On unbiased and higher-order large-scale structure statistics: Covariance Matrices and Minkowski Functionals

Martha Lippich Golobart

München 2021

On unbiased and higher-order large-scale structure statistics: Covariance Matrices and Minkowski Functionals

Martha Lippich Golobart

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Martha Lippich Golobart
aus Genf

München, den 4. Februar 2021

Erstgutachter: Prof. Dr. Ralf Bender
Zweitgutachter: Prof. Dr. Jochen Weller
Tag der mündlichen Prüfung: 28. April 2021

Zusammenfassung

Die Untersuchung der großräumigen Verteilung von Galaxien hat wesentlich zu unserem heutigen Verständnis von der Zusammensetzung und der Entwicklung unseres Universums beigetragen. Die kommende Generation von Galaxien-Durchmusterungen wird es uns ermöglichen, die kosmische Expansionsgeschichte und das Strukturwachstum mit bislang unerreichter Präzision zu messen, und dadurch das kosmologische Standardmodell und seine möglichen Erweiterungen noch genauer zu erforschen. Ziel dieser Arbeit ist es neue Einblicke in Verfahren für die Gewinnung kosmologischer Information aus der großräumigen Struktur zu gewähren, die für zukünftige Analysen der "Klumpungseigenschaften" von Galaxien (Galaxy Clustering) von Nutzen sein könnten.

Die Kovarianzmatrix von Clustering-Messungen ist ein wesentlicher Bestandteil von Analysen, der Voraussetzung für die unverfälschte Bestimmung kosmologischer Parameter ist. Die Verwendung einer großen Anzahl von Mock-Katalogen, die aus Simulationen konstruiert werden, gilt als der zuverlässigste Ansatz für die Schätzung der Kovarianzmatrix. Die Durchführung einer großen Anzahl von vollständigen kosmologischen Simulationen ist jedoch mit einem hohen Rechenaufwand verbunden. Der erste Teil der Arbeit beschreibt einen gründlichen Vergleich von Kovarianzmatrizen, die mit sieben unterschiedlichen approximativen Methoden zur gravitativen Strukturbildung hergeleitet werden. Das umfasst prädiktive Methoden, die die Entwicklung des Dichtefelds der dunklen Materie deterministisch vorhersagen (ICE-COLA, Peak Patch, Pinocchio), Methoden, die eine vorherige Kalibrierung mit vollständigen N-Körpersimulationen (Patchy und Halogen) erfordern, und zwei einfachere Verfahren, die auf der Annahme einer log-normalen oder normalen Wahrscheinlichkeitsdichtefunktion des Dichtefelds der Dunklen Materie basieren. Der Vergleich bezieht sich auf Messungen der anisotropen Zwei-Punkt-Korrelationsfunktion, eines der erfolgreichsten und am weitesten verbreiteten Mittel für die statistische Analyse des Galaxy Clustering.

Die mit den approximativen Methoden erhaltenen Kovarianmatrizen werden mit Referenz-Kovarianzmatrizen verglichen, die aus einem Satz kosmologischer N-Körpersimulationen hergeleitet werden. Insbesondere wird untersucht, wie geeignet die Kovarianzmatrizen sind, die mithilfe der approximativen Methoden geschätzt worden sind, um die Parameterbestimmung basierend auf den Referenz-Kovarianzmatrizen aus N-Körpersimulationen zu reproduzieren. Die Ergebnisse zeigen, dass alle approximativen Methoden die Referenzergebnisse mit einer Genauigkeit von 5% für unsere untere Massenschwelle und von 10% für unsere obere Massenschwelle der Dunkle-Materie-Halos nachbilden können, ohne dass

eine Methode signifikant besser ist als die anderen.

Gegenstand des zweiten Teils der Arbeit ist ein mögliches Verfahren zur Gewinnung kosmologischer Information aus der Galaxienverteilung, das über Zwei-Punkt-Statistiken hinausgeht. Dazu werden Minkowski-Funktionale untersucht, die die Geometrie und Topologie des kosmischen Dichtefelds charakterisieren und komprimierte Informationen von Korrelationen höherer Ordnung enthalten. Der erste Schritt ist die Implementierung einer robusten und genauen Methode zur Schätzung der Minkowski-Funktionale von dreidimensionalen Punktverteilungen, insbesondere aus Simulationen mit periodischen Randbedingungen. Der daraus resultierende Code mit dem Namen MEDUSA berechnet die Minkowski-Funktionale von triangulierten Oberflächen gleicher Dichte, die aus der Delaunay-Tessellation der Eingabepunktmenge konstruiert werden.

Nach der gründlichen Validierung des Codes mit Testpunktmengen folgt die Anwendung auf synthetische Kataloge aus verschiedenen N-Körpersimulationen. Die aus den Katalogen berechneten Minkowski-Funktionale weisen klare Abweichungen von den Vorhersagen für ein Gaußsches Dichtefeld auf. Dies ist aufgrund der nichtlinearen Gravitationsentwicklung des Dichtefelds zu erwarten, die zu Korrelationen höherer Ordnung in der Galaxienverteilung führt. Die Analyse der im Rotverschiebungsraum gemessenen Minkowski-Funktionale zeigt, dass die Rotverschiebungsraum-Verzerrungen die Schätzung der Minkowski-Funktionale signifikant beeinflusst. Der Effekt der Verzerrungen kann jedoch erheblich verringert werden, wenn die Messungen als Funktion des Anteils des ausgefüllten Volumens anstelle der Dichteschwelle ausgedrückt werden. Auch die Alcock-Paczynski(AP)-Verzerrungen beeinflussen die Messungen der Minkowski-Funktionale. Ihr Effekt kann modelliert werden, indem die Messungen mit geeigneten Potenzen des isotropen AP-Parameters q skaliert werden, der von der volumengemittelten Distanz $D_{\mathrm{V}}(z)$ abhängt. Die außerdiagonalen Elemente der Kovarianzmatrix der Minkowski-Funktionale sind zum Teil stark korreliert, und müssen daher für zukünftige Analysen berücksichtigt werden.

Schließlich wird ein neuartiger Ansatz getestet, genannt *Evolution Mapping*, um die Abhängigkeit der Minkowski-Funktionale von den kosmologischen Parametern zu beschreiben, die die Entwicklung des linearen Wachstumsfaktors mit der Rotverschiebungs beeinflussen. Die zugrundeliegende Idee besteht darin, den Einfluss dieser Parameter anhand von σ_{12}, der Dispersion des über Kugeln mit einem Radius von 12 Mpc gemittelten, linear nach heute extrapolierten Dichtekontrastes, zu charakterisieren. Die Analyse zeigt, dass dieser Ansatz für die Minkowski-Funktionale mit hoher Genauigkeit gültig ist und als Ausgangspunkt für die Entwicklung eines simulationsbasierten Modells für die Minkowski-Funktionale von Nicht-Gaußschen-Dichtefeldern dienen kann.

Abstract

The study of the large-scale galaxy clustering has significantly contributed to our present-day understanding of the composition and the evolution of our Universe. The upcoming generation of galaxy redshift surveys will open the opportunity to further probe our standard cosmological model and its possible extensions by measuring the cosmic expansion and growth of structure histories to unprecedented precision. This thesis aims to provide new insights into the challenges of capturing the information encoded in the large-scale structure, which might be useful for future galaxy clustering analyses.

The covariance matrix of the clustering measurements is an essential ingredient to derive unbiased constraints on cosmological parameters. Using large numbers of mock catalogues from simulations is considered the most reliable approach to estimate the covariance matrix. However, running large numbers of full cosmological simulations is computationally very expensive. The first part of the thesis presents a detailed comparison of the covariance matrices inferred from seven state-of-the-art approximate methods for gravitational structure formation. These include predictive methods that follow the evolution of the underlying matter density field deterministically (ICE-COLA, Peak Patch, Pinocchio), methods that require a prior calibration with full N-body simulations (Patchy and Halogen), and two simpler recipes based on assuming log-normal or Gaussian shapes of the full density probability distribution function. The comparison focuses on measurements of the anisotropic two-point correlation function, one of the most successful and widespread tools for the statistical analysis of galaxy clustering.

We compare the covariance estimates obtained from the approximate methods against reference covariances inferred from a set of cosmological N-body simulations. In particular, we examine the performance of the covariance matrices from the approximate methods at reproducing parameter constraints from the analysis based on the N-body simulations. Our results show that all approximate methods can recover the results from the N-body simulations with an accuracy of 5% for our lower halo mass threshold, and of 10% for our higher halo mass threshold, with no method clearly outperforming the others.

With the goal of extracting information beyond two-point statistics, the second part of the thesis considers Minkowski functionals, which characterize the geometry and topology of the cosmic density field and contain compressed higher-order information. In this context, we implement a robust and accurate method for estimating the Minkowski functionals of three-dimensional point distributions, in particular of samples from simulations with periodic boundary conditions. The resulting code, dubbed MEDUSA, computes the

Minkowski functionals of triangulated isodensity surfaces that are constructed from the Delaunay tessellation of the input point sample.

After the thorough validation of the code on test samples, we apply it to synthetic catalogues from different N-body simulations. The resulting Minkowski functionals exhibit clear non-Gaussian signatures. These are expected due to the non-linear gravitational evolution of the density field, which leads to higher-order correlations in the galaxy distribution. The analysis of the Minkowski functionals measured in redshift space indicates that the redshift-space distortions significantly change the Minkowski functional estimates, but their impact is considerably reduced if the measurements are expressed as a function of the volume-filling fraction instead of the density threshold. The Minkowski functional measurements are also sensitive to Alcock-Paczynski (AP) distortions. Their effect can be modelled by scaling the measurements with suitable powers of the isotropic AP parameter q, which depends on the volume-averaged distance $D_{\rm V}(z)$. The covariance matrix of the Minkowski functional measurements has significant off-diagonal structure, which needs to be taken into account for future analyses.

Finally, we test a novel approach to describe the cosmology dependence of the Minkowski functionals, dubbed *evolution mapping*, in which the impact of a large number of cosmological parameters that affect the redshift evolution of the linear growth factor is characterized by the value of σ_{12}, the linear-theory rms mass fluctuation in spheres of radius 12 Mpc. Our analysis shows that this approach is valid for the Minkowski functionals with high accuracy and can serve as a starting point to develop a simulation-based model for the Minkowski functionals of non-Gaussian density fields.

Für meine Eltern

"*There is no harmony in the Universe*
We have to get acquainted to this idea that
there is no real harmony as we have conceived it
But when I say this, I say this all full of admiration"
Werner Herzog
In Takeshi's Cashew "There Is No Harmony"

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern cosmology has advanced our understanding of the composition and the evolution of our Universe to an extraordinary accuracy. A rich variety of observations, including measurements of the cosmic microwave background, Type Ia supernovae and the large-scale clustering of galaxies, have established the standard ΛCDM model of cosmology.

According to this concordance model, the Universe is mainly composed of two constituents that cannot be directly observed. The first one is a mysterious dark energy component that accounts for the accelerated expansion of the Universe and can be described by the cosmological constant Λ. The second one is attributed to cold dark matter (CDM), an additional mass component that only interacts gravitationally, but not electromagnetically. Furthermore, the ΛCDM model provides the general framework to describe the evolution of cosmic structure. Under the effect of gravity and cosmic expansion, small inhomogeneities grow into the large-scale structure that today can be observed as galaxy filaments, galaxy clusters and vast almost devoid regions.

The statistical analysis of the large-scale structure traced by galaxies is one of the primary tools of observational cosmology. For this purpose, during the last decades galaxy redshift surveys have mapped the three-dimensional galaxy distribution in increasingly larger cosmic volumes, providing catalogues with the angular positions and the redshifts of the observed galaxies (e.g. Huchra et al., 1983; York et al., 2000; Colless et al., 2001; Jones et al., 2004; The Dark Energy Survey Collaboration, 2005; Hill et al., 2008; Drinkwater et al., 2010).

An outstanding example is the Baryon Oscillation Spectroscopic Survey (BOSS) of the Sloan Digital Sky Survey (SDSS) III, which delivered the largest galaxy catalogue with spectroscopic redshifts to date (Dawson et al., 2013). The footprint of the survey covers approximately 10 000 deg^2 of the sky, containing more than 1.5 million galaxies in a redshift range of $0.15 < z < 0.75$ and 150 000 quasars at a mean redshift of $z \simeq 2.5$ (Alam et al., 2015). The extended BOSS program (eBOSS, Dawson et al., 2016) further added ∼550 000 galaxies in a redshift range of $0.6 < z < 1.1$ and ∼340 000 quasars in a redshift range of $0.8 < z < 2.2$ (Ross et al., 2020).

The most commonly used methods to extract the cosmological information encoded in the galaxy distribution are two-point statistics, more specifically the two-point correla-

tion function and its Fourier transform, the power spectrum. One of the most remarkable features in the measured two-point statistics is the signature of the baryon acoustic oscillations (BAO). These oscillations arise from sound waves propagating in the plasma of the early Universe and imprint a characteristic scale in the clustering pattern of galaxies. BAO distance measurements employ this characteristic scale as a standard ruler, in order to infer constraints on the expansion history of the Universe (Blake & Glazebrook, 2003; Linder, 2003; Cole et al., 2005; Eisenstein et al., 2005). Further information can be retrieved from distortions in the measured clustering statistic that are caused by the peculiar velocities of the galaxies. These so-called redshift-space distortions (RSD) can be used to probe the growth-rate of cosmic structures (Guzzo et al., 2008). Measurements of the full shape of the two-point statistics provide complementary information to the BAO distance measurements and simultaneously constrain the growth rate (Percival et al., 2001; Sánchez et al., 2006; Chuang & Wang, 2012; Sánchez et al., 2013).

The clustering analyses of the past surveys have confirmed and constrained the ΛCDM model with an increasing precision. In particular, the most recent surveys turned the large-scale structure analysis into a precision scientific discipline (e.g., Alam et al., 2017; Sánchez et al., 2017; Grieb et al., 2017; eBOSS Collaboration et al., 2020; Hou et al., 2021).

Despite the success of the ΛCDM model, which has been well-tested not only by galaxy clustering measurements, but a large variety of complimentary cosmological probes, including the CMB, gravitational lensing, supernovae and galaxy clusters, many open questions remain. The nature of dark energy or the origin of the accelerated cosmic expansion is one of the most intriguing questions of modern physics.

The upcoming new generation of surveys sampling several millions of galaxies in huge cosmic volumes, will allow us to measure the expansion history of the Universe and the growth of cosmic structure to unprecedented precision and accuracy, and will hopefully bring us closer to unravel the mystery of cosmic expansion. Apart from this, these surveys will also explore other important physical problems, such as the neutrino mass and the physics of inflation.

The ESA mission *Euclid*, which is planned for launch in 2022, is expected to make significant progress in this direction. *Euclid* will observe up to 50 million galaxies with near-infrared spectroscopy in a redshift range of $0.7 < z < 2.1$ covering 15 000 deg^2 of the extragalactic sky (Laureijs et al., 2011). A second upcoming ground-based survey will be conducted by the Dark Energy Spectroscopic Instrument (DESI Collaboration et al., 2016), which will contain 30 million galaxy and quasar redshifts starting at $z \sim 0.2$ and up to $z > 2$ in a footprint of 14 000 deg^2.

These future surveys will provide an abundance of cosmological information. In order to reliably extract the largest possible amount of this information, two key aspects need to be further explored. First, the large amount of high-quality data will significantly reduce the statistical errors of large-scale structure measurements. Consequently, the systematic errors introduced by the analysis methods might become the largest error source. For this reason, it is important to examine all components of the systematic error budget. Secondly, as previous analyses have shown, the underlying matter density field traced by the galaxies is not simply Gaussian distributed, and thus, it cannot be completely characterized by two-

point statistics. Complementing standard two-point analyses by higher-order statistics will be essential to fully exploit the cosmological information encoded in the large-scale structure. The aim of this thesis is to make contributions to both aspects that might serve to extract unbiased and higher-order information from the large-scale structure in the clustering analyses of future surveys.

The first part of this thesis is devoted to the covariance matrix, one of the key ingredients in the analysis of clustering measurements. In particular, I focus on the covariance matrices of two-point correlation function measurements. Commonly, these covariance matrices are estimated from mock catalogues based on simulations that are designed to reproduce the observed galaxy clustering properties. Due to the finite number of mock catalogues, the estimation of the covariance matrix is affected by noise, which must be considered as potential systematic error and must be propagated into the final cosmological constraints (Taylor et al., 2013; Dodelson & Schneider, 2013; Percival et al., 2014; Sellentin & Heavens, 2016). To satisfy the precision requirements of future surveys, it might be necessary to produce an unfeasibly large number of simulations. Approximate methods for gravitational structure formation and evolution can represent a viable alternative to full cosmological simulations, since they allow for a faster generation of mock catalogues. Before applying these methods to clustering analyses, they must be thoroughly tested, in order to avoid introducing systematic errors or biases on the final parameter constraints. To this end, I conduct a detailed comparison of several state-of-the-art approximate methods with regard to their accuracy in reproducing the covariance matrix estimates from full cosmological simulations.

The second part of this thesis addresses the topic of higher-order statistics and, more specifically, considers Minkowski functionals. The direct approach to extract the information encoded in the large-scale galaxy distribution beyond two-point statistics is to compute higher-order N-point functions, such as the three-, four- ore five-point correlation functions and their Fourier transforms. However, already the estimation and the analysis of the three-point correlation function and its Fourier transform the bispectrum is challenging (e.g., Marín et al., 2013; Gil-Marín et al., 2015; Gil-Marín et al., 2017; Slepian et al., 2017b,a; Pearson & Samushia, 2018), the analysis of higher-order N-point functions is still unfeasible.

An alternative approach is to use statistics that contain compressed higher-order information, such as the Minkowski functionals, which characterize the geometry and topology of the galaxy density field. The idea to use Minkowski functionals for the large-scale structure analysis dates back to the 1990s (Mecke et al., 1994) and since then there have been several cosmological studies on Minkowski functionals. Two different main methods were introduced to determine Minkowski functionals from the galaxy distribution. One of them is based on the Germ-Grain model, where the Minkowski functionals are estimated on intersecting spheres inflated around the galaxies. The Germ-Grain Minkowski functionals have been used to study several galaxy and galaxy cluster catalogues in the last three decades (e.g., Mecke et al., 1994; Kerscher et al., 1997, 1998, 2001; Wiegand et al., 2014; Wiegand & Eisenstein, 2017).

Alternatively, the Minkowski functionals can be estimated from excursion sets, i.e.

isodensity surfaces, constructed from the galaxy distribution. The isodensity MFs have long known analytic expressions for Gaussian density fields (Tomita, 1990; Schmalzing & Buchert, 1997; Matsubara, 2003), and they are commonly evaluated using calculation methods from differential or integral geometry. They also were applied to different galaxy surveys (e.g., Hikage et al., 2003; Park et al., 2005; Gott et al., 2009; James et al., 2009; Choi et al., 2010; Zhang et al., 2010; Blake et al., 2014). Most analyses were based on the Gaussian predictions, which are sensitive to the two-point statistics only, and often only considered the genus, one of the Minkowski functionals encompassing the topological information.

Although the studies of the germ-grain and isodensity Minkowski functionals of the galaxy density field led to several interesting cosmological results, they could not compete with the large number of the highly advanced analyses based on two-point statistics. In the light of upcoming surveys, however, it becomes promising to reconsider Minkowski functionals as complementary tools to the standard two-point clustering statistics. Driven by this long-term goal, we developed MEDUSA, a new implementation of a robust and accurate method to estimate the isodensity Minkowski functionals of three-dimensional galaxy distributions based on Delaunay tessellations. The first applications of MEDUSA to synthetic catalogues already allows us to study several key aspects that will be relevant for future analyses of Minkowski functionals, and thus will also be important for extracting unbiased information.

The thesis is organised as follows: In the following Chapter 2, I depict the key theoretical concepts of cosmology and the statistical analysis of galaxy clustering relevant for this work. In Chapter 3, I describe the covariance matrix comparison project including the methodology, the considered state-of-the-art approximate methods and the analysis of the inferred covariance matrices. In Chapter 4, I present the basic algorithm to measure Minkowski functionals implemented in MEDUSA, the various validation tests and the first applications to synthetic catalogues. The latter allows exploring the non-Gaussian features, some relevant distortions expected from real galaxy surveys and the covariance matrices of the Minkowski functional measurements. Finally, I analyse the novel approach of evolution mapping for the Minkowski functionals that could be used as the basis for modelling the Minkowski functionals of non-Gaussian density fields. In the last Chapter 5, I discuss the main results and possible future perspectives.

Chapter 2

Theory of the Cosmological Large-Scale Structure

This chapter gives an overview of the principal theoretical concepts in cosmology with an emphasis on the statistical analysis of galaxy clustering. The analysis of the large-scale structure (LSS) of the Universe is a mature field which covers a broad range of approaches and topics. Here, I focus on the aspects more relevant to my work on two-point statistics and Minkowski functionals.

Section 2.1 is dedicated to the homogeneous ΛCDM universe and explains cosmic distances. The subsequent sections describe the evolution of cosmic density perturbations on the smooth background Universe from their seeds (Section 2.2) to the late-time non-linear structure formation (Section 2.5). In this context, Section 2.3 introduces two-point clustering statistics, which fully characterize the initial Gaussian density field. An important technique for the prediction of the non-linear evolution of the matter density field are N-body simulations, whose basic concepts are summarized in Section 2.6. Section 2.7 presents Minkowski functionals as a useful set of statistics beyond two-point correlations. The last Section 2.8 addresses important aspects of clustering observations encompassing galaxy bias, redshift-space distortions and Alcock-Paczynski distortions.

Most of the sections, besides the introduction to Minkowski functionals, are largely inspired by the textbook by Dodelson & Schmidt (2020) and the lecture notes on "The Formation and Evolution of Cosmic Structures" by Sánchez (in prep.). For a more detailed description of the concepts introduced here and a more complete overview over observational cosmology the reader is referred to these references.

2.1 The homogeneous universe

2.1.1 The background evolution

In our current physical understanding, gravity and cosmic expansion are the main drivers of the evolution of the large-scale structure in the Universe. The theory of General Relativity

provides the framework to describe cosmology. In General Relativity, gravity is attributed to the curvature of the four-dimensional space-time, characterized by the metric $g_{\mu\nu}$[1]. The Einstein field equations, which form the core of this theory, relate the space-time geometry encoded in the Einstein tensor $G_{\mu\nu}$ to the matter distribution given by the energy-momentum tensor $T_{\mu\nu}$[2],

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{2.1}$$

G is the Newtonian gravitational constant. Λ is the *cosmological constant*, which is essential for the description of the accelerated cosmic expansion. It can be interpreted as part of the space-time geometry on the left-hand side of the equation or it can be associated to the energy-momentum tensor as so-called *dark energy* on the right-hand side.

The solution of the Einstein field equations for our Universe is based on a fundamental hypothesis: the *cosmological principle* that states that the Universe is spatially homogeneous and isotropic on large scales. This clearly can only hold for the very large scales of hundreds of megaparsecs, since for smaller scales large non-uniformly distributed density fluctuations such as galaxies, filaments and voids are observed. However, modelling the evolution of the large-scale background cosmology is also crucial for the study of smaller scales, since it allows to describe the structure in the Universe as perturbations on the homogeneous and isotropic background.

