at the beginning of this chapter, the constraints on the cosmological parameters required the definition of the precision matrix, that is the inverse of the whole covariance matrix. For this reason we are interested in the analysis of the off-diagonal parts of the power spectrum covariance matrix.

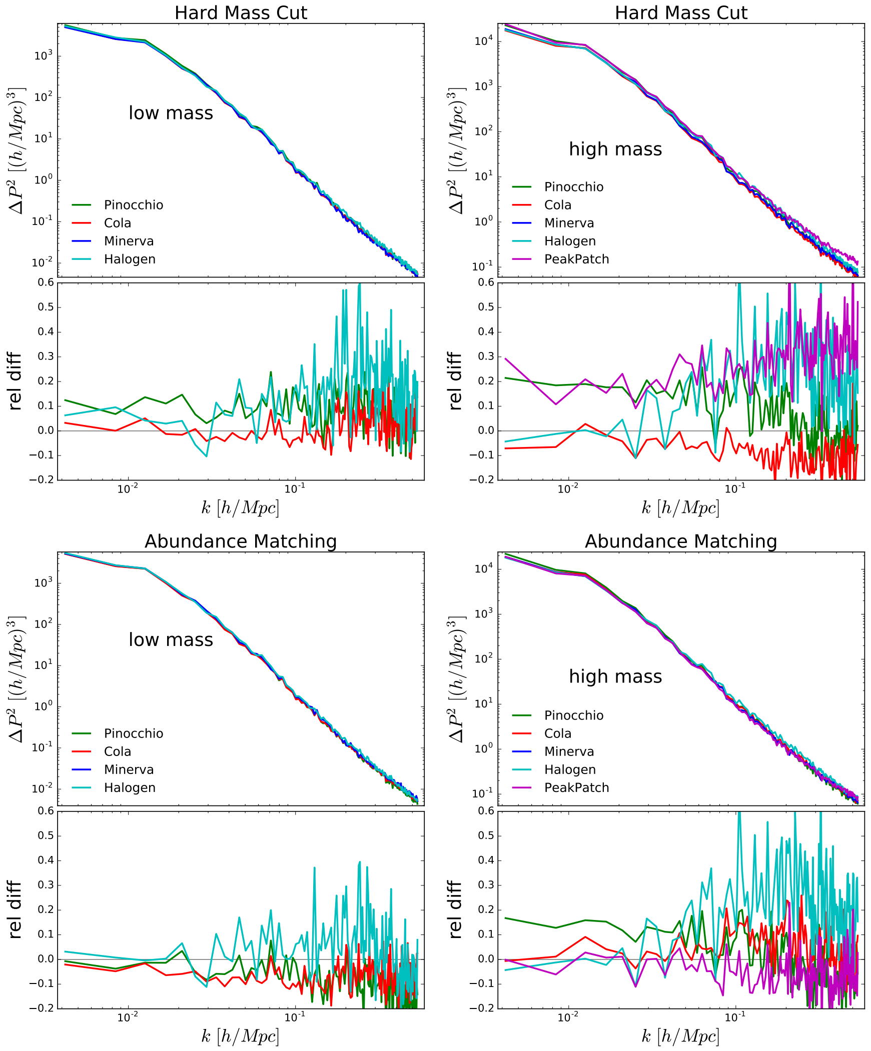


Figure 6.5: Power spectrum variance in real space. On the left low mass bins, on the right high mass bins. In the lower panels the relative difference of the approximated methods with respect to Minerva.

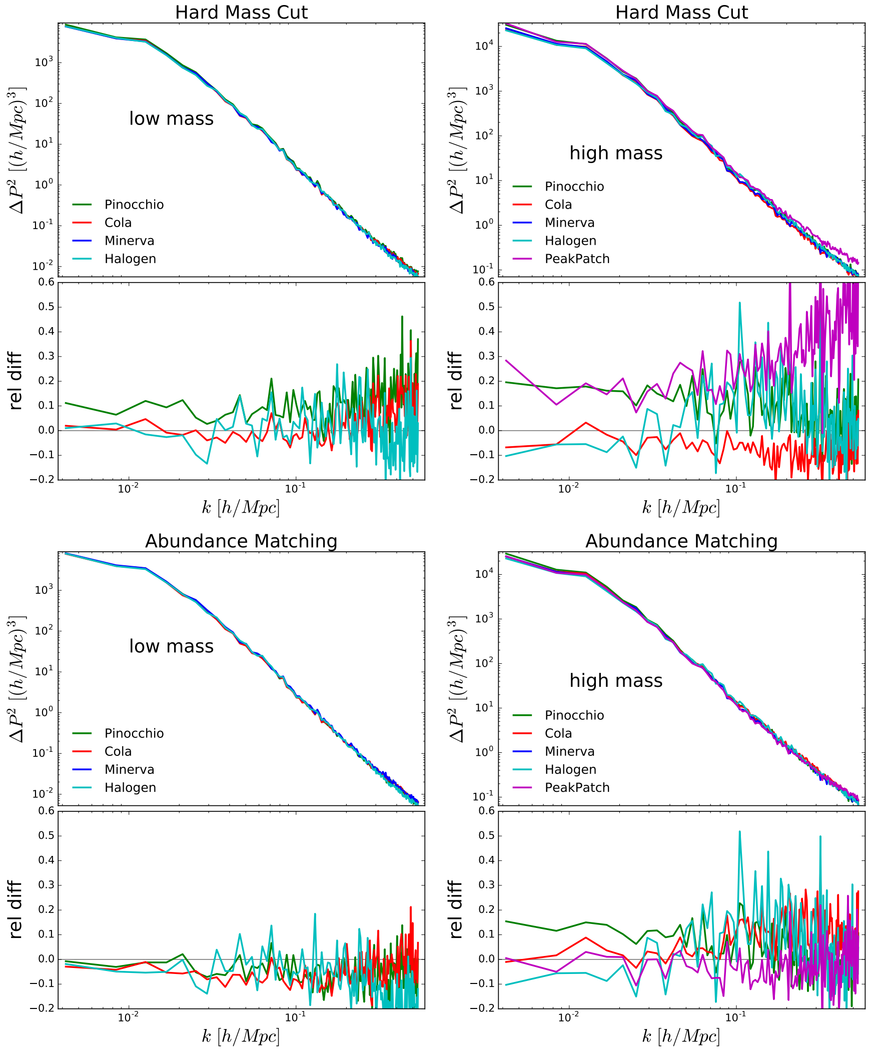


Figure 6.6: As figure 6.5 but in redshift space: monopole.

There are different ways to represent the power spectrum covariance matrix, but because we want to highlight the off-diagonal structure of the matrix, we use the cross-correlation coefficient, a quantity that we have also used in the foreground analysis (see section 5.4.2):

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} \ , \tag{6.10}$$

where $C_{ij}$ is the full covariance matrix and $C_{ii}$ is the variance of the matrix. In this way when $i = j$ we are looking at the diagonal element of the matrix and $r_{ij} = 1$. When we move away from the diagonal, $i \neq j$, we are looking at the off-diagonal terms.

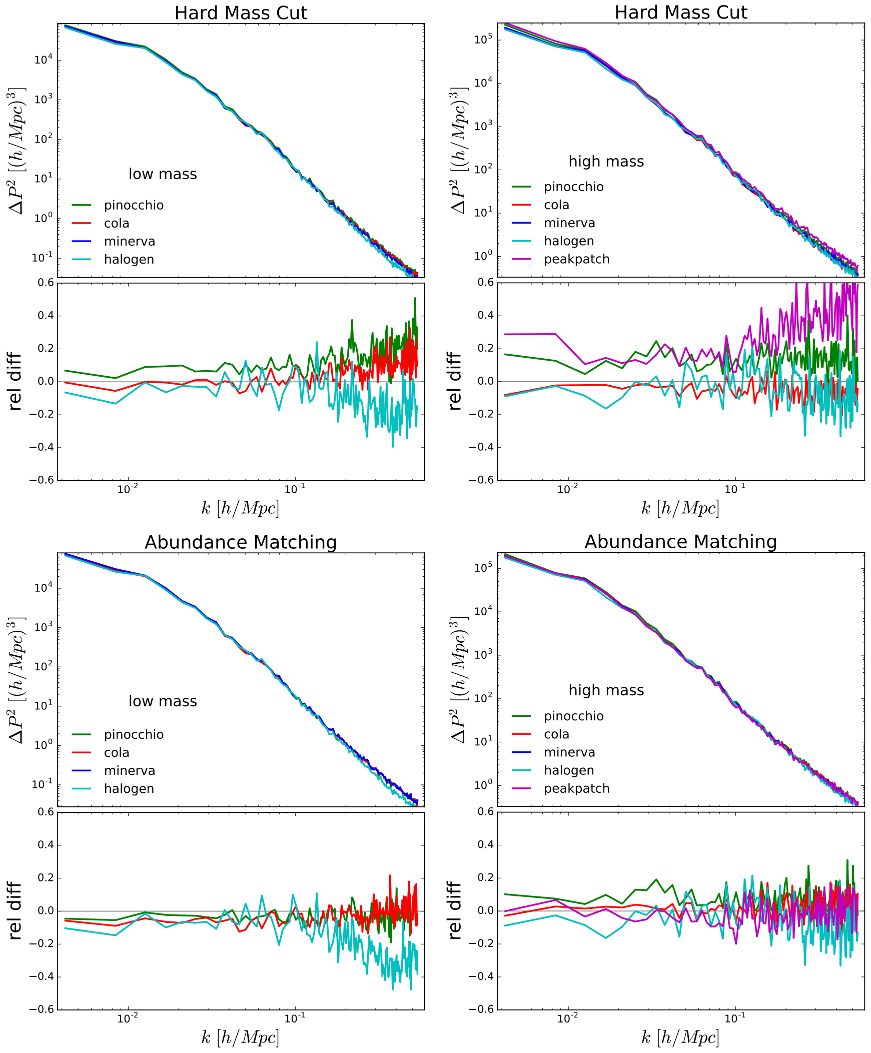


Figure 6.7: As figure 6.5 but in redshift space: quadrupole.

We have evaluated the cross-correlation coefficient for the real power spectrum and for all the power spectrum multipoles in redshift space, but we describe only the case in real space because the difference between all the cases are negligible and they give us the same information. The real space power spectrum cross-correlation coefficient, for the abuncance matching case, is showed in figure 6.9.

As for the power spectrum and its variance we show the two halo selection cases and the two mass bins. Figure 6.9 is composed by four panels and in each of them we have fixed the value of one of the wavenumber, $k_i$ and we have varied the other wavenumber, $k_j$. We have chosen the values for $k_i$ to investigate the behavior of the elements of the covariance matrix

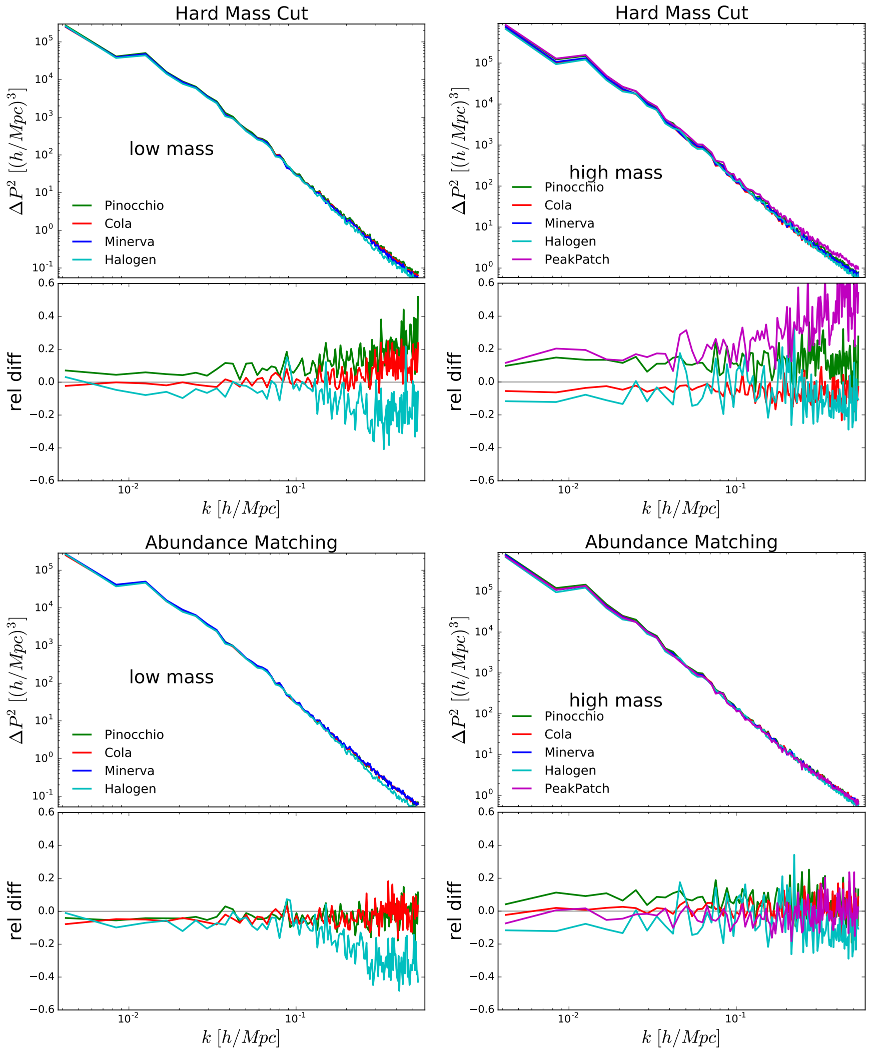


Figure 6.8: As figure 6.5 but in redshift space: hexadecapole.

from large to small scales. Looking at the covariance matrix structure, it appears not modified by different sample selections and mass binning. The number of realizations we can use (300) to evaluate the covariance matrices is not large enough to proceed with an accurate analysis of the off-diagonal elements of the matrix, but we can still look at noise and what we observe is that the approximate methods and the N-body simulation cross-correlation coefficients are characterized by the same noise. It is after all true that for some particular scale ($k_i \sim 0.3\,h\,\mathrm{Mpc}^{-1}$), looking at the high mass sample, we can identify an enhancement of the amplitude of particular off-diagonal elements of the covariance matrix both using the approximate methods and N-body simulations. We will further comment on these two last results in the discussion section.

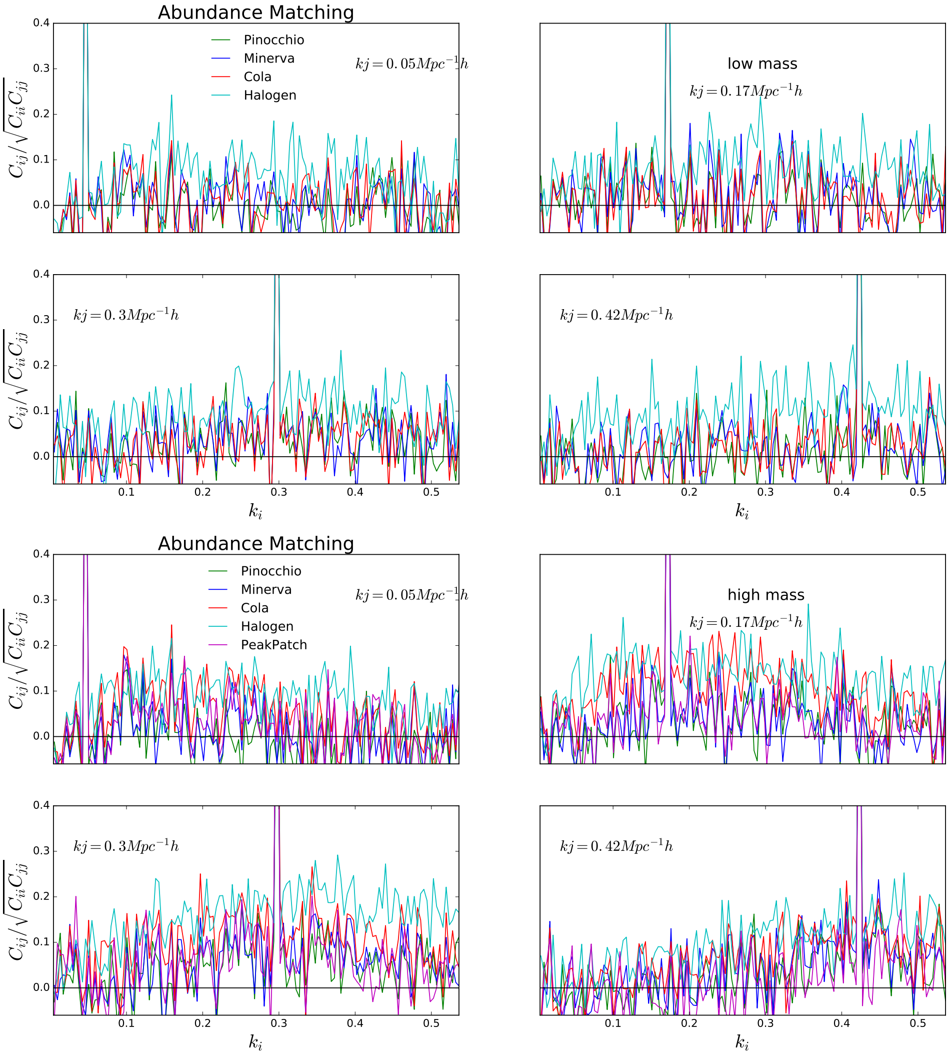


Figure 6.9: Power spectrum cross-correlation coefficient in real space. On the top low mass bins, on the bottom high mass bins.

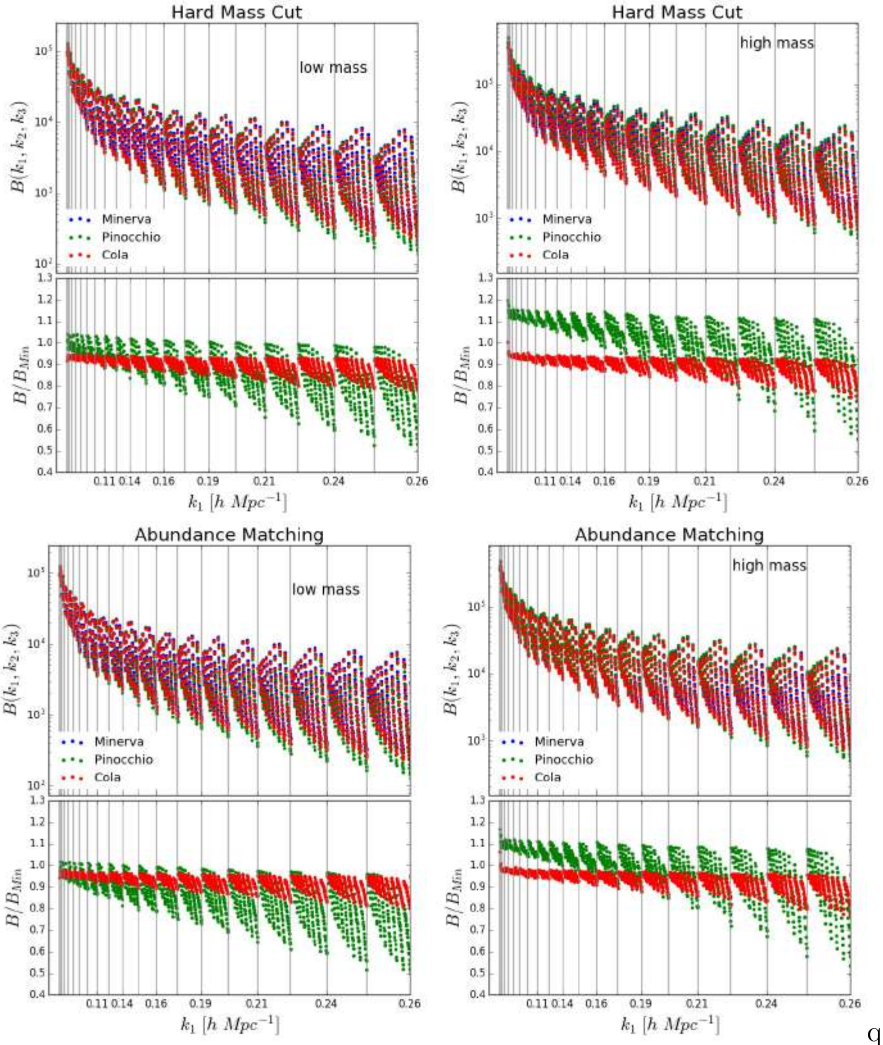
### 6.2.2 Results: bispectrum

In our test, in addition to the two-point statistics, we consider as well higher-order correlation functions, as the bispectrum. We study this quantity because the power spectrum covariance matrix is affected by non-Gaussianity, that are induced by non-linear correlators.

In this particular case we analyze the bispectrum and its covariance matrix. The selection we make on the simulated catalogs is the same of the case of power spectrum analysis (see table 6.1 and 6.2), but we have access to only the bispectrum measures on PINOCCHIO and COLA catalogs for the approximate methods, and of course the measures on Minerva. Moreover we show only real space analysis.

In figure 6.10 we plot the average bispectrum in real space. The power spectrum depends on only one wavenumber, so its representation is trivial, but the bispectrum is a function of three wavenumbers with the condition, given by the Dirac delta function (eq. 2.60), to form a closed triangle. For this reason its representation is more complex. We want to study how well the different triangular configurations are reproduced by the approximate methods compared with the N-body simulation. For this reason we decide to plot all the possible triangular configurations given by different values of $k_1$, $k_2$ and $k_3$, showing on the x-axis one wavenumber only the value of one vertex $k_1$. In this way, each point represents one particular bispectrum value for a particular triangle configuration. Points between vertical gray lines have the same $k_1$ and span all allowed values of $k_2$ and $k_3$. In the lower panels we show the ratio, configuration by configuration, between the approximate methods and Minerva. Looking at the low mass bin for the “hard” mass selection, we can see that on large scales ($k < 0.11\,h\,\mathrm{Mpc}^{-1}$) PINOCCHIO (green dots) shows slightly larger differences than COLA, but it shows a larger number of configurations that are in better agreement with Minerva; COLA shows smaller disagreements, but a greater part of its configurations underestimate the simulation by 10%. A smaller scales ($k > 0.11\,h\,\mathrm{Mpc}^{-1}$) PINOCCHIO and COLA show a different behaviour with PINOCCHIO that is less precise to reproduce the simulation; This differences are larger than those one observed on large scales. The number of configurations that agree with simulation is smaller, and for $k > 0.16\,h\,\mathrm{Mpc}^{-1}$ a considerable portion of points is below one, meaning that an increasing number of triangular combinations reproduce Minerva with an accuracy that can be lower than 20% ($k \sim 0.26\,h\,\mathrm{Mpc}^{-1}$). COLA appears to be more under control than PINOCCHIO; the differences with Minerva vary from $\sim$ 10% up to a max of $\sim$ 20%. For the higher mass bin, up to $k \sim 0.16$, a large number of PINOCCHIO configurations overestimates Minerva by $\sim$ 10%. For smaller scales the spread increases, so that certain configurations agree better, but for $k > 0.21\,h\,\mathrm{Mpc}^{-1}$ PINOCCHIO is larger or smaller with respect to Minerva by 10%. COLA behavior is similar to the previous case with an increasing accuracy on large scales, with a difference smaller than 10%. For the abundance matching case, the situation appears to be unchanged for both the mass bins, with a few per-cent improvement, showed by COLA on large scales.