Under this assumption of homogeneity and isotropy, the most general solution for the metric tensor $g_{\mu\nu}$ is the spatially maximally symmetric *Friedmann-Lemaître-Robertson-Walker* (FLWR) metric. For the FLRW metric, the line element, which corresponds to the infinitesimal distance in the four space-time coordinates, has the form

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu = -\,\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right], \tag{2.2}$$

where t is the time coordinate and r, θ and ϕ are the spatial spherical coordinates. The time evolution of the spatial part of the metric is entirely absorbed into the *scale factor* $a(t)$, which allows for uniform spatial expansion or contraction. The constant K specifies the curvature of spatial hypersurfaces of the Universe and can be classified as flat (*Euclidean*) for $K = 0$, open (hyperbolic) $K < 0$ or closed (elliptical) $K > 0$.

The idealized matter and energy distribution in the Universe, which is homogeneous and isotropic, can be modelled as a perfect fluid. For a perfect fluid with pressure p and energy density ρ in its own rest-frame the energy-momentum tensor is given by

$$T_{\mu\nu} = (p+\rho)U_\mu U_\nu + p g_{\mu\nu}, \tag{2.3}$$

where U_μ is its four-velocity.

Inserting the perfect-fluid form for the energy momentum tensor (2.3) and the FLRW metric (2.2) into the field equations (2.1), yields the *Friedmann equations*,

[1] Greek indices denote the four space-time components from 0 to 3 (time and spatial coordinates), Latin indices the three spatial components from 1 to 3.

[2] Assuming $c = 1$ throughout the work.

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\sum_i \rho_i + \frac{\Lambda}{3} - \frac{K}{a^2}, \tag{2.4}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\sum_i (\rho_i + 3p_i) + \frac{\Lambda}{3}. \tag{2.5}$$

The Friedmann equations describe the time evolution of the scale factor $a(t)$ as a function of the curvature K, the cosmological constant Λ, the energy densities ρ_i and the pressure p_i of the different constituents of the Universe. The rate of cosmic expansion is also called *Hubble parameter* H,

$$H \equiv \frac{\dot{a}}{a}, \tag{2.6}$$

and today's value is denoted as $H_0 \equiv H(t_0)$. The scale factor today is set to $a(t_0) = 1$. For historical reasons the Hubble constant is often expressed by the parameter h, $H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

Another frequently used quantity is the critical density,

$$\rho_\mathrm{c} = \frac{3H_0^2}{8\pi G} = 2.773 \cdot 10^{11}\,\mathrm{M_\odot}\,h^{-1}(\,\mathrm{Mpc}\,h^{-1})^{-3}, \tag{2.7}$$

which, according to equation (2.4), is the total energy density that leads to a flat Universe today ($K = 0$). The energy content of the different species in the Universe is often expressed in terms of the energy density normalized by the critical density,

$$\Omega_i = \frac{\rho_i(t_0)}{\rho_\mathrm{c}}. \tag{2.8}$$

The density parameter of the component associated with the cosmological constant Λ can also be expressed as

$$\Omega_\Lambda = \frac{\Lambda}{3H_0^2}. \tag{2.9}$$

The density parameter of the curvature K is defined as

$$\Omega_K = -\frac{K}{H_0^2}. \tag{2.10}$$

When so defined, the energy densities Ω_i depend on the value of h, according to equation (2.7). In some cases, it is useful to consider the physical density parameters,

$$w_i = \Omega_i h^2 = \frac{8\pi G}{3H_{100}^2}\rho_i, \tag{2.11}$$

which are defined with respect to the constant $H_{100} = 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, and therefore do not depend on the value of h. From equation (2.8) it follows that $\sum_i \Omega_i = 1$, and from equation (2.11) that $\sum_i w_i = h^2$.

To derive the evolution of each density component, it is necessary to assume a relation between the pressure and the density in equation (2.5). In the perfect fluid approach, the pressure is related to the density by a linear equation of state

$$p_i = w_i \rho_i, \tag{2.12}$$

with a constant equation-of-state parameter w_i. Using this relation and expressing the Friedmann equations in terms of the density parameters, the evolution of each density component can be written as

$$\rho_i(t) = \Omega_i \, \rho_\mathrm{c} \, a(t)^{-3(1+w_i)}. \tag{2.13}$$

This means that the evolution of the energy density of a specific constituent of the Universe is determined by its equation-of-state parameter w_i. A larger w_i leads to a faster dilution and therefore different constituents will dominate the evolution of the scale factor at different cosmic times.

The definitions in equations (2.9) and (2.12) imply a constant equation-of-state parameter for the dark energy density, $w_{DE} = w_\Lambda = -1$. A standard parametrization to test deviations from a constant $w_{DE} = -1$ is

$$w_{DE}(a) = w_0 + w_a(1 - a). \tag{2.14}$$

2.1.2 The ΛCDM Universe

There are four constituents with a specific w_i in our Universe that can contribute to the total energy content, $\Omega_\mathrm{m} + \Omega_\mathrm{r} + \Omega_\Lambda + \Omega_K = 1$. The constraints on the energy densities that are specified in the following come from the base ΛCDM analysis from *Planck* Collaboration et al. (2020) if not otherwise stated. The *Planck* results were obtained from the analysis of the cosmic microwave background (CMB), consisting of photons that decoupled from the photon-baryon fluid in the early Universe, and representing the most precise measurements at present. The different constituents are:

- *Matter*: Only 5% of the total energy content is in the form of ordinary matter, which encompasses the non-relativistic particles of the standard particle physics model and is dubbed *baryonic matter* ($\Omega_\mathrm{b} = 0.0493 \pm 0.0022$). From CMB measurements and a wide range of other observations from galactic to cosmological scales it is known that there must be an additional matter component which predominantly interacts gravitationally, namely dark matter. More specifically, this component is referred to as *cold dark matter* (CDM) because it can be modelled as a collisionless fluid with a small velocity dispersion, and has a energy density $\Omega_\mathrm{c} = 0.2645 \pm 0.0033$. Both matter species together make up around 30% of the total energy density ($\Omega_\mathrm{m} = 0.3153 \pm 0.0073$). On large scales matter can be modelled as dust with $w_\mathrm{m} = 0$ and hence $\rho_\mathrm{m} \propto a^{-3}$.

- *Radiation*: The contribution of radiation to the total energy density today is negligible, since it can be modelled as fluid with $w_\mathrm{r} = 1/3$ and the energy density decreases rapidly with $\rho_\mathrm{r} \propto a^{-4}$. The photon density can be estimated from the temperature of the cosmic microwave background $T_0 = 2.726 \pm 0.001\,\mathrm{K}$ (Fixsen, 2009), such that $\Omega_\gamma = 5.54 \cdot 10^{-5}$. It should be noted that neutrinos also behave as radiation as long as they are relativistic, which is true for the early Universe before they transition from relativistic to non-relativistic.

- *Dark Energy*: There is strong observational evidence for the accelerated expansion of the Universe. This requires an additional energy component with a negative pressure taking up $\sim 70\%$ of the total energy content. This component can be interpreted as the cosmological constant Λ, which corresponds to $w_\Lambda = -1$ with an exponentially increasing scale factor, $a(t) \propto \exp(Ht)$. This means that Ω_Λ is the dominant component in the present-day Universe. The current constraints from the base ΛCDM analysis from the *Planck* collaboration are $\Omega_\Lambda = 0.6847 \pm 0.0073$ and $H_0 = 67.36 \pm 0.0073$. Late-time measurements in the local Universe tend to find higher values of H_0, pointing to a tension of 4σ to 6σ (Verde et al., 2019) and it is an open question of actively ongoing research whether there is a Hubble tension between measurements based on the early and the late-time Universe. Current analyses which allow deviations of w_DE from -1, such as the parametrization in equation (2.14), find values consistent with the cosmological constant model.

- *Curvature*: Current measurements of the spatial curvature indicate that the Universe is extremely close to flat. For example, the *Planck* Collaboration et al. (2020) find $\Omega_K = 0.0007 \pm 0.0019$ when extending the baseline model by a free Ω_K and combining the information from *Planck* power spectra, *Planck* lensing and the analysis of baryon acoustic oscillations, which will be introduced in Section 2.4.

Together these components define the standard cosmological model - the flat ΛCDM model - which is the simplest model consistent with cosmological measurements, and their corresponding evolution under the hypothesis of homogeneity and isotropy.

2.1.3 Distances in the expanding Universe

The distance in an expanding universe has no unique definition, and a further challenge is that it cannot be measured directly but has to be inferred from observed redshifts and angles. The following definitions of cosmological distances are crucial in order to understand the origin of distortions in large-scale structure measurements that will be described in Section 2.8.

The *cosmological redshift*, z, results from the increase of the wavelength, λ, of the light due to the cosmic expansion between the time of its emission at a distant source and its observation,

$$1 + z = \frac{\lambda_\mathrm{obs}}{\lambda_\mathrm{em}} = \frac{1}{a(t)}, \tag{2.15}$$

where the scale factor at the time of observation is set to today's value $a(t_0) = 1$. For cosmological observations, the redshift is a useful quantity, since it can be used to infer radial distances and it is directly related to the scale factor and the lookback time that elapsed since the light was emitted at the source.

The metric in equation (2.2) was defined in *comoving* coordinates, where the effect of the expansion is factored out, and is common in cosmology to express measurements, simulations or theoretical models in those coordinates. For a Euclidean Universe, the relation between the *comoving distance*, D_c, to an object and its redshift is

$$D_\mathrm{c}(z) = \int_z^0 \frac{\mathrm{d}z'}{H(z')}, \tag{2.16}$$

and $H(z) = \sqrt{\Omega_\mathrm{r}(1+z)^4 + \Omega_\mathrm{m}(1+z)^3 + \Omega_\Lambda}$ based on the previous two sections.

The physical distance of an object at constant comoving coordinates changes with time due to the expansion. This distance is called *proper distance*, D_p, and for a Euclidean Universe it is related to the comoving distance by

$$D_\mathrm{p}(t) = a(t) D_\mathrm{c}. \tag{2.17}$$

While the proper distance is not directly measurable in observations, it is possible to instead determine the so-called *angular diameter distance*, D_A. For an object of physical size, δl, and measured angular diameter, $\delta\phi$, the angular diameter distance is

$$D_\mathrm{A}(z) \equiv \frac{\delta l}{\delta \phi}. \tag{2.18}$$

The *comoving angular distance* is given by $D_\mathrm{M}(z) = D_\mathrm{A}(z)(1+z)$, and for a flat Universe $D_\mathrm{M}(z) = D_\mathrm{c}(z)$.

In the following sections, we will also use the *volume-averaged distance*, D_V, which is a combination of the comoving angular distance and the radial comoving distance,

$$D_\mathrm{V}(z) = \left(D_\mathrm{M}^2 \frac{z}{H(z)} \right)^{1/3}. \tag{2.19}$$

2.2 Seeds of cosmic structure

The observations of the large-scale structure in the Universe show clear deviations from perfect homogeneity and isotropy that cannot be solely described by general relativity and the cosmological principle. Already in the cosmic microwave background (CMB) at $z \approx 1100$ small temperature fluctuations of the order $\Delta T/\bar{T} \approx 10^{-5}$ can be observed. The origin of these small fluctuations can be well explained by the mechanisms of inflation.

Inflation refers to a brief phase of rapid accelerated expansion that the Universe underwent during a very early epoch, even before the generation of known matter. During this phase the Universe expanded by at least a factor of e^{60} and quantum fluctuations were

blown up to cosmological scales, thereby becoming classical by decoherence. In the most simple inflationary scenario, this expansion is driven by a single scalar field, $\varphi(t, \boldsymbol{x})$, called inflaton. Quantum fluctuations of the field lead to slightly different values in different regions, $\boldsymbol{x}$, at a given time, t,

$$\varphi(t, \boldsymbol{x}) = \bar{\varphi}(t) + \delta\varphi(t, \boldsymbol{x}), \tag{2.20}$$

where $\bar{\varphi}(t)$ is the mean value of the field. While, on average, the fluctuations $\delta\varphi$ in the different regions of the field add up to give a zero mean value, the variance, i.e. the average of the square of the fluctuations is non-zero. Since inflation ends at a determined value of the field, φ_{end}, there are patches in the Universe where inflation lasts slightly shorter or longer than the average due to these inflaton fluctuations. The final density of these patches depends on the time they have experienced inflation, and therefore the time fluctuations propagate into *density fluctuations*, $\delta = \delta\rho/\bar{\rho}$.[3]

The further growth of the density fluctuations is determined by the comoving Hubble scale, $(aH)^{-1}$. During inflation, the Hubble scale decreases due to the accelerated expansion and the fluctuations loose causal contact. The fluctuations remain 'frozen-in' outside the causal horizon, also referred to as Hubble horizon, until the Hubble scale increases again in the subsequent cosmic epochs. More and more fluctuations re-enter the Hubble horizon and can finally grow.

Inflation does not only provide a theoretical explanation for the origin of density fluctuations that evolve into the cosmic structure observed today, but also solves several other cosmological problems. For example, it explains why the geometry of the Universe today is so close to flat, and why the regions of the CMB that are not in causal contact with each other have almost the same temperature. There has been no direct observational evidence of inflation yet, but a few very useful predictions have been derived from it and tested thoroughly in observations.

The most important prediction for the following study is that the distribution of these density fluctuations is *Gaussian* due to their quantum mechanical origin and is characterized by a single power spectrum $P(k) = 2\sigma_k^2$, where σ_k^2 is the variance of the distribution in Fourier space. This primordial power spectrum is nearly scale-invariant and can be described as a power law with *spectral index*,

$$n_{\mathrm{s}} \equiv \frac{\mathrm{d}\ln P(k)}{\mathrm{d}\ln k}. \tag{2.21}$$

Furthermore, the density fluctuations are predicted to be adiabatic, implying that the number densities of the different species in the Universe are changed by the same factor under expansion or compression. The CMB temperature fluctuations then can be associated to radiation density fluctuations that in turn are linked to the matter density fluctuations,

$$\frac{\delta T}{\bar{T}} \propto \frac{\delta\rho_\gamma}{\bar{\rho}_\gamma} \propto \frac{\delta\rho_{\mathrm{m}}}{\bar{\rho}_{\mathrm{m}}}. \tag{2.22}$$

[3]This is the shortened and simplified picture. The full physical description is based on cosmological perturbation theory, including metric and density perturbations and taking into account gauge freedom i.e. freedom of coordinate choice. The lengthy in-depth treatment goes beyond the scope of this section.

These are the matter density fluctuations that originate from inflation and, during the subsequent epochs, grow into the large-scale structure observed today.

2.3 Gaussian random fields and two-point statistics

The study of the cosmic large-scale structure is based on the *matter density fluctuation field*, $\delta(\boldsymbol{x})$,

$$\delta(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x}) - \bar{\rho}}{\bar{\rho}}, \tag{2.23}$$

where $\bar{\rho}$ is the mean density and $\delta(\boldsymbol{x})$ is also referred to as the density contrast.

Due to the quantum mechanical origin of the density fluctuations, the density field at position $\boldsymbol{x}$ is a realization of a Gaussian random process. The probability of the density field taking some value $\delta(\boldsymbol{x})$ in a given interval interval, $P(\delta(\boldsymbol{x}))\,\mathrm{d}\delta$, is described by the Gaussian probability distribution function with zero mean and variance, σ^2, given by

$$P(\delta(\boldsymbol{x})) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\delta^2(\boldsymbol{x})}{2\sigma^2}\right). \tag{2.24}$$

The cosmic density field at any number of locations $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ forms a *Gaussian random field* with N random variables $\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2), ..., \delta(\boldsymbol{x}_N)$ drawn from the joint probability distribution function that is the multivariate Gaussian distribution

$$P(\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2), ..., \delta(\boldsymbol{x}_N)) = \frac{1}{(2\pi|\xi|)^{N/2}} \exp\left(-\frac{1}{2}\sum_{i,j=1}^{N} \delta(\boldsymbol{x}_i)(\xi_{ij})^{-1}\delta(\boldsymbol{x}_j)\right). \tag{2.25}$$

Here ξ is the covariance or the *two-point correlation function* of the density fluctuations. The variance $\xi_{ii} = \langle\delta(\boldsymbol{x}_i)\delta(\boldsymbol{x}_i)\rangle = \sigma^2$ is the same as in the single-variate case of equation (2.24). It can easily be shown that all odd moments of the probability density function vanish, i.e. $\langle\delta(\boldsymbol{x}_i)\delta(\boldsymbol{x}_j)\delta(\boldsymbol{x}_k)\rangle = 0$, and thus the density fluctuation field is completely specified by its two-point correlation function. Since the density field is homogeneous and isotropic, the two-point correlation function only depends on the distance between two positions, $r = |\boldsymbol{r}| = |\boldsymbol{x}_i - \boldsymbol{x}_j|$,

$$\xi(r) = \langle\delta(\boldsymbol{x})\delta(\boldsymbol{x} + \boldsymbol{r})\rangle. \tag{2.26}$$

The two-point correlation function of a discrete set of points can also be understood as the excess probability of finding pairs at a separation r compared to a homogeneous distribution of points. The probability of finding pairs in two volume elements $\mathrm{d}V_1$ and $\mathrm{d}V_2$ separated by a distance r is given by

$$\mathrm{d}P_{12} = \langle\rho_1\rho_2\rangle\,\mathrm{d}V_1\,\mathrm{d}V_2 \tag{2.27}$$

$$= \bar{\rho}^2(1 + \xi(r))\,\mathrm{d}V_1\,\mathrm{d}V_2. \tag{2.28}$$

It is often convenient to work in Fourier space, since the Fourier modes of the density fluctuation field $\delta(\boldsymbol{k})$ evolve independently, as long as their amplitude remains small $|\delta| \ll 1$. Analogous to the correlation function in configuration space, the *power spectrum* is defined as the covariance of the density fluctuations in Fourier space

$$\langle \delta(\boldsymbol{k}) \delta^*(\boldsymbol{k}') \rangle = (2\pi)^3 \delta_{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}') P(\boldsymbol{k}), \tag{2.29}$$

where δ_{D} denotes the 3D Dirac delta function. The power spectrum is the Fourier transform of the correlation function and vice versa,

$$P(\boldsymbol{k}) = \int \xi(\boldsymbol{r})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}^3 r, \tag{2.30}$$

$$\xi(\boldsymbol{r}) = \frac{1}{(2\pi)^3} \int P(\boldsymbol{k})\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}^3 k. \tag{2.31}$$

The Fourier modes of an isotropic and homogeneous density field depend only on the absolute wavenumber $k = |\boldsymbol{k}|$. The relation between correlation function $\xi(r)$ and power spectrum $P(k)$ can then be expressed as

$$\xi(r) = \frac{1}{2\pi^2} \int P(k) j_0(kr) k^2\, \mathrm{d}k, \tag{2.32}$$

where j_0 is the spherical Bessel function of the first kind with $j_0(x) = \sin(x)/x$.

2.4 The linear evolution of density fluctuations

2.4.1 Equations of motion and growth rate

The small matter density fluctuations that originate from inflation are amplified by gravity and finally form the observed large-scale structure.

Right after inflation, all density fluctuations are still outside the horizon and cannot grow, as they are not causally connected. During the subsequent radiation and matter dominated epochs, however, more density fluctuations re-enter the Hubble horizon, and start to evolve. The evolution of the density fluctuations can be derived from the linearised Einstein equations using general relativistic perturbation theory. Since the full treatment is outside the scope of this work, this section focuses on the evolution of matter density fluctuations on scales much smaller than the Hubble horizon.

On these scales, the equations of motion of matter can be in principle described by those of Newtonian gravity, but taking into account the background expansion for the particle positions and momenta and for the gravitational potential. Based on this notion, the equations of motion for the density contrast, $\delta(\boldsymbol{x}, \tau)$, of the pressureless and non-relativistic cold dark matter and its comoving peculiar velocity field, $\boldsymbol{v}(\boldsymbol{x}, \tau)$, are the *Poisson equation*,

$$\nabla^2 \Phi(\boldsymbol{x}, \tau) = \frac{3}{2} \Omega_m(\tau) \mathcal{H}^2(\tau) \delta(\boldsymbol{x}, \tau), \tag{2.33}$$

the *continuity equation*,

$$\frac{\partial \delta(\boldsymbol{x},\tau)}{\partial \tau} + \nabla \cdot \{[1+\delta(\boldsymbol{x},\tau)]\boldsymbol{v}(\boldsymbol{x},\tau)\} = 0, \tag{2.34}$$

and the *Euler equation*,

$$\frac{\partial \boldsymbol{v}(\boldsymbol{x},\tau)}{\partial \tau} + \mathcal{H}(\tau)\boldsymbol{v}(\boldsymbol{x},\tau) + \boldsymbol{v}(\boldsymbol{x},\tau)\cdot\nabla\boldsymbol{v}(\boldsymbol{x},\tau) = -\nabla\Phi(\boldsymbol{x},\tau), \tag{2.35}$$

where τ is the conformal time with $\mathrm{d}t = a(\tau)\,\mathrm{d}\tau$, $\mathcal{H} = Ha$ is the conformal expansion rate and $\Phi(\boldsymbol{x},\tau)$ the gravitational potential sourced by the density field $\rho(\boldsymbol{x},\tau)$.

In the linear regime where $|\delta| \ll 1$ and $|\boldsymbol{v}| \ll 1$, the equations (2.34)-(2.35) can be linearised,

$$\frac{\partial \delta(\boldsymbol{x},\tau)}{\partial \tau} + \theta(\boldsymbol{x},\tau) = 0, \tag{2.36}$$

$$\frac{\partial \boldsymbol{v}(\boldsymbol{x},\tau)}{\partial \tau} + \mathcal{H}(\tau)\boldsymbol{v}(\boldsymbol{x},\tau) = -\nabla\Phi(\boldsymbol{x},\tau), \tag{2.37}$$

where $\theta(\boldsymbol{x},\tau) \equiv \nabla\cdot\boldsymbol{v}(\boldsymbol{x},\tau)$ is the velocity divergence. Combining these two equations and inserting equation (2.33), one finds a second order differential equation for $\delta(\boldsymbol{x},\tau)$ alone,

$$\frac{\partial^2 \delta(\boldsymbol{x},\tau)}{\partial^2 \tau} + \mathcal{H}(\tau)\frac{\partial \delta(\boldsymbol{x},\tau)}{\partial \tau} - \frac{3}{2}\Omega_m(\tau)\mathcal{H}^2(\tau)\delta(\boldsymbol{x},\tau) = 0. \tag{2.38}$$

From the growing mode solution of this equation, the linear growth of matter density fluctuations is obtained as

$$\delta(\boldsymbol{x},\tau) = D_1(\tau)\delta(\boldsymbol{x}), \tag{2.39}$$

where the *linear growth factor* D_1, as a function of the scale factor a, is given by

$$D_1(a) = \frac{5\Omega_\mathrm{m}}{2}\frac{H(a)}{H_0}\int_0^a \frac{\mathrm{d}a_1}{(a_1 H(a_1)/H_0)^3}. \tag{2.40}$$

2.4.2 The linear power spectrum

The linear matter density field can be best studied by means of the power spectrum $P(k)$ of equation (2.29), since the Fourier modes $\delta(k)$ of the density field evolve independently if $\delta(k) \ll 1$.