As we have already highlighted for the power spectrum, we are interested in the variance and the covariance of the correlators. In figure 6.11 we plot the bispectrum variance squared divided by the bispectrum squared for all the triangular configurations and in the lower panels we show the squared



q

Figure 6.10: Average real space bispectrum over 300 realizations. On the left low mass bins, on the right high mass bins. Each dot represents the bispectrum for a particular triangular configuration with that particular $k_1$. In the lower panels the relative difference of the approximated methods with respect to Minerva.

ratio between the variance obtained with approximate methods and Minerva. Looking at the "hard" mass selection, COLA bispectrum variance following the same behavior of the correspondent bispectrum, showing a disagreement with simulation variance larger than 20% for a large number of triangular configurations; PINOCCHIO shows a smaller differences for the most part not larger than 20%. In the higher mass bin, PINOCCHIO variance shows an enhancement of $\sim 20\%$ for the majority of the triangular configurations while

COLA shows the opposite behavior, underestimating the simulation by not more than 20%. The situation we show for the abundance matching case is very different from the previous case: for the low mass bin the bispectrum variance of both COLA and PINOCCHIO shows disagreements smaller than the other cases. All the triangular configurations well reproduce the simulation with a difference smaller than $\sim 10\%$ at large scales ($k < 0.14\,h\,\mathrm{Mpc}^{-1}$) and smaller than 20% on smaller scales. COLA differences appeared to be a little reduced with respect to PINOCCHIO, but overall the two methods agree very well between each other and with simulation. Also for the high mass bin we can improve the results compared with the "hard" mass cut case; both PINOCCHIO and COLA show improvement in reproducing the simulation variance, keeping the differences smaller than 20%. As we have already noticed, PINOCCHIO tends to overestimates the bispectrum variance obtained from simulations, while COLA tends to underestimates it.

The last plot we show, figure 6.12, concerns the analysis of the bispectrum cross-correlation coefficient, defined in the same way of eq. 6.10. The main difference is that the power spectrum covariance depends on two wavenumbers we have called $k_i$ and $k_j$, while the bispectrum covariance matrix depends on six wavenumbers, so that in this case the subscript $i$ correspond to one triangular configuration, defined by $(k_{i,1}, k_{i,2}, k_{i,3})$, and the subscript $j$ to another triangular configuration, defined by $(k_{j,1}, k_{j,2}, k_{j,3})$.

We plot the cross-correlation coefficient in analogy with what we have done for the power spectrum. Each panel is a slice of the bispectrum covariance matrix obtained fixing one particular triangular configuration $i$ and varying the others configurations $j$. On the x-axis we plot the three wavenumbers values, $k_{j,1}, k_{j,2}, k_{j,3}$, of the the various triangular configurations. From the plot we can study the correlation of the error between different triangular configurations. We have chosen the covariance row to show the structure of the covariance matrix at different scales. As we have already noticed in the power spectrum analysis, the information given by the different mass bins and matching procedures is the same. It is clear that the approximate methods are capable to reproduce the features of the bispectrum cross-correlation coefficient obtained with simulation. As for the power spectrum, the noise in the off-diagonal elements of the covariance evaluated with N-body simulation, is well reproduced by PINOCCHIO and COLA.

## 6.3 Discussion

In this section we want to summarize the main results of this chapter, stressing what we have learned in the analysis of the approximate methods. As we have already noticed, the accuracy of the power spectrum and bispectrum evaluated on samples obtained with approximate methods is lower when compared with the measures on full N-body simulation. This is due to the

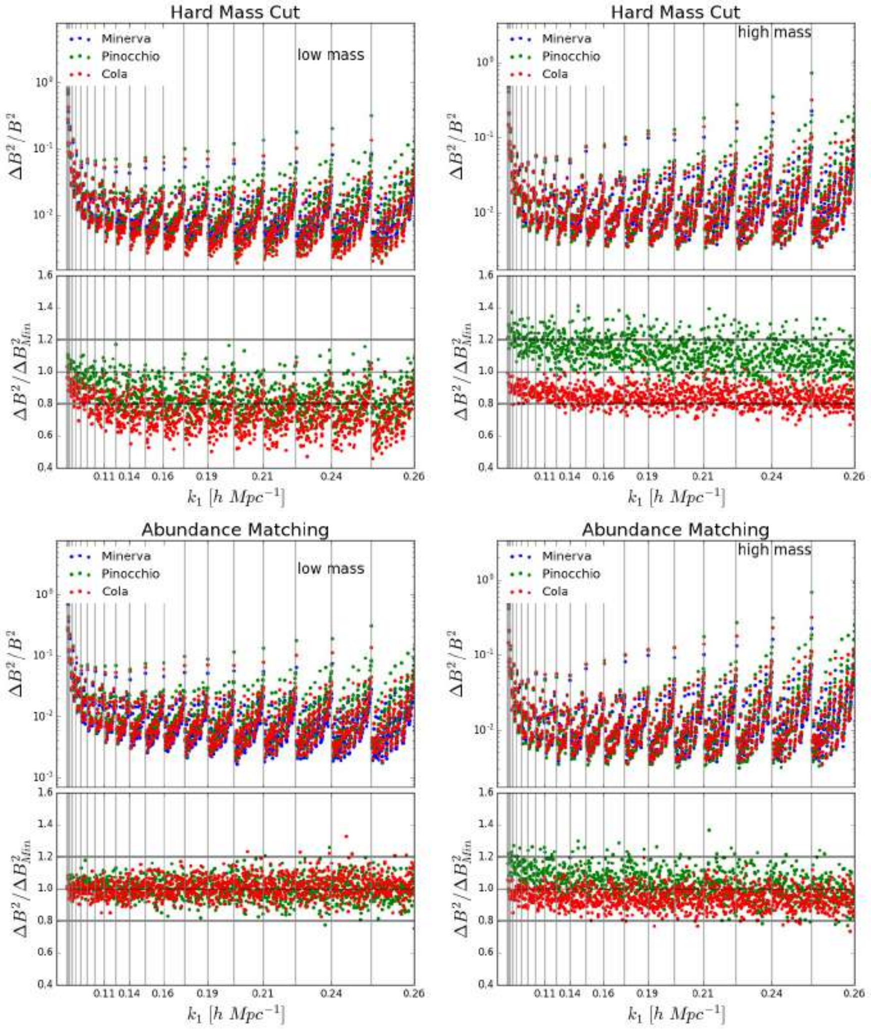


Figure 6.11: Bispectrum variance in real space. On the left low mass bins, on the right high mass bins. In the lower panels the relative difference of the approximated methods with respect to Minerva.

approximations that characterize the fast methods and that do not allow us to well reproduce the non-linear scales.

What we observe, looking at the average power spectrum, is in fact a lack of power in the description of the small scales, $k > 0.1\,h\,\mathrm{Mpc}^{-1}$; on these scales the simulations cannot be reproduced with an accuracy smaller than 20%, apart from the hexadecapole that is a quantity very noisy and very close to zero on these scales. We observe a similar behavior for the average bispectrum, but because $B \sim P^2$, for this statistics, the accuracy of the approximate methods appears to be smaller when compared with the power spectrum. The bispectrum is a more complex quantity and we

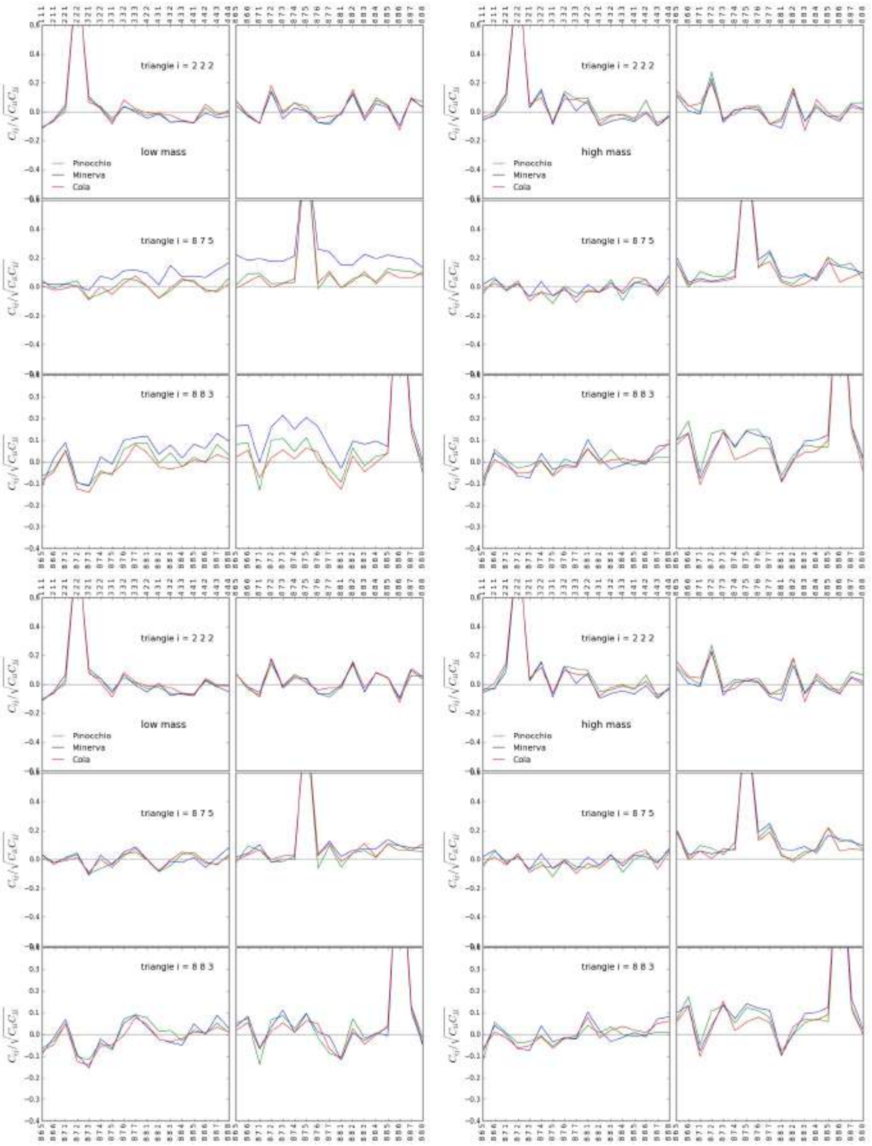


Figure 6.12: Bispectrum cross-correlation coefficient in real space. On the left low mass bins, on the right high mass bins.

have to take in account that there are different triangular configurations, so different combinations of the three wavenumbers. Looking at large scales ($k < 0.11\,h\,\mathrm{Mpc}^{-1}$) the majority of the configurations well reproduce the simulation, but going to smaller scales, a significant part of the total triangular configurations systematically overestimates or underestimates the simulation. It is worth to say that for the power spectrum and bispectrum analyses, COLA is slightly better than the other codes in reproducing the simulation, but we can expect this high accuracy by COLA, because it use a procedure very similar to an N-body simulation: it solves the large scales using LPT and the small scales using an N-body simulation (see section 4.2). The lack of accuracy on small scales is not the main issue, because the aim of this project is not to use the approximate methods to describe the statistics of clustering, but to accurate evaluate their covariance matrices. What we hope is to reproduce the errors on the power spectrum and bispectrum better than spectra themselfs. This is what we observe looking at the variances: for the power spectrum we are capable to reproduce the variance also on small scales, with a accuracy higher than that observed for the power spectrum; moreover all the codes agree better between each other, so that they appear to reproduce the same error an all scales. This is true also looking at the bispectrum variance: the spread of the different triangular configurations is smaller and constant an all scales; almost all the configurations coming from simulations are reproduced with a accuracy higher than the one of the bispectrum itself. Constraints on the cosmological model requires the modeling of the full covariance matrix, so in the last part of the section on power spectrum and bispectrum we have showed how the off-diagonal elements of the metrices behave. Because of the small number of realizations we have used, we cannot to really see the correlations of the error on different scales, but what we can conclude is just as important: the noise in the off-diagonal terms that we observe from the measures on the simulation is the same that we observe using approximate methods; not only the level of the noise is the same, but also the correlation of the noise is very well reproduced. In conclusion the approximate methods can reproduce the features of the covariances matrix obtained with a full N-body simulation. It is worth to mention that for the power spectrum case, looking at particular scales we can see small enhancements of the particular off-diagonal terms of the covariance matrix that are not due only to statistical noise. Increasing the number of realizations we can really lower this noise and study the correlation between different scales. We will show how a large ensemble of realizations affects the structure of the covariance matrix in chapter 8.

# Chapter 7

# Galaxy power spectrum covariance matrix

In the previous chapters we have stressed the necessity to have a large set of simulated galaxy catalogs for the accurate evaluation of the power spectrum covariance matrix. We have also described various techniques aimed at determining the covariance matrix with a limited number of simulations, facing the problem from another perspectives. All these techniques are characterized by some kind of analytic approximations or they depend on a certain number of free parameters.

One could consider, at same time, a full analytic prediction for the power spectrum covariance matrix that includes the correlation of modes on the non-linear scales that determines non-Gaussianity. (Scoccimarro et al., 1999; Hu and White, 2001; Cooray and Hu, 2001). A great effort was put into place in the last years to model the matter power and its covariance matrix (Takahashi et al., 2009; Sato et al., 2011; Takada and Hu, 2013; Mohammed and Seljak, 2014; Blot et al., 2015, 2016). As we have already highlighted, the aim of this project is to go beyond the matter density field, therefore to model the *galaxy* power spectrum covariance. At same time it is worth to briefly describe the state of the art of the matter power spectrum covariance matrix analyses. This is the purpose of the first section. This kind of introduction will help us to point out the differences between what is already been done and what we want to achieve with our analysis.

An accurate description of a real galaxy survey requires the modeling of the galaxy density field, that is the observable we have access to. The modeling of the galaxy power spectrum covariance matrix is more complex because of galaxy bias and discreetness effects. When realistic effects are taken in account, such as selection function features or systematic effects, like those described in chapter 5, an analytic estimation is too complex and numerical methods are required. For this reason we do not expect the analytic predictions to replace numerical evaluations of the covariance matrix, but they can help to reduce the number of realizations needed for its evaluation (see section 6.1).

In the following sections we describe how to include the shot-noise contribution in the power spectrum covariance matrix. Then we define the bias model we use and how this definition enters in the theoretical predictions. In the results section we show the comparison between the analytic predictions and the measurements of the power spectrum covariance matrix obtained using 10’000 PINOCCHIO simulated galaxy catalogs, highlighting the terms that mainly contribute to the variance and to the off-diagonal elements of the matrix. In the last section we discuss the results and we describe how to improve them.

## 7.1 The state of the art

In section 6.1 we have already described some procedures to model the matter power spectrum covariance. A full power spectrum matter covariance matrix prediction is given by Bertolini et al. (2016) and Mohammed et al. (2017).

In Mohammed et al. (2017) the authors use perturbation theory at 1-loop order and they compare the prediction with simulations. Their matter power spectrum covariance model includes the non-Gaussian part given by the trispectrum considering both the modes outside and inside the survey.

To show how the different components contribute they decompose the non-Gaussian part of total covariance matrix in three terms:

$$\mathbf{Cov}_{ij}^{\mathrm{Full}} = \mathbf{Cov}_{ij}^{\mathrm{G}} + \left[\mathbf{Cov}_{ij}^{\mathrm{BC}} + \mathbf{Cov}_{ij}^{\mathrm{Tree-level}} + \mathbf{Cov}_{ij}^{\mathrm{1-loop}}\right] , \tag{7.1}$$

where $\mathbf{Cov}_{ij}^{\mathrm{G}}$ is the diagonal Gaussian contribution, $\mathbf{Cov}_{ij}^{\mathrm{BC}}$ is a tree-level contribution that is responsible for the beat coupling (BC) or super-sample covariance (SSC), e.g. (Hamilton et al., 2006; Sefusatti et al., 2006; de Putter et al., 2012), $\mathbf{Cov}_{ij}^{\mathrm{Tree-level}}$ is the term proportional to the tree-level trispectrum and $\mathbf{Cov}_{ij}^{\mathrm{1-loop}}$ is the 1-loop contribution from the 1-loop trispectrum that they consider in their model. The BC term is given by the tree-level trispectrum in the squeezed limit that can be modeled as the response of the matter power spectrum to the change of the background density:

$$\mathbf{Cov}_{ij}^{\mathrm{BC}} = \sigma_b^2 \frac{\partial P(k_i)}{\partial \delta_b} \frac{\partial P(k_j)}{\partial \delta_b} , \tag{7.2}$$

where $\sigma_b^2$ is the variance of the mean density field in the survey window.

The tree-level trispectrum is given by (Scoccimarro et al., 1999):

$$\begin{aligned}\bar{T}(k_i,k_j) &= \int_{k_i}\frac{d^3k_1}{V_s(k_i)}\int_{k_i}\frac{d^3k_2}{V_s(k_j)}\Big(12F_3^{(s)}(\mathbf{k}_1,-\mathbf{k}_1,\mathbf{k}_2)P_L(\mathbf{k}_1)^2P_L(\mathbf{k}_2)\Big)\\ &+ 8F_2^{(s)}(\mathbf{k}_1-\mathbf{k}_2,\mathbf{k}_2)^2P_L(\mathbf{k}_2)^2P_L(\mathbf{k}_2-\mathbf{k}_1)\\ &+ 8F_2^{(s)}(\mathbf{k}_1-\mathbf{k}_2,\mathbf{k}_2)F_2^{(s)}(\mathbf{k}_2-\mathbf{k}_1,\mathbf{k}_1)P_L(\mathbf{k}_1)P_L(\mathbf{k}_2-\mathbf{k}_1) \quad (7.3)\\ &+ \{\mathbf{k}_1\leftrightarrow\mathbf{k}_2\}\ . \quad (7.4)\end{aligned}$$

A full calculation on all the 1-loop contribution has been recently presented in Bertolini et al. (2016); in the specific case of Mohammed et al. (2017) the 1-loop contribution that they consider has the following form:

$$\mathbf{Cov}_{ij}^{1-loop} = \frac{1}{V\pi^2}\int dq\ P_L^2(q)q^2\mathbf{V}(q,k_i)\mathbf{V}(q,k_j)\ , \tag{7.5}$$

where $\mathbf{V}(\mathbf{q},\mathbf{k})$ is the normalized functional derivative given by Nishimichi et al. (2016):

$$\mathbf{V}(\mathbf{q},\mathbf{k}) = \frac{P_{\rm L}(q)}{\Delta^2(q)}\Big\langle\frac{\delta P_{\text{1-loop}}(\mathbf{k})}{\delta P_{\rm L}(\mathbf{k})}\Big\rangle_\Omega\ , \tag{7.6}$$

where $\delta$ stays for the functional derivative, $\Delta^2(q) = 4\pi q^3P(q)/(2\pi)^3$ and $\langle ...\rangle$ is the angle averaging.

The covariance matrix, eq. 7.1, evaluated in this way is compared with a large set of N-body simulations (Blot et al., 2015) (see section 3 of Mohammed et al. (2017)).

In figure 7.1 they show a quantity that we will use in the next section, defined as the full covariance matrix normalized to the power spectrum:

$$c_{ij} \equiv \frac{C_{ij}}{P_{tot}(k_i)\,P_{tot}(k_j)} \tag{7.7}$$

where $P_{tot}$ include the shot-noise contributions.

All the three components are showed, tree-level term with a green dashed line, one-loop term with a blue dotted line and the BC term with a yellow dashed line. The predictions are compared with the simulations, the solid black line; the cyan solid line is the only Gaussian part of the matrix. They are capable to reproduce the results from simulations to about 10% in the quasi linear regime. The BC term ends up to be the most relevant one, while the tree-level trispectrum is important on large scales. The smaller scales are dominated by the 1-loop term. They find that the agreement with simulations is better when the BC is included even without the additional non-linear correction. The non-Gaussian part of the matrix is dominated by one single eigenmode so that the non-Gaussian response has always the same shape, but its amplitude can varies. If one eigenmode dominates, an alternative approach could be the following: the non-Gaussian covariance

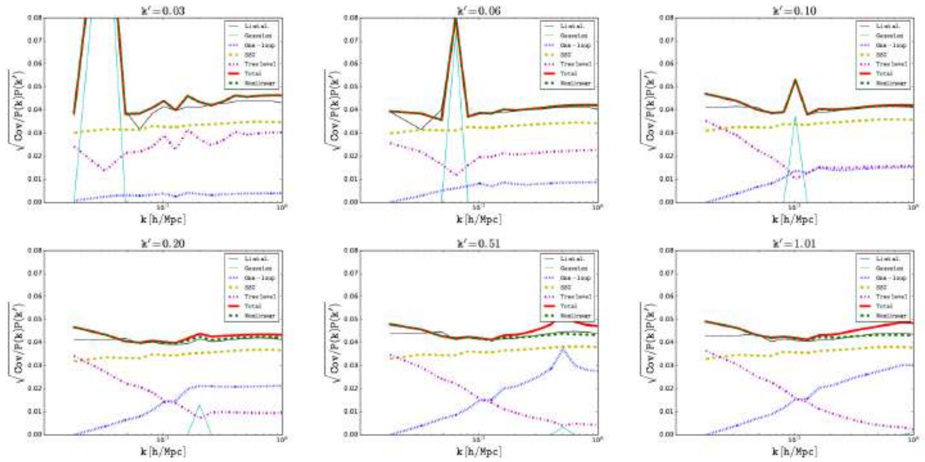


Figure 7.1: Comparison from, (Mohammed et al., 2017), of the analytic model with simulations from (Li et al., 2014a,b) at redshift 0.0.

is ignored and the eigenvector response is included as a fictitious external parameter in the analysis. The only issue with this procedure is that the parameter would be quite degenerate with other cosmological parameters. They show also results at high redshift where the non-linear contribution is suppressed relative to the BC one and the agreement with simulations is improved.

As they pointed out in the conclusion of the paper, this model could be applied to galaxy clustering surveys only if the bias, the shot-noise and the redshift space distortions terms are included.

In the next sections we describe our model for the galaxy power spectrum covariance, studying two of these additional contributions: the bias and the shot-noise terms.

## 7.2 Covariance of the galaxy power spectrum

We want to obtain a theoretical prediction for the covariance of the power spectrum of biased objects, halos in this case, for the simple case of cubic periodic box in real space. The estimator for the power spectrum we use is defined in eq. 3.2. The estimator for the shot-noise-corrected power spectrum can then be written as

$$\hat{P}(k) = \hat{P}_{tot} - \hat{P}_{SN} \ . \tag{7.8}$$

It is important to stress that $\hat{P}_{SN}$ is the shot-noise of the *single realization.* This is a fundamental point because it impacts in the way we treat the shot-noise terms in the the covariance matrix prediction.