The evolution of the linear power spectrum with time or scale factor can be described by the squared growth factor $D_1(a)^2$ from equation (2.40). The shape of the power spectrum depends on the primordial power spectrum from equation (2.21), which is characterized by the scale factor n_s. During the radiation and matter-dominated epoch, the shape of the primordial power spectrum is modified as the density fluctuations grow. The change in shape is determined by the horizon entry of each density mode $\delta(k)$, which depends on the wavelength k, and is encoded in the transfer function, $T(k)$. Calculating $T(k)$ is

commonly done by using the publicly available, fast codes CAMB (Lewis et al., 2000) or CLASS (Blas et al., 2011), which accurately integrate the underlying system of coupled Boltzmann equations in a perturbed FLRW metric numerically.

Putting all these ingredients together, the full expression for the linear power spectrum, $P_{\rm lin}$, is given by

$$P_{\rm lin}(k,a) = A_{\rm s}\left[\frac{D_1(a)}{D_1(a_0)}\right]^2 T^2(k)\left(\frac{k}{k_0}\right)^{n_{\rm s}}, \tag{2.41}$$

where the spectral amplitude A_s is a free parameter that usually sets the global amplitude of the fluctuations at $z = 0$ for an arbitrary pivot scale, k_0. The linear matter power spectrum results in a shape similar to that shown in the left panel of Fig. 2.1. It was computed using the Boltzmann solver CLASS at a redshift $z = 0.57$ with an input cosmology matching the one of the Minerva simulations, which will be described in Section 3.4. The linear two-point correlation function is the Fourier transform of the linear power spectrum, as defined in equation (2.32), and shown on the right panel of the figure.

An interesting feature is visible for the power spectrum as well as for the correlation function: a series of wiggles modulate the amplitude of the power spectrum, whereas the correlation function exhibits a broad peak at a scale of $\sim$155 Mpc. These are the imprints of the *baryon acoustic oscillations (BAO)*. To explain the physical origin of BAO, we have to go back to the early Universe. Before electrons and nuclei form atoms, baryons and photons are tightly coupled due to Thomson scattering and form the so-called baryon-photon plasma. During this phase, density fluctuations in the dark matter component can already grow at a logarithmic rate. In the baryon-photon plasma, however, the radiation pressure prevents the growth of density fluctuations and the interplay between gravity and radiation pressure produces acoustic waves. When the Universe becomes cold enough at a temperature of roughly $T \sim 3000$ K, nuclei and electrons form atoms during the so-called recombination. As a result, the mean free path of photons becomes larger and they can decouple from the baryons, forming the previously mentioned CMB. After the baryons are released from the photons, they fall into the potential wells of the dark matter distribution. From that point on, it is possible study the evolution of the total matter density fluctuations with the same physical description. Although the contribution of the baryonic matter component to the gravitational potential is small, it has a non-negligible impact on the dark matter density fluctuations. The imprint of BAO in the baryonic matter distribution leads to an excess of clustering in the dark matter distribution at the scale of $r_{\rm s}(z_{\rm d})$, which is the *sound horizon* at the drag redshift $z_{\rm d}$ corresponding to the time when the baryons were released from the photons. The broad peak in the correlation function and the oscillatory pattern in the power spectrum reflect the size of the sound horizon.

To complete the description of the linear matter power spectrum, a further parameter is defined that is commonly used to characterize its amplitude. The variance of the smoothed linear density fluctuations within a sphere of radius R is given by

$$\sigma_R^2(a) = \int \frac{{\rm d}^3k}{(2\pi)^3}\hat{W}_R^2(k)P_{\rm lin}(k,a), \tag{2.42}$$

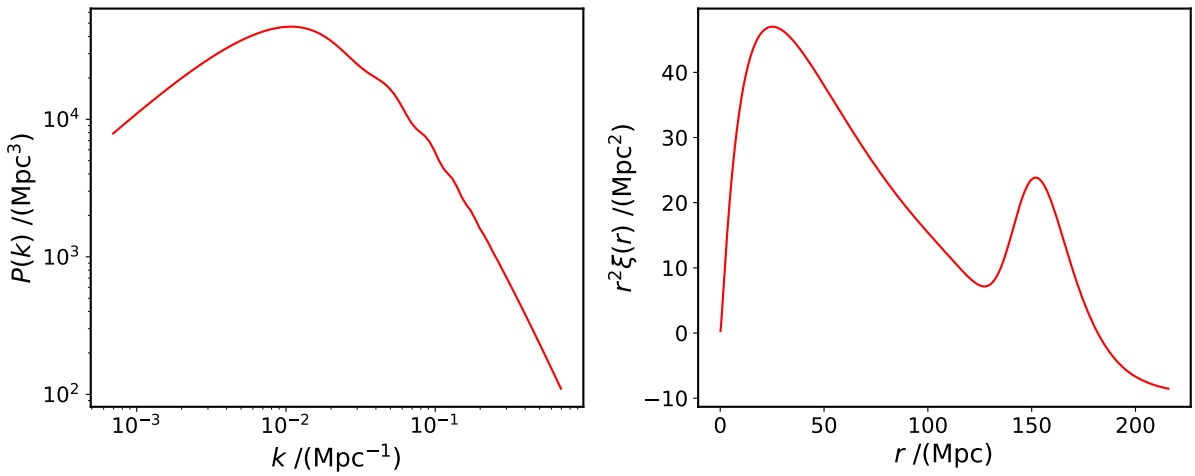

Figure 2.1: The linear matter power spectrum (left panel) and two-point correlation function (right panel) at redshift $z = 0.57$ for the ΛCDM cosmology of the Minerva simulations, which are described in Section 3.4. The linear theory predictions were computed using the Python binding for the Boltzmann solver CLASS of the NBODYKIT package (Hand et al., 2018). The BAO signature is visible as a series of wiggles in the power spectrum and as the peak of the correlation function at $r \sim 155\,\mathrm{Mpc}$.

where $\hat{W}_R$ is the Fourier transform of the spherical top hat used for smoothing. Traditionally, $\sigma_8 = \sigma_8(a = 1)$ at a scale of $R = 8h^{-1}\,\mathrm{Mpc}$ has been used in large-scale structure analyses. In order to avoid a dependence on h, a very recent work by Sánchez (2020) proposes to use the reference scale $R = 12\,\mathrm{Mpc}$ that has a similar value as σ_8 for $h \simeq 0.67$ (more details are given in Section 2.8.2).

2.5 Non-linear gravitational evolution

The linear matter power spectrum and correlation function discussed in the previous section give only a good theoretical description if the density fluctuations are very small, $|\delta| \ll 1$. This is the case for the early Universe and very large scales. For the late Universe, at the redshifts and scales that are relevant for the following work, linear theory is not valid anymore. To illustrate this, Fig. 2.2 shows the measurement of the dark matter two-point correlation function from the Minerva simulations, which will be described in Section 3.4, at a redshift of $z = 0.57$ in comparison to the corresponding linear prediction. The non-linear growth of structure changes the shape of the correlation function, most notably as a damping and broadening of the BAO feature.

The most common approaches to model the non-linear gravitational evolution of density fluctuations are based on perturbation theory. This section gives a brief summary of standard cosmological perturbation theory (SPT), which is the foundation of the model dubbed gRPT used in the following work, and describes the idea of gRPT. The concepts

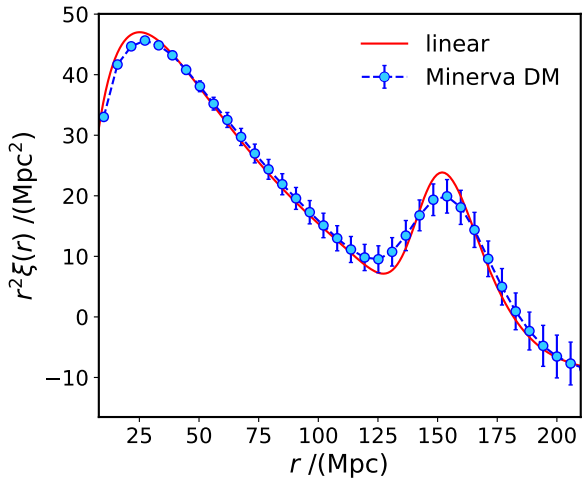

Figure 2.2: The dark matter correlation function measured from the Minerva simulations (see Section 3.4) at redshift $z = 0.57$ compared against the corresponding linear theory prediction, which is described in the previous Section 2.4. The broadening of the BAO feature in the measurement is an obvious signature of the non-linear growth of the density fluctuations.

of Lagrangian perturbation theory and the spherical collapse of dark matter halos are also introduced, since they will be relevant for the approximate methods of the comparison project in Chapter 3. The subsections 2.5.1 and 2.5.3 are largely based on Bernardeau et al. (2002). For a more detailed treatment the reader is referred to this review.

2.5.1 Standard Perturbation Theory

The main idea of perturbation theory is to model the evolution of density fluctuations beyond the linear regime by expanding the density contrast and velocity fields as series of powers of the linear solutions from Section 2.4.1. For this purpose, it is convenient to work in Fourier space, since the derivatives in the Poisson, continuity and Euler equations (2.33)-(2.35) become multiplications and the coupling of different modes can be studied.

The equations of motion (2.33)-(2.35) can be written as two coupled equations in Fourier space,

$$\frac{\partial\delta(\boldsymbol{k},\tau)}{\partial\tau} + \theta(\boldsymbol{k},\tau) = -\int \frac{\mathrm{d}^3k_1}{(2\pi)^3}\frac{\mathrm{d}^3k_2}{(2\pi)^3}(2\pi)^3\delta_\mathrm{D}(\boldsymbol{k}-\boldsymbol{k}_{12})\alpha(\boldsymbol{k}_1\boldsymbol{k}_2)\theta(\boldsymbol{k}_1,\tau)\delta(\boldsymbol{k}_2,\tau), \tag{2.43}$$

$$\frac{\partial\theta(\boldsymbol{k},\tau)}{\partial\tau} + \mathcal{H}(\tau)\theta(\boldsymbol{k},\tau) + \frac{3}{2}\Omega_m(\tau)\mathcal{H}^2(\tau)\delta(\boldsymbol{k},\tau) = \\ -\int \frac{\mathrm{d}^3k_1}{(2\pi)^3}\frac{\mathrm{d}^3k_2}{(2\pi)^3}(2\pi)^3\delta_\mathrm{D}(\boldsymbol{k}-\boldsymbol{k}_{12})\beta(\boldsymbol{k}_1\boldsymbol{k}_2)\theta(\boldsymbol{k}_1,\tau)\theta(\boldsymbol{k}_2,\tau), \tag{2.44}$$

where $\boldsymbol{k}_{12} = \boldsymbol{k}_1 + \boldsymbol{k}_2$. The kernel functions,

$$\alpha(\boldsymbol{k}_1, \boldsymbol{k}_2) = \frac{\boldsymbol{k}_{12} \cdot \boldsymbol{k}_1}{k_1^2}, \qquad \beta(\boldsymbol{k}_1, \boldsymbol{k}_2) = \frac{k_{12}^2 (\boldsymbol{k}_1 \cdot \boldsymbol{k}_2)}{2 k_1^2 k_2^2}, \tag{2.45}$$

describe the coupling between different modes. Unlike the linear evolution where each mode evolves independently, the non-linear evolution introduces couplings between modes with different wavelengths k.

The perturbation theory approach relies on the assumption that the non-linear density and velocity field can be expanded as

$$\delta(\boldsymbol{k}, \tau) = \sum_{n=1}^{\infty} \delta^{(n)}(\boldsymbol{k}, \tau), \qquad \theta(\boldsymbol{k}, \tau) = \sum_{n=1}^{\infty} \theta^{(n)}(\boldsymbol{k}, \tau). \tag{2.46}$$

At first order, $\delta^{(1)}$ corresponds to the linear density contrast from equation (2.39). The higher-order terms $\delta^{(n)}$ are proportional to the corresponding power of the linear density contrast, $\delta^{(n)} \propto (\delta_{\rm lin})^n$.

A general solution for equations (2.43) and (2.44) at $n-$th order can be obtained from the ansatz

$$\delta^{(n)}(\boldsymbol{k}, \tau) = \int \frac{\mathrm{d}^3 k_1 \cdots \mathrm{d}^3 k_n}{(2\pi)^{2n-3}} \delta_{\rm D} \left(\boldsymbol{k} - \sum_{i=1}^{n} \boldsymbol{k}_i \right) F_n(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n, \tau) \delta^{(1)}(\boldsymbol{k}_1, \tau) \cdots \delta^{(1)}(\boldsymbol{k}_n, \tau), \tag{2.47}$$

$$\theta^{(n)}(\boldsymbol{k}, \tau) = -\mathcal{H}(\tau) f(\tau) \int \frac{\mathrm{d}^3 k_1 \cdots \mathrm{d}^3 k_n}{(2\pi)^{2n-3}} \delta_{\rm D} \left(\boldsymbol{k} - \sum_{i=1}^{n} \boldsymbol{k}_i \right) \times G_n(\boldsymbol{k}_1, ..., \boldsymbol{k}_n, \tau) \delta^{(1)}(\boldsymbol{k}_1, \tau) \cdots \delta^{(1)}(\boldsymbol{k}_n, \tau), \tag{2.48}$$

where F_n and G_n are the perturbation theory kernels (cf. equations (12.40) in Dodelson & Schmidt (2020) and (41)-(42) in Bernardeau et al. (2002)). At first order, $F_1 = G_1 = 1$ and the linear theory solution without mode-coupling is recovered. The higher-order kernels F_n and G_n can be computed iteratively using the recursion relations from Bernardeau et al. (2002).

Analogous to the density field, also the power spectrum can be expanded in a series of $\delta(\boldsymbol{k}, \tau)$,

$$\begin{aligned} \langle \delta(\boldsymbol{k}, \tau) \delta(\boldsymbol{k'}, \tau) \rangle = \langle \delta^{(1)}(\boldsymbol{k}, \tau) \delta^{(1)}(\boldsymbol{k'}, \tau) \rangle + \langle \delta^{(1)}(\boldsymbol{k}, \tau) \delta^{(3)}(\boldsymbol{k'}, \tau) \rangle \\ + \langle \delta^{(3)}(\boldsymbol{k}, \tau) \delta^{(1)}(\boldsymbol{k'}, \tau) \rangle + \langle \delta^{(2)}(\boldsymbol{k}, \tau) \delta^{(2)}(\boldsymbol{k'}, \tau) \rangle + \ldots, \end{aligned} \tag{2.49}$$

where the odd moments vanish due to the Gaussian initial conditions. This equation can be rewritten as a sum of the linear power spectrum and the so-called 'loop' corrections,

$$P(k, a) = P_{\rm lin}(k, a) + P^{(22)}(k, a) + 2P^{(13)}(k, a) + \ldots. \tag{2.50}$$

The linear order contribution, $P_{\rm lin}(k,a) = D_1^2(a)/D_1^2(a=1)P_{\rm lin}(k)$, is called 'tree-level'. The different combinations that add up to the next even power of the density field are called 'loop' corrections. The contributions to the one-loop correction of the power spectrum are

$$P^{(22)}(k,a) = 2\int \frac{{\rm d}^3q}{(2\pi)^3}[F_2(\boldsymbol{q},(\boldsymbol{k}-\boldsymbol{q}))]^2 P_{\rm lin}(q,a)P_{\rm lin}(|\boldsymbol{k}-\boldsymbol{q}|,a), \tag{2.51}$$

$$P^{(13)}(k,a) = 3P_{\rm lin}(k,a)\int \frac{{\rm d}^3q}{(2\pi)^3}[F_3(\boldsymbol{q},-\boldsymbol{q},\boldsymbol{k})]P_{\rm lin}(q,a), \tag{2.52}$$

where F_2 and F_3 are the corresponding perturbation theory kernels. In the same way, higher-loop corrections can be found.

2.5.2 Renormalized Perturbation Theory

The expansion series of SPT cannot be efficiently used for predicting the non-linear power spectrum at the redshifts and the small scales required by state-of-the-art galaxy clustering analysis. The reason is that the SPT loop corrections can each have negative and positive contributions that can cancel each other out, and higher-order loop corrections can have the same magnitudes as lower-order loop corrections. This makes it difficult to truncate the expansion at a specific order in the expansion.

Renormalized Perturbation Theory (RPT, Crocce & Scoccimarro, 2006) improves the convergence by re-organizing the terms of the perturbative expansion into two main contributions as

$$P(k,a) = P_{\rm lin}(k)G(k,a)^2 + P_{\rm mc}(k,a). \tag{2.53}$$

The first term sums up all orders of the perturbative expansion proportional to the linear power spectrum $P_{\rm lin}(k)$. The renormalized propagator $G(k,a)$ measures how much power is directly linked to the linear initial conditions as a function of scale and redshift. At 0-th order or at very large scales (corresponding to small k), $G(k,a)$ corresponds to the linear growth factor $D_1(a)$. For increasing k it decays approximately as a Gaussian with constant variance σ_v^2. The second term $P_{\rm mc}$ sums up the power from the different mode couplings that are ordered according to the number of initial modes coupled. At the 1-loop correction, $P_{\rm mc}(k,a)$ corresponds to $P^{(22)}(k,a)$ from equation (2.51).

The advantage of the reorganized expansion of RPT is that all loop terms are positive, the successive terms dominate the expansion series at increasingly smaller scales, and their amplitudes become successively smaller. This makes it easier to find the specific order to truncate the expansion series for a given accuracy.

A further improvement in the resummation of the mode-coupling terms can be made by imposing Galilean-invariance in the equations of motion and in that way making it consistent with the resummation of the propagator $G(k,a)$ (Crocce et al., in prep.). This approach, dubbed Galilean-invariant renormalized perturbation theory (gRPT), leads to an even better convergence of the expansion series. Already the 1-loop gRPT expansion predicts the non-linear power spectrum up to a scale $k_{\rm max}$ with an accuracy suitable for

high-precision galaxy clustering analyses. RPT would require a much larger number of terms to achieve the same accuracy.

The gRPT modelling has been successfully applied for several cosmological analyses: the Fourier-space analysis of the SDSS BOSS DR12 galaxy samples with a $k_{\rm max} \sim 0.25\,h\,{\rm Mpc}^{-1}$ by Grieb et al. (2017), the configuration-space analyses of the same samples by Sánchez et al. (2017) and by Salazar-Albornoz et al. (2017) with a minimum distance scale of $s_{\rm min} = 20\,h^{-1}{\rm Mpc}$, and the configuration-space analysis of the SDSS eBOSS DR14 quasar sample with the same $s_{\rm min}$ by Hou et al. (2018). In the present work, all predictions of the non-linear power spectrum are based on gRPT.

2.5.3 Lagrangian perturbation theory

A different approach to model the evolved density field at a given comoving spatial position $\boldsymbol{x}$ is to follow the trajectories of particles. The initial Lagrangian position $\boldsymbol{q}$ is mapped through the displacement field, $\boldsymbol{\Psi}(\boldsymbol{q},\tau)$, to the corresponding Eulerian position $\boldsymbol{x}$ at time τ,

$$\boldsymbol{x}(\boldsymbol{q},\tau) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q},\tau). \tag{2.54}$$

The equation of motion for the particle trajectory $\boldsymbol{x}(\tau)$ in the expanding universe is

$$\frac{{\rm d}^2\boldsymbol{x}}{{\rm d}\tau^2} + \mathcal{H}(\tau)\frac{{\rm d}\boldsymbol{x}}{{\rm d}\tau} = -\nabla\Phi, \tag{2.55}$$

which corresponds to the linear Euler equation (2.37). Due to mass conservation, the density field in Lagrangian coordinates is related to the density field in Eulerian coordinates by

$$\bar{\rho}(\tau)\,{\rm d}^3q = \bar{\rho}(\tau)(1+\delta(\boldsymbol{x},\tau))\,{\rm d}^3x. \tag{2.56}$$

The Jacobian of the transformation between Eulerian and Lagarangian space is then given by

$$\mathcal{J}(\mathbf{q},\tau) = \left|\frac{{\rm d}^3x}{{\rm d}^3q}\right| = \frac{1}{1+\delta(\boldsymbol{x},\tau)} = \det\left(\delta_{ij} + \Psi_{i,j}\right), \tag{2.57}$$

where $\Psi_{i,j} \equiv \partial\Psi_i/\partial\boldsymbol{q}_j$.

Taking the divergence of equation (2.55) and inserting the Jacobian from (2.57) and the Poisson equation, the following equation can be derived,

$$\mathcal{J}(\boldsymbol{q},\tau)\nabla\cdot\left(\frac{{\rm d}^2\boldsymbol{\Psi}(\boldsymbol{q},\tau)}{{\rm d}\tau^2} + \mathcal{H}(\tau)\frac{{\rm d}\boldsymbol{\Psi}(\boldsymbol{q},\tau)}{{\rm d}\tau}\right) = \frac{3}{2}\mathcal{H}^2(\tau)\Omega_{\rm m}(\tau)\left(\mathcal{J}(\boldsymbol{q},\tau)-1\right), \tag{2.58}$$

where the gradient is still in Eulerian coordinates $\boldsymbol{x}$. This equation (2.58) describes the evolution of the displacement field $\boldsymbol{\Psi}$.

The central idea of Lagarangian perturbation theory is to solve this equation perturbatively by expanding $\boldsymbol{\Psi}(\mathbf{q},\tau)$ as

$$\boldsymbol{\Psi}(\boldsymbol{q},\tau) = \sum_{n=1}^{\infty}\boldsymbol{\Psi}^{(n)}(\boldsymbol{q},\tau), \tag{2.59}$$

and from this to obtain predictions for the evolved density field and corresponding power spectrum. The linear solution, called Zel'dovich Approximation, provides a simple and intuitive model to study structure formation. Second-order Lagrangian perturbation theory (2LPT) is often used to generate the initial density field for simulations at a high redshifts. Many approximate methods of gravitational evolution and structure formation, such as the methods that will be compared in this work in Chapter 3, are based on second-order or third-order Lagrangian perturbation theory (3LPT), or other approaches derived from Lagrangian perturbation theory. A limit for the validity of the Lagrangian approach is the orbit-crossing (or also called shell-crossing) that occurs when particles from different initial positions $\boldsymbol{q}$ get to the same Eulerian position $\boldsymbol{x}$. For the redshifts and scales relevant for the galaxy clustering analysis, Lagrangian perturbation theory cannot predict the non-linear matter power spectrum with sufficient accuracy.

2.5.4 Spherical collapse

A further, very simplistic model for the non-linear growth of structure is the gravitational collapse of spherically symmetric perturbations. The interesting aspect about this approach is that it provides an analytic solution for the formation of dark matter halos. The starting point is a spherical matter perturbation with a physical radius r and a slightly higher density, ρ_m, than the otherwise homogeneous universe with density $\bar{\rho}_m$. The corresponding matter density contrast is then given by

$$\delta(t) = \frac{M(< r(t))}{\bar{\rho}_m 4\pi r(t)^3/3} - 1, \tag{2.60}$$

where $M(< r(t))$ is the enclosed mass. This spherical region evolves like a FLRW universe with a higher density $\rho_m > \bar{\rho}_m$ and the evolution of its radius r is given by the Newtonian equation of motion extended with the repulsive effect from the expansion of the Universe,

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -G\frac{M(< r(t))}{r(t)^2} + \frac{8\pi G}{3}\rho_\Lambda r(t). \tag{2.61}$$

The spherical region initially expands due to the background Hubble expansion, but its self-gravitational attraction slows down the expansion leading eventually to collapse. If the density inside the spherical region is sufficiently high, such that $\Omega_\mathrm{m} \approx 1$, the expansion term can be neglected and a parametric solution can be found,

$$\begin{aligned} r &= r_\mathrm{max}(1 - \cos\theta), \\ t &= t_\mathrm{max}(\theta - \sin\theta), \end{aligned} \tag{2.62}$$

with the constants $r_\mathrm{max}^3 = GMt_\mathrm{max}^2$ and a phase parameter θ. The spherical region collapses at $\theta = 2\pi$ and the collapse time is

$$t_\mathrm{sc} = 2\pi t_\mathrm{max}. \tag{2.63}$$

The density at collapse is $\delta \to \infty$, but it is typically assumed that a real perturbation reaches virial equilibrium at this point and forms a bound *dark matter halo*. At a given time t, the critical initial density needed for collapse can be computed from this approach. Denser perturbations will have collapsed, while less dense ones have not. A more detailed description of spherical collapse can be found in Bertschinger (1985).

The spherical collapse model is the simplest treatment of the formation of collapsed structures, and this approach has been further improved and generalized. In particular, the homogeneous ellipsoidal collapse model provides a way to compute accurate collapse times for halo formation using triaxial perturbations and Lagarangian perturbation theory. For a detailed description the reader is referred to Monaco (1997).

2.6 N-body simulations

The previous section presented different ways of modelling the non-linear evolution of the dark matter density fluctuations. The second major pillar in the large-scale structure analysis are N-body simulations, which evolve the density field numerically. N-body simulations can precisely predict the density field up to the highly non-linear regime where arbitrary, overdense regions collapse under their own gravity and known PT models break down.

Cosmological N-body simulations have a wide range of applications. They are used for different aspects of the large-structure analysis such as covariance matrix estimation, which will be described in Section 3.3, for the generation of mock observations of real galaxy surveys, and to test theory predictions against simulations and vice versa. A complete description of N-body simulations would be a work on its own, here I only summarize the main ideas based on Bertschinger (1998), Dodelson & Schmidt (2020), Bernardeau et al. (2002) and the lecture notes by Blot (2020).

N-body simulations solve the gravitational dynamics of a system of N particles numerically. For cosmological N-body simulations, these particles do not represent physical particles, but are test particles for elements of the discretized phase-space of the cold dark matter density field. The particles are specified by their position, velocity and mass. Modern N-body techniques are able to simulate systems of order 10^9 "dark matter particles" with typically equal masses of $m_\mathrm{p} \approx 10^{10}\mathrm{M}_\odot$ in boxes with Gpc side lengths. The coordinates of the particles are comoving and periodic boundary conditions are employed, in order to create a representative volume of the expanding universe.

The Gaussian initial conditions are randomly drawn according to a linear input power spectrum that is computed with specific values for the cosmological parameters, most commonly using CAMB or CLASS. The initial positions of the particles are generated using 2LPT, which allows starting the simulation at a redshift of $z \leq 70$.

The gravitational force acting on the particles is computed by solving the Poisson equation (2.33) numerically at discrete time steps. Evaluating the force on each particle by summing directly over the contributions from all neighbours scales as N^2, and is therefore infeasible for the large number of particles needed for cosmological simulations.

There are two main alternatives to efficiently compute the gravitational forces: *tree methods* and *particle-mesh (PM)* approaches. Tree algorithms perform a hierarchical decomposition of the particle system according to their relative distances. The gravitational forces are expanded in multipoles and the contributions of long range interactions from distant regions are approximated by the lowest multipoles. In the particle-mesh method, a grid is superimposed to the particle distribution and the particle masses are interpolated to densities of grid cells. The Poisson equation is solved through a Fast Fourier Transform on the grid and the forces are interpolated back to the particle positions. The accuracy of this approach can be increased if the grid is adaptively refined in high density regions. Both types of algorithms require a smoothing of the force on very small scales, in order to avoid unphysical artefacts due to direct particle encounters, since the particles do not represent actual DM particles. The computational cost of these algorithms scales approximately as $N \log N$.

The N-body simulations considered in this work were performed with the massively parallelized GADGET code by Springel (2005), which is widely applied for cosmological studies. GADGET is based on a hybrid approach which uses a tree algorithm for the short-range and a PM algorithm for the long-range gravitational force.

The output of the simulations are so-called *snapshots* corresponding to the collection of the particle positions and velocities at specific redshifts, which usually lie in the range of $0 \leq z \leq 2$ for galaxy clustering analyses. The positions of the particles can be used to compute the statistics of the evolved dark matter density field, such as the power spectrum and correlation function, at the redshift of the snapshot.

Finally, one can also further post-process the snapshots by searching for gravitationally bound dark matter halos and populating the halos with synthetic galaxies. There are several algorithms to identify halos from simulations. A very popular one, which was first proposed by Davis et al. (1985), is the *Friends-of-Friends (FoF)* algorithm, which groups all particles that are separated by distance less than a given linking length into halos. A potential weak point of this simple approach is that the identified halos are not necessarily gravitationally bound structures. This problem can be alleviated by using more sophisticated versions of FoF algorithms such as SubFind (Springel et al., 2001) and ROCKSTAR (Behroozi et al., 2013), which will described for the corresponding halo catalogues in Chapters 3 and 4. For the generation of synthetic galaxy catalogues the halos can be populated using *halo occupation distribution (HOD)* models. More details on a HOD galaxy catalogue will be given in Section 4.4.1.

2.7 Beyond two-point statistics: Minkowski functionals

The non-linear gravitational evolution of the density contrast generates higher-order correlations leading to deviations from a Gaussian distribution. Consequently, these higher-order correlations contain valuable information on the growth of structure.

The most direct approach to access this information is to directly measure the higher-order correlations. For example, the third-order correlation in Fourier space corresponds to the *bispectrum*, $B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$, which is defined as

$$\langle \delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2)\delta(\boldsymbol{k}_3)\rangle = (2\pi)^3 \delta_D(\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3) B(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}_3), \tag{2.64}$$

and encompasses all possible triplets $(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}_3)$. Measuring all possible triplets is a complex task, as well as modelling the underlying theory predictions from perturbation theory. Present-day surveys allow for accurate analyses of the bispectrum and its Fourier transform, the three-point function (e.g. Gil-Marín et al., 2017; Slepian et al., 2017b,a; Pearson & Samushia, 2018).

A full characterization of the density field, however, would require measuring also the four-, five-point and ultimately an infinite hierarchy of N-point correlations. Measuring, modelling and analysing N-point functions of higher order than three-point exceeds our current possibilities. An alternative approach is to use statistics that encode compressed higher-order information complementary to two-point statistics. One of the most promising statistics to explore the non-Gaussian information in the matter density field is the set of Minkowski functionals.

2.7.1 Definition and properties

Minkowski functionals derive from the theory of convex sets in integral geometry and were introduced to large-scale structure analysis by Mecke et al. (1994). According to Hadwiger's Theorem (Hadwiger, 1957) in d-dimensional Euclidean space E the global morphological properties satisfying motional invariance, continuity and additivity of a suitable set $Q \subseteq E$ can be fully characterized by a linear combination of $d+1$ Minkowski functionals.

In this work, we consider the excursion sets of the three-dimensional density field that are obtained by applying a given density threshold, ρ_{th} (or equivalently δ_{th}). Points with a density $\rho(\boldsymbol{x}) > \rho_{\mathrm{th}}$ are considered to be inside the isodensity surfaces encompassing the excursion sets. For such a three-dimensional set, there are four Minkowski functionals:

(i) the *surface area* S,

(ii) the *volume* V enclosed by the surface,

(iii) the *integrated mean curvature* C of the surface,

$$C = \frac{1}{2} \oint\!\!\!\!\!\!\int_S \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \mathrm{d}S,$$

where R_1 and R_2 are the principal radii of curvature at a given point on the surface,

(iv) the *integrated Gaussian curvature* of the surface,

$$\chi = \frac{1}{2\pi} \oint\!\!\!\!\!\!\int_S \left(\frac{1}{R_1 R_2} \right) \mathrm{d}S, \tag{2.65}$$

also known as *Euler-characteristic*.

The Minkowski functionals do not only provide direct information about the geometry of the isodensity surface, but also contain topological information. The Euler characteristic is equivalent to a fundamental quantity in topology, the *genus*,

$$G = 1 - \chi/2. \tag{2.66}$$

The genus characterizes the topology of the isodensity surface by counting the number of holes and isolated regions,

$$G = 1 + \text{number of holes} - \text{number of isolated regions}.$$

A hole is commonly referred to as tunnel in large-scale structure. For example, the genus of a sphere or ellipsoid is $G = 0$, the genus of a torus with one handle is $G = 1$, and the genus of an eyeglasses frame is $G = 2$. Closed multiply-connected surfaces always have $G > 0$. The genus measures the connectivity of the isodensity surface and is invariant under continuous deformations such as stretching, compression or rotation. Therefore, it can offer interesting insights into the evolution of the density field, which are complementary to the information inferred from standard two-point statistics.

In large-scale structure analysis, we aim to study the global Minkowski functionals of the three-dimensional matter or galaxy density field by summing over the local Minkowski functionals of all excursion sets. It is often suitable to rescale the global MFs by the total volume considered, $V_{\rm tot}$ and work in terms of the Minkowski functional densities. The surface, curvature, and genus densities will be denoted by s, c, and g, respectively. The rescaled volume functional is also called the *volume-filling fraction*,

$$f_V = V/V_{\rm tot}. \tag{2.67}$$

A different approach to study Minkowski functionals of the three-dimensional galaxy distribution is used by the so-called Germ-Grain models. Theses models consider the Minkowski functionals from intersecting spheres inflated around the input galaxy sample (see, e.g. Mecke et al., 1994; Schmalzing et al., 1996; Kerscher, 2000, for a comprehensive overview). An interesting aspect of this definition is that the Minkowski functionals can be expressed as sums over integrals of the N-point correlation functions (Schmalzing et al., 1999b; Wiegand et al., 2014). However, the focus of this work lies on isodensity Minkowski functionals, since they are more directly linked to the underlying density field.

2.7.2 Theory predictions for a Gaussian density field

For Gaussian random fields, isodensity Minkowski functionals have known analytical predictions, which depend on the power spectrum of the distribution. According to Tomita's formula, the Minkowski functional densities for a Gaussian density field in three dimension

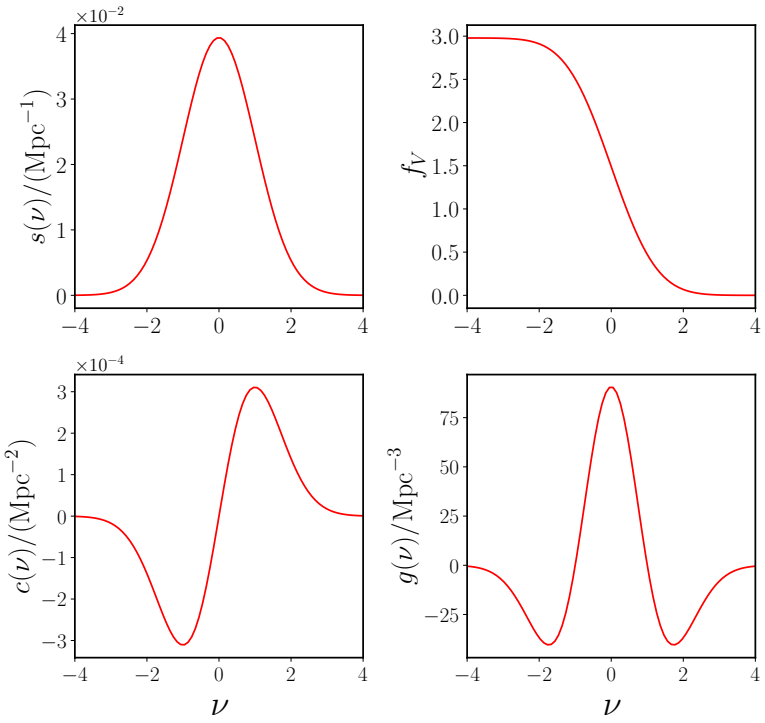

Figure 2.3: Theory predictions for the Minkowski functionals of a Gaussian density field with the same linear power spectrum as the Minerva simulations (see Section 3.4) as a function of the normalized density threshold ν at redshift $z = 0.57$.

are given by (Tomita, 1990; Schmalzing & Buchert, 1997; Matsubara, 2003):

$$f_V(\nu) = \frac{1}{2} - \frac{1}{2}\Phi\left(\frac{\nu}{\sqrt{2}}\right), \tag{2.68}$$

$$\text{where } \Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{d}t \exp\left(-t^2\right) \text{ denotes the error function,} \tag{2.69}$$

$$s(\nu) = \frac{2}{\lambda_c}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\nu^2}{2}\right), \tag{2.70}$$

$$c(\nu) = \frac{\sqrt{2\pi}}{\lambda_c^2}\nu\exp\left(-\frac{\nu^2}{2}\right), \tag{2.71}$$

$$g(\nu) = \frac{1}{\lambda_c^3\sqrt{2\pi}}(1-\nu^2)\exp\left(-\frac{\nu^2}{2}\right). \tag{2.72}$$

These predictions are defined as functions of the threshold ν, which corresponds to a certain value of the smoothed density field normalized by its standard deviation, $\nu = \delta/\sigma_0$, such that is has zero mean, $\langle\nu\rangle = 0$, and unit variance, $\langle\nu^2\rangle = 1$. The smoothing is done by applying a smoothing kernel W_R with smoothing length R to the raw, unsmoothed density

field,

$$\delta(\mathbf{x}) = \int \mathrm{d}^3x' W_R(|\mathbf{x} - \mathbf{x}'|)\delta_{\mathrm{raw}}(\mathbf{x}'). \tag{2.73}$$

Most commonly, the smoothing is performed with a Gaussian kernel and will be discussed in more detail in Chapter 4.

The parameter λ_c can be derived from the value of the correlation function $\xi(0)$ and its second derivative $\xi''(0)$ at zero separation by

$$\lambda_\mathrm{c} = \sqrt{\frac{2\pi\xi(0)}{|\xi''(0)|}}. \tag{2.74}$$

The values for $\xi(0)$ and $\xi''(0)$ can be directly computed from the power spectrum of the smoothed distribution as

$$\xi(0) = \langle\delta^2\rangle = \sigma_0^2, \tag{2.75}$$

$$|\xi''(0)| = \langle|\nabla\delta|^2\rangle = \sigma_1^2, \text{ and} \tag{2.76}$$

$$\sigma_j^2(R) = \int \frac{k^2\,\mathrm{d}k}{2\pi^2} k^{2j} P(k)\hat{W}(k)^2, \tag{2.77}$$

where $\hat{W}(k)$ is the Fourier transform of the smoothing filter. As can be seen from these equations, the amplitudes of the surface area, curvature and genus densities depend on the underlying power spectrum. The volume-filling fraction, however, is insensitive to the power spectrum. Fig. 2.3 shows the theory predictions for a Gaussian density field at redshift $z = 0.57$ with the same linear ΛCDM power spectrum as the Minerva simulations, which will be described in more detail in Section 3.4. A deviation from Gaussianity, which can be caused by the non-linear gravitational evolution, leads to discrepancies of the measured Minkowski functionals with the corresponding Gaussian predictions. The most notable discrepancy manifests in an asymmetry of the genus curve (Matsubara, 1994).

For the comparison of measurements to theory predictions, it is often convenient to express the Minkowski functional densities as functions of f_V instead of the density threshold ν. In particular, the Minkowski functional densities expressed as functions of f_V are expected to be invariant under any local monotonic transformation (Codis et al., 2013). For a Gaussian density field, the prediction of the volume-filling fraction f_V in equation (2.68) can be easily inverted and the predictions of the other Minkowski functional densities can then be written in terms of f_V.

The Gaussian theory predictions are useful for searching for deviations from Gaussianity and extracting non-Gaussian information, which is equivalent to higher-order information. In the recent years, new theoretical predictions for the Minkowski functionals of weakly non-Gaussian fields have been derived (Pogosyan et al., 2009; Matsubara, 2010; Gay et al., 2012; Codis et al., 2013; Matsubara & Kuriki, 2020; Matsubara et al., 2020). Since the formulae for the Minkowski functionals of non-Gaussian density fields are lengthy and their application to large-scale structure analysis has practically not been explored yet, I will not discuss them here. The Minkowski functionals of the non-linear galaxy field will be one of the main topics of Chapter 4.

2.8 Galaxy clustering in redshift space

In large-scale structure analysis, the underlying matter density field can only be probed indirectly. This section describes the three main problems associated with the interpretation of the observed galaxy clustering.

2.8.1 Galaxy bias

The first problem is how to connect the galaxy distribution to the theory predictions for the matter density field. Present-day galaxy formation models indicate that galaxies form and reside inside dark matter halos. The dark matter halos in turn form through the gravitational collapse of the matter. The spherical collapse model, which is discussed in Section 2.5.4, gives a simplified explanation for the formation of gravitationally bound halos. Since the halos form at the peaks of the matter density field, the halo clustering is not identical to that of the underlying matter density field, and the relation between both can be described by the so-called *halo bias*.

The galaxy density contrast on the other hand is not exactly the same as the halo density contrast due to the complex and non-linear process of galaxy formation. More massive halos, for example, can host multiple galaxies. In general, galaxies are biased tracers of the underlying matter density field.

Without making any assumptions about the formation, the density contrast of tracers such as galaxies and halos can be related to the underlying matter density contrast by a perturbative bias expansion, as

$$\delta_{\mathrm{g}}(\mathbf{x}) = \sum_k \frac{b_k}{k!}\delta(\mathbf{x})^k, \tag{2.78}$$

where the coefficients b_k are the bias parameters. Here, I show the expansion for the galaxy density field δ_{g}, but the same framework can be applied to the halo density field δ_{h}. On large scales, this relation can be approximated by the linear galaxy bias, $\delta_{\mathrm{g}} \approx b_1\delta$. Including also the second-order bias b_2 results in a more accurate modelling of the galaxy or halo two-point statistics.

Equation (2.78) assumes that the galaxy bias only depends on the local matter density contrast. The non-linear gravitational evolution, however, can introduce a nonlocal galaxy bias, which depends on the velocity divergence and gravitational potentials. This work follows the bias expansion by Chan & Scoccimarro (2012), which was also used in the recent analyses of the BOSS and eBOSS surveys by Sánchez et al. (2017), Grieb et al. (2017), Salazar-Albornoz et al. (2017) and Hou et al. (2018),

$$\delta_{\mathrm{g}} = b_1\delta + \frac{b_2}{2}\delta^2 + \gamma_2\mathcal{G}_2 + \gamma_3^-\Delta_3\mathcal{G} + \ldots, \tag{2.79}$$

where γ_2 and γ_3^- are nonlocal bias parameters, and $\mathcal{G}_2$ is the so-called second 'Galileon' operator of the normalized density and velocity potentials Φ and Φ_v, which is defined as

$$\mathcal{G}_2[\Phi_v] = \left(\nabla_{ij}\Phi_v\right)^2 - \left(\nabla^2\Phi_v\right)^2, \tag{2.80}$$

and

$$\Delta_3\mathcal{G} = \mathcal{G}_2[\Phi] - \mathcal{G}_2[\Phi_v]. \tag{2.81}$$

Under the assumption of a local bias in Lagrangian coordinates, the nonlocal bias parameters can be expressed in terms of the local linear bias b_1. Sánchez et al. (2017) and Grieb et al. (2017) found that fixing γ_2 to the local Lagrangian relation is a good approximation for the bias modelling of the two-points statistics that are measured from the Minerva N-body simulations, which will be described in Section 3.4. According to Chan & Scoccimarro (2012), the local Lagrangian relation for γ_2 is

$$\gamma_2 = -\frac{2}{7}(b_1 - 1). \tag{2.82}$$

The full and rather long expressions for the galaxy power spectrum that follow from the full bias expansion of equation (2.79) can be found in Sánchez et al. (2017).

The implications of galaxy bias for Minkowski functionals have not been studied in detail yet. The local bias, however, corresponds to a local monotonic transformation of the density field, for which Codis et al. (2013) showed that the Minkowski functional densities as function of f_V are expected to be invariant.

2.8.2 Redshift-space distortions

The distance of a galaxy is inferred from the observed redshift. The measured redshift, however, does not only contain the cosmological redshift due to expansion of the Universe, which is described in Section 2.1.3, but also an additional component due to the peculiar velocity of the galaxy. The measured distance $\boldsymbol{s}$ in redshift space differs from the true distance $\boldsymbol{r}$ in real space by the peculiar velocity along the line of sight $v_\parallel$,

$$\boldsymbol{s} = \boldsymbol{r} + \frac{v_\parallel}{aH(a)}\widehat{\boldsymbol{n}} \ \text{(in comoving Mpc)}, \tag{2.83}$$

where $v_\parallel = \boldsymbol{v}\cdot\widehat{\boldsymbol{n}}$, $\widehat{\boldsymbol{n}} = \boldsymbol{r}/|\boldsymbol{r}|$ and $|\boldsymbol{v}| \ll c$. Hence, the galaxy appears displaced along the line of sight. Although this effect impedes the exact measurement of the true distance to the galaxy, it contains useful information on the growth of structure, since the evolution of the velocity field is tightly connected to the matter density field. During this whole section, the distant observer approximation is used, i.e. assuming that the considered objects are far away and separated by a small angle.

Linear redshift-space distortions

In the linear regime, the relation between the divergence of the velocity field and the matter density field can be obtained from the linearised continuity equation (2.36) and the growing mode solution for the matter density field, $\delta(\boldsymbol{x}, a) = D_1(a)\delta(\boldsymbol{x})$ of equation (2.39),

$$\nabla\cdot\boldsymbol{v}(\boldsymbol{x}, a) = -aH(a)f(a)\delta(\boldsymbol{x}, a), \tag{2.84}$$

where the dimensionless linear growth factor, f, is defined as

$$f \equiv \frac{a}{D_1}\frac{\mathrm{d}D_1}{\mathrm{d}a} = \frac{\mathrm{d}\ln D_1}{\mathrm{d}\ln a}. \tag{2.85}$$

Taking the number conservation between redshift and real space into account, $\rho^{\mathrm{s}}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \rho^{\mathrm{r}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}$, the relation between the redshift and real space density contrast is given by

$$(1+\delta^{\mathrm{s}}(\boldsymbol{s}))\,\mathrm{d}^3 s = (1+\delta^{\mathrm{r}}(\boldsymbol{r}))\,\mathrm{d}^3 r, \tag{2.86}$$

where s and r are the superscripts for redshift and real space, respectively. The Jacobian, $J = |\partial r_i/\partial s_i|$, that maps between redshift and real space, has the form

$$J = \left(1 + \frac{1}{aH}\frac{\partial v_\parallel}{\partial r}\right)^{-1} \approx \left(1 - \frac{1}{aH}\frac{\partial v_\parallel}{\partial r}\right). \tag{2.87}$$

Using equation (2.84) for the velocity field in Fourier space, where the derivative is transformed as $\partial/\partial r_i \to ik_i$, finally leads to

$$\delta^s(\boldsymbol{k}) = \delta^r(\boldsymbol{k})(1 + f(z)\mu_k^2), \tag{2.88}$$

where $\mu_k = k_\parallel/k$ is the cosine of the angle between the line of sight and wave vector k. The growth factor f imprinted in the redshift-space density field is a valuable source of cosmological information. According to Linder & Cahn (2007), it can be parametrized by $f = \Omega_m(z)^\gamma$. General relativity predicts for a ΛCDM universe $\gamma = 0.55$, and therefore f can either probe general relativity or the value of Ω_m.