### 7.2.1 Biased tracers

We consider an approximation for the second contribution in the r.h.s of the equation 7.7, consisting in the linear theory and tree-level PT predictions respectively for the power spectrum $P$ and the higher-order correlators, the bispectrum $B$ and the trispectrum $T$. We assume the following bias model for the halo overdensity $\delta_h$ (Chan et al., 2012):

$$\delta_h = b_1\delta + \frac{1}{2}b_2\delta^2 + \gamma_2\mathcal{G}_2 + \frac{1}{6}\tilde{b}_3\delta^3 + \mathcal{O}(\delta^4) \ . \tag{7.9}$$

with $\delta$ being the matter overdensity and where $\mathcal{G}_2$ defined by eq. 2.93.

In eq.7.9, $\tilde{b}_3$ is to be interpreted as an effective cubic bias correction since we are ignoring all non-local correction beyond the quadratic one. We choose to write, and compute, all contributions with their explicit dependence on bias parameters. This is more practical since, factoring-out the bias, what is left is the same bias for all halo masses and needs to be computed once. In addition one can highlight the relative importance of each parameter.

We consider the simple linear approximation for the halo power spectrum

$$P_g(k) = b_1^2\, P_L(k) \tag{7.10}$$

$P_L(k)$ being the linear matter power spectrum, but we will explain later that non-linear correction are needed to study the off-diagonal term of the covariance matrix in a consistent way.

As we have already pointed out in eq. 8.18, for the bispectrum we have

$$\begin{aligned} B_g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = b_1^3\, B_m(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \quad + & \quad b_1^2\, b_2\, P_L(k_1)\, P_L(k_2) + \text{perm.} + \\ & 2\, b_1^2\, \gamma_2\, \Sigma_2(\mathbf{k}_1,\mathbf{k}_2)\, P_L(k_1)\, P_L(k_2) + \text{perm.} \end{aligned} \tag{7.11}$$

where the matter bispectrum is defined by eq. 2.61 and $\Sigma_2(\mathbf{k}_1,\mathbf{k}_2) \equiv \cos_{12}\theta - 1$. In the same way of the galaxy bispectrum we can obtain the expression for the galaxy trispectrum:

$$\begin{aligned} & T_g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) = b_1^4\, T_m(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) \\ + \quad & b_1^3 b_2[P_L(k_1)B_m(k_3,k_4,k_{34}) + 11 \text{ perm.}] \\ + \quad & 2b_1^3\gamma_2[\Sigma_2(\mathbf{k}_1,\mathbf{k}_{34})P_L(k_1)B_m(k_3,k_4,k_{34}) + 11 \text{ perm.}] \\ + \quad & b_1^2 b_2^2\,[P_L(k_1)P_L(k_2)P_L(k_{13}) + 11 \text{ perm.}] \\ + \quad & 4b_1^2\gamma_2^2\,[\Sigma(\mathbf{k}_1,-\mathbf{k}_{13})\Sigma(\mathbf{k}_2,-\mathbf{k}_{24})P_L(k_1)P_L(k_2)P_L(k_{13}) + 11 \text{ perm.}] \\ + \quad & b_1^2 b_2\gamma_2\,\{[\Sigma(\mathbf{k}_1,-\mathbf{k}_{13}) + \Sigma(\mathbf{k}_2,-\mathbf{k}_{24})]\,P_L(k_1)P_L(k_2)P_L(k_{13}) + 11 \text{ perm.}\} \\ + \quad & b_1^3\tilde{b}_3 P_L(k_1)P_L(k_2)P_L(k_3) + 3 \text{ perm.} \\ . \quad & \end{aligned} \tag{7.12}$$

We denote the contributions in the expressions above as

$$T_g = T_{g,b_1} + T_{g,b_2} + T_{g,\gamma_2} + T_{g,b_2^2} + T_{g,\gamma_2^2} + T_{g,b_2\gamma2} + T_{g,\tilde{b}_3} \ , \tag{7.13}$$

where the notation should be clear enough.

### 7.2.2 Power spectrum covariance and shot-noise

The covariance of $\hat{P}$ should account, in addition to the terms depending on $\hat{P}_{tot}$, for the contribution due to the shot-noise correction. In fact, we have

$$\begin{aligned} C_{ij} &\equiv \langle \hat{P}(k_i)\hat{P}(k_j)\rangle - \langle \hat{P}(k_i)\rangle\langle \hat{P}(k_j)\rangle \\ &= \langle \hat{P}_{tot}(k_i)\hat{P}_{tot}(k_j)\rangle - \langle \hat{P}_{tot}(k_i)\ \hat{P}_{SN}\rangle - \langle \hat{P}_{tot}(k_j)\ \hat{P}_{SN}\rangle + \langle \hat{P}^2_{SN}\rangle\ . \end{aligned} \tag{7.14}$$

Since

$$\langle \hat{P}^2_{SN}\rangle \equiv \langle \hat{P}_{SN}\rangle\langle \hat{P}_{SN}\rangle + \Delta P^2_{SN}\ , \tag{7.15}$$

where we denoted with $\Delta P^2_{SN} = \mathrm{Var}(\hat{P}_{SN})$ the variance of the shot-noise contribution, we have

$$\langle \hat{P}_{tot}(k_i)\ \hat{P}_{SN}\rangle = \langle \hat{P}_{tot}(k_i)\rangle\langle \hat{P}_{SN}\rangle + \Delta P^2_{SN} \tag{7.16}$$

so that, introducing the covariance $C^{tot}_{ij} \equiv \langle \hat{P}_{tot}(k_i)\hat{P}_{tot}(k_j)\rangle - \langle \hat{P}_{tot}(k_i)\rangle\langle \hat{P}_{tot}(k_j)\rangle$ of the total power spectrum estimator $\hat{P}_{tot}(k)$, we obtain

$$\begin{aligned} C_{ij} &= C^{tot}_{ij} - \langle \hat{P}_{tot}(k_i)\rangle\langle \hat{P}_{SN}\rangle - 2\ \Delta P^2_{SN} - \langle \hat{P}_{tot}(k_j)\rangle\langle \hat{P}_{SN}\rangle + \Delta P^2_{SN} + \langle \hat{P}_{SN}\rangle^2 \\ &\quad - \left\{ - \left[ \langle \hat{P}_{tot}(k_i)\rangle + \langle \hat{P}_{tot}(k_j)\rangle \right] \langle \hat{P}_{SN}\rangle + \langle \hat{P}_{SN}\rangle^2 \right\} \\ &= C^{tot}_{ij} - \Delta P^2_{SN} \end{aligned} \tag{7.17}$$

We can estimate this quantity in terms of the variance in the total number of objects $N$, assumed to be a Poisson variable:

$$\Delta P^2_{SN} \simeq \frac{1}{k_f^6}\frac{1}{\langle N\rangle^4}\Delta N^2 = \frac{1}{k_f^6}\frac{1}{\langle N\rangle^3} = k_f^3\, P^3_{SN}\ . \tag{7.18}$$

The covariance for the estimator $\hat{P}_{tot}$ can be expressed as

$$\begin{aligned} C^{tot}_{ij} &\equiv \langle \hat{P}_{tot}(k_i)\hat{P}_{tot}(k_j)\rangle - \langle \hat{P}_{tot}(k_i)\rangle\langle \hat{P}_{tot}(k_j)\rangle \\ &\simeq \delta_{ij}\ \frac{2}{N_{k_i}} P^2_{tot,g}(k_i) + k_f^3 \widetilde{T}_{tot,g}(k_i,k_j) \end{aligned} \tag{7.19}$$

where the first term is the Gaussian contribution while the second is the average of the trispectrum $T_g(\mathbf{k}_i, -\mathbf{k}_i, \mathbf{k}_j, -\mathbf{k}_j)$ over the angle $\theta$ between the vectors $\mathbf{k}_i$ and $\mathbf{k}_j$, given by

$$\widetilde{T}_g(k_i,k_j) \equiv \frac{1}{2}\int_{-1}^{+1} d\cos\theta\, T_g(\mathbf{k}_i, -\mathbf{k}_i, \mathbf{k}_j, -\mathbf{k}_j)\ . \tag{7.20}$$

We then obtain

$$
\begin{aligned}
& \widetilde{T}_g(k_i,k_j) = b_1^4 \widetilde{T}_m(k_i,k_j) \\
+ \quad & 4b_1^3 b_2 \left[P_L(k_i) + P_L(k_j)\right] \widetilde{B}_m(k_i,k_j) \\
+ \quad & 8b_1^3 \gamma_2 \left\{ P_L(k_i) \int_{-1}^{1} d\cos\theta B_m(\mathbf{k}_i,\mathbf{k}_j) \left[\Sigma_2(\mathbf{k}_i,-\mathbf{k}_{ij}) + \Sigma_2(\mathbf{k}_j,-\mathbf{k}_{ij})\right] \right\} \\
+ \quad & 2b_1^2 b_2^2 \left[P_L(k_i) + P_L(k_j)\right]^2 \widetilde{P}_L(k_i,k_j) \\
+ \quad & 8b_1^2 \gamma_2^2 \left\{ P_L^2(k_i) \int_{-1}^{1} d\cos\theta_{ij} \Sigma_2^2(\mathbf{k}_i,-\mathbf{k}_{ij}) P_L(k_{ij}) \right. \\
+ \quad & 2P_L(k_i)P_L(k_j) \int_{-1}^{1} d\cos\theta_{ij} \Sigma_2(\mathbf{k}_i,-\mathbf{k}_{ij}) \Sigma_2(\mathbf{k}_j,\mathbf{k}_{ij}) P_L(k_{ij}) \\
+ \quad & \left. P_L^2(k_j) \int_{-1}^{1} d\cos\theta_{ij} \Sigma_2^2(\mathbf{k}_j,-\mathbf{k}_{ij}) P_L(k_{ij}) \ \right\} \\
+ \quad & 4b_1^2 b_2 \gamma_2 \left\{ P_L^2(k_i) \int_{-1}^{1} d\cos\theta_{ij} \Sigma_2(\mathbf{k}_i,-\mathbf{k}_{ij}) P_L(k_{ij}) \right. \\
+ \quad & P_L(k_i)P_L(k_j) \int_{-1}^{1} d\cos\theta_{ij} \left[\Sigma_2(\mathbf{k}_i,-\mathbf{k}_{ij}) + \Sigma_2(\mathbf{k}_j,\mathbf{k}_{ij})\right] P_L(k_{ij}) \\
+ \quad & \left. P_L^2(k_j) \int_{-1}^{1} d\cos\theta_{ij} \Sigma_2(\mathbf{k}_j,-\mathbf{k}_{ij}) P_L(k_{ij}) \ \right\} \\
+ \quad & 2b_1^3 \widetilde{b}_3 P_L(k_i) P_L(k_j) \left[P_L(k_i) + P_L(k_j)\right] \ . \qquad (7.21)
\end{aligned}
$$

where $\widetilde{B}_g(k_i,k_j)$ is the average of the bispectrum $B(\mathbf{k}_{ij},\mathbf{k}_i,\mathbf{k}_j)$ over the angle $\theta$ between the vectors $\mathbf{k}_i$ and $\mathbf{k}_j$, given by:

$$
\begin{aligned}
\widetilde{B}_g(k_i,k_j) &= b_1^3, \widetilde{B}_m(k_i,k_j) \\
&+ b_1^2 b_2 \left[P_L(k_i)P_L(k_j) + P_L(k_i)\widetilde{P}_L(k_i,k_j) + P_L(k_j)\widetilde{P}_L(k_i,k_j)\right] \\
&+ 2b_1^2 \gamma_2 \left[ -\frac{1}{3} P_L(k_i)P_L(k_j) + P_L(k_i) \int d\cos\theta_{ij} \Sigma_2(\mathbf{k}_i,-\mathbf{k}_{ij}) P_L(\mathbf{k}_{ij}) \right. \\
& \left. + P_L(k_j) \int d\cos\theta_{ij} \Sigma_2(\mathbf{k}_j,-\mathbf{k}_{ij}) P_L(\mathbf{k}_{ij}) \right] \ . \qquad (7.22)
\end{aligned}
$$

All correlation functions appearing in the expression for $C_{ij}^{tot}$ include shot-noise contributions, so

$$
\begin{aligned}
P_{tot}(k) &= P_g(k) + P_{SN} \, , \qquad (7.23) \\
T_{tot}(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) &= T_g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) + P_{SN}\left[B_g(k_{12},k_3,k_4) + 5 \text{ perm.}\right] \\
& + P_{SN}^2 [P_g(k_1) + P_g(k_2) + P_g(k_3) + \\
& P_g(k_4) + P_g(k_{12}) + P_g(k_{13}) + P_g(k_{14})] + P_{SN}^3 . (7.24)
\end{aligned}
$$

In particular,

$$\begin{aligned} T_{\rm tot}(\mathbf{q},-\mathbf{q},\mathbf{p},-\mathbf{p}) &= T_g(\mathbf{q},-\mathbf{q},\mathbf{p},-\mathbf{p}) \\ &+ 2\,P_{SN}\left[B_g(\mathbf{q}+\mathbf{p},-\mathbf{q},-\mathbf{p})+B_g(\mathbf{q}-\mathbf{p},-\mathbf{q},\mathbf{p})\right] \\ &= P_{SN}^2\left[2\,P_g(q)+2\,P_g(p)+P_g(|\mathbf{q}+\mathbf{p}|)+P_g(|\mathbf{q}-\mathbf{p}|)\right] \\ & +P_{SN}^3 \end{aligned} \qquad (7.25)$$

so that

$$\begin{aligned} \widetilde{T}_{tot}(k_i,k_j) &\simeq \frac{1}{2}\int_{-1}^{+1} d\cos\theta\, T_g(\mathbf{k}_i,-\mathbf{k}_i,\mathbf{k}_j,-\mathbf{k}_j) \\ &+ 4\,P_{SN}\,\frac{1}{2}\int_{-1}^{+1} d\cos\theta\, B_g(\mathbf{k}_{ij},-\mathbf{k}_i,-\mathbf{k}_j) \\ &+ 2\,P_{SN}^2\,\left[P_g(k_i)+\,P_g(k_j)\right]+2\,P_{SN}^2\frac{1}{2}\int_{-1}^{+1} d\cos\theta\, P_g(k_{ij})+k_f^3\,P_{SN}^3 \\ &\equiv \widetilde{T}_g(k_i,k_j)+4\,P_{SN}\,\widetilde{B}_g(k_i,k_j) \\ & +2\,P_{SN}^2\,\left[P_g(k_i)+\,P_g(k_j)+\widetilde{P}_g(k_i,k_j)\right]+P_{SN}^3\,, \end{aligned} \qquad (7.26)$$

where $k_{ij}=k_{ij}(k_i,k_j,\theta)=|\mathbf{k}_{ij}|$ and where we further defined

$$\begin{aligned} \widetilde{B}(k_i,k_j) &\equiv \frac{1}{2}\int_{-1}^{1} d\cos\theta\; B(k_{ij},k_i,k_j) \\ &= \frac{1}{2\,k_i\,k_j}\int_{|k_i-k_j|}^{k_i+k_j} dq\,q\,B(q,k_i,k_j)\ , \end{aligned} \qquad (7.27)$$

and

$$\begin{aligned} \widetilde{P}(k_i,k_j) &\equiv \frac{1}{2}\int_{-1}^{1} d\cos\theta\; P(k_{ij}) \\ &= \frac{1}{2\,k_i\,k_j}\int_{|k_i-k_j|}^{k_i+k_j} dq\,q\,P(q)\ . \end{aligned} \qquad (7.28)$$

We can therefore express the covariance $C_{ij}$ with all explicit shot-noise contributions as:

$$\begin{aligned} C_{ij} &= \delta_{ij}\,\frac{2}{N_{k_i}}P_{tot}^2(k_i)+k_f^3\,\widetilde{T}_{tot}(k_i,k_j)+k_f^3\,P_{SN}^3-\Delta P_{SN}^2 \\ &\simeq \delta_{ij}\,\frac{2}{N_{k_i}}\left[P_g(k_i)+P_{SN}\right]^2 \\ & +k_f^3\,\left\{\widetilde{T}_g(k_i,k_j)+4\,P_{SN}\,\widetilde{B}_g(k_i,k_j)\right. \\ & \left. +2\,P_{SN}^2\,\left[P_g(k_i)+\,P_g(k_j)+\widetilde{P}_g(k_i,k_j)\right]\right\}\ , \end{aligned} \qquad (7.29)$$

where we assumed $\Delta P_{SN}^2 \simeq k_f^3 \, P_{SN}^3$. In particular, the power spectrum variance reads:

$$\begin{aligned} \Delta P_g^2(k) &= \frac{2}{N_k}\left[P_g(k)+P_{SN}\right]^2 + k_f^3 \, \left\{\widetilde{T}_g(k,k) + 4\,P_{SN}\,\widetilde{B}_g(k,k)\right. \\ & \left. +2\,P_{SN}^2\left[2\,P_g(k)+\widetilde{P}_g(k,k)\right]\right\} \end{aligned} \tag{7.30}$$

We are particularly interested in a theoretical description of the off-diagonal contributions to the covariance matrix $C_{ij}$. A robust prediction for the Gaussian, diagonal component can simply be given in terms of the measured, total power spectrum $P_{tot}$. An interesting quantity, therefore, it is the *reduced*, adimensional covariance already defined in eq. 7.7, that in our case reads:

$$\begin{aligned} c_{ij} &\equiv \frac{C_{ij}}{P_{tot}(k_i)\,P_{tot}(k_j)} \\ &= \delta_{ij}\,\frac{2}{N_{k_i}} + k_f^3\frac{\widetilde{T}_g(k_i,k_j)+4\,P_{SN}\,\widetilde{B}_g(k_i,k_j)+2\,P_{SN}^2\left[P_g(k_i)+\,P_g(k_j)+\widetilde{P}_g(k_i,k_j)\right]}{\left[P_g(k_i)+P_{SN}\right]\left[P_g(k_j)+P_{SN}\right]}\,. \end{aligned} \tag{7.31}$$

In order to be consistent with the tree-level approximation for the trispectrum, we would need to include 1-loop corrections to $P_g(k)$, eq. 7.10 in the denominator of the reduced covariance matrix defined above.

The diagonal corresponds to

$$\begin{aligned} \frac{\Delta P^2(k)}{P_{tot}^2(k)} &= \frac{2}{N_k} \\ &+ \quad k_f^3\frac{\widetilde{T}_g(k,k)+4\,P_{SN}\,\widetilde{B}_g(k,k)+2\,P_{SN}^2\left[2\,P_g(k)+\widetilde{P}_g(k,k)\right]}{\left[P_g(k)+P_{SN}\right]^2}\,. \end{aligned} \tag{7.32}$$

## 7.3 Results

In this section we show the comparison between the theoretical prediction and the measures obtained from the simulated galaxy catalogs. We use the 10,000 PINOCCHIO catalogs we have already described in section 5.3.1. We consider all the halos above the mass threshold given by 40 times the halo particle mass $M_p = 2.67 \times 10^{11} M_\odot$.

In figure 7.2, we show the average measured power spectrum with and without shot-noise (red and green solid lines), compared with the linear (cyan dashed line) and non-linear (blue dashed line) power spectrum in Standard Perturbation Theory (SPT) with Gaussian initial conditions, eq. 2.57; to compare the average power spectrum measured on the simulated catalogs, the theory power spectra account for a shot-noise contribution obtained as the average shot-noise over all the 10,000 realizations. The horizontal gray line is the average shot-noise. In the lower panel we plot the ratio between the two predictions and the total measured power spectrum. The power spectrum measured from the PINOCCHIO realizations matches with the prediction up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$ where the differences are of 5%. A comparision of the measured power spectrum with the expected one from theory is fundamental for the following analysis of the power spectrum covariance matrix: the Gaussian part of eq. 7.29 is defined directly using the measured power spectrum, so we have to be sure that it well reproduces the theory on the scales of interest.

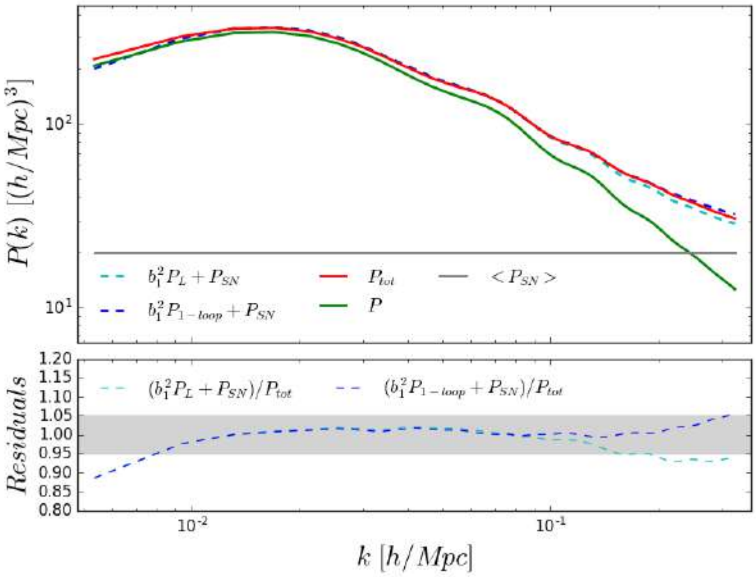


Figure 7.2: Comparison of the averaged power spectrum over 10'000 realizations with (solid red line) and without (solid green line) shot-noise, against the theory predictions in linear theory (dashed cyan line) and non-linear theory (dashed blue line). Theory includes the average shot-noise over the 10,000 realizations, plotted with a solid gray line. In the lower panel the residuals is given by the ratio of the theory predictions with respect the measured power spectrum including shot-noise.

Our aim is to model the off-diagonal elements of the covariance matrix that take into account the non-Gaussianity. First of all we can check if our model is capable to reproduce the diagonal of the matrix. We expect the major contribution comes from the Gaussian part, but the diagonal is also modified by non-Gaussian corrections, eq. 7.30. In figure 7.3 we plot, in the upper panel, the power spectrum variance divided by the total power spectrum, in the middle panel the ratio between the measured variance with respect the analytic prediction and in the lower panel the contributions given by the main trispectrum terms to the total variance prediction. In the upper panel the solid red line is the measured variance, the solid purple line the Gaussian prediction while the solid, the dashed and the dotted lines are the most relevant contributions coming from the galaxy trispectrum and its shot-noise corrections:

$$\begin{aligned} T_{SN1}(k_i,k_j) &\equiv 4P_{SN}\,\widetilde{B}_g(k_i,k_j) & (7.33)\\ T_{SN2}(k_i,k_j) &\equiv 2P_{SN}^2\left[P_g(k_i)+\,P_g(k_j)+\widetilde{P}_g(k_i,k_j)\right]\,, & (7.34)\end{aligned}$$

where $\widetilde{B}_g(k_i,k_j)$ is defined in eq. 7.22. The solid black line is the full predicted variance. It seems clear that on the scales we are investigating the Gaussian term is dominant, while the trispectrum corrections are negligible.

In the middle panel the red solid line shows the ratio between our measured variance and the predictions. As we have already mentioned, the variance is dominated by the Gaussian term so the large scales ($k < 0.01\,h\,\mathrm{Mpc}^{-1}$) difference is due to sample variance, while on the intermediate scales we are able to match the prediction with an error smaller than 10%. For scales smaller than $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$ t he decreasing trend seems to indicate an initial effect of the non-Gaussian corrections. Even if the trispectrum terms are smaller than the Gaussian one, they represent the largest additional contributions to the diagonal. In the lower panel we see how the term proportional to the matter trispectrum (solid green line) is decreasing on scales smaller than $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$, while the two shot-noise corrections to the trispectrum, eq. 7.34, begin to gaining power at same scales.

We do not plot the other non-Gaussian contributions of eq.7.21 because they are smaller than these three terms. For this analysis, we neglect the terms proportional to $b_3$ because we expect to be negligible with respect the other contributions.

In eq. 7.31 we have defined the adimensional covariance matrix to investigate the non-Gaussian correction to the Gaussian contribution simply given by $2/N_k$. In figure 7.4 we plot the reduced covariance matrix for particular values of $k_i$ varying $k_j$. From this figure we can study the accuracy of the predictions of the non-Gaussian terms and how they affect the structure of the covariance matrix. The solid red line is the measured reduced covariance, the solid black line the total prediction including the Gaussian part and all the trispectrum terms; the solid green line is the term proportional to $b_1^4 T$ while the dashed and the dotted green lines are the trispectrum shot-noise terms. When $k_i = k_j$ the reduced covariance matrix shows a peak corresponding to the Gaussian contribution plus the small, as we have seen in figure 7.3, corrections to the diagonal. We expect the black line to reproduce the observed covariance matrix, but we observe that our prediction overestimate our measurements on all the scales. The main reason appears

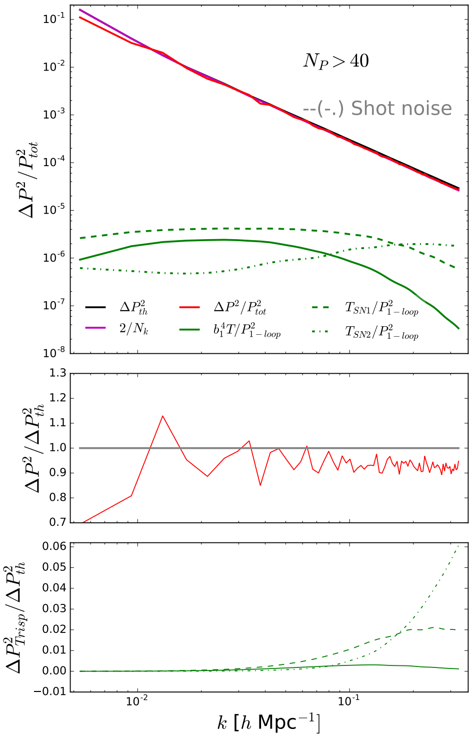


Figure 7.3: Comparison of the diagonal of the power spectrum covariance matrix obtained from 10'000 measured power spectrum (solid red line), against the Gaussian prediction (solid magenta line) and the full prediction (solid black line). In green are showed the main trispectrum contributions: the solid line represents the term proportional the $b_1^4$, the dashed and the dotted lines are the two shot-noise terms in the trispectrum. In the middle panel it is plotted the ratio between the measured variance and the full prediction, while in the lower panel the it is showed the ratio of the three trispectrum terms with respect to the full theory.

to be the overestimation of the shot-noise contributions to the trispectrum that produce an enhancement of the off-diagonal elements of the matrix.

## 7.4 Discussion

In this chapter we have described our analytic model for the full galaxy power spectrum covariance matrix. We have considered a bias model that includes the linear bias, the quadratic bias, the first non-local bias and an effective bias $b_3$ to model other non-linear corrections in the galaxy density field. From the analysis of the shot-noise contribution, we have discovered that it is fundamental to correct the measured power spectrum on each

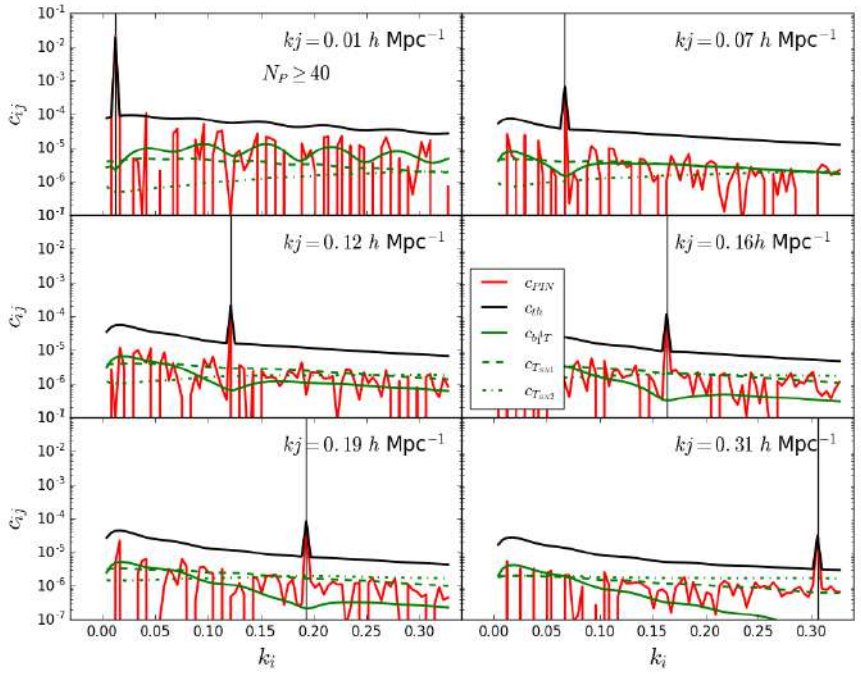


Figure 7.4: Adimensional reduced covariance matrix obtained varying $k_i$ for particular values of $k_j$. The color code is the same of figure 7.3.

realization with the shot-noise of that particular realization and not with a constant value across the realizations. The power spectrum estimator is modified when we consider this kind of shot-noise term; the consequence is a correction term in the power spectrum covariance matrix (see eq. 7.17).

Looking at the diagonal of the covariance matrix, we can say that it is dominated by sample variance at very large scales and by the Gaussian term for scale up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$. The main non-Gaussian corrections, coming from the first term of the galaxy trispectrum and from its shot-noise corrections are negligible on these scales, but the shot-noise contributions begin to get larger at scales smaller than $k = 0.3\,h\,\mathrm{Mpc}^{-1}$. The measured variance is well reproduced by the prediction.

To study the off-diagonal terms of the covariance matrix we have defined the adimensional covariance matrix as the full covariance normalized by the total power spectrum. We have divided the Gaussian part by the measured power spectrum, while the non-Gaussian part is divided by the 1-loop power spectrum. This quantity allow us to quantify how relevant are the non-Gaussian corrections to the Gaussian term. As for the variance there were the most relevant terms are those one we have described above. Contrary to the variance our full prediction is not capable to describe the features of the measured covariance matrix due to our modeling; we observe an overestimation on all the scales.

To get a more reliable and accurate description of the off-diagonal elements of the covariance matrix we will proceed with two main analyses:

- we have considered the simple case of Poissonian shot-noise; this is an approximation that probably is not valid in our case and it is known that non-Poissonian correction can be modeled and can be non-negligible, e.g. (Hamaus et al., 2010). The more important non-Gaussian terms that influence the off-diagonal covariance matrix depends on shot-noise, so a better modeling of this quantity will improve the modeling of all the covariance structure;

- the only measured quantity in the full covariance matrix prediction is the galaxy power spectrum, but it is also possible to obtain a measurements of the averaged trispectrum. A measurements of the trispectrum directly from the simulation can be compared with our predictions for the galaxy trispectrum.

We belive that these these two furhter analyses will provide insights on how to improve the agreement between model predictions and observations.

# Chapter 8

# Toward an analysis of the bispectrum

When we look at the early Universe (z$\sim$1100) we observe a Gaussian distribution; all the information we need to describe this field is given by the two-point statistics the power spectrum or the two-point correlation function, all the higher-order correlators are zero. Gravitational instability amplifies the initial Gaussian perturbations allowing the formation of structures we observe today. Gravity is an intrinsic non-linear process and these non-linearities give rise to non-zero higher-order statistics, that have to be take into account for a correct clustering analyisis.

We focus on the bispectrum that is the three-point function in Fourier space. As we have already mentioned in section 2.3.2, the bispectrum can be used to break some degeneracy in the galaxy bias parameters (Fry, 1994; Matarrese et al., 1997; Scoccimarro et al., 1998), but it is also a powerful tool to obtain constraints on cosmological parameters.

In this chapter we procede with the analysis of the galaxy bispectrum with the aim of constrain the galaxy bias parameters. The first part of the chapter is devoted to a short introduction to the bispectrum evaluation: as we have pointed out, the correct definition of the bispectrim for all the triangular configurations requires a large number of simulated galaxy catalogs; the number of simulations required cannot be obtained with full N-body simulations, so one alternative is to reduce the number of triangular configurations without losing too much information. In section 8.1.1 we breafly describe some quantities that can make possible to reduce the number of the triangular configurations needed to define the bispectrum. The other alternative, is to take advantage of the large set of realizations (10,000) obtained with PINOCCHIO that allow us to work with the full bispectrum. In this context we describe the model that we use for the galaxy bispectrum and the analysis we have carried out for the estimation of the galaxy bias parameters. In the last part of the chapter we show our results putting in evidence the qualities and the limitation of our model. The last section is devoted to a discussion on the results and on the future improvements that have to be done.

## 8.1 The bispectrum challenge

In the first chapter we have introduced the 2-point and 3-point functions in Fourier space. It is useful to write the two expressions to make some comparison and understand why the bispectrum analysis is challenging:

$$\begin{aligned}\langle \delta_{\mathbf{k}_1}\delta_{\mathbf{k}_2}\rangle &= \delta_D^{(3)}(\mathbf{k}_1+\mathbf{k}_2)P(k)\\ \langle \delta_{\mathbf{k}_1}\delta_{\mathbf{k}_2}\delta_{\mathbf{k}_3}\rangle &= \delta_D^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)B(k_1,k_2,k_3)\ .\end{aligned}$$

We assume the spatial isotropy so that the power spectrum depends on one wavenumber only: $k = |\mathbf{k}_1| = |\mathbf{k}_2|$. The bispectrum is a more complex object because it depends on three wavenumbers with the only condition that the three k's form a closed triangle. There are many configurations of triangles that we have to take into account and this number increases faster than the case of the power spectrum. Our aim is to use the bispectrum as a cosmological probe and this means that we have to compute its covariance matrix. The smaller are the scales we want to investigate, the larger is the number of triangular configurations and the larger is the size of the bispectrum covariance matrix.

It is useful to show an example to stress that the evaluation of the bispectrum and its covariance matrix requires a large number of realizations, because of the large number of triangular configurations: considering a periodic box of side $3.5\,h^{-1}$ Gpc looking at scale up to $k_{\rm max} \sim 0.3\,h\,{\rm Mpc}^{-1}$ there are ~1000 triangle configurations using a bin-width of $6k_f$, with $k_f = 2\pi/L$ the fundamental frequency of the box of side $L$. For this type of configuration in the SDSS-BOSS collaboration bispectrum analysis (Gil-Marín et al., 2017), they use ~2000 simulations. This number has to be considered as a "lower limit", in the sense that is the minimal request in terms of number of simulations to get a well defined bispectrum on those scales. The investigation of smaller scales than $k = 0.3\,h\,{\rm Mpc}^{-1}$ require a larger number of realizations.

### 8.1.1 Alternative bispectrum estimators

It is possible to reduce the number of triangles using different bispectrum estimators, that are defined as the average of particular set of triangular configurations. These quantities allow us to reduce the size of the covariance matrix, but the price to pay is a loss in the information contained in it. In this chapter we briefly describe some of these aggregations, to distinguish them from the full *bispectrum*.

The **integrated bispectrum** was first proposed by Chiang et al. (2014). This quantity considers particular configurations with one wavenumber, $k_i$, much smaller than the other two. These configurations are called "squeezed" configurations. Assuming $k_3 \ll k_1$, $k_2$ the 3-point function can be expressed as the correlation between the single long-wavelength mode $\delta(\mathbf{k}_3)$ and the 2-point function $\langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\rangle$. The cubic survey volume is divided in $N_s$ sub-volumes centered at positions $\mathbf{r}_L$; the integrated bispectrum is defined as:

$$iB(k) \equiv \int \frac{d^2k}{4\pi}\ \langle P(\mathbf{k},\mathbf{r}_L)\bar{\delta}(\mathbf{r}_L)\rangle_{N_s}\ , \tag{8.1}$$

where $P(\mathbf{k},\mathbf{r}_L)$ is the power spectrum computed in each sub-volume and $\bar{\delta}(\mathbf{r}_L)$ is the average overdensity in the sub-volume; the expectation is taken over all sub-volumes. Following Chiang et al. (2014) we can expand the power spectrum in powers of $\bar{\delta}(\mathbf{r}_L)$; at leading order the average in eq. 8.1 is:

$$\langle P(\mathbf{k},\mathbf{r}_L)\bar{\delta}(\mathbf{r}_L)\rangle_{N_s} \approx \left.\frac{\mathrm{d}\ \ln\, P(\mathbf{k})}{\mathrm{d}\bar{\delta}}\right|_{\bar{\delta}} P(\mathbf{k})\sigma_L^2\ , \tag{8.2}$$

with $\sigma_L^2 \equiv \langle\bar{\delta}^2(\mathbf{r}_L)\rangle_{N_s}$. Eq. 8.2 tells us that the integrated bispectrum describes variation of the power spectrum in response to changes in the large-scale overdensity. The power spectrum and its variance can be directly measured so that any new information in the integrated bispectrum is contained in the normalized component:

$$ib(k) \equiv \frac{iB(k)}{P(k)\sigma_L^2} \approx \left.\frac{\mathrm{d}\ \ln P(k)}{\mathrm{d}\bar{\delta}}\right|_{\bar{\delta}=0}\ , \tag{8.3}$$

where d $\ln P(k)/\mathrm{d}\bar{\delta}$ is the linear response function that can approximate $ib(k)$ on large scales.

The **Line Correlation Function (LCF)** is an estimator based on the property of the bispectrum to encode information on both amplitude and phases. The line correlation function was for the first time proposed by Obreschkow et al. (2013) to measure a subset of 3-point *phase correlation* of the density field; in particular this subset accounts for collinear configurations. The definition of the LCF is given, as for the integrated bispectrum, by an average of the phase field smoothed on a scale r:

$$l(r) \equiv \frac{V^3}{(2\pi)^9}\Big(\frac{r^3}{V}\Big)^{3/2} \int \frac{d^2\mathbf{r}}{4\pi}\ \langle\epsilon_r(\mathbf{x})\epsilon_r(\mathbf{x}+\mathbf{r})\epsilon_r(\mathbf{x}-\mathbf{r})\rangle\ , \tag{8.4}$$

where $\frac{V^3}{(2\pi)^9}$ is a volume regularization, $\epsilon_r(\mathbf{x})$ is the real phase field defined as

$$\epsilon_r(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3}\ \epsilon(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}}W(k|r)\ , \tag{8.5}$$

with $W(k|r)$ the Fourier transform of the smoothing window function and $\epsilon(k) \equiv \delta(\mathbf{k})/|\delta(\mathbf{k})|$. From the works by Wolstenhulme et al. (2015) and Eggemeier and Smith (2017) we know that the Fourier transform of the 3-point phase function $\langle\epsilon_r(\mathbf{x})\epsilon_r(\mathbf{x}+\mathbf{r})\epsilon_r(\mathbf{x}-\mathbf{r})\rangle$: using eq. 8.5 into 8.4 we can obtain the three-point function of $\epsilon_r(\mathbf{k})$ that is associated to the bispectrum; at lowest order we can write (Wolstenhulme et al., 2015):

$$\langle\epsilon_r(\mathbf{k}_1)\epsilon_r(\mathbf{k}_2)\epsilon_r(\mathbf{k}_3)\rangle \approx \frac{(2\pi)^3}{V}\Big(\frac{\sqrt{\pi}}{2}\Big)\frac{B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)}{\sqrt{V\ P(\mathbf{k}_1)\ P(\mathbf{k}_2)\ P(\mathbf{k}_3)}}\times\delta_D(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \tag{8.6}$$

therefore the LCF contains some fraction of the information on the bispectrum. We stress that is only a fraction of the information because $l(r)$ is an average over specific collinear configurations, so it represents a compression.

The **modal bispectrum** is equivalent to the usual bispectrum except that we exchange the Fourier basis $e^{i\mathbf{k}\cdot\mathbf{x}}$ for a set of alternative modes that are adapted to the structure of the bispectrum. Following Fergusson and Shellard (2009); Regan et al. (2010), we can define the basis function $Q_n$ so that:

$$B(k_1,k_2,k_3) \approx B_{\rm modal}(k_1,k_2,k_3) \equiv \frac{1}{w(k_1,k_2,k_3)}\sum_{n=0}^{n_{\rm max}-1}\beta_n^Q Q_n(k_1,k_2,k_3)\ , \tag{8.7}$$

where $\beta_n^Q$ are the "modal coefficient" and $w(k_1,k_2,k_3)$ is a weight function. $Q_n$ can be factorized in three functions depending respectively on $k_1$, $k_2$ and $k_3$; this factorization make easier to evaluate the amplitude of the bispectrum in the limit of weak non-Gaussianity (Schmittfull et al., 2013a).

Eq. 8.7 can be seen as an expansion of the bispectrum over a set of configurations picked out by the corresponding $Q_n$. $n_{\rm max}$ is the number of triangles used to represent the bispectrum and to reduce the number of mode it is expected to be $n_{\rm max} \ll N_{\rm triangles}$.

Finally, the **quadratic estimator** developed by Schmittfull et al. (2015) describes the tree-level bispectrum as three separate components, which can be Legendre decomposed in three terms. The first one is the *squared density*:

$$\delta^2(\mathbf{x}) = \int\frac{d^3\mathbf{k}}{(2\pi)^3}\,e^{i\mathbf{k}\cdot\mathbf{x}}\int\frac{d^3\mathbf{q}}{(2\pi)^3}\,\mathsf{P}_0(\mu)\delta(\mathbf{q})\delta(\mathbf{k}-\mathbf{q})\ , \tag{8.8}$$

the second is *shift-term*:

$$-\mathbf{\Psi}(\mathbf{x})\nabla\cdot\delta(\mathbf{x}) = -\int\frac{d^3\mathbf{k}}{(2\pi)^3}\,e^{i\mathbf{k}\cdot\mathbf{x}}\int\frac{d^3\mathbf{q}}{(2\pi)^3}\,F_2^1(q,|\mathbf{k}-\mathbf{q}|)\mathsf{P}_1(\mu)\delta(\mathbf{q})\delta(\mathbf{k}-\mathbf{q})\ , \tag{8.9}$$

with $\mathbf{\Psi}(\mathbf{k}) = -\frac{i\mathbf{k}}{k^2}\delta(\mathbf{k})$ and $F_2^1$ is the symmetric part of the kernel defined in eq. 2.55, and the third is the *tidal-term*

$$s^2(\mathbf{x}) = \int\frac{d^3\mathbf{k}}{(2\pi)^3}\,e^{i\mathbf{k}\cdot\mathbf{x}}\int\frac{d^3\mathbf{q}}{(2\pi)^3}\,\mathsf{P}_2(\mu)\delta(\mathbf{q})\delta(\mathbf{k}-\mathbf{q})\ , \tag{8.10}$$

The Legendre polynomial in eqs. 8.8, 8.9, 8.10 are:

$$\mathsf{P}_0(\mu) = 1 \tag{8.11}$$

$$\mathsf{P}_1(\mu) = \mu \tag{8.12}$$

$$\mathsf{P}_2(\mu) = \frac{3}{2}\Big(\mu^2 - \frac{1}{3}\Big) \tag{8.13}$$

and $\mu$ is the cosine between $\mathbf{q}$ and $\mathbf{k}-\mathbf{q}$. In the limit where the tree-level theory applies, one optimal way to write the bispectrum estimator is using the three terms we have defined above. In particular the procedure consists in computing the cross-spectra between the quadratic fields and the density. Following the notation of Schmittfull et al. (2015) we have:

$$\hat{P}_{\delta^2,\delta}(k) \sim \sum_{\mathbf{k},|\mathbf{k}|=k} [\delta^2](\mathbf{k})\delta(-\mathbf{k}) \tag{8.14}$$

$$\hat{P}_{-\Psi^i\partial_i\delta,\delta}(k) \sim \sum_{\mathbf{k},|\mathbf{k}|=k} [-\Psi^i\partial_i\delta](\mathbf{k})\delta(-\mathbf{k}) \tag{8.15}$$

$$\hat{P}_{s^2,\delta}(k) \sim \sum_{\mathbf{k},|\mathbf{k}|=k} [s^2](\mathbf{k})\delta(-\mathbf{k}) \ , \tag{8.16}$$

where they define a general quadratic field

$$D[\delta](\mathbf{k}) = \int \frac{d^3\mathbf{q}}{(2\pi)^3}\, D(\mathbf{q},\mathbf{k}-\mathbf{q})\delta(\mathbf{q})\delta(\mathbf{k}-\mathbf{q}) \tag{8.17}$$

with a kernel $D(\mathbf{q},\mathbf{k}-\mathbf{q})$ given by one of the Legendre polinomials defined above. Eq. 8.15 carries almost the total bispectrum information on the linear bias $b_1$; eqs. 8.14 and 8.16 can improve the constraints on $b_2$ and on the non-local bias $b_{s^2}$. From the results of Schmittfull et al. (2015) these cross-spectra contain the same constraining power on bias parameters as a full Fourier bispectrum. From the comparison with the theoretical expectation this procedure works up to $k \lesssim 0.9\, h\, \mathrm{Mpc}^{-1}$ at $z = 0.55$. The cross-spectra depend only on a single wavenumber so that the modeling of the covariance can be simplified.

The *modal bispectrum* can well approximate the Fourier bispectrum by $\sim 50$ modes (Schmittfull et al., 2013b; Lazanu et al., 2016), resulting in one of the best estimator.

Our analysis is based on the usual bispectrum, because we want to use as much as possible information carried by the bispectrum. In this case the covariance matrix can be estimated only using fast approximate methods that allow us to produce a large set of synthetic galaxy catalogs.