For galaxies or halos, equation (2.88) becomes

$$\delta^s_{\mathrm{g}}(\boldsymbol{k}) = \delta^r_{\mathrm{g}}(\boldsymbol{k})(1 + \beta\mu_k^2), \tag{2.89}$$

where $\beta = f/b_1$, and assuming that there is no velocity bias ($\boldsymbol{v}_{\mathrm{g}} = \boldsymbol{v}$). The corresponding linear power spectrum is then given by

$$P^s_{\mathrm{g}}(k,\mu_k) = b_1^2(1+\beta\mu_k^2)^2 P(k). \tag{2.90}$$

The *Kaiser factor* $(1+\beta\mu_k^2)^2$ describes the squashing along the line of sight in the clustering pattern that is produced by coherent bulk flows of matter towards overdense regions. Due to this effect, the power spectrum in redshift space is anisotropic.

From equations (2.89) and (2.90) the growth factor f can only be extracted in combination with the linear bias b_1. An alternative is to rewrite equation (2.90) according to Sánchez (2020) as

$$P_{\mathrm{g}}(k,\mu_k,z) = \left(b_1\sigma_{12}(z) + f\sigma_{12}(z)\mu_k^2\right)^2 \frac{P(k)}{\sigma_{12}^2(z)}, \tag{2.91}$$

where σ_{12}, the linear-theory rms mass fluctuation in a sphere of radius 12 Mpc, has been defined in equation (2.42). Expressing the redshift-space power spectrum in this way has

the advantage that the last factor only depends on the shape of the isotropic matter power spectrum and not on the amplitude of the fluctuations. Hence, the anisotropy and the amplitude of the galaxy power spectrum are described by $b\sigma_{12}(z)$ and $f\sigma_{12}(z)$. So far, it has been standard in galaxy clustering analyses to rewrite equation (2.91) in terms of σ_8 instead of σ_{12} and to constrain the *growth rate* $f\sigma_8(z)$.

A full derivation for the linear redshift-space distortions can be found in Kaiser (1987) in Fourier space, and Hamilton (1992) in configuration space.

The non-linear redshift-space power spectrum

The non-linear gravitational evolution induces a non-linear coupling between the density and velocity fields. Furthermore, the random motions of galaxies within virialized structures lead to an elongation of the clustering pattern on small scales. Both effects are not captured by the linear model for redshift-space distortions.

Many models of non-linear redshift-space distortions that are applied to present-day galaxy clustering analyses are based on the ansatz

$$P_{\mathrm{g}}(k,\mu_k) = F_{\mathrm{FoG}} P_{\mathrm{novir}}(k,\mu_k), \tag{2.92}$$

where the fingers-of-god factor F_{FoG} describes the small-scale redshift-space distortions and P_{novir} includes the effect of the non-linear bulk flow of matter on larger scales.

According to Scoccimarro (2004) and Taruya et al. (2010) the non-linear anisotropic power spectrum P_{novir} can be decomposed into

$$P_{\mathrm{novir}}(k,\mu_k) = P^{(1)}_{\mathrm{novir}}(k,\mu_k) + P^{(2)}_{\mathrm{novir}}(k,\mu_k) + P^{(3)}_{\mathrm{novir}}(k,\mu_k). \tag{2.93}$$

The first term $P^{(1)}_{\mathrm{novir}}$ corresponds to a non-linear version of the Kaiser formula of equation (2.90),

$$P^{(1)}_{\mathrm{novir}}(k,\mu_k) = P_{\mathrm{gg}}(k) + 2f\mu_k^2 P_{\mathrm{g}\theta}(k) + f^2\mu_k^4 P_{\theta\theta}(k), \tag{2.94}$$

where $P_{\mathrm{gg}}(k) = \langle \delta_{\mathrm{g}}(\boldsymbol{k})\delta_{\mathrm{g}}(\boldsymbol{k}')\rangle$, $P_{\mathrm{g}\theta}(k) = \langle \delta_{\mathrm{g}}(\boldsymbol{k})\theta(\boldsymbol{k}')\rangle$ and $P_{\theta\theta}(k) = \langle \theta(\boldsymbol{k})\theta(\boldsymbol{k}')\rangle$ are the galaxy-galaxy, galaxy-velocity and velocity-velocity power spectra.[4] All these power spectra can be computed using gRPT and the bias expansion of equation (2.79).

The second term is given by

$$P^{(2)}_{\mathrm{novir}}(k,\mu_k) = \int \mathrm{d}^3q \frac{q_z}{q^2}[B_\sigma(\boldsymbol{q},\boldsymbol{k}-\boldsymbol{q},-\boldsymbol{k}) - [B_\sigma(\boldsymbol{q},\boldsymbol{k},-\boldsymbol{k}-\boldsymbol{q})], \tag{2.95}$$

where B_σ is the cross bispectrum between the density and velocity field,

$$\langle \theta(\boldsymbol{k}_1)\sigma(\boldsymbol{k}_2)\sigma(\boldsymbol{k}_3)\rangle = (2\pi)^3\delta_{\mathrm{D}}(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3)B_\sigma(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3), \tag{2.96}$$

[4] θ is the subscript for the velocity divergence defined in Section 2.5.1

with $\sigma(\boldsymbol{k}) = \delta_{\mathrm{g}}(\boldsymbol{k}) + f(k_\parallel/k^2)\theta(\boldsymbol{k})$. Finally, the third contribution to P_{novir} is

$$P^{(3)}_{\mathrm{novir}}(k,\mu_k) = \int \mathrm{d}^3q \frac{q_\parallel(k_\parallel - q_\parallel)}{q^2(\boldsymbol{k}-\boldsymbol{q})^2}(b_1 + f\mu_q^2)(b_1 + f\mu_{k-q}^2)P_{\delta\theta}(k-q)P_{\delta\theta}(q). \tag{2.97}$$

The last two terms can be evaluated with standard tree-level perturbation theory.

The factor F_{FoG} models the redshift-space distortions on small scales, which are dominated by the random motions of galaxies inside virialized structures. The resulting elongation of the clustering pattern along the line of sight is called "Fingers of God" (FoG) effect. It can be well described by the following functional form,

$$F_{\mathrm{FoG}} \equiv W_\infty(k,\mu_k) = \frac{1}{\sqrt{1 + f^2\mu_k^2k^2a_{\mathrm{vir}}}} \exp\left(\frac{-f^2\mu_k^2k^2\sigma_v^2}{1 + f^2\mu_k^2k^2a_{\mathrm{vir}}}\right), \tag{2.98}$$

where σ_v is the one-dimensional linear velocity dispersion and a_{vir} encodes the kurtosis of the small-scale velocity distribution. If the clustering of halos instead of galaxies is considered, the FoG factor should be neglected. The model for P_{novir} can also be applied for halo clustering measurements.

The model of the redshift-space galaxy power spectrum described here was used in the analyses by Sánchez et al. (2017), Grieb et al. (2017), Salazar-Albornoz et al. (2017) and Hou et al. (2018). A more detailed explanation can be found in Sánchez et al. (2017).

A model for the galaxy two-point correlation function in redshift space, $\xi(s,\mu)$, can be obtained by the Fourier transform of the predictions for the power spectrum. More details on the measurement of $\xi(s,\mu)$ will be described in Chapter 3.

While the analysis of redshift-space distortions is highly advanced for the two-point clustering statistics, the impact of redshift-space distortions on Minkowski functionals is not understood and modelled in such great detail yet. I will discuss the effect of redshift-space distortions on the Minkowski functionals as part of the original work of this thesis in Chapter 4.

2.8.3 Alcock-Paczynski distortions

In addition to bias and redshift-space distortions, a further complication needs to be taken into account. In order to convert the measured redshift into a comoving distance, a fiducial cosmology has to be assumed. If the fiducial cosmology differs from the true underlying cosmology, the inferred distance deviates from the true one. This leads to the the so-called Alcock-Paczynski (AP) distortions (Alcock & Paczynski, 1979) in the components parallel and perpendicular to the line-of-sight, $s_\parallel$ and $s_\perp$, of the separation vector $\boldsymbol{s}$ between two galaxies,

$$s_\parallel = q_\parallel s'_\parallel, \qquad s_\perp = q_\perp s'_\perp, \tag{2.99}$$

where the primes denote the quantities in the fiducial cosmology.

The *geometric distortion parameters* $q_\parallel$ and $q_\perp$ are given by the ratios of the angular diameter distance D_M and the Hubble parameter H in the true and fiducial cosmologies,

$$q_\parallel = \frac{H'(z)}{H(z)}, \tag{2.100}$$

$$q_\perp = \frac{D_\mathrm{M}(z)}{D'_\mathrm{M}(z)}. \tag{2.101}$$

For a flat ΛCDM universe with a negligibly small radiation density Ω_r, the distortion parameters are directly sensitive to Ω_m, since in that case $D_\mathrm{M}(z) = D_\mathrm{c}(z)$, and $H(z) = \sqrt{\Omega_\mathrm{m}(1+z)^3 + (1-\Omega_\mathrm{m})}$ (see Sections 2.1.1 and 2.1.3).

As described in the previous Section 2.8.2, measurements of the anisotropic two-point statistics are usually expressed as functions of the separation $s = \sqrt{s_\perp^2 + s_\parallel^2}$ and the cosine of the angle between the separation vector $\boldsymbol{s}$ and line of sight, denoted as μ. The rescaling of s and μ due to the AP distortions can then be obtained as

$$s = s'\sqrt{q_\parallel^2(\mu')^2 + q_\perp^2(1-(\mu')^2)}, \tag{2.102}$$

$$\mu = \frac{q_\parallel \mu'}{\sqrt{q_\parallel^2(\mu')^2 + q_\perp^2(1-(\mu')^2)}}. \tag{2.103}$$

The anisotropy in the measurements of the two-dimensional two-point statistics allows to separate the information on the Hubble parameter $H(z)$ and the angular diameter distance $D_\mathrm{M}(z)$. However, for the case of a volume-averaged measurement, such as the Minkowski functionals, the isotropic volume element changes as

$$\mathrm{d}^3 s = q^3 \mathrm{d}^3 s' = \left(q_\perp^2 q_\parallel\right) \mathrm{d}^3 s' = \left(\frac{D_\mathrm{V}(z_m)}{D'_\mathrm{V}(z_m)}\right)^3 \mathrm{d}^3 s', \tag{2.104}$$

where q is dubbed *isotropic AP parameter*. Hence, it is only possible to obtain information on the volume-averaged distance D_V, which is defined in equation (2.19). Similar to the redshift-space distortions, I will discuss the effect of the Alcock-Paczynski distortions on Minkowski functionals as part of the original work in Chapter 4.

BAO distance measurements can constrain the angular diameter distance $D_\mathrm{M}(z)$ and and the Hubble parameter $H(z)$ relative to the sound horizon scale at the drag redshift, r_d (see Section 2.4.2). In order to obtain comparable results, the geometric parameters inferred from the analysis of full-shape clustering measurements are usually rescaled by the ratios of the sound horizon in the fiducial and true cosmologies,

$$\alpha_\parallel = q_\parallel \frac{r'_\mathrm{d}}{r_\mathrm{d}}, \quad \alpha_\perp = q_\perp \frac{r'_\mathrm{d}}{r_\mathrm{d}}, \tag{2.105}$$

and $\alpha_\parallel$ and $\alpha_\perp$ are referred to as *Alcock-Paczynski parameters* or *BAO shift parameters*.

Chapter 3

Covariance Matrix Comparison

This chapter focuses on the covariance matrix, one of the most important requisites to extract unbiased cosmological information from clustering measurements. As already pointed out in the introduction in Chapter 1, future galaxy surveys covering large cosmological volumes with millions of galaxies will allow for unprecedented statistical precision. To achieve this high precision, it is essential to identify all components of the systematic error budget affecting cosmological analyses and if possible, reduce the associated uncertainties. In this regard, the robust estimation of the covariance matrix plays a key role in the inference of cosmological parameters.

In galaxy clustering analyses, the covariance matrix is typically computed from a set of mock catalogues. In order to reach the level of statistical precision needed for future surveys, a number of several thousand mock catalogues might have to be generated. Since N-body simulations can reproduce the non-linear structure formation with high accuracy (see Section 2.6), they are the ideal choice for the construction of mock catalogues. However, N-body simulations are expensive in terms of run-time and memory. The construction of the large number of mock catalogues that is presumable necessary for the covariance estimates of future surveys might therefore be infeasible. During the last decades, several approximate methods for gravitational dynamics have been developed that enable a faster generation of mock catalogues. The nIFTy comparison project by Chuang et al. (2015) compared several approximate methods regarding their ability to reproduce clustering statistics, more specifically the two-point correlation function, power spectrum and bispectrum, of halo samples drawn from N-body simulations.

Here, we extend those efforts and perform a comparison of the covariance matrices inferred from different approximate methods. In particular, the focus lies on the performance of the different covariance matrices at reproducing parameter constraints obtained using N-body simulations. To this end, seven state-of-the-art approximate methods and recipes and a reference N-body simulation are considered. The following analysis is based on halo samples, in order to obtain results that do not depend on a specific recipe for populating the halos with galaxies.

The work presented in this chapter is part of a set of comparison projects with the goal of analysing the covariance matrices of two-point correlation function (Lippich et al., 2019),

power spectrum (Blot et al., 2019) and bispectrum (Colavincenzo et al., 2019) measurements. After the comparison project, I also contributed to a follow-up project on a further novel approximate method that was developed in the meantime (Balaguera-Antolínez et al., 2020). Here, I will focus on the work based on correlation function measurements, since I led the corresponding analysis. One may also note that the covariance matrix in configuration space has a more interesting structure than in Fourier space, where it is almost diagonal.

This chapter is structured as follows. The first part of the chapter gives an overview of the general methodology of correlation function measurements (Section 3.1), standard likelihood analysis (Section 3.2) and covariance matrix estimation (Section 3.3), which are the basis for inferring parameter constraints in the subsequent performance tests. Section 3.4 introduces the reference halo catalogues from the Minerva N-body simulations. The different approximate methods and recipes included in our comparison are described in Section 3.5. The halo samples that we consider are defined in Section 3.6. The methodology of the performance tests is presented in Section 3.7. We compare the clustering properties of the different halo samples in Section 3.8, the corresponding covariance matrices in Section 3.9, and finally the performance of the different covariance matrices in Section 3.10. The last Section 3.11 further discusses the results from the performance tests and puts them into the context of the joint comparison project.

The study presented here has been published as Lippich et al. (2019). I carried out the majority of the analysis and wrote most of the text, with significant advisory contributions from the authors in order of their appearance in the authors list. The authors in alphabetical order contributed with the raw halo catalogues generated by the different approximate methods. Sections 3.1 to 3.4 present a more detailed description of the underlying methodology than in the publication. Sections 3.5 to 3.11 reproduce the corresponding sections of the publication, adapted such the format, references and section titles match the thesis format.

3.1 Clustering measurements in configuration space

In Section 2.3 the two-point correlation function was introduced to describe the probability of finding an excess of pairs compared to a homogeneous distribution. The probability of finding halo pairs separated by the distance s in a direction specified by μ, which is the cosine of the angle between the separation vector $\mathbf{s}$ and the line of sight as defined in Section 2.8.2, is given by

$$DD(s,\mu) = \frac{N_{\mathrm{pairs}}(s,\mu)}{N_{\mathrm{tot}}}, \tag{3.1}$$

where $N_{\mathrm{tot}} = N_{\mathrm{h}}(N_{\mathrm{h}}-1)/2$ is the total number of halo pairs. The probability of finding halo pairs in a random distribution is denoted by $RR(s,\mu)$.

In the present work, all measurements are performed in simulation boxes with periodic boundary conditions. In order to obtain redshift-space measurements, one Cartesian axis,

here the z-axis, is treated as the line-of-sight. The halo positions are then distorted by adding the halo velocities parallel to this axis.

Due to the periodic boundary conditions, the normalized random pair counts can be computed directly as the ratio of the volume of a shell $\mathrm{d}V$ at a radius s and the total volume of the simulation box, $RR(s) = \mathrm{d}V(s)/V_{\mathrm{tot}}$, where

$$\mathrm{d}V(s) = \frac{4\pi}{3}[(s + \mathrm{d}s)^3 - s^3]. \tag{3.2}$$

According to equation (2.28) the two probabilities are related by

$$DD(s, \mu) = RR(s, \mu)(1 + \xi(s, \mu)). \tag{3.3}$$

This equation directly yields an expression for computing the two-point correlation function $\xi(s, \mu)$ from pair counts, which is also known as the *natural estimator*,

$$\xi(s, \mu) = \frac{DD(s, \mu)}{RR(s, \mu)} - 1. \tag{3.4}$$

There are alternative estimators that lead to a lower variance in the correlation function estimation of real galaxy surveys. If the random counts can be calculated using equation (3.2), however, they reduce to the natural estimator. Hence, all two-point correlation function measurements in this work are based on equation (3.4).

The measurement of the full two-dimensional correlation function $\xi(s, \mu)$ is typically associated with a large number of data points and a low signal-to-noise ratio leading to complications in its further analysis. Therefore, in most cosmological analyses the information in $\xi(s, \mu)$ is compressed into a small number of functions. A standard approach is to decompose $\xi(s, \mu)$ into *Legendre multipoles*, $\xi_\ell(s)$, given by

$$\xi_\ell(s) = \frac{2\ell + 1}{2} \int_{-1}^{1} L_\ell(\mu)\xi(\mu, s)\, \mathrm{d}\mu, \tag{3.5}$$

where $L_\ell(\mu)$ denotes the Legendre polynomial of order ℓ. Due to the symmetry of $\xi(\mu, s)$ with respect to μ, all multipoles with odd ℓ are zero. For current clustering analyses, the largest amount of information can be extracted by considering the multipoles with $\ell = 0, 2, 4$, called monopole, quadrupole and hexadecapole, respectively.

An alternative is the *clustering wedges* statistic (Kazin et al., 2012), which corresponds to the average of the full two-dimensional correlation function over wide bins in μ, that is

$$\xi_{\mathrm{w},i}(s) = \frac{1}{\Delta\mu} \int_{(i-1)/n}^{i/n} \xi(\mu, s)\, \mathrm{d}\mu, \tag{3.6}$$

where $\xi_{\mathrm{w},i}$ denotes each individual clustering wedge, and n represents the total number of wedges. In this study, we follow the analysis by Sánchez et al. (2017) and divide the μ range from 0 to 1 into three equal-width intervals, $i = 1, 2, 3$.

For the measurements of the Legendre multipoles and clustering wedges, we consider scales in the range 20 h^{-1}Mpc $\leq s \leq$ 160 h^{-1}Mpc throughout this chapter, and implement a binning scheme with $ds = 10\,h^{-1}$Mpc for the following analysis. For illustration purposes, we also use a binning of $ds = 5\,h^{-1}$Mpc for figures showing correlation function measurements. Since we have three multipoles $\ell = 0, 2, 4$ and three μ wedges, the dimension of the total data vector, $\boldsymbol{\xi}$, for each case, containing all the measured statistics is $N_{\rm b} = 42$ for the binning with $ds = 10\,h^{-1}$Mpc (and $N_{\rm b} = 84$ for $ds = 5\,h^{-1}$Mpc). Note that here we use the common units of h^{-1}Mpc so that the joint comparison project is directly comparable to previous clustering analyses. For the future we advocate the use of Mpc units (see Sánchez, 2020).

3.2 Standard likelihood analysis

To infer the cosmological parameters that best describe the clustering measurements, we follow a Bayesian approach and perform fits that maximize the likelihood function. The *likelihood*, $\mathcal{L}$, is defined as the probability of the measured data given a specific parameter set, $\boldsymbol{\theta}$, of the considered theory model. In many cases the likelihood can be well approximated by a multivariate-Gaussian, $\mathcal{L} \propto \exp(-\chi^2(\boldsymbol{\theta})/2)$. This is motivated by the central limit theorem stating that the sum of many independent and identically distributed random variables is Gaussian. Here we explore the Gaussian likelihood in the form,

$$-2\ln\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{\theta}) = (\boldsymbol{\xi} - \boldsymbol{\xi}_{\mathbf{theo}}(\boldsymbol{\theta}))^t \boldsymbol{\Psi}, (\boldsymbol{\xi} - \boldsymbol{\xi}_{\mathbf{theo}}(\boldsymbol{\theta})) \tag{3.7}$$

where the expression on the right-hand side of the equation corresponds to the standard χ^2, and $\boldsymbol{\xi}_{\mathbf{theo}}$ represents the theoretical model of the measured statistics, which here corresponds to the Legendre multipoles or clustering wedges $\boldsymbol{\xi}$. An important ingredient of the likelihood is the precision matrix $\boldsymbol{\Psi}$, which is the inverse of the covariance matrix, $\boldsymbol{\Psi} = \mathbf{C}^{-1}$. The estimation of the covariance matrix is the topic of the next Section 3.3.

For the theoretical model, we adopt the prediction of the anisotropic non-linear power spectrum based on gRPT that is described in Sections 2.5.2 and 2.8. We transfer the theory power spectrum to predictions for the correlation function multipoles and wedges using a Fourier transform and equations (3.5) and (3.6). Since here we analyse halo samples, we do not include the fingers-of-God factor, $W_\infty(k, \mu)$, of equation (2.98).

In galaxy clustering analyses, it has been common to fix the cosmological parameters for the prediction of the non-linear matter power spectrum to those of a 'template' cosmology and only use the anisotropic information to constrain $D_M(z)/r_{\rm d}$, $H(z)r_{\rm d}$ and $f\sigma_8(z)$. In that way, there are less free parameters and one can obtain cosmological constraints with a smaller computational cost. As the aim of this work is to compare the performance of covariance matrices from several different methods and not to obtain the most accurate parameter constraints from real data, we also follow this approach and set the template parameters to those of the Minerva simulation. In total, our parameter space contains *six free parameters*: the Alcock-Paczynski parameters $\alpha_\parallel$ and $\alpha_\perp$ from equation (2.105), the growth rate $f\sigma_8$ from the RSD model in Section 2.8.2, the nuisance parameters associated

with the linear and quadratic local bias, b_1 and b_2, and the non-local bias γ_3^- from the bias model in Section 2.8.1.

We perform fits varying these six parameters by means of the Monte Carlo Markov Chains (MCMC) technique and based on the likelihood of equation (3.7). The best-fitting value and corresponding error of a parameter are inferred from the marginalised mean and dispersion of the resulting MCMC chains.