## 8.2 The bispectrum model

In section 2.3.2 we have stressed that the relation between the matter density field and the galaxy density field is not trivial. To relate the two fields in the following analysis, we consider the linear bias $b_1$, the quadratic bias $b_2$ and the non-local bias $\gamma_2$ (see eq. 2.91). The tree-level galaxy bispectrum model, eq. 2.61, depends on these three parameters:

$$\begin{aligned} B_g(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) &= b_1^3 B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \\ &+ b_1^2 b_2\Big[P_L(k_1)P_L(k_2)+P_L(k_2)P_L(k_3)+P_L(k_1)P_L(k_3)\Big] \\ &+ b_1^2\gamma_2\Big[\Sigma_{2,12}P_L(k_1)P_L(k_2)+\Sigma_{2,23}P_L(k_2)P_L(k_3) \\ &\quad +\Sigma_{2,13}P_L(k_1)P_L(k_3)\Big]\ , \end{aligned} \tag{8.18}$$

where is $B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)$ is the matter bispectrum, eq. 2.61, and we introduce the kernel:

$$\Sigma_2(\mathbf{q}_1,\mathbf{q}_2) \equiv \cos^2\theta_{12} - 1 \tag{8.19}$$

with $\cos\theta_{12} \equiv \hat{q}_1 \cdot \hat{q}_2$. To keep the model as easy as possible we include the lowest non-linear and non-local corrections: we consider the quadratic bias $b_2$ to include the non-linear corrections and the non-local bias $\gamma_2$ to account also for the non-locality.

In all the analysis that we will present in the following, the bispectrum depends only on the three galaxy bias parameters; the cosmology dependence is not included. The set of simulations we use is the same described in section 6.2 with fixed cosmology.

### 8.2.1 Likelihood analysis

We assume that the parameter distribution follows a multi-variate Gaussian likelihood:

$$\mathcal{L}(b_1,b_2,\gamma_2) = \frac{1}{\sqrt{2\pi}|\mathrm{det}C|^{1/2}}\exp\Big[-\frac{1}{2}\sum_{ij}\Big(\bar{B}_i - B_{g,i}\Big)C_{ij}^{-1}\Big(\bar{B}_j - B_{j,g}\Big)\Big]\ , \tag{8.20}$$

where $|\mathrm{det}C|$ is the determinant of the bispectrum covariance matrix, the sum $\sum_{ij}$ runs over all the triangles, $\bar{B}_i = \bar{B}_i(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)$ is the average bispectrum over 300 realizations, $B_{g,i} = B_{g,i}(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)$ is given by eq. 8.18 and $C_{ij}^{-1}$ is the inverse of the bispectrum covariance matrix, with:

$$C_{ij} = \langle \delta B_i \delta B_j\rangle = \langle (B_i - \bar{B}_i)(B_j - \bar{B}_j)\rangle\ , \tag{8.21}$$

where $\langle ...\rangle$ is the average on all the realizations.

Given the likelihood of the parameters it is possible to proceed with a Maximum Likelihood Estimation (MLE): finding the set of values that maximize the likelihood corresponds to find the best model to represent to observed data. The bispectrum model, given by eq. 8.18, depends explicitly on the bias parameters. This allows to compute analytically the values that maximize the likelihood.

From now on we will work with the natural logarithm of the likelihood to simplify the calculation:

$$\ln\,\mathcal{L}(b_1,b_2,\gamma_2)=-\frac{1}{2}\sum_{ij}\Big(\bar{B}_i-B_{g,i}\Big)C_{ij}^{-1}\Big(\bar{B}_j-B_{j,g}\Big)+\ln A\ , \tag{8.22}$$

where A is the normalization factor.

Maximazing the likelihood function with respect the three bias parameters we obtain three equations for $b_1$, $b_2$ and $\gamma_2$. We numerically solve the equation to find the value of the parameters in function of the scale that maximize the likelihood.

### 8.2.2 Fisher matrix analysis

Considering the set of parameters that maximize the likelihood ($b_1^*$,$b_2^*$,$\gamma_2^*$) we make a Taylor expansion of the logarithm of the likelihood around the maximum:

$$\begin{aligned}\ln\,\mathcal{L}(\mathrm{b}_1,\mathrm{b}_2,\gamma_2) &= \ln\,\mathcal{L}(\mathrm{b}_1^*,\mathrm{b}_2^*,\gamma_2^*)+\Big(\frac{\partial\ln\,\mathcal{L}}{\partial\theta_\mathrm{i}}\Big)_{\theta=\theta^*}(\theta_\mathrm{i}-\theta^*)\\ &+ \frac{1}{2}\Big(\frac{\partial^2\ln\,\mathcal{L}}{\partial\theta_i\partial\theta_j}\Big)_{\theta=\theta^*}(\theta_i-\theta^*)(\theta_j-\theta^*)\ ,\end{aligned} \tag{8.23}$$

where we omit the sum on $i,j=1,...,3$ and $\theta_i=(b_1,b_2,\gamma_2)$. By definition the first derivative of the likelihood, evaluated to the maximum, vanishes:

$$\ln\,\mathcal{L}(\mathrm{b}_1,\mathrm{b}_2,\gamma_2)=\ln\,\mathcal{L}(\mathrm{b}_1^*,\mathrm{b}_2^*,\gamma_2^*)+\frac{1}{2}\Big(\frac{\partial^2\ln\,\mathcal{L}}{\partial\theta_\mathrm{i}\partial\theta_\mathrm{j}}\Big)_{\theta=\theta^*}(\theta_\mathrm{i}-\theta^*)(\theta_\mathrm{j}-\theta^*)\ . \tag{8.24}$$

We can define the Hessian matrix in the flowing way:

$$H_{ij}\equiv-\frac{\partial^2\ln\,\mathcal{L}}{\partial\theta_i\partial\theta_j}\ . \tag{8.25}$$

This matrix contains the information we need on the parameters error. The Fisher matrix is given by:

$$F_{ij}\equiv-\langle\frac{\partial^2\ln\,\mathcal{L}}{\partial\theta_i\partial\theta_j}\rangle=\langle H_{ij}\rangle\ . \tag{8.26}$$

where $\langle ...\rangle$ denotes an ensemble average over observational data. The error matrix is given by the inversion of the Hessian matrix:

$$F_{ij}^{-1}\leq\sigma_{ij}^2\ , \tag{8.27}$$

where $F_{ij}$ is given by eq. 8.26.

In eq. 8.27 we used an inequality between the Fisher matrix and the error matrix, named *Cramer-Rao inequality*: the Fisher matrix approach always gives an underestimate of the errors (Verde, 2010).

In general if the data errors are Gaussian distributed then it is correct to use a multi-variate Gaussian distribution. If the data are not Gaussianly distributed it is always possible to re-bin the data so that in each bin there is a superposition of a set of independent measurements (central limit theorem); after the re-binning the final distribution will be bettere approximated by a multi-variate Gaussian. It is worth to stress that the data can be Gaussianly distributed, but the parameters can follow a distribution that is not described by a multi-variate Gaussian: for this to be true it is required that the model depends linearly on the parameters. Looking at eq. 8.18 it seems that we are not in this case, but we can always re-define the parameters in a way to have a linear dependent model; for example one can define:

$$\begin{aligned} \mathcal{P}_1 &= b_1^3 \\ \mathcal{P}_2 &= b_1^2 b_2 \\ \mathcal{P}_3 &= b_1^3 \gamma_2 \ , \end{aligned} \tag{8.28}$$

so that eq. 8.18 becomes:

$$\begin{aligned} B_g(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) &= \mathcal{P}_1 B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \\ &+ \mathcal{P}_2 \Big[ P_L(k_1)P_L(k_2) + P_L(k_2)P_L(k_3) + P_L(k_1)P_L(k_3) \Big] \\ &+ \mathcal{P}_3 \Big[ \Sigma_{2,12} P_L(k_1)P_L(k_2) + \Sigma_{2,23} P_L(k_2)P_L(k_3) \\ &\quad + \Sigma_{2,13} P_L(k_1)P_L(k_3) \Big] \ ; \end{aligned} \tag{8.29}$$

In this case the model for the bispectrum will depends, linearly, on three parameters $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$. In this sense in our analysis we use the likelihood is Gaussian, therefore we assume that the Fisher matrix and the error matrix are equal.

The covariance matrix of the parameter is:

$$[F]^{-1} = [C] = \begin{bmatrix} \sigma_{b_1}^2 & \sigma_{b_1 b_2} & \sigma_{b_1 \gamma_2} \\ \sigma_{b_2 b_1} & \sigma_{b_2}^2 & \sigma_{b_2 \gamma_2} \\ \sigma_{\gamma_2 b_1} & \sigma_{\gamma_2 b_2} & \sigma_{\gamma_2}^2 \end{bmatrix} \tag{8.30}$$

The marginalized errors on the parameters are given by $\sigma_{b_1}$,$\sigma_{b_2}$ and $\sigma_{\gamma_2}$. After we get the parameter error matrix from the Fisher analysis, we can plot the error ellipses corresponding to the error on two parameters after marginalizing on the third. The marginalization process produces three new covariance matrices:

$$\begin{aligned} C_{b_1b_2} &= \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1b_2} \\ \sigma_{b_1b_2} & \sigma^2_{b_2} \end{bmatrix} \\ C_{b_1\gamma_2} &= \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1\gamma_2} \\ \sigma_{b_1\gamma_2} & \sigma^2_{\gamma_2} \end{bmatrix} \\ C_{b_2\gamma_2} &= \begin{bmatrix} \sigma^2_{b_2} & \sigma_{b_2\gamma_2} \\ \sigma_{b_2\gamma_2} & \sigma^2_{\gamma_2}\,. \end{bmatrix} \end{aligned} \tag{8.31}$$

The ellipse parameters depend on these covariance matrices; for a general matrix of the type:

$$C = \begin{bmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y\,. \end{bmatrix} \tag{8.32}$$

the eigenvalues are:

$$\lambda_1 = \frac{\sigma^2_x + \sigma^2_y}{2} + \sqrt{\frac{(\sigma^2_x - \sigma^2_y)^2}{4} + \sigma^2_{xy}} \tag{8.33}$$

$$\lambda_2 = \frac{\sigma^2_x + \sigma^2_y}{2} - \sqrt{\frac{(\sigma^2_x - \sigma^2_y)^2}{4} + \sigma^2_{xy}} \ : \tag{8.34}$$

the semi-axes of the ellipse parallel to the x-axis are given by:

$$a = \begin{cases} \sqrt{\max(\lambda_1, \lambda_2)} & \text{if } \sigma_\mathrm{x} > \sigma_\mathrm{y} \\ \sqrt{\min(\lambda_1, \lambda_2)} & \text{if } \sigma_\mathrm{y} > \sigma_\mathrm{x}\ , \end{cases} \tag{8.35}$$

while the semi-axis parallel to the y-axis is:

$$b = \begin{cases} \sqrt{\min(\lambda_1, \lambda_2)} & \text{if } \sigma_\mathrm{x} > \sigma_\mathrm{y} \\ \sqrt{\max(\lambda_1, \lambda_2)} & \text{if } \sigma_\mathrm{y} > \sigma_\mathrm{x}\ . \end{cases} \tag{8.36}$$

The inclination of the ellipse is:

$$\tan\theta = \frac{2\sigma_{xy}}{\sigma^2_x - \sigma^2_y}\ . \tag{8.37}$$

The semi-axes a and b should be multiplied by a rescaling factor $\alpha$ to account for the different confidence levels (table 8.1).

The Fisher matrix analysis allow us to determine the parameters covariance matrices, so that the error for each of the parameters is associated to its mean value. In the next sections we study how the parameter likelihood changes when we consider only the diagonal of the bispectrum covariance matrix or the full covariance matrix.

Table 8.1: Confidence Ellipses

| $\sigma$ | CL | $\Delta\chi^2$ | $\alpha = \sqrt{\Delta\chi^2}$ |
|---|---|---|---|
| 1 | 68.3 % | 2.3 | 1.52 |
| 2 | 95.4 % | 6.17 | 2.48 |
| 3 | 99.7 % | 11.8 | 3.44 |

### 8.2.3 Bispectrum variance and covariance

In the first part of this chapter we have pointed out the need of having a large number of realizations to build a reliable bispectrum covariance matrix. Using the 300 N-body realizations, that we have described in chapter 6, we are able to estimate the bispectrum variance, while the off-diagonal elements of the covariance matrix are too noisy to extract physical information. On the other hand using the 10,000 realizations made with PINOCCHIO, all the noise is lowered and what remains is the physical correlation between specific triangle configurations. In figure 8.1 we show a density plot of the bispectrum cross-correlation coefficient; the squares are red when the cross-correlation coefficient is 1, while they are blue when it is -1, all the other colors are for values between -1 and 1. On the x- and y-axes there are the different triangular configurations, represented as a triplet of values, as already done in the bispectrum comparison in section 6.2.2. On the left the case with 300 N-body simulations, on the right the case with 10,000 PINOCCHIO realizations. We derive the bias parameters values using both the 300 N-body realizations

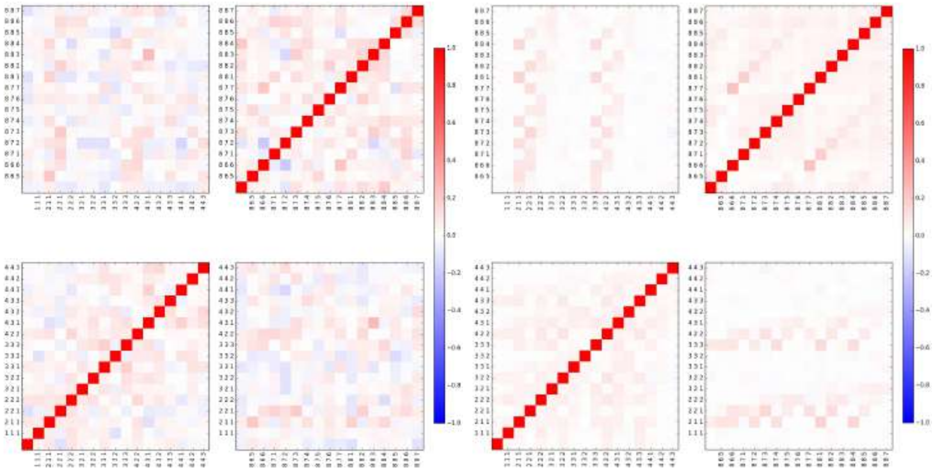


Figure 8.1: Cross-correlation coefficient defined by eq. 8.39. On the left the case with 300 realizations, on the right the case with 10,000 realizations.

and the 10,000 PINOCCHIO.

In the first case, we know that we can trust only of the diagonal of the covariance matrix. In this case we use a likelihood that depends only on the bispectrum variance $\Delta B_i^2$:

$$\ln \mathcal{L}(b_1, b_2, \gamma_2) \propto -\frac{1}{2} \sum_i \frac{\left(\bar{B}_i - B_{g,i}\right)^2}{\Delta B_i^2} \ . \tag{8.38}$$

In the second case we have to define the covariance matrix in the following way: we start from the PINOCCHIO cross-correlation coefficient

$$r_{ij}^{\rm PIN} = \frac{C_{ij}^{\rm PIN}}{\sqrt{C_{ii}^{\rm PIN} C_{jj}^{\rm PIN}}} \ , \tag{8.39}$$

where $C_{ij}^{\rm PIN}$ is the covariance matrix and $C_{ii}^{\rm PIN}$ the diagonal of the matrix, the variance. This quantity gives information on the off-diagonal part of the matrix as shown in figure 8.1. Eq. 8.39 is a general relation and for the 300 Minerva simulations we can write:

$$C_{ij}^{\rm N-body} = r_{ij}^{\rm N-body} \sqrt{C_{ii}^{\rm N-body} C_{jj}^{\rm N-body}} \ . \tag{8.40}$$

We want to built our final covariance matrix using the diagonal coming from the N-body simulations and the off-diagonal elements from the 10,000 PINOCCHIO. To this purpose we start from eq. 8.40 and we replace the cross-correlation from the N-body with that computed using the PINOCCHIO, eq. 8.39. The final expression for the covariance matrix is:

$$C_{ij} = r_{ij}^{\rm PIN} \sqrt{C_{ii}^{\rm N-body} C_{jj}^{\rm N-body}} \ , \tag{8.41}$$

It is worth to use a matricial language to express the covariance matrix with all its elements. In this way the cross-correlation in eq. 8.41 becomes:

$$r_{ij}^{\rm PIN} \equiv \mathbf{r} = \begin{bmatrix} \mathbf{r}_{11} & \mathbf{r}_{12} \\ \mathbf{r}_{21} & \mathbf{r}_{22} \end{bmatrix} \ , \tag{8.42}$$

with $\mathbf{r}_{11}$ and $\mathbf{r}_{22}$ equal to one by definition of cross-correlation coefficient; The variance from the simulations can be thought as a diagonal matrix as:

$$\sqrt{C_{ii}^{\rm N-body}} \equiv \mathbf{D} = \begin{bmatrix} \sqrt{\mathbf{d}_1} & 0 \\ 0 & \sqrt{\mathbf{d}_2} \end{bmatrix} \ . \tag{8.43}$$

Using this language we can write our covariance matrix in eq. 8.41 as:

$$\mathbf{C} = \mathbf{D} \ \mathbf{r} \ \mathbf{D} \ . \tag{8.44}$$

In the expression of the likelihood, eq. 8.20, enters the inverse of the covariance matrix:

$$\mathbf{C}^{-1} = \mathbf{D}^{-1}\,\mathbf{r}^{-1}\,\mathbf{D}^{-1}\ . \tag{8.45}$$

It useful to compute how the signal-to-noise ratio changes when all the elements of the covariance matrix are included. Moreover we can compare the information coming from power spectrum only, bispectrum only and from the covariance that include the cross-correlation between power spectrum and bispectrum:

$$C_{ij}^{\rm tot} = \begin{pmatrix} \langle \delta P_i \delta P_j \rangle & \langle \delta P_j \delta B_{j_1,j_2,j_3} \rangle \\ \\ \langle \delta P_j \delta B_{j_1,j_2,j_3} \rangle & \langle \delta B_{i_1,i_2,i_3} \delta B_{j_1,j_2,j_3} \rangle \end{pmatrix} \tag{8.46}$$

The generic definition for the signal-to-noise, for a generic quantity $X$, is the following:

$$\left(\frac{S}{N}\right)^2_X = \sum_{i,j} X_i (C_{ij}^{-1})^{\rm X} X_j\ ; \tag{8.47}$$

for example for the power spectrum we have:

$$\left(\frac{S}{N}\right)^2_P = \sum_{i=1}^{N_k} \sum_{i=j}^{N_k} P_i (C_{ij}^{-1})^{\rm P} P_j\ , \tag{8.48}$$

with $N_k$ the number of modes. In figure 8.2 we show the signal-to-noise, as function of $k$ for the power spectrum (blue lines), the bispectrum (green lines) and power spectrum plus bispectrum (red lines) for two different halo samples. The measurements are done on the 10,000 PINOCCHIO galaxy catalogs. The solid lines are for the signal-to-noise computed using all the element of the covariance matrix, while the dashed line comes only from the variance (eq. 8.47 with $i = j$). In the lower panel we plot the ratio of all the different components with respect the Gaussian prediction (dotted black line) given by:

$$\left(\frac{S}{N}\right)^2_{\rm Gauss} = \frac{N_k^2}{2}\ . \tag{8.49}$$

As we expect the bispectrum only signal-to-noise computed with all the bispectrum covariance matrix is smaller than that one computed with only the variance for $k > 0.02\,h\,{\rm Mpc}^{-1}$. At scale $k \sim 0.07\,h\,{\rm Mpc}^{-1}$, the smaller scale we consider in our analysis, the total signal-to-noise (solid red) is higher when all the covariance matrix is considered with respect to the power spectrum alone; this means that we can gather more information by considering the power spectrum and the bispectrum together including the cross-correlation between the two.

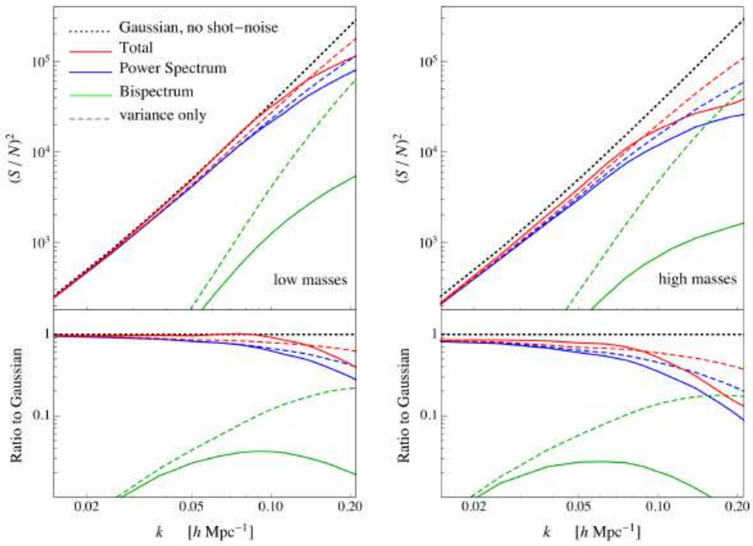


Figure 8.2: Signal-to-Noise as function of k for two different mass bins. The main panel shows the signal-to-noise for the power spectrum (blue lines), for the bispectrum (green lines) and for the power spectrum plus bispectrum (red lines). The dotted lines give information when only the variance is used while the solid ones include the whole covariance matrix. The red lines include in the covariance the cross-correlation between power spectrum and bispectrum. The black dashed line gives the signal-to-noise Gaussian prediction without shot-noise. In the lower panel is plotted the ratio between all of the signal-to-noise terms and the Gaussian prediction.

### 8.2.4 Power spectrum prior

In section 8.2.1 we use the bispectrum to find the bias values to maximize the likelihood. In that case we do not assume any prior. It is useful to compare the results found in that case with those obtained by assuming a prior for the linear bias using the power spectrum. At tree-level the galaxy power spectrum is related to the matter power spectrum through $b_1$:

$$P_g(k) = b_1^2 P(k) \ . \tag{8.50}$$

The likelihood for $b_1$ is trivial:

$$\ln \mathcal{L}^{\mathrm{P}}(\mathrm{b}_1) \propto -\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N_k}} \frac{(\bar{\mathrm{P}}_\mathrm{i} - \mathrm{P}_{\mathrm{g,i}})^2}{\Delta \mathrm{P}_\mathrm{i}^2} \tag{8.51}$$

where $\bar{P}$ is the average power spectrum over 300 N-body simulations and $\Delta P^2$ is the error on the mean. The maximization of eq. 8.51 is done analytically computing the derivative with respect to $b_1$:

$$\frac{\partial \ln \mathcal{L}^{\mathrm{P}}}{\partial b_1} = 0 \qquad \rightarrow \qquad b_1^* \; . \tag{8.52}$$

We use $b_1^*$ as prior in the bispectrum likelihood that depends only on $b_2$ and $\gamma_2$. Eq. 8.22 reads:

$$\ln \mathcal{L}(\mathrm{b}_1^*, \mathrm{b}_2, \gamma_2) \propto -\frac{1}{2} \sum_{\mathrm{ij}} \Big( \bar{\mathrm{B}}_{\mathrm{i}} - \mathrm{B}_{\mathrm{g,i}}(\mathrm{b}_1^*, \mathrm{b}_2, \gamma_2) \Big) \mathrm{C}_{\mathrm{ij}}^{-1} \Big( \bar{\mathrm{B}}_{\mathrm{j}} - \mathrm{B}_{\mathrm{j,g}}(\mathrm{b}_1^*, \mathrm{b}_2, \gamma_2) \Big) \; . \tag{8.53}$$

At this point the values for $b_2$ and $\gamma_2$, that maximize the likelihood, are given by putting to zero the derivatives with respect to $b_2$ and to $\gamma_2$, of eq. 8.53, equal to zero. The final set of bias values is therefore $(b_1^*, b_2^{**}, \gamma_2^{**})$.

## 8.3 Results

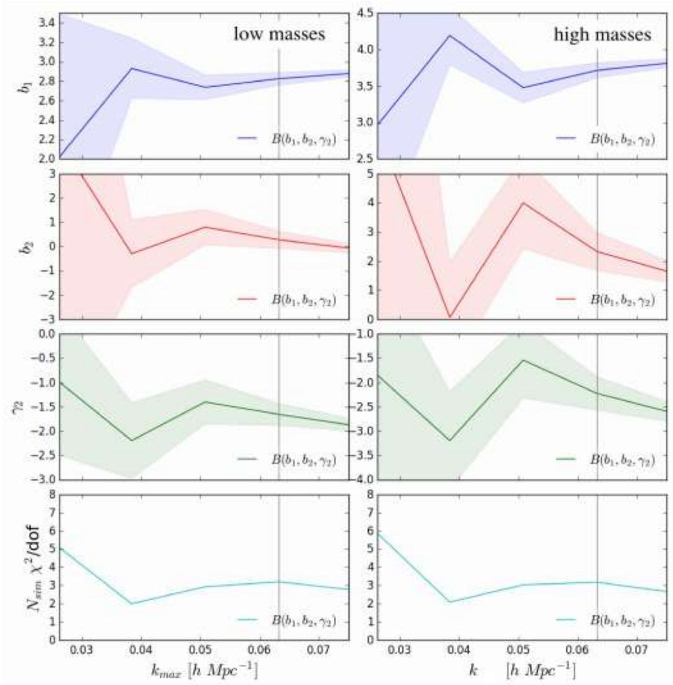


Figure 8.3: Galaxy bias parameters $b_1$ (blue line), $b_2$ (red line) and $\gamma_2$ (green line) for two different mass bins obtained using the bispectrum without prior; in the likelihood of the parameters enters only the bispectrum variance. The shaded area is the $1\sigma$ error obtained from Fisher analysis. The last panel (cyan line) shows the $\chi^2$ per degree of freedom. The vertical gray line corresponds to $k = 0.062\, h\,\mathrm{Mpc}^{-1}$.

In this section we show the results on the constraints of the galaxy bias parameters and their error ellipses. The analyses we have done concern the dependence of these parameters on the maximum value for the wavenumber $k = k_{max}$ that can be predicted and their errors.

Figure 8.3 shows the result assuming no prior and deriving all of the three parameters using only the bispectrum and its variance from the 300 N-body simulations, eq. 8.38. In the two series of plots are shown the values for $b_1$, $b_2$ and $\gamma_2$ for two mass bins, obtained using a MLE. The shaded area is the marginalized $1\sigma$ error computed using the Fisher analysis. In the last panel (cyan line) we plot the reduced $\chi^2$ defined as $-2\ln\mathcal{L}/\text{d.o.f}$, where the d.o.f are the number of triangles in the k-shell. We stop at $k \sim 0.075\, h\,\text{Mpc}^{-1}$ because for smaller scales the $\chi^2$ takes unacceptable values because the model for the bispectrum we are using is not accurate enough to describe those smaller scales; moreover we expect that the values of the parameters at smaller scales should change. In figure 8.4 we show the same results, but compared with the determination of the parameters using a prior for the linear bias. The light colors and the cyan dashed line for the $\chi^2$ are the values assuming $b_1$ from the power spectrum. The shaded area is again the $1\sigma$ error from Fisher analysis. Up to the scale we investigate the two determinations are

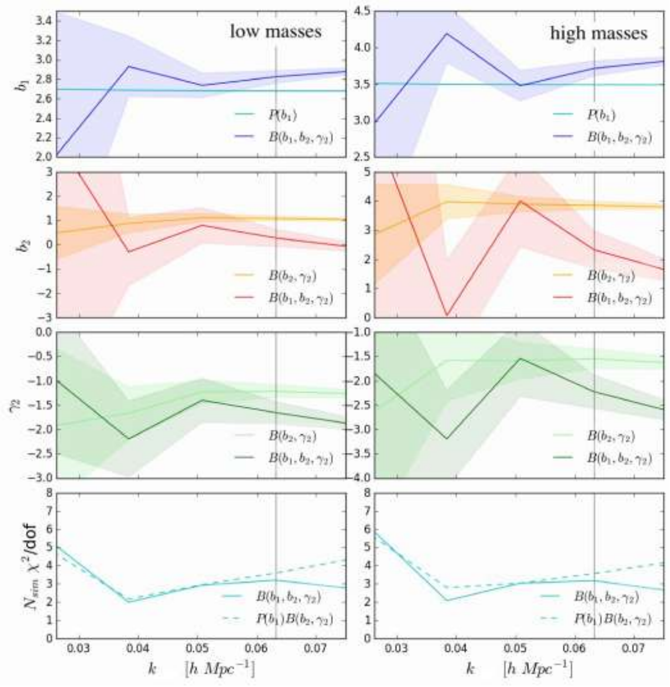


Figure 8.4: The same as figure 8.3, but the values computed using only the bispectrum are compared with those ones computed using $b_1$ as prior. The likelihood of the parameters is diagonal.

compatible within $1-2\sigma$

The agreement starts to fail at $k \sim 0.065\,h\,\mathrm{Mpc}^{-1}$ and at same scale the $\chi^2$ starts to diverge.

Moving to the case where all the bispectrum covariance is used in the likelihood estimation, eq. 8.22, we do not really see any substantial change either in the biases determination or in the $\chi^2$ (figure 8.5). This is true for both the two cases of prior assumption on $b_1$ and no-prior. From the Fisher analysis we obtain the error ellipses we show in figure 8.6. On the left we plot the ellipses obtained using the diagonal likelihood, while on the right the one obtained using all the bispectrum covariance. The red and the blue ellipses are respectively the $1\sigma$ and $2\sigma$ marginalized error. In the plane $\gamma_2$-$b_2$ we plot also the ellipses coming assuming the linear bias from the power spectrum. The estimation is done at $k \sim 0.06\,h\,\mathrm{Mpc}^{-1}$ that is the smallest scale at which we can trust the theoretical model. The difference in the two ellipses in the the plane $\gamma_2$-$b_2$ plane is the same we observe in figures (8.4, 8.5) looking at the determination of $b_2$, $\gamma_2$ with or without using prior information.

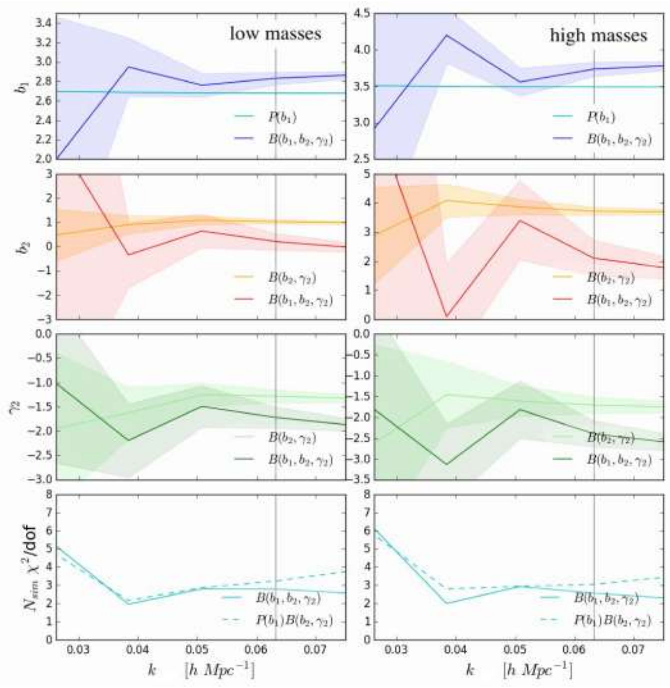


Figure 8.5: The same of figure 8.4, but using the whole covariance in the likelihood.

As we already notice, the error ellipses do not show significant changes when a full likelihood is used in placed of a diagonal likelihood, because, as we will stress in the discussion section, the effect of the off-diagonal elements

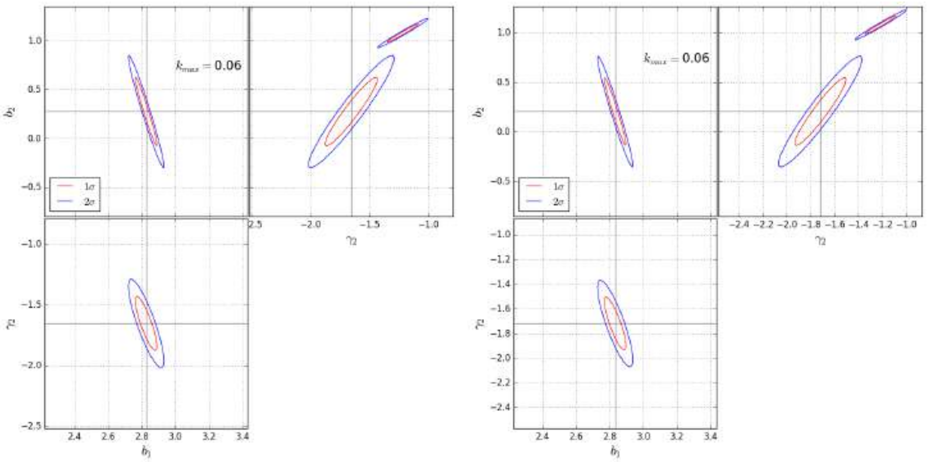


Figure 8.6: Error ellipse for $b_1$, $b_2$ and $\gamma_2$ at $k = 0.06\,h\,\mathrm{Mpc}^{-1}$. The red and the blue ellipse are respectively the $1\sigma$ and $2\sigma$ error. In the plane $\gamma_2$-$b_2$ it is showed the constraint for $b_2$ and $\gamma_2$ assuming $b_1$ fixed by the power spectrum. On the left the errors computed using the diagonal likelihood while on the right the whole bispectrum likelihood.

of the covariance matrix are negligible at the scale we are studying.

## 8.4 Discussion

In this chapter we have described the main challenges in the evaluation of the bispectrum and its covariance matrix; various bispectrum proxies have been described as possible alternative the the full Fourier bispectrum to reduce the number of modes, therefore the size of the covariance matrix.

The analysis we have carried out consists in the evaluation of the Fourier bispectrum. The problem of having a large covariance matrix, is managed using fast approximated methods that allow us to produce a large number of synthetic galaxy catalogs.

The galaxy bispectrum model we have decided to use for this analysis, depends on three bias parameters: the linear bias $b_1$, the quadratic bias $b_2$ and the non-local bias $\gamma_2$. We assume a Gaussian distribution for these parameters and we proceed with a MLE to derive their values. The error on the parameters are computed using a Fisher matrix analysis.

The likelihood analysis consisted of two parts: in the first part, using only the bispectrum covariance matrix measured over 300 N-body simulations, we have derived the parameters with a diagonal likelihood, because the small number of realizations that we had did not allow to use the off-diagonal elements of the covariance matrix that are highly noisy; in the second part we have computed the covariance matrix using 10,000 realizations, made with PINOCCHIO; we have showed that with this number of realizations it is possible to lower the noise that affects the off-diagonal elements of the matrix. With all the elements of the matrix available, we have estimated the parameters from the full likelihood. All these derivations can be done analytically because the bispectrum model depends only on the galaxy bias parameters.

In the section of the results, we have showed a comparison between the bias values derived using the diagonal likelihood and the full likelihood, as function of $k$. Moreover we have also obtained the linear bias from the power spectrum and we have used this value as prior for the determination of $b_2$ and $\gamma_2$.

At this level we cannot draw any definitive conclusion on the values of the galaxy bias parameters and their errors. It seems clear that the galaxy bispectrum model we are using is not accurate enough to describe scales smaller than $k \sim 0.07\,h\,\mathrm{Mpc}^{-1}$ and in particular that the effects of the bispectrum covariance matrix are not important at the scales we have investigated. We could expect this behavior by looking at figure 8.2: the bispectrum only signal-to-noise, at $k \sim 0.06\,h\,\mathrm{Mpc}^{-1}$, when using only the variance is quite similar to the signal-to-noise computed using the bispectrum covariance. Probably an analysis at smaller scales would have showed the differences in considering the off-diagonal elements of the matrix. What we can observe is that on these scales, the additional non-linear corrections are quite negligible.

One way to push the analysis on more non-linear scales and to increase the constraining power is to include, in the bispectrum model, the 1-loop corrections; moreover from the signal-to-noise analysis we know that considering the whole covariance matrix, including the cross-correlation between the power spectrum and the bispectrum eq. 8.46, can increase the information we get, even at the scales we have already investigated. The challenge we have to tackle in this case concerns the inversion of such covariance matrix that appears more complicated than the bispectrum or power spectrum only covariance matrix.

The last, but not least, point is to extend the analysis to derive constraints on cosmological parameters. These can be done measuring the power spectrum and the bispectrum for a set of cosmological models and then running a Monte Carlo Markov Chain (Robert, 2004; Gamerman and Lopes, 2006) on the grid of cosmological parameters.

# Chapter 9

# Conclusions

This Ph.D Thesis was devoted to model the covariance matrices of clustering measures. The estimation of the covariance matrices is one of the main requirements to improve the knowledge on the history and evolution of the Universe through the determination of cosmological parameters.

The structures we observe today are the result of the growth of initial fluctuations in the cosmic density field due to gravitation instability. The properties of the cosmological model define the characteristics of the spectrum of perturbations. Constraints on cosmological parameters are obtained from different observations; in chapter 2 we focused in particular on the baryonic acoustic oscillations, that provides constraints on the dark energy density parameters and the dark energy equation of state, and on the redshift-space distortions, that are used to extract statistical information on the large-scale peculiar velocity field traced by galaxies and, from this, on the rate of growth of density perturbations.

Future surveys will observe millions of galaxies, so that the error budget will be dominated by systematics. Chapter 5, is focused on the effects of the visibility mask on the clustering statistics. The visibility mask includes all the foreground effects due to galactic extinction and stellar contamination as well as instrumental or survey features.

The main results of this part of the Thesis can be summarized as follows:

- the subtraction of the mask is not perfect so we have to consider the presence of a residual foreground that can mimic the large-scale structures, affecting the accuracy in the determination of cosmological parameters;
- the foreground residual modeling is of fundamental importance to accurately sample large scales beyond the BAO scale;
- we measured a non-negligible coupling between different scales that

affects the structure of the power spectrum covariance matrix;

- the off-diagonal elements of the power spectrum covariance matrix are non-negligible. It is not correct to separate the contributions due to the cosmic variance alone and the cosmology-independent covariance because of the coupling term between the mask and the cosmological signal.

In chapter 6 we focused on the results we have obtained from the "comparison project" of approximate methods within the Euclid collaboration galaxy clustering working group. The main purposes of this project are to prove that the covariance matrices obtained using fast methods are unbiased and that they can be used, instead of N-body simulations, to quantify the errors on cosmological parameters. We have identified a number of interesting results:

- the average quantities obtained from the approximate methods do not reproduce the same quantities computed on simulations on small scales because of the nature of the approximate methods;

- the diagonal of the covariance matrices is well reproduced also at smaller scales with the required precision of 1%;

- the structure of the covariance matrix is well reproduced, even if, at this level, we have compared only the noise in the off-diagonal elements of the matrices;

- from forecast analyses we have checked that the errors obtained from a covariance matrix estimated from N-body simulations are indistinguishable from those obtained using approximate methods.

The large set of simulated galaxy catalogs obtained from fast approximate methods allows us to evaluate covariance matrices of clustering to be compared with analytic predictions.

In chapter 7 we focused on the inclusion of the bias and shot-noise terms in the prediction of the power spectrum covariance matrix. From the comparison of our analytic prediction with the power spectrum covariance evaluated using 10,000 simulated galaxy catalogs generated with PINOCCHIO, we have drawn the following conclusions:

- looking at the variance we have found that it is well reproduced by the Gaussian term, with small differences on large scales due to sample variance;

- for the variance the main trispectrum contributions are the two shot-noise terms and the first term of the galaxy trispectrum: these are larger than all the other non-Gaussian corrections, but they appear to be still negligible at $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$;

- the analytic prediction for the off-diagonal elements of the matrix overestimates the measurements, showing an enhancement of the non-Gaussian contributions of the power spectrum covariance matrix.

We repeatedly stressed that even if the primordial fluctuations are Gaussian, non-Gaussianity arise due to non-linear higher-order correlators: three and four point functions.

In chapter 8 we presented an analysis of the bispectrum and its covariance matrix aimed at constraining the linear, the quadratic and the non-local galaxy bias. From the analysis we have carried out, we obtained the following main results:

- at the scale we are looking ($k \sim 0.07\,h\,\mathrm{Mpc}^{-1}$) the bias parameters estimation is not affected by a high level of non-linearity: the off-diagonal terms of the bispectrum covariance matrix do not affect the parameters and their errors;

- the model for the bispectrum we have used is not accurate enough to describe scales smaller than $k \sim 0.07\,h\,\mathrm{Mpc}^{-1}$;

In all the analyses we have presented in this Thesis we have emphasized that a precise estimation of the clustering covariance matrices is mandatory to maximize the cosmological information to be extracted from large galaxy surveys. In particular, our attention was focused on the Euclid space mission. In this context it is of primary importance to have an accurate and realistic model for the systematic errors; we have started with a toy-model for a generic visibility mask, but we want to extend our future analyses toward a more realistic situation: the measurements of the foreground from cross-correlation of different redshift bins. We expect that the angular cross-correlation of different redshift bins is nearly vanishing, so that the observed power spectrum will directly give information on the mask power spectrum; we have to take into account that there are also other contributions to the cross-correlation, apart from the foreground contamination, given by gravitational lensing and by catastrophic redshift errors. This kind of analysis will require to generate halo catalogs on the light cones with PINOCCHIO, to perform an abundance matching of halos with a luminosity function of H$\alpha$ emitters and finally to apply a realistic galactic extinction map, e.g. Planck's maps.

The study of systematic errors as well as all the studies that involve the evaluation of the covariance matrices of clustering require to have access to a large number of realizations of the Universe: in this respect, approximate methods offer a viable approach. Apart from testing other types of techniques, we plan to make forecasts for the determination of the cosmological parameters also using the bispectrum covariance matrix. Besides resorting to Fisher matrix analyses we aim to include a large number of cosmological parameters; in this case we will use numerical methods, like Monte Carlo Markov Chain, for their determination.

From the analytic side, for what concerns the prediction for the galaxy power spectrum covariance matrix, a more accurate modeling of the shot-noise contributions to the trispectrum should solve the overestimation of the off-diagonal elements of the covariance matrix; moreover we plan to test the predictions of the galaxy trispectrum against a direct measurements of the trispectrum on simulations. After the off-diagonal terms are recovered, we will introduce a selection function and then we will move to redshift-space. In the first step we will use a spherical selection function that allows us to give an analytic prediction for the covariance thanks to the simplified geometry; then we will consider more realistic selection functions: in these more realistic cases the analytic description need to be supported by numerical computations. In any case we expect that the power spectrum covariance matrix will present new terms due to the coupling of modes outside and inside the selection function. These terms could be larger than the trispectrum corrections we have already included in our model.

Looking at higher-order statistics, the bispectrum, we aim to increase the precision in the bias determination including the information coming from the cross-correlation of power spectrum and bispectrum. Clearly to extend the analysis to smaller scales we need a bispectrum model where the 1-loop corrections are implemented at the bias level; this bispectrum model has to be used along with a full bispectrum covariance matrix, evaluated on 10,000 realizations, that includes the cross-correlations with the power spectrum. Using such covariance matrix we will face the problem of a more complex inversion of the matrix, as required in the likelihood function: an accurate eigenvalues analysis will be required in order to invert the matrix in the smartest way possible.

As already highlighted the bispectrum is also useful to constrain cosmological parameters: we will extend the analysis we have carried out, measuring the bispectrum for a set of cosmological models and then running a Monte Carlo Markov Chain to constrain such cosmological parameters.

From all the analyses and results we have obtained with this Thesis, it appears to be clear that an accurate characterization of the theoretical and observational systematics will be fundamental for the estimation of the clustering covariance matrix in preparation for next future galaxy surveys that will allow us to understand the nature of the constituents of the Universe and the physics that govern its evolution.

# Bibliography

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., Ho, S., Kinemuchi, K., Kirkby, D., Kitaura, F., Malanushenko, E., Malanushenko, V., Maraston, C., McBride, C. K., Nichol, R. C., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Pellejero-Ibanez, M., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B. A., Rodríguez-Torres, S. A., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martín, J. A., Saito, S., Salazar-Albornoz, S., Samushia, L., Sánchez, A. G., Satpathy, S., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Slosar, A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Vazquez, J. A., Verde, L., Wake, D. A., Wang, Y., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yèche, C., Zehavi, I., Zhai, Z., and Zhao, G.-B. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. Mon. Not. R. Astron. Soc., 470:2617–2652.

Allen, S. W., Evrard, A. E., and Mantz, A. B. (2011). Cosmological Parameters from Observations of Galaxy Clusters. Annu. Rev. Astron. Astrophys, 49:409–470.

Anderson, L., Aubourg, É., Bailey, S., Beutler, F., Bhardwaj, V., Blanton, M., Bolton, A. S., Brinkmann, J., Brownstein, J. R., Burden, A., Chuang, C.-H., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gunn, J. E., Guo, H., Ho, S., Honscheid, K., Howlett, C., Kirkby, D., Lupton, R. H., Manera, M., Maraston, C., McBride, C. K., Mena, O., Montesano, F., Nichol, R. C., Nuza, S. E., Olmstead, M. D., Padmanabhan, N., Palanque-Delabrouille, N., Parejko, J., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B., Roe, N. A., Ross, A. J., Ross, N. P., Sabiu, C. G., Saito, S., Samushia, L., Sánchez, A. G., Schlegel, D. J., Schneider, D. P., Scoccola, C. G., Seo, H.-J., Skibba, R. A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Verde, L., Wake, D. A., Weaver, B. A., Weinberg, D. H., White, M., Xu, X., Yèche, C., Zehavi, I., and Zhao, G.-B. (2014). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples. Mon. Not. R. Astron. Soc., 441:24–62.

Anderson, L., Aubourg, E., Bailey, S., Bizyaev, D., Blanton, M., Bolton, A. S., Brinkmann, J., Brownstein, J. R., Burden, A., Cuesta, A. J., da Costa, L. A. N., Dawson, K. S., de Putter, R., Eisenstein, D. J., Gunn, J. E., Guo, H., Hamilton, J.-C., Harding, P., Ho, S., Honscheid, K., Kazin, E., Kirkby, D., Kneib, J.-P., Labatie, A., Loomis, C., Lupton, R. H., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., McBride, C. K., Mehta, K. T., Mena, O., Montesano, F., Muna, D., Nichol, R. C., Nuza, S. E., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J., Pâris, I., Percival, W. J., Petitjean, P., Prada, F., Reid, B., Roe, N. A., Ross, A. J., Ross, N. P., Samushia, L., Sánchez, A. G., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Skibba, R. A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Verde, L., Wagner, C., Wake, D. A., Weaver, B. A., Weinberg, D. H., White, M., Xu, X., Yèche, C., Zehavi, I., and Zhao, G.-B. (2012). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. Mon. Not. R. Astron. Soc., 427:3435–3467.