Note that in practice we fit for the geometric distortion parameters $q_\parallel$ and $q_\perp$, as h is fixed, and quote our results in terms of $\alpha_\parallel$ and $\alpha_\perp$, as discussed in Section 2.8.3. Also, we quote our results using the traditional growth rate $f\sigma_8$, because we conducted this analysis as part of the joint comparison project before Sánchez (2020) was published, with the intention of having comparable results to previous analyses. Since h is fixed, the results on $f\sigma_8$ can be directly expressed in terms of $f\sigma_{12}$ by multiplying the resulting values of $f\sigma_8$ by the ratio of σ_{12}/σ_8 for the considered redshift and cosmology. For that reason, all our results remain valid, however, we endorse the use of $f\sigma_{12}$ and Mpc units in the future (see also Section 2.8.2).

3.3 Covariance matrix estimation from mocks

The key ingredient in the likelihood is the covariance matrix $\mathbf{C}$. There are three main approaches to estimate the covariance of the data vector. One possibility is to model the covariance analytically. This has the advantage that the covariance matrix estimate is not affected by noise. The challenge of this approach, however, is to model the covariance in the non-linear regime, and in particular to include the complex masks of real galaxy surveys. A way to circumvent this problem is to estimate the covariance directly from the data, for example by means of Jackknife estimates. Besides noise, the downside of such method is that it can introduce biases in the covariance estimates that are difficult to capture.

Therefore, the most popular choice for galaxy clustering analyses is to estimate the covariance matrix from a set of N_s mock catalogues from simulations as

$$C_{ij} = \frac{1}{N_\mathrm{s}-1}\sum_{k=1}^{N_\mathrm{s}}(\xi_i^k - \bar{\xi}_i)(\xi_j^k - \bar{\xi}_j), \tag{3.8}$$

where $\bar{\xi}_i = \frac{1}{N_\mathrm{s}}\sum_k \xi_i^k$ is the mean value of the measurements at the i-th bin and ξ_i^k is the corresponding measurement from the k-th mock. In order to obtain an invertible covariance matrix, and hence an estimate for the precision matrix $\mathbf{\Psi}$ in the likelihood, the number of mocks should be significantly larger than the number of measurement bins, $N_\mathrm{s} \gg N_\mathrm{b}$. The covariance estimate from mocks tends to be less affected by biases than estimates from the data, and does not require any assumptions regarding the properties of the true covariance matrix, as it is the case for theoretical estimates. Furthermore, survey masks can be easily included, in order to generate mock observations that reproduce the properties of a given survey.

However, the finite number of mocks introduces noise in the covariance estimates that must be propagated into the final parameter constraints as additional uncertainty (Taylor

et al., 2013; Dodelson & Schneider, 2013; Percival et al., 2014; Sellentin & Heavens, 2016). Following Dodelson & Schneider (2013), this additional uncertainty can be approximated by

$$\frac{\sigma_{\rm extra}}{\sigma_{\rm ideal}} \approx 1 + \frac{N_{\rm b} - N_{\rm p}}{2(N_{\rm s} - N_{\rm b})}, \tag{3.9}$$

where $\sigma_{\rm extra}$ is the parameter variance inferred from the noisy covariance matrix, $\sigma_{\rm ideal}$ is the ideal variance without noise, $N_{\rm s}$ the number of simulations, $N_{\rm b}$ the number of bins, $N_{\rm p}$ the number of parameters and we assume $N_{\rm s} \gg N_{\rm b} \gg N_{\rm p}$. We consider the following example where to goal is to limit the additional uncertainty to 2%. Already for our set up with $N_{\rm b} = 48$ and $N_{\rm p} = 6$, we would need more than 1000 simulations according to equation (3.9). A real galaxy clustering analysis would ideally have smaller and consequently more bins, vary all cosmological parameters and fit a number of redshift slices simultaneously. The control of this additional error would require an even larger number of independent realizations, with $N_{\rm s}$ in the range of a few thousands.

For the new generation of galaxy surveys with large volumes and multiple redshift bins, the construction of mock catalogues in such a number will be challenging and might need to rely on approximate N-body methods. The goal of our analysis is to test the impact of the covariance estimates $\mathbf{C}$ from different approximate methods on parameter constraints.

There are several techniques that can help to reduce the required number of realizations, such as resampling the phases of N-body simulations (Hamilton et al., 2006; Schneider et al., 2011), shrinkage (Pope & Szapudi, 2008), calibrating the non-Gaussian contributions of an empirical model against N-body simulations (O'Connell et al., 2016), or covariance tapering (Paz & Sánchez, 2015). However, even after applying such methods, the generation of multiple N-body simulations with the required number-density and volume for the clustering analysis of future surveys would be extremely demanding.

3.4 The Minerva N-body simulations and halo catalogues

We compare the performance of the approximate methods against reference N-body simulations called Minerva. The simulations were performed using GADGET-3, the third version of the GADGET code (Springel, 2005) that was introduced in Section 2.6. They consist of 300 independent realizations of the same cosmology, corresponding to the best-fitting flat ΛCDM model from the WMAP+BOSS DR9 analysis of Sánchez et al. (2013). This cosmology is characterized by the total matter and baryon densities $\Omega_{\rm m} = 0.285$ and $\Omega_{\rm b} = 0.046$, a Hubble constant of $H_0 = 69.5\,{\rm km\,s^{-1}Mpc^{-1}}$, a scalar spectral index $n_{\rm s} = 0.968$, and a linear-theory rms mass fluctuation in spheres of radius $12\,{\rm Mpc}$, $\sigma_{12} = 0.805$ (Sánchez, 2020, c.f. Section 2.4). The first set of 100 realizations, which is described in Grieb et al. (2016), was used in the BOSS analyses by Sánchez et al. (2017) and Grieb et al. (2017). For this analysis, 200 new independent realizations were added, which were generated with the same set-up as the first simulations. The initial conditions were

derived from second-order Lagrangian perturbation theory with an input power spectrum computed by CAMB (Lewis et al., 2000, c.f. Section 2.4) and the simulations were started at a redshift $z_{\rm ini} = 63$. Each realization simulates the evolution of 1000^3 dark-matter (DM) particles in a cubic box of side length $L = 1.5h^{-1}$Gpc with periodic boundary conditions. For the approximate methods included in the following analysis, we use exactly the same initial conditions for each realization as in the Minerva simulations and the same box size.

The positions and velocities of the evolved DM particles were stored in five snapshots at $z \in \{2.0, 1.0, 0.57, 0.3, 0.0\}$. For each snapshot, halos were identified with a Friends-of-friends (FoF) algorithm with a linking length of 0.2 of the mean inter-particle separation. The FoF halos were then further processed with the unbinding procedure provided by the SUBFIND code (Springel et al., 2001), such that particles with positive total energy are removed and halos that were artificially linked by FoF are separated. In the following, we use the so identified halos at a snapshot of $z = 1.0$ as our reference catalogues. Given the particle mass resolution of the Minerva simulations, the minimum halo mass is $2.667 \times 10^{12}\, h^{-1}{\rm M}_\odot$.

3.5 Approximate methods for covariance matrix estimates

3.5.1 Methods included in the comparison

In this comparison project[1], we included covariance matrices inferred from different approximate methods and recipes, which we compared to the estimates obtained from the set of reference N-body simulations of the previous Section 3.4. Approximate methods have recently been revived by high-precision cosmology, due to the need of producing a large number of realizations to compute covariance matrices of clustering measurements. This topic has been reviewed by Monaco (2016), where methods have been roughly divided into two broad classes. "Lagrangian" methods, as N-body simulations, are applied to a grid of particles subject to a perturbation field. They reconstruct the Lagrangian patches that collapse into dark matter halos, and then displace them to their Eulerian positions at the output redshift, typically with Lagrangian Perturbation Theory (LPT, see Section 2.5.3). ICE-COLA, Peak Patch and Pinocchio fall in this class. These methods are predictive, in the sense that, after some cosmology-independent calibration of their free parameters (that can be thought at the same level as the linking length of friends-of-friends halo finders), they give their best reproduction of halo masses and clustering without any further tuning. This approach can be demanding in terms of computing resources and can have high memory requirements. In particular, ICE-COLA belongs to the class of Particle-Mesh codes; these are in fact N-body codes that converge to the true solution (at least on large scales) for sufficiently small time-steps. As such, Particle-Mesh

[1]The text of this Section 3.5 has significant contributions from the authors providing the methods, in particular by P. Monaco

codes are expected to be more accurate than other approximate methods, at the expense of higher computational costs.

The second class of "bias-based" methods is based on the idea of creating a mildly non-linear density field using some version of LPT, and then populate the density field with halos that follow a given mass function and a specified bias model. The parameters of the bias model must be calibrated on a simulation, so as to reproduce halo clustering as accurately as possible. The point of strength of these methods is their very low computational cost and memory requirement, that makes it possible to generate thousands of realizations in a simple workstation, and to push the mass limit to very low masses. This is however achieved at the cost of lower predictivity, and need of recalibration when the sample selection changes. HALOGEN and PATCHY fall in this category.

In the following, we will refer to the two classes as "predictive" and "calibrated" models. All approximate methods used here have been applied to the same set of 300 initial conditions (ICs) of the reference N-body simulations, so as to be subject to the same sample variance; as a consequence, the comparison, though limited to a relatively small number of realizations, is not affected by sample variance.

Additionally, we included in the comparison two simple recipes for the shape of the PDF of the density fluctuations, a Gaussian analytic model that is only valid in linear theory and a log-normal model. The latter was implemented by generating 1000 catalogues of "halos" that Poisson-sample a log-normal density field; in this test case we do not match the ICs with the reference simulations, and use a higher number of realizations to lower sample variance.

3.5.2 Predictive methods: ICE-COLA, PEAK PATCH, PINOCCHIO

ICE-COLA

COLA (Tassev et al., 2013) is a method to speed up N-body simulations by incorporating a theoretical modelling of the dynamics into the N-body solver and using a low resolution numerical integration. It starts by computing the initial conditions using second-order Lagrangian Perturbation Theory (2LPT, see Crocce et al. 2006). Then, it evolves particles along their 2LPT trajectories and adds a residual displacement with respect to the 2LPT path, which is integrated numerically using the N-body solver. Mathematically, the displacement field x is decomposed into the LPT component $x_{\rm LPT}$ and the residual displacement $x_{\rm res}$ as

$$x_{\rm res}(t) \equiv x(t) - x_{\rm LPT}(t). \tag{3.10}$$

In a dark matter-only simulation, the equation of motion relates the acceleration to the Newtonian potential Φ, and omitting some constants it can be written as: $\partial_t^2 x(t) = -\nabla\Phi(t)$. Using equation (3.10), the equation of motion reads

$$\partial_t^2 x_{\rm res}(t) = -\nabla\Phi(t) - \partial_t^2 x_{\rm LPT}(t). \tag{3.11}$$

COLA uses a Particle-Mesh method to compute the gradient of the potential at the position x (first term of the right hand side), it subtracts the acceleration corresponding

to the LPT trajectory and finally the time derivatives on the left hand side are discretized and integrated numerically using few time steps. The 2LPT ensures convergence of the dynamics at large scales, where its solution is exact, and the numerical integration solves the dynamics at small non-linear scales. Halos can be correctly identified running a FoF algorithm on the dark matter density field, and halo masses, positions and velocities are recovered with accuracy enough to build mock halo catalogues.

ICE-COLA (Izard et al., 2016, 2018) is a modification of the parallel version of COLA developed in Koda et al. (2016) that produces all-sky light cone catalogues on-the-fly. Izard et al. (2016) presented an optimal configuration for the production of accurate mock halo catalogues and Izard et al. (2018) explains the light cone production and the modelling of weak lensing observables.

Mock halo catalogues were produced with ICE-COLA placing 30 time steps between an initial redshift of $z_i = 19$ and $z = 0$[2] and forces were computed in a grid with a cell size 3 times smaller than the mean inter-particle separation distance. For the FoF algorithm, a linking length of $b = 0.2$ was used. Each simulation reached redshift 0 and used 200 cores for 20 minutes in the MareNostrum3 supercomputer at the Barcelona Supercomputing Center [3], consuming a total of 20 CPU khrs for the 300 realizations.

Peak Patch

From each of the 300 initial density field maps of the Minerva suite, we generate halo catalogues following the peak patch approach initially introduced by Bond & Myers (1996). In particular, we use a new massively parallel implementation of the peak patch algorithm to create efficient and accurate realizations of the positions and peculiar velocities of dark matter halos (Stein et al., 2019). The peak patch approach is essentially a Lagrangian space halo finder that associates halos with the largest regions that have just collapsed by a given time. The pipeline can be separated into four subprocesses: (1) the generation of a random linear density field with the same phases and power spectrum as the Minerva simulations; (2) identification of collapsed regions using the homogeneous ellipsoidal collapse approximation; (3) exclusion and merging of the collapsed regions in Lagrangian space; and (4) assignment of displacements to these halos using second order Lagrangian perturbation theory.

The identification of collapsed regions is a key step of the algorithm. The determination of whether any given region will have collapsed or not is made by approximating it as an homogeneous ellipsoid, the fate of which is determined completely by the principal axes of the deformation tensor of the linear displacement field (i.e. the strain) averaged over the region. In principle, the process of finding these local mass peaks would involve measuring the strain at every point in space, smoothed on every scale. However, experimentation has shown that equivalent results can be obtained by measuring the strain around density peaks found on a range of scales[4]. This is done by smoothing the field on a series of logarithmically spaced scales with a top-hat kernel, from a minimum radius of $R_{f,\min} = 2a_{\rm latt}$, where $a_{\rm latt}$

[2]The time steps were linearly distributed with the scale factor.

[3]`http:www.bsc.es`.

is the lattice spacing, to a maximum radius of $R_{f,\max} = 40\,\mathrm{Mpc}$, with a ratio of 1.2. For each candidate peak, we then find the largest radius for which a homogeneous ellipsoid with the measured mean strain would collapse by the redshift of interest. If a candidate peak has no radius for which a homogeneous ellipsoid with the measured strain would have collapsed, then that point is thrown out. Each candidate point is then stored as a peak patch at its location with its radius. We then proceed down through the filter bank to all scales and repeat this procedure for each scale, resulting in a list of peak patches which we refer to as the unmerged catalogue.

The next step is to account for exclusion, an essential step to avoid double counting of matter, since distinct halos should not overlap, by definition. We choose here to use binary exclusion (Bond & Myers, 1996). Binary exclusion starts from a ranked list of candidate peak patches sorted by mass or, equivalently, Lagrangian peak patch radius. For each patch we consider every other less massive patch that overlaps it. If the smaller patch is outside of the larger one, then the radius of the two patches is reduced until they are just touching. If the center of the smaller patch is inside the large one, then that patch is removed from the list. This process is repeated until the least massive remaining patch is reached.

Finally, we move halos according to 2LPT using displacements computed at the scale of the halo.

This method is very fast: each realization ran typically in 97 seconds on 64 cores of the GPC supercomputer at the SciNet HPC Consortium in Toronto (1.72 hours in total) . It allows to get accurate – and fast – halo catalogues without any calibration, achieving high precision on the mass function typically for masses above a few 10^{13} $\mathrm{M}_\odot$.

Pinocchio

The PINpointing Orbit Crossing Collapsed HIerarchical Objects (Pinocchio) code (Monaco et al., 2002) is based on the following algorithm.

A linear density contrast field is generated in Fourier space, in a way similar to N-body simulations. As a matter of fact, the code version used here implements the same loop in k-space as the initial condition generator (N-GenIC) used for the simulations, so the same realization is produced just by providing the code with the same random seed. The density is then smoothed using several smoothing radii. For each smoothing radius, the code computes the time at which each grid point ("particle") is expected to get to the highly non-linear regime. The dynamics of grid points, as mass elements, is treated as the collapse of a homogeneous ellipsoid, whose tidal tensor is given by the Hessian of the potential at that point. Collapse is defined as the time at which the ellipsoid collapses on the first axis, going through orbit crossing and into the highly non-linear regime; this

[4]This is not to say that a halo found on a given scale corresponds to a peak in the density smoothed on that scale, however, which is only the case when the strain is isotropic and the collapse is spherical. Thus, the use of density peaks as centers for strain measurements and ellipsoidal collapse calculations in the algorithm is only an optimization, to avoid wasting computations measuring the properties of regions of Lagrangian space that are unlikely to collapse in the first place.

is a difference with respect to PEAK PATCH, where the collapse of extended structures is modelled. The equations for ellipsoidal collapse are solved using third-order Lagrangian Perturbation Theory (3LPT). Following the ideas behind excursion-sets theory, for each particle we consider the earliest collapse time as obtained by varying the smoothing radius.

Collapsed particles are then grouped together using an algorithm that mimics the hierarchical assembly of halos: particles are addressed in chronological order of collapse time; when a particle collapses the six nearest neighbours in the Lagrangian space are checked, if none has collapsed yet then the particle is a peak of the inverse collapse time (defined as $F = 1/D_c$, where $D_c = D(t_c)$ is the growth rate at the collapse time) and it becomes a new halo of one particle. If the collapsed particle is touching (in Lagrangian space) a halo, then both the particle and the halo are displaced using LPT, and if they get "near enough" the particle is accreted to the halo, otherwise it is considered as a "filament" particle, belonging to the filamentary network of particles that have suffered orbit crossing but do not belong to halos. If a particle touches two halos, then their merging is decided by moving them and checking whether they get again "near enough". Here "near enough" implies a parametrization that is well explained in the original papers (see Munari et al., 2017, for the latest calibration). This results in the construction of halos together with their merger histories, obtained with continuous time sampling. Halos are then moved to the final position using 3LPT. The so-produced halos have discrete masses, proportional to the particle mass M_p, as the halos found in N-body simulations. To ease the procedure of number density matching described below in Section 3.6, halo masses were made continuous using the following procedure. It is assumed that a halo of N particles has a mass that is distributed between $N \times M_p$ and $(N+1) \times M_p$, and the distribution is obtained by interpolating the mass function as a power law between two values computed in successive bins of width M_p. This procedure guarantees that the cumulative mass function of halos of mass $> N \times M_p$ does not change, but it does affect the differential mass function.

We use the latest code version presented in Munari et al. (2017), where the advantage of using 3LPT is demonstrated. No further calibration was required before starting the runs. That paper presents scaling tests of the massively parallel version V4.1 and timings. The 300 runs were produced in the GALILEO@CINECA Tier-1 facility, each run required about 8 minutes on 48 cores.

3.5.3 Calibrated methods: HALOGEN, PATCHY

HALOGEN

HALOGEN (Avila et al., 2015) is an approximate method designed to generate halo catalogues with the correct two-point correlation function as a function of mass. It constructs the catalogues following four simple steps:

- Generate a 2LPT dark matter field, and distribute their particles on a grid with cell size $l_{\rm cell}$.
- Draw halo masses M_h from an input Halo Mass Function (HMF).

- Place the halo masses (from top to bottom) in the cells with a probability that depends on the cell density and the halo mass $P \propto \rho_{\rm cell}^{\alpha(M_h)}$. Within cells we choose random particles to assign the halo position. We further ensure mass conservation within cells and avoid halo overlap.

- Assign halo velocities from the particle velocities, with a velocity bias factor: $\mathbf{v}_{\rm halo} = f_{\rm vel}(M_h) \cdot \mathbf{v}_{\rm part}$

Following the study in (Avila et al., 2015), we fix the cell size at $l_{\rm cell} = 5\,h^{-1}$Mpc. In this paper we take the input HMF from the mean of the 300 Minerva simulations, but in other studies analytical HMF have been used. The parameter $\alpha(M_{\rm h})$ controls the clustering as a function of halo mass and has been calibrated using the two-point function from the Minerva simulations in logarithmic mass bins ($M_{\rm h} = 1.06 \times 10^{13}$, 2.0×10^{13}, 4.0×10^{13}, 8.0×10^{13}, $1.6 \times 10^{14}\,h^{-1}M_{\odot}$). The factor $f(M_{\rm h})$ is also tuned to match the variance of the halo velocities from the N-body simulations.

Halogen is a code that advocates for the simplicity and low needs of computing resources. The fact that it does not resolve halos (i.e. using a halo finder), allows to probe low halo masses while keeping low the computing resources. This has the disadvantage of needing to introduce free parameters. However, Halogen only needs one clustering parameter α and one velocity parameter $f_{\rm vel}$, making the fitting procedure simple.

Patchy

The Patchy code (Kitaura et al., 2014, 2015) relies on modelling the large-scale density field with an efficient approximate gravity solver, which is populated with the halo density field using a non-linear, scale dependent, and stochastic biasing description. Although it can be applied to directly paint the galaxy distribution on the density mesh (see Kitaura et al., 2016).

The gravity solver used in this work is based on Augmented Lagrangian Perturbation Theory (ALPT, Kitaura & Heß, 2013), fed with the same initial conditions as those implemented in the Minerva simulations. In the ALPT model, 2LPT is modified by employing a spherical collapse model on small comoving scales, splitting the displacement field into a long and a short range component. Better results can in principle be obtained using a particle mesh gravity solver at a higher computational cost (see Vakili et al., 2017).

Once the dark matter density field is computed, a deterministic bias relating it to the expected number density of halos is applied. This deterministic bias model consists of a threshold, an exponential cut-off, and a power-law bias relation. The number density is fixed by construction using the appropriate normalization of the bias expression.

The Patchy code then associates the number of halos in each cell by sampling from a negative binomial distribution modelling the deviation from Poissonity with an additional stochastic bias parameter.

In order to provide peculiar velocities, these are split into a coherent and a quasi-virialised component. The coherent flow is obtained from ALPT and the dispersion term is sampled from a Gaussian distribution assuming a power law with the local density.

The masses are associated to the halos by means of the HADRON code (Zhao et al., 2015). In this approach, the masses coming from the N-body simulation are classified in different density bins and in different cosmic web types (knots, filaments, sheets and voids) and their distribution information is extracted. Then HADRON uses this information to assign masses to halos belonging to mock catalogues. This information is independent of initial conditions, meaning it will be the same for each of the 300 Minerva realizations.

We used the MCMC python wrapper published by Vakili et al. (2017) to infer the values of the bias parameters from Minerva simulations using one of the 300 random realizations. Once these parameters are fixed one can produce all of the other mock catalogues without further fitting. The PATCHY mocks were produced using a down-sampled white noise of 500^3 instead of the 1000^3 original Minerva ones with an effective cell side resolution of $3\,h^{-1}$Mpc to produce the dark matter field.

3.5.4 Models of the density PDF: Log-normal and Gaussian distribution

Log-normal distribution

The log-normal mocks were produced using the public code presented in Agrawal et al. (2017), which models the matter and halo density fields as log-normal fields, and generates the velocity field from the matter density field, using the linear continuity equation.

To generate a log-normal field $\delta(\boldsymbol{x})$, a Gaussian field $G(\boldsymbol{x})$ is first generated, which is related to the log-normal field as $\delta(\boldsymbol{x}) = e^{-\sigma_G^2+G(\boldsymbol{x})} - 1$ (Coles & Jones, 1991). The pre-factor with the variance σ_G^2 of the Gaussian field $G(\boldsymbol{x})$, ensures that the mean of $\delta(\boldsymbol{x})$ vanishes. Because different Fourier modes of a Gaussian field are uncorrelated, the Gaussian field $G(\boldsymbol{x})$ is generated in Fourier space. The power spectrum of $G(\boldsymbol{x})$ is found by Fourier transforming its correlation function $\xi^G(r)$, which is related to the correlation function $\xi(r)$ of the log-normal field $\delta(\boldsymbol{x})$ as $\xi^G(r) = \ln[1 + \xi(r)]$ (Coles & Jones, 1991). Having generated the Gaussian field $G(\boldsymbol{x})$, the code transforms it to the log-normal field $\delta(\boldsymbol{x})$ using the variance σ_G^2 measured from $G(\boldsymbol{x})$ in all cells.