Avila, S., Murray, S. G., Knebe, A., Power, C., Robotham, A. S. G., and Garcia-Bellido, J. (2015). HALOGEN: a tool for fast generation of mock halo catalogues. Mon. Not. R. Astron. Soc., 450:1856–1867.

Baldauf, T., Seljak, U., Desjacques, V., and McDonald, P. (2012). Evidence for quadratic tidal tensor bias from the halo bispectrum. Phys. Rev. D, 86(8):083540.

Barrow, J. D., Bhavsar, S. P., and Sonoda, D. H. (1984). A bootstrap resampling analysis of galaxy clustering. Mon. Not. R. Astron. Soc., 210:19P–23P.

Baumgart, D. J. and Fry, J. N. (1991). Fourier spectra of three-dimensional data. Astrophys. J., 375:25–34.

Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolta, M. R., Odegard, N., Peiris, H. V., Verde, L., and Weiland, J. L. (2003). First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results. Astrophys. J. Suppl., 148:1–27.

Bernardeau, F., Colombi, S., Gaztañaga, E., and Scoccimarro, R. (2002). Large-scale structure of the Universe and cosmological perturbation theory. Phys. Rep., 367:1–248.

Bernstein, G. M. (1994). The variance of correlation function estimates. Astrophys. J., 424:569–577.

Berry, M., Ivezić, Ž., Sesar, B., Jurić, M., Schlafly, E. F., Bellovary, J., Finkbeiner, D., Vrbanec, D., Beers, T. C., Brooks, K. J., Schneider, D. P., Gibson, R. R., Kimball, A., Jones, L., Yoachim, P., Krughoff, S., Connolly, A. J., Loebman, S., Bond, N. A., Schlegel, D., Dalcanton, J., Yanny, B., Majewski, S. R., Knapp, G. R., Gunn, J. E., Allyn Smith, J., Fukugita, M., Kent, S., Barentine, J., Krzesinski, J., and Long, D. (2012). The Milky Way Tomography with Sloan Digital Sky Survey. IV. Dissecting Dust. Astrophys. J., 757:166.

Bertolini, D., Schutz, K., Solon, M. P., Walsh, J. R., and Zurek, K. M. (2016). Non-Gaussian covariance of the matter power spectrum in the effective field theory of large scale structure. Phys. Rev. D, 93(12):123505.

Bertschinger, E. (1998). Simulations of Structure Formation in the Universe. Annu. Rev. Astron. Astrophys, 36:599–654.

Betoule, M., Kessler, R., Guy, J., Mosher, J., Hardin, D., Biswas, R., Astier, P., El-Hage, P., Konig, M., Kuhlmann, S., Marriner, J., Pain, R., Regnault, N., Balland, C., Bassett, B. A., Brown, P. J., Campbell, H., Carlberg, R. G., Cellier-Holzem, F., Cinabro, D., Conley, A., D'Andrea, C. B., DePoy, D. L., Doi, M., Ellis, R. S., Fabbro, S., Filippenko, A. V., Foley, R. J., Frieman, J. A., Fouchez, D., Galbany, L., Goobar, A., Gupta, R. R., Hill, G. J., Hlozek, R., Hogan, C. J., Hook, I. M., Howell, D. A., Jha, S. W., Le Guillou, L., Leloudas, G., Lidman, C., Marshall, J. L., Möller, A., Mourão, A. M., Neveu, J., Nichol, R., Olmstead, M. D., Palanque-Delabrouille, N., Perlmutter, S., Prieto, J. L., Pritchet, C. J., Richmond, M., Riess, A. G., Ruhlmann-Kleider, V., Sako, M., Schahmaneche, K., Schneider, D. P., Smith, M., Sollerman, J., Sullivan, M., Walton, N. A., and Wheeler, C. J. (2014). Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. Astron. Astrophys., 568:A22.

Beutler, F., Blake, C., Colless, M., Jones, D. H., Staveley-Smith, L., Campbell, L., Parker, Q., Saunders, W., and Watson, F. (2011). The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant. Mon. Not. R. Astron. Soc., 416:3017–3032.

Beutler, F., Blake, C., Colless, M., Jones, D. H., Staveley-Smith, L., Poole, G. B., Campbell, L., Parker, Q., Saunders, W., and Watson, F. (2012). The 6dF Galaxy Survey: z=0 measurements of the growth rate and sigma8. Mon. Not. R. Astron. Soc., 423:3430–3444.

Blake, C., Baldry, I. K., Bland-Hawthorn, J., Christodoulou, L., Colless, M., Conselice, C., Driver, S. P., Hopkins, A. M., Liske, J., Loveday, J., Norberg, P., Peacock, J. A., Poole, G. B., and Robotham, A. S. G. (2013). Galaxy And Mass Assembly (GAMA): improved cosmic growth measurements using multiple tracers of large-scale structure. Mon. Not. R. Astron. Soc., 436:3089–3105.

Blake, C., Brough, S., Colless, M., Contreras, C., Couch, W., Croom, S., Croton, D., Davis, T. M., Drinkwater, M. J., Forster, K., Gilbank, D., Gladders, M., Glazebrook, K., Jelliffe, B., Jurek, R. J., Li, I.-h., Madore, B., Martin, D. C., Pimbblet, K., Poole, G. B., Pracy, M., Sharp, R., Wisnioski, E., Woods, D., Wyder, T. K., and Yee, H. K. C. (2012). The WiggleZ Dark Energy Survey: joint measurements of the expansion and growth history at z $<$ 1. Mon. Not. R. Astron. Soc., 425:405–414.

Blake, C., Kazin, E. A., Beutler, F., Davis, T. M., Parkinson, D., Brough, S., Colless, M., Contreras, C., Couch, W., Croom, S., Croton, D., Drinkwater, M. J., Forster, K., Gilbank, D., Gladders, M., Glazebrook, K., Jelliffe, B., Jurek, R. J., Li, I.-H., Madore, B., Martin, D. C., Pimbblet, K., Poole, G. B., Pracy, M., Sharp, R., Wisnioski, E., Woods, D., Wyder, T. K., and Yee, H. K. C. (2011). The WiggleZ Dark Energy Survey: mapping the distance-redshift relation with baryon acoustic oscillations. Mon. Not. R. Astron. Soc., 418:1707–1724.

Blot, L., Corasaniti, P. S., Alimi, J.-M., Reverdy, V., and Rasera, Y. (2015). Matter power spectrum covariance matrix from the DEUS-PUR ΛCDM simulations: mass resolution and non-Gaussian errors. Mon. Not. R. Astron. Soc., 446:1756–1764.

Blot, L., Corasaniti, P. S., Amendola, L., and Kitching, T. D. (2016). Non-linear matter power spectrum covariance matrix errors and cosmological parameter uncertainties. Mon. Not. R. Astron. Soc., 458:4462–4470.

Borgani, S. and Guzzo, L. (2001). X-ray clusters of galaxies as tracers of structure in the Universe. Nature, 409:39–45.

Burden, A., Padmanabhan, N., Cahn, R. N., White, M. J., and Samushia, L. (2016). Mitigating the Impact of the DESI Fiber Assignment on Galaxy Clustering. *ArXiv e-prints*.

Catelan, P. (1995). Lagrangian dynamics in non-flat universes and non-linear gravitational evolution. Mon. Not. R. Astron. Soc., 276:115–124.

Chan, K. C. and Scoccimarro, R. (2012). Halo sampling, local bias, and loop corrections. Phys. Rev. D, 86(10):103519.

Chan, K. C., Scoccimarro, R., and Sheth, R. K. (2012). Gravity and large-scale nonlocal bias. Phys. Rev. D, 85(8):083509.

Chiang, C.-T., Wagner, C., Schmidt, F., and Komatsu, E. (2014). Position-dependent power spectrum of the large-scale structure: a novel method to measure the squeezed-limit bispectrum. J. Cosmol. Astropart. Phys., 5:048.

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., and Yepes, G. (2015). EZmocks: extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics. Mon. Not. R. Astron. Soc., 446:2621–2628.

Colavincenzo, M., Monaco, P., Sefusatti, E., and Borgani, S. (2017). Uncertainty in the visibility mask of a survey and its effects on the clustering of biased tracers. J. Cosmol. Astropart. Phys., 3:052.

Coles, P. and Lucchin, F. (1995). *Cosmology: the origin and evolution of cosmic structure*. John Wiley.

Cooray, A. and Hu, W. (2001). Power Spectrum Covariance of Weak Gravitational Lensing. Astrophys. J., 554:56–66.

Cooray, A. and Sheth, R. (2002). Halo models of large scale structure. Phys. Rep., 372:1–129.

Corbelli, E. and Salucci, P. (2000). The extended rotation curve and the dark matter halo of M33. Mon. Not. R. Astron. Soc., 311:441–447.

Croton, D. J., Springel, V., White, S. D. M., De Lucia, G., Frenk, C. S., Gao, L., Jenkins, A., Kauffmann, G., Navarro, J. F., and Yoshida, N. (2006). The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. Mon. Not. R. Astron. Soc., 365:11–28.

Cuesta, A. J., Vargas-Magaña, M., Beutler, F., Bolton, A. S., Brownstein, J. R., Eisenstein, D. J., Gil-Marín, H., Ho, S., McBride, C. K., Maraston, C., Padmanabhan, N., Percival, W. J., Reid, B. A., Ross, A. J., Ross, N. P., Sánchez, A. G., Schlegel, D. J., Schneider, D. P., Thomas, D., Tinker, J., Tojeiro, R., Verde, L., and White, M. (2016). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the correlation function of LOWZ and CMASS galaxies in Data Release 12. Mon. Not. R. Astron. Soc., 457:1770–1785.

de la Torre, S., Guzzo, L., Peacock, J. A., Branchini, E., Iovino, A., Granett, B. R., Abbas, U., Adami, C., Arnouts, S., Bel, J., Bolzonella, M., Bottini, D., Cappi, A., Coupon, J., Cucciati, O., Davidzon, I., De Lucia, G., Fritz, A., Franzetti, P., Fumana, M., Garilli, B., Ilbert, O., Krywult, J., Le Brun, V., Le Fèvre, O., Maccagni, D., Małek, K., Marulli, F., McCracken, H. J., Moscardini, L., Paioro, L., Percival, W. J., Polletta, M., Pollo, A., Schlagenhaufer, H., Scodeggio, M., Tasca, L. A. M., Tojeiro, R., Vergani, D., Zanichelli, A., Burden, A., Di Porto, C., Marchetti, A., Marinoni, C., Mellier, Y., Monaco, P., Nichol, R. C., Phleps, S., Wolk, M., and Zamorani, G. (2013). The VIMOS Public Extragalactic Redshift Survey (VIPERS) . Galaxy clustering and redshift-space distortions at z = 0.8 in the first data release. Astron. Astrophys., 557:A54.

de Putter, R., Wagner, C., Mena, O., Verde, L., and Percival, W. J. (2012). Thinking outside the box: effects of modes larger than the survey on matter power spectrum covariance. J. Cosmol. Astropart. Phys., 4:019.

Dodelson, S. (2003). *Modern cosmology.* Academic Press, San Diego, CA.

Dodelson, S. and Schneider, M. D. (2013). The effect of covariance estimator error on cosmological parameter constraints. Phys. Rev. D, 88(6):063537.

Drinkwater, M. J., Jurek, R. J., Blake, C., Woods, D., Pimbblet, K. A., Glazebrook, K., Sharp, R., Pracy, M. B., Brough, S., Colless, M., Couch, W. J., Croom, S. M., Davis, T. M., Forbes, D., Forster, K., Gilbank, D. G., Gladders, M., Jelliffe, B., Jones, N., Li, I.-H., Madore, B., Martin, D. C., Poole, G. B., Small, T., Wisnioski, E., Wyder, T., and Yee, H. K. C. (2010). The WiggleZ Dark Energy Survey: survey design and first data release. Mon. Not. R. Astron. Soc., 401:1429–1452.

Eggemeier, A. and Smith, R. E. (2017). Cosmology with phase statistics: parameter forecasts and detectability of BAO. Mon. Not. R. Astron. Soc., 466:2496–2516.

Einasto, J., Kaasik, A., and Saar, E. (1974). Dynamic evidence on massive coronas of galaxies. Nature, 250:309–310.

Eisenstein, D. J. and Hu, W. (1998). Baryonic Features in the Matter Transfer Function. Astrophys. J., 496:605–614.

Eisenstein, D. J., Seo, H.-J., and White, M. (2007). On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter. Astrophys. J., 664:660–674.

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castander, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., Frieman, J. A., Glazebrook, K., Gunn, J. E., Hendry, J. S., Hennessy, G., Ivezić, Z., Kent, S., Knapp, G. R., Lin, H., Loh, Y.-S., Lupton, R. H., Margon, B., McKay, T. A., Meiksin, A., Munn, J. A., Pope, A., Richmond, M. W., Schlegel, D., Schneider, D. P., Shimasaku, K., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Tucker, D. L., Yanny, B., and York, D. G. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. Astrophys. J., 633:560–574.

Feldman, H. A., Kaiser, N., and Peacock, J. A. (1994). Power-spectrum analysis of three-dimensional redshift surveys. Astrophys. J., 426:23–37.

Feng, Y., Chu, M.-Y., Seljak, U., and McDonald, P. (2016). FASTPM: a new scheme for fast simulations of dark matter and haloes. Mon. Not. R. Astron. Soc., 463:2273–2286.

Fergusson, J. R. and Shellard, E. P. S. (2009). Shape of primordial non-Gaussianity and the CMB bispectrum. Phys. Rev. D, 80(4):043510.

Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Kennicutt, Jr., R. C., Ford, H. C., Graham, J. A., Huchra, J. P., Hughes, S. M. G., Illingworth, G. D., Macri, L. M., and Stetson, P. B. (2001). Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant. Astrophys. J., 553:47–72.

Fry, J. N. (1994). Gravity, bias, and the galaxy three-point correlation function. *Physical Review Letters*, 73:215–219.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC Press.

Gil-Marín, H., Noreña, J., Verde, L., Percival, W. J., Wagner, C., Manera, M., and Schneider, D. P. (2015). The power spectrum and bispectrum of SDSS DR11 BOSS galaxies - I. Bias and gravity. Mon. Not. R. Astron. Soc., 451:539–580.

Gil-Marín, H., Percival, W. J., Verde, L., Brownstein, J. R., Chuang, C.-H., Kitaura, F.-S., Rodríguez-Torres, S. A., and Olmstead, M. D. (2017). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies. Mon. Not. R. Astron. Soc., 465:1757–1788.

Goroff, M. H., Grinstein, B., Rey, S.-J., and Wise, M. B. (1986). Coupling of modes of cosmological mass density fluctuations. Astrophys. J., 311:6–14.

Green, J., Schechter, P., Baltay, C., Bean, R., Bennett, D., Brown, R., Conselice, C., Donahue, M., Fan, X., Gaudi, B. S., Hirata, C., Kalirai, J., Lauer, T., Nichol, B., Padmanabhan, N., Perlmutter, S., Rauscher, B., Rhodes, J., Roellig, T., Stern, D., Sumi, T., Tanner, A., Wang, Y., Weinberg, D., Wright, E., Gehrels, N., Sambruna, R., Traub, W., Anderson, J., Cook, K., Garnavich, P., Hillenbrand, L., Ivezic, Z., Kerins, E., Lunine, J., McDonald, P., Penny, M., Phillips, M., Rieke, G., Riess, A., van der Marel, R., Barry, R. K., Cheng, E., Content, D., Cutri, R., Goullioud, R., Grady, K., Helou, G., Jackson, C., Kruk, J., Melton, M., Peddie, C., Rioux, N., and Seiffert, M. (2012). Wide-Field InfraRed Survey Telescope (WFIRST) Final Report. *ArXiv e-prints*.

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., and Dalla Vecchia, C. (2016). Gaussian covariance matrices for anisotropic galaxy clustering measurements. Mon. Not. R. Astron. Soc., 457:1577–1592.

Guzzo, L., Pierleoni, M., Meneux, B., Branchini, E., Le Fèvre, O., Marinoni, C., Garilli, B., Blaizot, J., De Lucia, G., Pollo, A., McCracken, H. J., Bottini, D., Le Brun, V., Maccagni, D., Picat, J. P., Scaramella, R., Scodeggio, M., Tresse, L., Vettolani, G., Zanichelli, A., Adami, C., Arnouts, S., Bardelli, S., Bolzonella, M., Bongiorno, A., Cappi, A., Charlot, S., Ciliegi, P., Contini, T., Cucciati, O., de la Torre, S., Dolag, K., Foucaud, S., Franzetti, P., Gavignaud, I., Ilbert, O., Iovino, A., Lamareille, F., Marano, B., Mazure, A., Memeo, P., Merighi, R., Moscardini, L., Paltani, S., Pellò, R., Perez-Montero, E., Pozzetti, L., Radovich, M., Vergani, D., Zamorani, G., and Zucca, E. (2008). A test of the nature of cosmic acceleration using galaxy redshift distortions. Nature, 451:541–544.

Guzzo, L., Scodeggio, M., Garilli, B., Granett, B. R., Fritz, A., Abbas, U., Adami, C., Arnouts, S., Bel, J., Bolzonella, M., Bottini, D., Branchini, E., Cappi, A., Coupon, J., Cucciati, O., Davidzon, I., De Lucia, G., de la Torre, S., Franzetti, P., Fumana, M., Hudelot, P., Ilbert, O., Iovino, A., Krywult, J., Le Brun, V., Le Fèvre, O., Maccagni, D., Małek, K., Marulli, F., McCracken, H. J., Paioro, L., Peacock, J. A., Polletta, M., Pollo, A., Schlagenhaufer, H., Tasca, L. A. M., Tojeiro, R., Vergani, D., Zamorani, G., Zanichelli, A., Burden, A., Di Porto, C., Marchetti, A., Marinoni, C., Mellier, Y., Moscardini, L., Nichol, R. C., Percival, W. J., Phleps, S., and Wolk, M. (2014). The VIMOS Public Extragalactic Redshift Survey (VIPERS). An unprecedented view of galaxies and large-scale structure at $0.5 < z < 1.2$. Astron. Astrophys., 566:A108.

Hamaus, N., Seljak, U., Desjacques, V., Smith, R. E., and Baldauf, T. (2010). Minimizing the stochasticity of halos in large-scale structure surveys. Phys. Rev. D, 82(4):043515.

Hamilton, A. J. S. (1993). Toward Better Ways to Measure the Galaxy Correlation Function. Astrophys. J., 417:19.

Hamilton, A. J. S. (1997). Towards optimal measurement of power spectra - I. Minimum variance pair weighting and the Fisher matrix. Mon. Not. R. Astron. Soc., 289:285–294.

Hamilton, A. J. S. (1998). Linear Redshift Distortions: a Review. In Hamilton, D., editor, *The Evolving Universe*, volume 231 of *Astrophysics and Space Science Library*, page 185.

Hamilton, A. J. S., Rimes, C. D., and Scoccimarro, R. (2006). On measuring the covariance matrix of the non-linear power spectrum from simulations. Mon. Not. R. Astron. Soc., 371:1188–1204.

Harnois-Déraps, J. and Pen, U.-L. (2012). Non-Gaussian error bars in galaxy surveys - I. Mon. Not. R. Astron. Soc., 423:2288–2307.

Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B. J., Lawrence, E., and Wagner, C. (2009). The Coyote Universe. II. Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. Astrophys. J., 705:156–174.

Heitmann, K., White, M., Wagner, C., Habib, S., and Higdon, D. (2010). The Coyote Universe. I. Precision Determination of the Nonlinear Matter Power Spectrum. Astrophys. J., 715:104–121.

Howlett, C., Ross, A. J., Samushia, L., Percival, W. J., and Manera, M. (2015). The clustering of the SDSS main galaxy sample - II. Mock galaxy catalogues and a measurement of the growth of structure from redshift space distortions at z = 0.15. Mon. Not. R. Astron. Soc., 449:848–866.

Hu, W. and White, M. (2001). Power Spectra Estimation for Weak Lensing. Astrophys. J., 554:67–73.

Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173.

Jain, B. and Bertschinger, E. (1994). Second-order power spectrum and nonlinear evolution at high redshift. Astrophys. J., 431:495–505.

Jelić, V., Zaroubi, S., Labropoulos, P., Thomas, R. M., Bernardi, G., Brentjens, M. A., de Bruyn, A. G., Ciardi, B., Harker, G., Koopmans, L. V. E., Pandey, V. N., Schaye, J., and Yatawatta, S. (2008). Foreground simulations for the LOFAR-epoch of reionization experiment. Mon. Not. R. Astron. Soc., 389:1319–1335.

Kaiser, N. (1987). Clustering in real space and in redshift space. Mon. Not. R. Astron. Soc., 227:1–21.

Kaufman, C. G., Schervish, M. J., and W., N. D. (2008). The cosmological simulation code gadget-2. *Am. Statist. Assoc.*, (15451555).

Kazin, E. A., Koda, J., Blake, C., Padmanabhan, N., Brough, S., Colless, M., Contreras, C., Couch, W., Croom, S., Croton, D. J., Davis, T. M., Drinkwater, M. J., Forster, K., Gilbank, D., Gladders, M., Glazebrook, K., Jelliffe, B., Jurek, R. J., Li, I.-h., Madore, B., Martin, D. C., Pimbblet, K., Poole, G. B., Pracy, M., Sharp, R., Wisnioski, E., Woods, D., Wyder, T. K., and Yee, H. K. C. (2014). The WiggleZ Dark Energy Survey: improved distance measurements to z = 1 with reconstruction of the baryonic acoustic feature. Mon. Not. R. Astron. Soc., 441:3524–3542.

Kiessling, A., Cacciato, M., Joachimi, B., Kirk, D., Kitching, T. D., Leonard, A., Mandelbaum, R., Schäfer, B. M., Sifón, C., Brown, M. L., and Rassat, A. (2015). Galaxy Alignments: Theory, Modelling Simulations. Space Science Reviews, 193:67–136.

Kitaura, F.-S. and Heß, S. (2013). Cosmological structure formation with augmented Lagrangian perturbation theory. Mon. Not. R. Astron. Soc., 435:L78–L82.

Kitaura, F.-S., Yepes, G., and Prada, F. (2014). Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing. Mon. Not. R. Astron. Soc., 439:L21–L25.

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., and Heß, S. (2016). MultiDark simulations: the story of dark matter halo concentrations and density profiles. Mon. Not. R. Astron. Soc., 457:4340–4359.