In practice, we use the measured real-space matter power spectrum from Minerva and Fourier transform it to get the matter correlation function. For halos we use the measured real-space correlation function. We then generate the Gaussian matter and halo fields with the same phases, so that the Gaussian fields are perfectly correlated with each other. Note however, that we use random realizations for these mocks, and so, these phases are not equal to those of the Minerva initial conditions. We then exponentiate the Gaussian fields, to get matter ($\delta_{\rm m}(\boldsymbol{x})$) and halo ($\delta_{\rm h}(\boldsymbol{x})$) density fields, following a log-normal distribution.

The expected number of halos in a cell is given as $N_{\rm h}(\boldsymbol{x}) = \bar{n}_{\rm h}[1 + \delta_{\rm h}(\boldsymbol{x})]V_{\rm cell}$, where $\bar{n}_{\rm h}$ is the mean number density of the halo sample from Minerva, $\delta_{\rm h}(\boldsymbol{x})$ is the halo density at position $\boldsymbol{x}$, and $V_{\rm cell}$ is the volume of the cell. However, this is not an integer. So, to obtain an integer number of halos from the halo density field, we draw a random number from a Poisson distribution with mean $N_{\rm h}(\boldsymbol{x})$, and populate halos randomly within the cell. The log-normal matter field is then used to generate the velocity field using the linear

continuity equation. Each halo in a cell is assigned the three-dimensional velocity of that cell.

Since the log-normal mocks use random phases, we generate 1000 realizations for each mass bin, with the real-space clustering and mean number density measured from Minerva as inputs. Also note, that because halos in this prescription correspond to just discrete points, we do not assign any mass to them. An effective bias relation can still be established using the cross-correlation between the halo and matter fields, or using the input clustering statistics (Agrawal et al. (2017)).

The key advantage of using this method is its speed. Once we had the target power spectrum of the matter and halo Gaussian fields, each realization of a 256^3 grid as in Minerva, was produced in 20 seconds using 16 cores at the RZG in Garching. The resulting catalogues agree perfectly with the Minerva realizations in their real-space clustering as expected. Because we use linear velocities, they also agree with the redshift-space predictions on large scales (Agrawal et al., 2017).

Gaussian distribution

A different approach to generating "mock" halo catalogues with fast approximate methods is to model the covariance matrix theoretically. As mentioned in Section 3.3, this has the advantage that the resulting estimate is free of noise. In this comparison project we included a simple theoretical model for the linear covariance of anisotropic galaxy clustering that is described in Grieb et al. (2016). Based on the assumption that the two-dimensional power spectrum $P(k,\mu)$ follows a Gaussian distribution and that the contributions from the trispectrum and super-sample covariance can be neglected, Grieb et al. (2016) derived the explicit formulae for the covariance of anisotropic clustering measurements in configuration and Fourier space. In particular, they obtain that the covariance between two Legendre multipoles of the correlation function of order ℓ and ℓ' (see Section 3.1) evaluated at the pair separations s_i and s_j, respectively, is given by

$$C_{\ell,\ell'}(s_i,s_j) = \frac{i^{\ell+\ell'}}{2\pi^2}\int_0^\infty k^2\sigma^2_{\ell\ell'}(k)\bar{j}_\ell(ks_i)\bar{j}_{\ell'}(ks_j)\,\mathrm{d}k, \tag{3.12}$$

where $\bar{j}_\ell(ks_i)$ is the bin-averaged spherical Bessel function as defined in equation A19 of Grieb et al. (2016), and

$$\sigma^2_{\ell\ell'}(k) \equiv \frac{(2\ell+1)\,(2\ell'+1)}{V_\mathrm{s}} \times \int_{-1}^{1}\left[P(k,\mu)+\frac{1}{\bar{n}}\right]^2 L_\ell(\mu)\,L_{\ell'}(\mu)\,\mathrm{d}\mu. \tag{3.13}$$

Here, $P(k,\mu)$ represents the two-dimensional power spectrum of the sample, V_s is its volume, and $\bar{n}$ corresponds to its mean number density.

Analogously, the covariance between two configuration-space clustering wedges μ and μ' (see Section 3.1) is given by

$$C_{\mu,\mu'}(s_i,s_j) = \sum_{\ell_1\ell_2}\frac{i^{\ell_1+\ell_2}}{2\pi^2}\bar{L}_{\ell_1,\mu}\bar{L}_{\ell_2,\mu'} \times \int_0^\infty k^2\sigma^2_{\ell_1\ell_2}(k)\bar{j}_{\ell_1}(ks_i)\bar{j}_{\ell_2}(ks_j)\,\mathrm{d}k, \tag{3.14}$$

where $\bar{L}_{\ell_1,\mu}$ represents the average of the Legendre polynomial of order ℓ within the corresponding μ-range of the clustering wedge. The covariance matrices derived from the Gaussian model have been tested against N-body simulations with periodic boundary conditions by Grieb et al. (2016), showing good agreement within the range of scales typically included in the analysis of galaxy redshift surveys ($s > 20\,h^{-1}$Mpc).

3.6 Halo samples

In this section we describe the criteria used to construct the halo samples on which we base our covariance matrix comparison.

We define two parent halo samples from the Minerva simulations by selecting halos with masses $M \geq 1.12 \times 10^{13} h^{-1} M_\odot$ and $M \geq 2.667 \times 10^{13} h^{-1} \mathrm{M}_\odot$, corresponding to 42 and 100 dark matter particles, respectively. We apply the same mass cuts to the catalogues produced by the approximated methods included in our comparison. We refer to the resulting samples as "mass1" and "mass2".

Note that the PATCHY and log-normal catalogues do not have mass information for individual objects and match the number density and bias of the parent samples from Minerva by construction. The Gaussian model predictions are also computed for the specific clustering amplitude and number density as the mass1 and mass2 samples. For the other approximate methods, the samples obtained by applying these mass thresholds do not reproduce the clustering and the shot noise of the corresponding samples from Minerva. These differences are in part caused by the different applied methods for identifying or assigning halos, e.g. PEAK PATCH uses spherical overdensities in Lagrangian space to define halo masses while most other methods are closer to FoF masses, as described in Section 3.5. Therefore, for the ICE-COLA, HALOGEN, PEAK PATCH and PINOCCHIO catalogues we also define samples by matching number density and clustering amplitude of the halo samples from Minerva. For the number-density-matched samples, we find the mass cuts where the total number of halos in the samples drawn from each approximate method best matches that of the two parent Minerva samples. We refer to these samples as "dens1" and "dens2". Analogously, we define bias-matched samples by identifying the mass thresholds for which the clustering amplitude in the catalogues drawn from the approximate methods best agrees with that of the mass1 and mass2 samples from Minerva. More concretely, we define the clustering-amplitude-matched samples by selecting the mass thresholds that minimize the difference between the mean correlation function measurements from the catalogues drawn from the approximate methods and the Minerva parent samples on scales $40\,h^{-1}\,\mathrm{Mpc} < s < 80\,h^{-1}\,\mathrm{Mpc}$. We refer to these samples as "bias1" and "bias2".

The mass thresholds defining the different samples, the number of particles corresponding to these limits, their halo number densities, and bias ratios with respect to the Minerva parent samples are listed in Table 4.1. Note that, as the halo masses of the PINOCCHIO and PEAK PATCH catalogues are made continuous for this analysis, the mass cuts defining

[5] As the halo masses corresponding to our low-mass threshold are not correctly resolved in the PEAK PATCH catalogues, only the high-mass threshold (mass2) is considered in this case.

Table 3.1: Overview of the different samples, including the mass limits, $M_{\rm lim}$, expressed in units of $h^{-1}{\rm M}_\odot$, the corresponding number of particles, $N_{\rm p}$, the mean number density, $\bar{n}$, and the bias ratio to the corresponding Minerva parent sample, $\langle(\xi_{\rm app}/\xi_{\rm Min})^{1/2}\rangle$. The sample names "mass", "dens", and "bias", indicate if the samples were constructed by matching the mass threshold, number density, or clustering amplitude of the parent halo samples from Minerva.

code	sample name	$M_{\rm lim}/\left(h^{-1}{\rm M}_\odot\right)$	N_p	$\bar{n}/\left(h^3{\rm Mpc}^{-3}\right)$	bias ratio
Minerva	mass1	1.12×10^{13}	42	2.12×10^{-4}	1.00
Minerva	mass2	2.67×10^{13}	100	5.42×10^{-5}	1.00
ICE-COLA	mass1	1.12×10^{13}	42	2.06×10^{-4}	0.99
ICE-COLA	dens1	1.09×10^{13}	41	2.12×10^{-4}	0.98
ICE-COLA	bias1	1.17×10^{13}	44	1.93×10^{-4}	1.00
ICE-COLA	mass2	2.67×10^{13}	100	5.81×10^{-5}	0.99
ICE-COLA	dens2, bias2	2.77×10^{13}	104	5.45×10^{-5}	1.00
HALOGEN	mass1, dens1, bias1	1.12×10^{13}	42	2.14×10^{-4}	1.00
HALOGEN	mass2, dens2	2.67×10^{13}	100	5.40×10^{-5}	0.98
HALOGEN	bias2	2.91×10^{13}	109	4.61×10^{-5}	1.00
PEAK PATCH[5]	mass2	2.67×10^{13}	100	4.45×10^{-5}	1.04
PEAK PATCH	dens2, bias2	2.35×10^{13}	88.3	5.44×10^{-5}	1.00
PINOCCHIO	mass1	1.12×10^{13}	42	1.95×10^{-4}	1.02
PINOCCHIO	dens1	1.04×10^{13}	39.1	2.15×10^{-4}	1.00
PINOCCHIO	bias1	1.06×10^{13}	39.9	2.09×10^{-4}	1.00
PINOCCHIO	mass2	2.67×10^{13}	100	5.35×10^{-5}	1.03
PINOCCHIO	dens2	2.63×10^{13}	98.6	5.48×10^{-5}	1.03
PINOCCHIO	bias2	2.42×10^{13}	90.7	6.27×10^{-5}	1.00

the density- and bias-matched samples do not correspond to an integer number of particles. Also note for the calibrated methods that the HALOGEN catalogue was calibrated using the input HMF from the mean of the 300 Minerva simulations in logarithmic mass bins for this analysis, whereas the PATCHY mass samples were calibrated for each mass cut individually. For the case of the HALOGEN catalogue, the selected high mass threshold lies nearly half way (in logarithmic scale) between two of the mass thresholds of the logarithmic input HMF. This explains why whereas for the first mass cut, bias and number density are matched by construction, that is not the case for the second mass cut. This has the effect that the bias2 sample of the HALOGEN catalogue has 15% fewer halos than the corresponding Minerva sample. Comparisons of the ratios of the number densities and bias of the different samples drawn from the approximate methods to the corresponding ones from Minerva are shown in Fig. 3.1. Since the catalogues drawn from log-normal and PATCHY match the number density and bias of the Minerva parent samples by construction, they are not included in the Table and figures.

In the following we refer to all samples corresponding to the first mass limit, mass1, dens1 and bias1, as "sample1", and the samples corresponding to the second mass limit, mass2, dens2 and bias2, as "sample2".

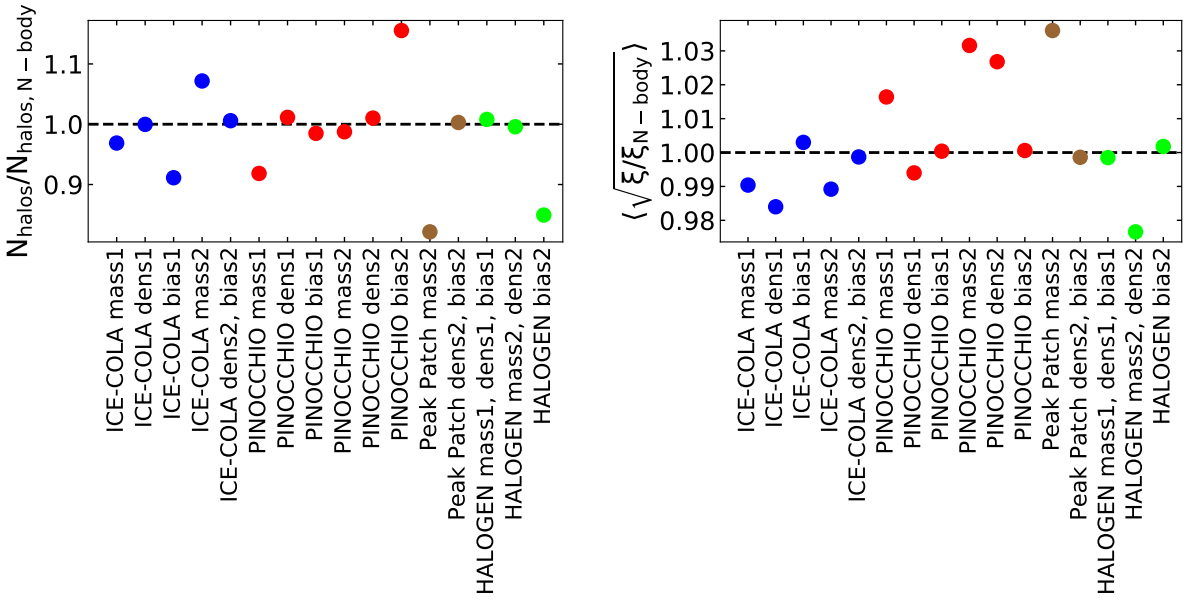

Figure 3.1: Ratios of the total halo number (left panel) and the clustering amplitude (right panel) of samples drawn from the approximate methods to the corresponding quantity in the Minerva parent samples. By definition, for the dens samples the halo number is matched to the corresponding N-body samples and therefore the corresponding ratio is close to one in the left panel, while the ratios from the bias samples are meant to be close to one in the right panel. For some cases two or three samples are represented with the same symbol, e.g. ICE-COLA dens2, bias2 which means that the ICE-COLA dens2 sample is the same as the ICE-COLA bias2 sample.

3.7 Methodology for performance tests

In order to asses the impact of using approximate methods to estimate $\mathbf{C}$, we perform fits based on the Gaussian likelihood approach of Section 3.2. The first step is the measurement of the multipoles and clustering wedges of the two-point correlation function of all the samples described in Section 3.6. For the halo samples of the reference N-body simulations, of the predictive and calibrated methods we compute the corresponding covariance matrices according to equation (3.8) with $N_s = 300$, for the log-normal samples with $N_s = 1000$. Since we only use the anisotropic information of the clustering measurements for the fits, for each halo sample we average the three separate estimates of $\mathbf{C}$ that can be obtained by treating each axis of the simulation boxes as the line-of-sight direction. This reduces the noise due to the small number of realizations in the final covariance estimates.

The aim of our performance tests is to compare the constraints obtained when $\mathbf{C}$ is estimated from the approximate methods described in Sec 3.5 to the corresponding results from N-body simulations. We focus on the cosmological information that can be recovered from fitting procedure described in Section 3.2 and is encoded in $\alpha_\perp$, $\alpha_\parallel$ and $f\sigma_8(z)$. This analysis set-up also matches that of the covariance matrix comparison in Fourier space of

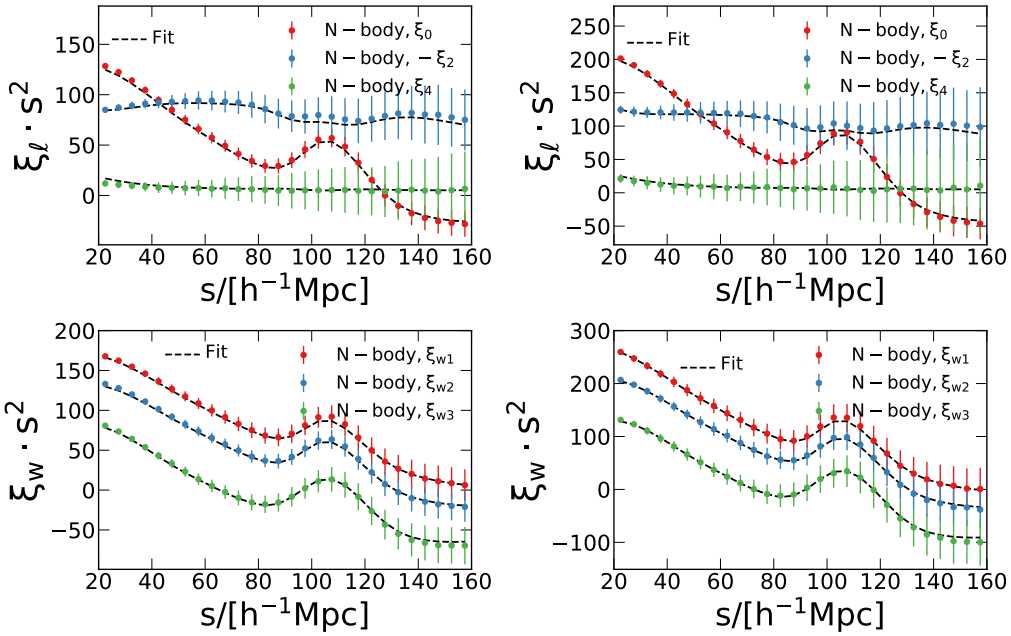

Figure 3.2: Comparison of the mean correlation function multipoles (upper panels) and clustering wedges (lower panels) of the mass1 and mass2 samples (left and right panels, respectively) drawn from our Minerva N-body simulations, and the model described in Section 3.2. The points with error bars show the simulation results and the dashed lines correspond to the fit to these measurements. The error bars on the measurements correspond to the dispersion inferred from the 300 Minerva realizations. In all cases, the model predictions show good agreement with the N-body measurements.

our companion paper (Blot et al., 2019).

In order to ensure that the model used for the fits has no impact on the covariance matrix comparison, we do not fit the measurements of the Legendre multipoles and wedges obtained from the N-body simulations. Instead, we use our baseline model of Section 3.2 to construct synthetic clustering measurements, which we then use for our fits. For this, we first fit the mean Legendre multipoles measured from the parent Minerva halo samples using our model and the N-body covariance matrices. We fix all cosmological parameters to their true values and only vary the bias parameters b_1, b_2, and γ_3^-. We then use the mean values of the parameters inferred from the fits, together with the true values of the cosmological parameters, to generate multipoles and clustering wedges of the correlation function using our baseline model. Fig. 3.2 shows the mean multipoles and clustering wedges measured from the Minerva halo sample for both mass cuts and the resulting fits. In all cases, our model gives a good description of the simulation results. The parameter values recovered from these fits were also used to compute the input power spectra when computing the Gaussian predictions of $\mathbf{C}$. As these synthetic data are perfectly described by our baseline model by construction, their analysis should recover the true values of the BAO parameters $\alpha_\parallel = \alpha_\perp = 1.0$, and the growth-rate parameter $f\sigma_8 = 0.4402$.

The comparison of the parameter values and their uncertainties recovered using different covariance matrices allows us to test the ability of the approximate methods described in Section 3.5 to reproduce the results obtained when $\mathbf{C}$ is inferred from full N-body simulations.

3.8 Comparison of correlation function measurements

In order to estimate the covariance matrices from all the samples introduced in Section 3.6, we first measure configuration-space Legendre multipoles and clustering wedges for each sample and in each realization as described in Section 3.1.

As an illustration of the agreement between the clustering measurements obtained from the approximate methods and the Minerva simulations, we focus here on two cases: i) the multipoles of the density-matched samples for the first mass cut (dens1 samples), and ii) the clustering wedges of the bias-matched samples for the second mass cut (bias2 samples). As described in Section 3.6, for PATCHY and the log-normal realizations, the density- and bias-matched samples are identical to the mass-matched samples by construction.

The upper panel of Fig. 3.3 shows the mean multipole measurements from all realizations for the dens1 samples obtained from the predictive methods ICE-COLA and PINOCCHIO (left panels) and the calibrated methods HALOGEN, PATCHY, and the log-normal recipe (right panels). The predictive methods are in excellent agreement with the measurements from the Minerva parent sample, showing only differences of less than 3% for the ICE-COLA monopole measurements on scales $< 40\,h^{-1}$Mpc. The monopole measurements obtained from the calibrated methods and the log-normal model are also in good agreement with the results from Minerva. However, the quadrupole and hexadecapole measurements obtained from HALOGEN and the log-normal samples exhibit deviations of more than 20% on scales $< 60\,h^{-1}$Mpc.

The lower panel of Fig. 3.3 shows the mean wedges measurements from all realizations for the bias2 samples obtained from the predictive methods ICE-COLA, PINOCCHIO and PEAK PATCH (left panels), and for the corresponding samples obtained from calibrated methods HALOGEN, PATCHY, and the log-normal recipe (right panels). Here we find that the measurements obtained from the predictive methods and the log-normal model agree well within the error bars with the corresponding Minerva measurements. We notice that the strongest deviations are present in the measurements of the transverse and parallel wedge from the HALOGEN samples, of up to 6% and 20% respectively on scales $< 60\,h^{-1}$Mpc. The measurements recovered from PATCHY show deviations ranging between 5% to 10% on small scales.