Komatsu, E., Smith, K. M., Dunkley, J., Bennett, C. L., Gold, B., Hinshaw, G., Jarosik, N., Larson, D., Nolta, M. R., Page, L., Spergel, D. N., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Tucker, G. S., Weiland, J. L., Wollack, E., and Wright, E. L. (2011). Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. Astrophys. J. Suppl., 192:18.

Landy, S. D., Szalay, A. S., and Broadhurst, T. J. (1998). The Pairwise Velocity Distribution of Galaxies in the Las Campanas Redshift Survey. Astrophys. J. Lett., 494:L133–L136.

Laureijs, R., Amiaux, J., Arduini, S., Auguères, J. ., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., and et al. (2011). Euclid Definition Study Report. *ArXiv e-prints.*

Lazanu, A., Giannantonio, T., Schmittfull, M., and Shellard, E. P. S. (2016). Matter bispectrum of large-scale structure: Three-dimensional comparison between theoretical models and numerical simulations. Phys. Rev. D, 93(8):083517.

Levi, M., Bebek, C., Beers, T., Blum, R., Cahn, R., Eisenstein, D., Flaugher, B., Honscheid, K., Kron, R., Lahav, O., McDonald, P., Roe, N., Schlegel, D., and representing the DESI collaboration (2013). The DESI Experiment, a whitepaper for Snowmass 2013. *ArXiv e-prints.*

Li, Y., Hu, W., and Takada, M. (2014a). Super-sample covariance in simulations. Phys. Rev. D, 89(8):083519.

Li, Y., Hu, W., and Takada, M. (2014b). Super-sample signal. Phys. Rev. D, 90(10):103530.

Liddle, A. (2003). *An introduction to modern cosmology; 2nd ed.* Wiley, Chichester.

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D., Asztalos, S. J., Axelrod, T. S., and et al. (2009). LSST Science Book, Version 2.0. *ArXiv e-prints.*

Manera, M., Scoccimarro, R., Percival, W. J., Samushia, L., McBride, C. K., Ross, A. J., Sheth, R. K., White, M., Reid, B. A., Sánchez, A. G., de Putter, R., Xu, X., Berlind, A. A., Brinkmann, J., Maraston, C., Nichol, B., Montesano, F., Padmanabhan, N., Skibba, R. A., Tojeiro, R., and Weaver, B. A. (2013). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues. Mon. Not. R. Astron. Soc., 428:1036–1054.

Matarrese, S., Verde, L., and Heavens, A. F. (1997). Large-scale bias in the Universe: bispectrum method. Mon. Not. R. Astron. Soc., 290:651–662.

Mather, J. C., Cheng, E. S., Cottingham, D. A., Eplee, Jr., R. E., Fixsen, D. J., Hewagama, T., Isaacman, R. B., Jensen, K. A., Meyer, S. S., Noerdlinger, P. D., Read, S. M., Rosen, L. P., Shafer, R. A., Wright, E. L., Bennett, C. L., Boggess, N. W., Hauser, M. G., Kelsall, T., Moseley, Jr., S. H., Silverberg, R. F., Smoot, G. F., Weiss, R., and Wilkinson, D. T. (1994). Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. Astrophys. J., 420:439–444.

Mo, H., van den Bosch, F. C., and White, S. (2010). *Galaxy Formation and Evolution.*

Mohammed, I. and Seljak, U. (2014). Analytic model for the matter power spectrum, its covariance matrix and baryonic effects. Mon. Not. R. Astron. Soc., 445:3382–3400.

Mohammed, I., Seljak, U., and Vlah, Z. (2017). Perturbative approach to covariance matrix of the matter power spectrum. Mon. Not. R. Astron. Soc., 466:780–797.

Monaco, P. (1997). A Lagrangian Dynamical Theory for the Mass Function of Cosmic Structures - I. Dynamics. Mon. Not. R. Astron. Soc., 287:753–770.

Monaco, P. (2016). Approximate Methods for the Generation of Dark Matter Halo Catalogs in the Age of Precision Cosmology. *Galaxies*, 4:53.

Monaco, P., Sefusatti, E., Borgani, S., Crocce, M., Fosalba, P., Sheth, R. K., and Theuns, T. (2013). An accurate tool for the fast generation of dark matter halo catalogues. Mon. Not. R. Astron. Soc., 433:2389–2402.

Monaco, P., Theuns, T., and Taffoni, G. (2002). The pinocchio algorithm: pinpointing orbit-crossing collapsed hierarchical objects in a linear density field. Mon. Not. R. Astron. Soc., 331:587–608.

Munari, E., Monaco, P., Sefusatti, E., Castorina, E., Mohammad, F. G., Anselmi, S., and Borgani, S. (2017). Improving fast generation of halo catalogues with higher order Lagrangian perturbation theory. Mon. Not. R. Astron. Soc., 465:4658–4677.

Nishimichi, T., Bernardeau, F., and Taruya, A. (2016). Response function of the large-scale structure of the universe to the small scale inhomogeneities. *Physics Letters B*, 762:247–252.

Obreschkow, D., Power, C., Bruderer, M., and Bonvin, C. (2013). A Robust Measure of Cosmic Structure beyond the Power Spectrum: Cosmic Filaments and the Temperature of Dark Matter. Astrophys. J., 762:115.

Oka, A., Saito, S., Nishimichi, T., Taruya, A., and Yamamoto, K. (2014). Simultaneous constraints on the growth of structure and cosmic expansion from the multipole power spectra of the SDSS DR7 LRG sample. Mon. Not. R. Astron. Soc., 439:2515–2530.

Ostriker, J. P., Peebles, P. J. E., and Yahil, A. (1974). The size and mass of galaxies, and the mass of the universe. Astrophys. J. Lett., 193:L1–L4.

Padmanabhan, N., Schlegel, D. J., Finkbeiner, D. P., Barentine, J. C., Blanton, M. R., Brewington, H. J., Gunn, J. E., Harvanek, M., Hogg, D. W., Ivezić, Ž., Johnston, D., Kent, S. M., Kleinman, S. J., Knapp, G. R., Krzesinski, J., Long, D., Neilsen, Jr., E. H., Nitta, A., Loomis, C., Lupton, R. H., Roweis, S., Snedden, S. A., Strauss, M. A., and Tucker, D. L. (2008). An Improved Photometric Calibration of the Sloan Digital Sky Survey Imaging Data. Astrophys. J., 674:1217–1233.

Paz, D. J. and Sánchez, A. G. (2015). Improving the precision matrix for precision cosmology. Mon. Not. R. Astron. Soc., 454:4326–4334.

Peacock, J. A., Cole, S., Norberg, P., Baugh, C. M., Bland-Hawthorn, J., Bridges, T., Cannon, R. D., Colless, M., Collins, C., Couch, W., Dalton, G., Deeley, K., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Percival, W. J., Peterson, B. A., Price, I., Sutherland, W., and Taylor, K. (2001). A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey. Nature, 410:169–173.

Peacock, J. A. and Dodds, S. J. (1994). Reconstructing the Linear Power Spectrum of Cosmological Mass Fluctuations. Mon. Not. R. Astron. Soc., 267:1020.

Pearson, D. W. and Samushia, L. (2016). Estimating the power spectrum covariance matrix with fewer mock samples. Mon. Not. R. Astron. Soc., 457:993–999.

Peebles, P. J. E. (1973). Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory. Astrophys. J., 185:413–440.

Peek, J. E. G. and Graves, G. J. (2010). A Correction to the Standard Galactic Reddening Map: Passive Galaxies as Standard Crayons. Astrophys. J., 719:415–424.

Percival, W. J., Burkey, D., Heavens, A., Taylor, A., Cole, S., Peacock, J. A., Baugh, C. M., Bland-Hawthorn, J., Bridges, T., Cannon, R., Colless, M., Collins, C., Couch, W., Dalton, G., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Norberg, P., Peterson, B. A., Sutherland, W., and Taylor, K. (2004). The 2dF Galaxy Redshift Survey: spherical harmonics analysis of fluctuations in the final catalogue. Mon. Not. R. Astron. Soc., 353:1201–1218.

Percival, W. J., Samushia, L., Ross, A. J., Shapiro, C., and Raccanelli, A. (2011). Redshift-space distortions. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, pages 5058–5067.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., Groom, D. E., Hook, I. M., Kim, A. G., Kim, M. Y., Lee, J. C., Nunes, N. J., Pain, R., Pennypacker, C. R., Quimby, R., Lidman, C., Ellis, R. S., Irwin, M., McMahon, R. G., Ruiz-Lapuente, P., Walton, N., Schaefer, B., Boyle, B. J., Filippenko, A. V., Matheson, T., Fruchter, A. S., Panagia, N., Newberg, H. J. M., Couch, W. J., and Project, T. S. C. (1999). Measurements of $\Omega$ and $\Lambda$ from 42 High-Redshift Supernovae. Astrophys. J., 517:565–586.

Pinol, L., Cahn, R. N., Hand, N., Seljak, U., and White, M. (2016). Imprint of DESI fiber assignment on the anisotropic power spectrum of emission line galaxies. *ArXiv e-prints.*

Planck Collaboration, Adam, R., Ade, P. A. R., Aghanim, N., Akrami, Y., Alves, M. I. R., Argüeso, F., Arnaud, M., Arroja, F., Ashdown, M., and et al. (2016a). Planck 2015 results. I. Overview of products and scientific results. Astron. Astrophys., 594:A1.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., Baccigalupi, C., Banday, A. J., and et al. (2014). Planck 2013 results. XIV. Zodiacal emission. Astron. Astrophys., 571:A14.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. (2015). Planck 2015 results. XIII. Cosmological parameters. *ArXiv e-prints.*

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. (2016b). Planck 2015 results. XIII. Cosmological parameters. Astron. Astrophys., 594:A13.

Planck Collaboration, Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Bartolo, N., and et al. (2016c). Planck 2015 results. XI. CMB power spectra, likelihoods, and robustness of parameters. Astron. Astrophys., 594:A11.

Pope, A. C. and Szapudi, I. (2008). Shrinkage estimation of the power spectrum covariance matrix. Mon. Not. R. Astron. Soc., 389:766–774.

Regan, D. M., Shellard, E. P. S., and Fergusson, J. R. (2010). General CMB and primordial trispectrum estimation. Phys. Rev. D, 82(2):023520.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., and Tonry, J. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. Astron. J., 116:1009–1038.

Robert, C. P. (2004). *Monte carlo methods.* Wiley Online Library.

Roberts, M. S. and Rots, A. H. (1973). Comparison of Rotation Curves of Different Galaxy Types. Astron. Astrophys., 26:483–485.

Ross, A. J., Beutler, F., Chuang, C.-H., Pellejero-Ibanez, M., Seo, H.-J., Vargas-Magaña, M., Cuesta, A. J., Percival, W. J., Burden, A., Sánchez, A. G., Grieb, J. N., Reid, B., Brownstein, J. R., Dawson, K. S., Eisenstein, D. J., Ho, S., Kitaura, F.-S., Nichol, R. C., Olmstead, M. D., Prada, F., Rodríguez-Torres, S. A., Saito, S., Salazar-Albornoz, S., Schneider, D. P., Thomas, D., Tinker, J., Tojeiro, R., Wang, Y., White, M., and Zhao, G.-b. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: observational systematics and baryon acoustic oscillations in the correlation function. Mon. Not. R. Astron. Soc., 464:1168–1191.

Ross, A. J., Ho, S., Cuesta, A. J., Tojeiro, R., Percival, W. J., Wake, D., Masters, K. L., Nichol, R. C., Myers, A. D., de Simoni, F., Seo, H. J., Hernández-Monteagudo, C., Crittenden, R., Blanton, M., Brinkmann, J., da Costa, L. A. N., Guo, H., Kazin, E., Maia, M. A. G., Maraston, C., Padmanabhan, N., Prada, F., Ramos, B., Sanchez, A., Schlafly, E. F., Schlegel, D. J., Schneider, D. P., Skibba, R., Thomas, D., Weaver, B. A., White, M., and Zehavi, I. (2011). Ameliorating systematic uncertainties in the angular clustering of galaxies: a study using the SDSS-III. Mon. Not. R. Astron. Soc., 417:1350–1373.

Ross, A. J., Percival, W. J., Sánchez, A. G., Samushia, L., Ho, S., Kazin, E., Manera, M., Reid, B., White, M., Tojeiro, R., McBride, C. K., Xu, X., Wake, D. A., Strauss, M. A., Montesano, F., Swanson, M. E. C., Bailey, S., Bolton, A. S., Dorta, A. M., Eisenstein, D. J., Guo, H., Hamilton, J.-C., Nichol, R. C., Padmanabhan, N., Prada, F., Schlegel, D. J., Magaña, M. V., Zehavi, I., Blanton, M., Bizyaev, D., Brewington, H., Cuesta, A. J., Malanushenko, E., Malanushenko, V., Oravetz, D., Parejko, J., Pan, K., Schneider, D. P., Shelden, A., Simmons, A., Snedden, S., and Zhao, G.-b. (2012). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics. Mon. Not. R. Astron. Soc., 424:564–590.

Rubin, V. C., Thonnard, N., and Ford, Jr., W. K. (1978). Extended rotation curves of high-luminosity spiral galaxies. IV - Systematic dynamical properties, SA through SC. Astrophys. J. Lett., 225:L107–L111.

Sánchez, A. G., Scoccimarro, R., Crocce, M., Grieb, J. N., Salazar-Albornoz, S., Dalla Vecchia, C., Lippich, M., Beutler, F., Brownstein, J. R., Chuang, C.-H., Eisenstein, D. J., Kitaura, F.-S., Olmstead, M. D., Percival, W. J., Prada, F., Rodríguez-Torres, S., Ross, A. J., Samushia, L., Seo, H.-J., Tinker, J., Tojeiro, R., Vargas-Magaña, M., Wang, Y., and Zhao, G.-B. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the configuration-space clustering wedges. Mon. Not. R. Astron. Soc., 464:1640–1658.

Santos, M. G., Cooray, A., and Knox, L. (2005). Multifrequency Analysis of 21 Centimeter Fluctuations from the Era of Reionization. Astrophys. J., 625:575–587.

Sato, M., Takada, M., Hamana, T., and Matsubara, T. (2011). Simulations of Wide-field Weak-lensing Surveys. II. Covariance Matrix of Real-space Correlation Functions. Astrophys. J., 734:76.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Schilizzi, R. T., Dewdney, P. E. F., and Lazio, T. J. W. (2008). The Square Kilometre Array. In *Ground-based and Airborne Telescopes II*, volume 7012 of Intern. Soc. Opt. Eng., page 70121I.

Schlegel, D., Abdalla, F., Abraham, T., Ahn, C., Allende Prieto, C., Annis, J., Aubourg, E., Azzaro, M., Baltay, S. B. C., Baugh, C., Bebek, C., Becerril, S., Blanton, M., Bolton, A., Bromley, B., Cahn, R., Carton, P. ., Cervantes-Cota, J. L., Chu, Y., Cortes, M., Dawson, K., Dey, A., Dickinson, M., Diehl, H. T., Doel, P., Ealet, A., Edelstein, J., Eppelle, D., Escoffier, S., Evrard, A., Faccioli, L., Frenk, C., Geha, M., Gerdes, D., Gondolo, P., Gonzalez-Arroyo, A., Grossan, B., Heckman, T., Heetderks, H., Ho, S., Honscheid, K., Huterer, D., Ilbert, O., Ivans, I., Jelinsky, P., Jing, Y., Joyce, D., Kennedy, R., Kent, S., Kieda, D., Kim, A., Kim, C., Kneib, J. ., Kong, X., Kosowsky, A., Krishnan, K., Lahav, O., Lampton, M., LeBohec, S., Le Brun, V., Levi, M., Li, C., Liang, M., Lim, H., Lin, W., Linder, E., Lorenzon, W., de la Macorra, A., Magneville, C., Malina, R., Marinoni, C., Martinez, V., Majewski, S., Matheson, T., McCloskey, R., McDonald, P., McKay, T., McMahon, J., Menard, B., Miralda-Escude, J., Modjaz, M., Montero-Dorta, A., Morales, I., Mostek, N., Newman, J., Nichol, R., Nugent, P., Olsen, K., Padmanabhan, N., Palanque-Delabrouille, N., Park, I., Peacock, J., Percival, W., Perlmutter, S., Peroux, C., Petitjean, P., Prada, F., Prieto, E., Prochaska, J., Reil, K., Rockosi, C., Roe, N., Rollinde, E., Roodman, A., Ross, N., Rudnick, G., Ruhlmann-Kleider, V., Sanchez, J., Sawyer, D., Schimd, C., Schubnell, M., Scoccimaro, R., Seljak, U., Seo, H., Sheldon, E., Sholl, M., Shulte-Ladbeck, R., Slosar, A., Smith, D. S., Smoot, G., Springer, W., Stril, A., Szalay, A. S., Tao, C., Tarle, G., Taylor, E., Tilquin, A., Tinker, J., Valdes, F., Wang, J., Wang, T., Weaver, B. A., Weinberg, D., White, M., Wood-Vasey, M., Yang, J., Yeche, X. Y. C., Zakamska, N., Zentner, A., Zhai, C., and Zhang, P. (2011). The BigBOSS Experiment. *ArXiv e-prints*.

Schlegel, D. J., Finkbeiner, D. P., and Davis, M. (1998). Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. Astrophys. J., 500:525–553.

Schmittfull, M., Baldauf, T., and Seljak, U. (2015). Near optimal bispectrum estimators for large-scale structure. Phys. Rev. D, 91(4):043530.

Schmittfull, M. M., Regan, D. M., and Shellard, E. P. S. (2013a). Fast estimation of gravitational and primordial bispectra in large scale structures. Phys. Rev. D, 88(6):063512.

Schmittfull, M. M., Regan, D. M., and Shellard, E. P. S. (2013b). Fast estimation of gravitational and primordial bispectra in large scale structures. Phys. Rev. D, 88(6):063512.

Scoccimarro, R., Colombi, S., Fry, J. N., Frieman, J. A., Hivon, E., and Melott, A. (1998). Nonlinear Evolution of the Bispectrum of Cosmological Perturbations. Astrophys. J., 496:586–604.

Scoccimarro, R. and Sheth, R. K. (2002). PTHALOS: a fast method for generating mock galaxy distributions. Mon. Not. R. Astron. Soc., 329:629–640.

Scoccimarro, R., Zaldarriaga, M., and Hui, L. (1999). Power Spectrum Correlations Induced by Nonlinear Clustering. Astrophys. J., 527:1–15.

Sefusatti, E., Crocce, M., Pueblas, S., and Scoccimarro, R. (2006). Cosmology and the bispectrum. Phys. Rev. D, 74(2):023522.

Sefusatti, E., Crocce, M., Scoccimarro, R., and Couchman, H. M. P. (2016). Accurate Estimators of Correlation Functions in Fourier Space. Mon. Not. R. Astron. Soc..

Sheth, R. K., Mo, H. J., and Tormen, G. (2001). Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. Mon. Not. R. Astron. Soc., 323:1–12.

Springel, V. (2005). The cosmological simulation code gadget-2. *Monthly Notices of the Royal Astronomical Society*, 364(4):1105–1134.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., and Pearce, F. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. Nature, 435:629–636.

Takada, M. and Hu, W. (2013). Power spectrum super-sample covariance. Phys. Rev. D, 87(12):123504.

Takahashi, R., Yoshida, N., Takada, M., Matsubara, T., Sugiyama, N., Kayo, I., Nishizawa, A. J., Nishimichi, T., Saito, S., and Taruya, A. (2009). Simulations of Baryon Acoustic Oscillations. II. Covariance Matrix of the Matter Power Spectrum. Astrophys. J., 700:479–490.

Tassev, S., Zaldarriaga, M., and Eisenstein, D. J. (2013). Solving large scale structure in ten easy steps with COLA. J. Cosmol. Astropart. Phys., 6:036.

Taylor, A., Joachimi, B., and Kitching, T. (2013). Putting the precision in precision cosmology: How accurate should your data covariance matrix be? Mon. Not. R. Astron. Soc., 432:1928–1946.

Taylor, A. N. and Hamilton, A. J. S. (1996). Non-linear cosmological power spectra in real and redshift space. Mon. Not. R. Astron. Soc., 282:767–778.

The Planck Collaboration (2006). The Scientific Programme of Planck. *ArXiv Astrophysics e-prints*.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Ann. Math. Statist.*

Verde, L. (2010). Statistical Methods in Cosmology. In Wolschin, G., editor, *Lecture Notes in Physics, Berlin Springer Verlag*, volume 800 of *Lecture Notes in Physics, Berlin Springer Verlag*, pages 147–177.

Verde, L., Heavens, A. F., Matarrese, S., and Moscardini, L. (1998). Large-scale bias in the Universe - II. Redshift-space bispectrum. Mon. Not. R. Astron. Soc., 300:747–756.

Wolstenhulme, R., Bonvin, C., and Obreschkow, D. (2015). Three-point Phase Correlations: A New Measure of Non-linear Large-scale Structure. Astrophys. J., 804:132.

Wolz, L., Abdalla, F. B., Blake, C., Shaw, J. R., Chapman, E., and Rawlings, S. (2014). The effect of foreground subtraction on cosmological measurements from intensity mapping. Mon. Not. R. Astron. Soc., 441:3271–3283.

Yamamoto, K., Bassett, B. A., and Nishioka, H. (2005). Dark Energy Reflections in the Redshift-Space Quadrupole. *Physical Review Letters*, 94(5):051301.

York, D. G., Adelman, J., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., Castander, F. J., Chen, B., Colestock, P. L., Connolly, A. J., Crocker, J. H., Csabai, I., Czarapata, P. C., Davis, J. E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B. R., Evans, M. L., Fan, X., Federwitz, G. R., Fiscelli, L., Friedman, S., Frieman, J. A., Fukugita, M., Gillespie, B., Gunn, J. E., Gurbani, V. K., de Haas, E., Haldeman, M., Harris, F. H., Hayes, J., Heckman, T. M., Hennessy, G. S., Hindsley, R. B., Holm, S., Holmgren, D. J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R. S. J., Kinney, E., Klaene, M., Kleinman, A. N., Kleinman, S., Knapp, G. R., Korienek, J., Kron, R. G., Kunszt, P. Z., Lamb, D. Q., Lee, B., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R. H., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., Meiksin, A., Merelli, A., Monet, D. G., Munn, J. A., Narayanan, V. K., Nash, T., Neilsen, E., Neswold, R., Newberg, H. J., Nichol, R. C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. L., Petravick, D., Pier, J. R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T. R., Richards, G. T., Richmond, M. W., Rivetta, C. H., Rockosi, C. M., Ruthmansdorfer, K., Sandford, D., Schlegel, D. J., Schneider, D. P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W. A., Smee, S., Smith, J. A., Snedden, S., Stone, R., Stoughton, C., Strauss, M. A., Stubbs, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D. H., Yanny, B., Yasuda, N., and SDSS Collaboration (2000). The Sloan Digital Sky Survey: Technical Summary. Astron. J., 120:1579–1587.

Zhao, C., Kitaura, F.-S., Chuang, C.-H., Prada, F., Yepes, G., and Tao, C. (2015). Halo mass distribution reconstruction across the cosmic web. Mon. Not. R. Astron. Soc., 451:4266–4276.