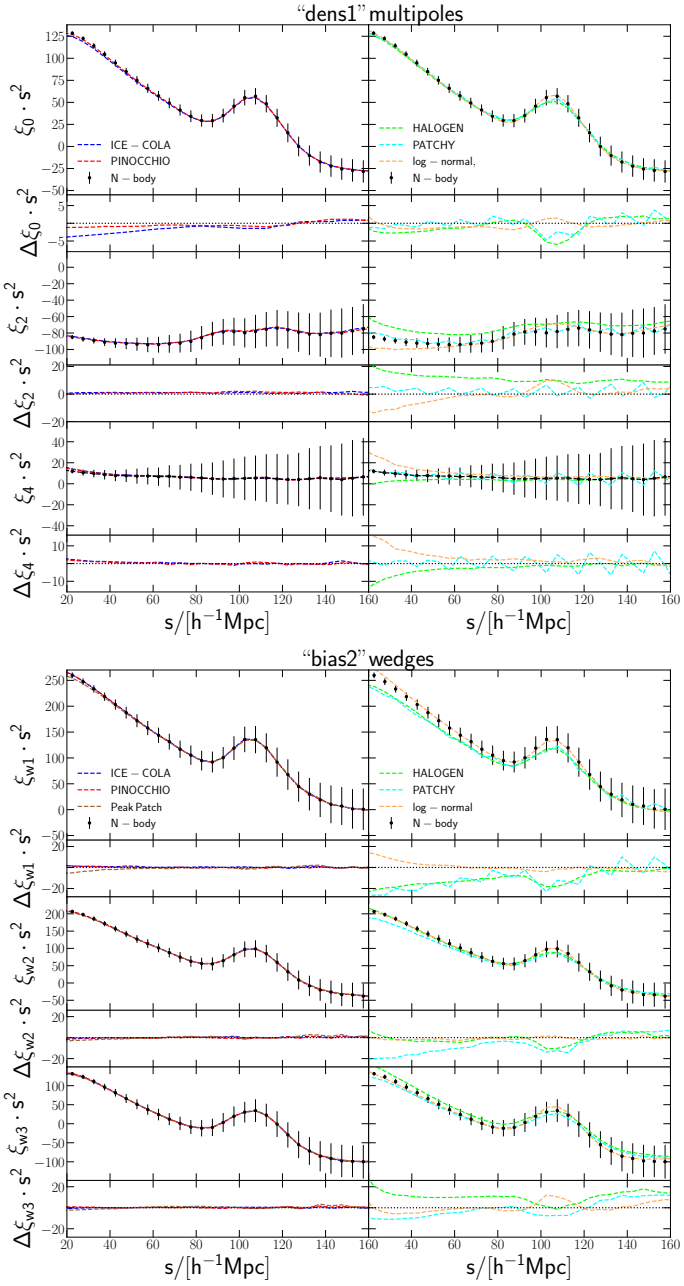

Figure 3.3: *Upper panel*: measurements of the mean multipoles for the density matched samples for the first mass cut (dens1 samples). The first, third and fifth row show the monople, quadrupole and hexadecapole, respectively. *Lower panel*: measurements of the mean clustering wedges for the bias matched samples for the second mass cut (bias2 samples). The first, third and fifth row show the transverse, intermediate and parallel wedge, respectively. Comparison of the measurements drawn from the results of the predictive methods ICE-COLA and PINOCCHIO (*left panels*) and the calibrated methods HALOGEN and PATCHY and the log-normal model (*right panels*) to the corresponding N-body parent sample. The error bars correspond to the dispersion of the results inferred from the 300 N-body catalogues. The remaining rows show the difference of the mean measurements drawn from the results of the approximate methods to the corresponding N-body measurement.

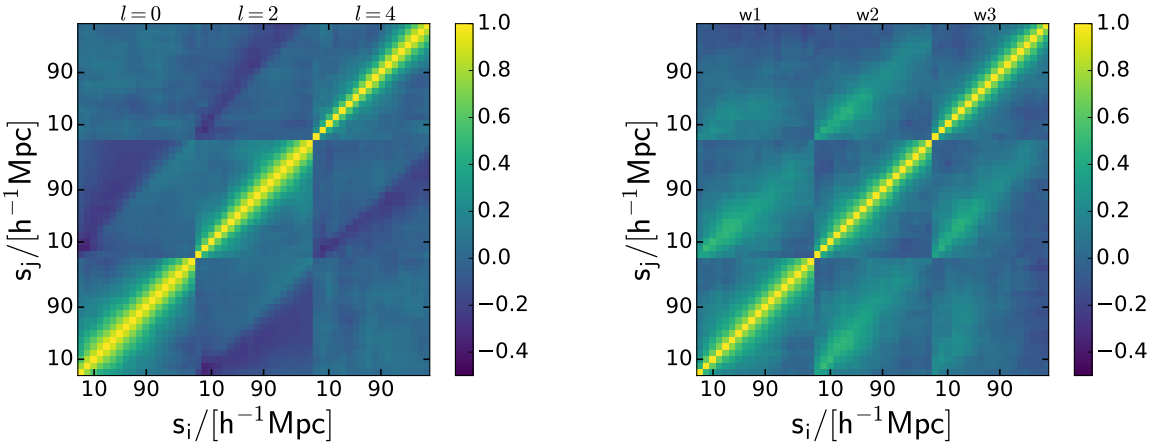

Figure 3.4: The full correlation matrix inferred from the multipoles of the N-body parent sample for the low-mass cut (mass1, left panel) and from the clustering wedges of the mass2 N-body parent sample (right panel).

3.9 Comparison of covariance matrix measurements

In this section we focus on the comparison of the covariance matrix estimates obtained from the different approximate methods, which we computed as described in Sections 3.3 and 3.7.

The structure of the off-diagonal elements of **C** of Legendre multipoles and clustering wedges measurements can be more clearly seen in the correlation matrix, defined as

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \tag{3.15}$$

Fig. 3.4 shows the correlation matrices of the multipoles inferred from the mass1 halo samples from Minerva (left panel) and the wedges of the mass2 samples (right panel).

The estimates of **R** obtained from the approximate methods are indistinguishable by eye from the ones inferred from the Minerva parent samples and therefore not shown here. Instead, we compare the variances and cuts through the correlation function matrices derived from the different samples. Fig. 3.5 shows the ratios of the variances drawn from the approximate methods with respect to those of the corresponding Minerva parent catalogues. We focus here on the same example cases as in Section 3.8: the multipoles measured from the dens1 samples, and the clustering wedges measured from the bias2 samples. We notice that in both cases the predictive methods perform better than the calibrated schemes and the PDF-based recipes. On average, the variance from Minerva is recovered within 10%, with a maximum difference of 20% for the variance of the monopole inferred from the PINOCCHIO dens1 sample at scales around $80\,h^{-1}$Mpc. The variances recovered from the other methods show larger deviations, in some cases up to 40%.

Fig. 3.6 shows cuts through the correlation matrix at $s_j = 105\,h^{-1}$Mpc for the same two example cases. The error bars for the measurements of the corresponding Minerva

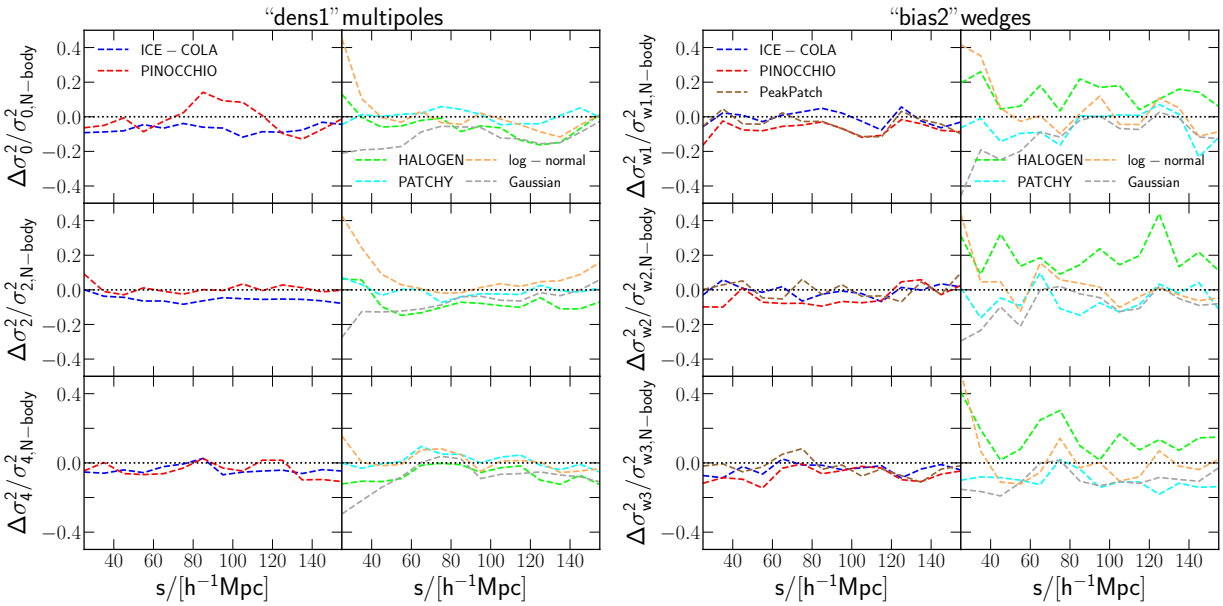

Figure 3.5: *Left panel*: Relative variance of the multipoles of the correlation function measurements from the density matched samples for the first mass cut (dens1 samples).The first, third and fifth row show the measurements for monopole, quadrupole and hexadecapole, respectively. *Right panel*: Relative variance of the clustering wedges of the two-point correlation function for the bias matched samples for the second mass cut (bias2 samples). The first, third and fifth row show the measurements for transverse, intermediate and parallel wedge, respectively. Comparison of the relative variance drawn from the results of the predictive methods ICE-COLA, PINOCCHIO, PEAK PATCH (*left*) and HALOGEN, PATCHY and the log-normal model (*right*) to the corresponding N-body parent sample.

parent samples are obtained from a jackknife estimate using the 300 Minerva mocks,

$$(\Delta M_{ij})^2 = \frac{N_\mathrm{S}-1}{N_\mathrm{S}} \sum_\mathrm{S} (M_{ij}^{(s)} - M_{ij})^2, \tag{3.16}$$

where $\mathbf{M}$ is the covariance matrix $\mathbf{C}$ or the correlation matrix $\mathbf{R}$ (for Fig. 3.6 we use $\mathbf{R}$). $\mathbf{M}^{(s)}$ is the covariance or correlation matrix which is obtained when leaving out the $s-$th realization,

$$M_{ij}^{(s)} = \frac{1}{N_\mathrm{S}-1} \sum_{r\neq s} (\xi_i^{(r)} - \bar{\xi}_i)(\xi_j^{(r)} - \bar{\xi}_j). \tag{3.17}$$

For the comparison of the cuts through the correlation matrices, all methods agree well with the corresponding N-body measurements with only very small differences. In order to quantify the discrepancies between the covariance and correlation matrices drawn from the approximate methods to the corresponding N-body measurements, we use a χ^2 approach.

Concretely, we compute χ^2 as

$$\chi^2 = \sum_i \sum_{j \geq i} \frac{(C_{ij,\text{approx}} - C_{ij,\text{Minerva}})^2}{\Delta C^2_{ij,\text{Minerva}}}, \tag{3.18}$$

and

$$\chi^2 = \sum_i \sum_{j > i} \frac{(R_{ij,\text{approx}} - R_{ij,\text{Minerva}})^2}{\Delta R^2_{ij,\text{Minerva}}}, \tag{3.19}$$

where the indices i and j run over the bins corresponding to the range of interest of $20 - 160\,h^{-1}$Mpc and $\Delta\ \mathbf{C}_{\text{Minerva}}$ and $\Delta\ \mathbf{R}_{\text{Minerva}}$ are the estimated errors from equation (3.16). If the approximate methods perfectly reproduce the expected covariances from the N-body simulations, the χ^2 obtained from the approximate methods should be $\chi^2 \approx 0$ for the predictive and calibrated methods. This is due to the fact that the simulation boxes of the predictive and calibrated methods match the initial conditions of Minerva and therefore the properties of the noise in the estimates of $\mathbf{C}$ should be very similar. For the covariance and correlation matrices obtained from the PDF-based predictions, we expect $\chi^2 \approx N(N-1)/2$ where N is the number of bins of the covariance or correlation matrix, since these predictions do not correspond to the same initial conditions. In table 3.2 we list the obtained relative χ^2-values,

$$\chi^2_{\text{rel}} = \frac{\chi^2}{N(N-1)/2}, \tag{3.20}$$

where $N = 42$, for all considered samples and clustering statistics. We notice that the χ^2-values are in most cases smaller for the predictive than the calibrated methods. Furthermore, the χ^2-values from the wedge measurements are overall smaller than the corresponding ones from the multipole measurements. Also, in most cases the χ^2-values obtained from the covariance matrices are slightly larger than the corresponding ones from the correlation matrices, indicating discrepancies in the variances obtained from the approximated methods.

The computed χ^2-values do not take the covariance between the different entries of $\mathbf{C}$ into account. In order to provide a more complete picture of how far the multipole and wedges distributions characterized by the different covariance matrices are, we also compute the Kullback-Leibler divergence (Kullback & Leibler, 1951; O'Connell et al., 2016). In our case (two multivariate normal distributions with the same means), the Kullback-Leibler divergence is given as

$$\begin{aligned} D_{\text{KL}}(\mathbf{C}_{\text{Minerva}} \parallel \mathbf{C}_{\text{approx}}) = &\frac{1}{2}\left(\text{tr}(\mathbf{C}^{-1}_{\text{approx}}\mathbf{C}_{\text{Minerva}})\right. \\ &\left. + \ln\left(\frac{\det \mathbf{C}_{\text{approx}}}{\det \mathbf{C}_{\text{Minerva}}}\right) - N\right). \end{aligned} \tag{3.21}$$

If the approximate methods perfectly reproduce the expected distributions from the N-body simulations, including the same noise, we expect $D_{\text{KL}} \approx 0$. In table 3.2 we list the obtained D_{KL} values. We find that the values for D_{KL} are closer to zero for the

Table 3.2: Values of the relative χ^2 for the covariance matrices $\mathbf{C}$ (equation 3.18), correlation matrices $\mathbf{R}$ (equation 3.19) and values for the the Kullback-Leibler divergence D_{KL} (equation 3.21) obtained from the approximate methods.

code	sample	χ^2_{rel} for $\mathbf{C}$ from ξ_{024}	χ^2_{rel} for $\mathbf{C}$ from ξ_{w}	χ^2_{rel} for $\mathbf{R}$ from ξ_{024}	χ^2_{rel} for $\mathbf{R}$ χ^2_{rel} from ξ_{w}	D_{KL} for ξ_{024}	D_{KL} for ξ_{w}
ICE-COLA	mass1	0.19	0.21	0.17	0.16	0.24	0.24
ICE-COLA	dens1	0.31	0.42	0.17	0.15	0.28	0.27
ICE-COLA	bias1	0.20	0.11	0.19	0.19	0.27	0.27
PINOCCHIO	mass1	0.48	0.51	0.27	0.26	0.33	0.33
PINOCCHIO	dens1	0.76	0.67	0.78	0.70	0.77	0.77
PINOCCHIO	bias1	0.23	0.20	0.24	0.22	0.28	0.29
HALOGEN	mass1	1.22	0.90	1.09	0.77	1.28	1.14
PATCHY	mass1	0.67	0.40	0.73	0.44	0.82	0.79
Gaussian	mass1	2.50	2.20	2.04	0.91	0.82	1.08
log-normal	mass1	1.76	1.09	1.31	0.97	0.96	0.98
ICE-COLA	mass2	0.40	0.36	0.38	0.33	0.43	0.45
ICE-COLA	dens2	0.36	0.23	0.35	0.27	0.28	0.28
PINOCCHIO	mass2	1.03	1.20	0.44	0.41	0.46	0.44
PINOCCHIO	dens2	0.81	0.83	0.44	0.40	0.41	0.40
PINOCCHIO	bias2	0.70	0.31	0.42	0.54	0.41	0.73
PEAK PATCH	mass2	1.84	2.02	0.69	0.69	1.05	1.03
PEAK PATCH	dens2	0.48	0.47	0.48	0.45	0.46	0.48
HALOGEN	mass2	1.77	1.32	1.70	1.29	1.07	1.07
HALOGEN	bias2	2.24	1.76	2.06	1.59	1.28	1.32
PATCHY	mass2	1.41	1.26	1.21	0.97	0.99	1.01
Gaussian	mass2	2.02	1.77	1.75	1.03	0.78	1.14
log-normal	mass2	2.27	2.57	1.64	1.88	1.02	1.07

predictive than for the other approximate methods. For the calibrated methods and for the distributions with different noise, obtained from the Gaussian and log-normal models, we find values $D_{\mathrm{KL}} \approx 1$.

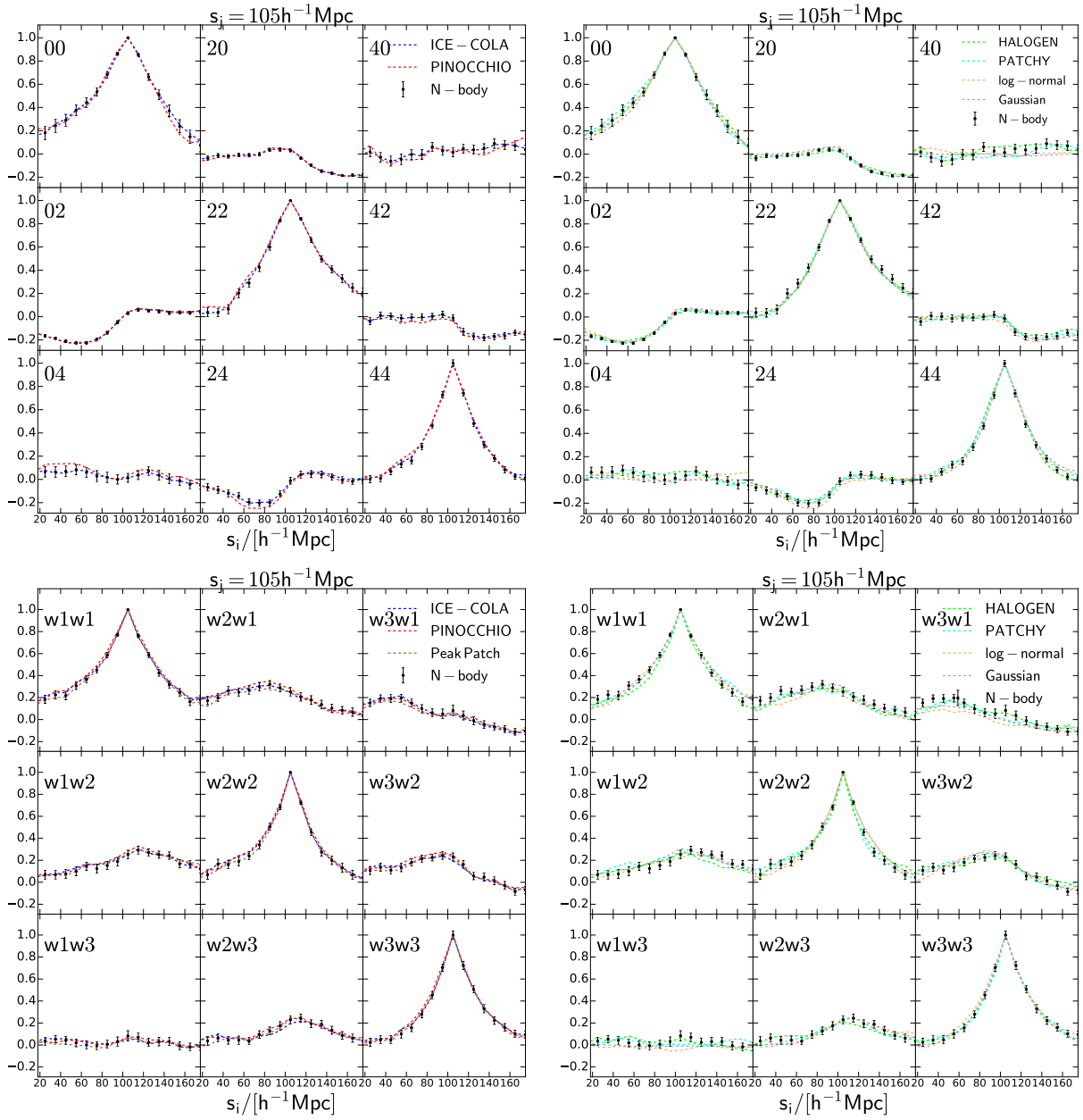

Figure 3.6: Cuts at $s_j = 105\,h^{-1}$Mpc through the correlation matrices for the two example cases drawn from the results of the approximate methods and the corresponding N-body parent sample. *Upper, left panel*: Correlation matrices measured from the multipoles of the correlation function drawn from dens1 samples from the predictive methods ICE-COLA and PINOCCHIO. *Upper, right panel*: Correlation matrices measured from the multipoles of the correlation function drawn from dens1 samples from the calibrated methods HALOGEN and PATCHY and the Gaussian and log-normal recipes. *Lower, left panel*: Correlation matrices measured from the clustering wedges of the correlation function drawn from the bias2 samples from the predictive methods ICE-COLA, PINOCCHIO and PEAK PATCH. *Lower, right panel*: Correlation matrices measured from the clustering wedges of the correlation function drawn from bias2 samples from the calibrated methods HALOGEN and PATCHY and the Gaussian and log-normal recipes. The error bars are obtained from a jackknife estimate using the 300 Minerva realizations.

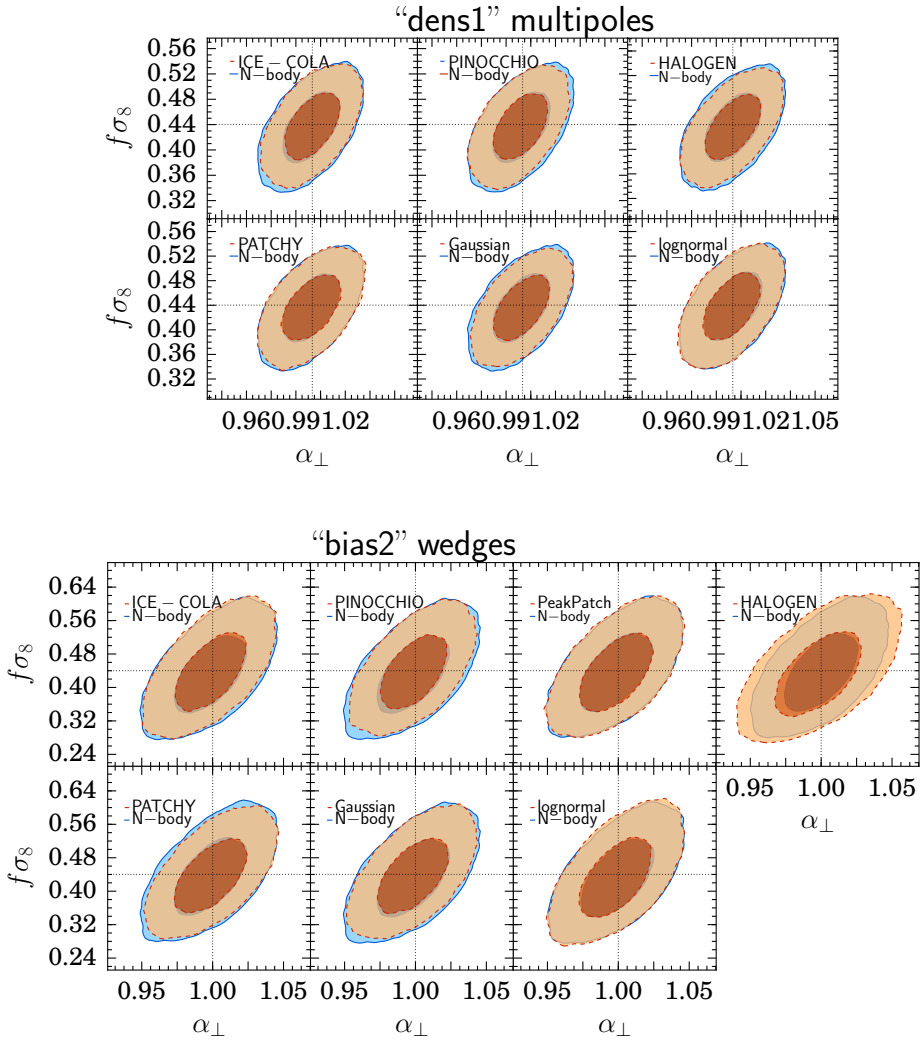

Figure 3.7: Comparison of the marginalised two-dimensional constraints in the $\alpha_\perp$-$f\sigma_8$ plane for the analysis of samples from the approximate methods with the corresponding constraints obtained from analysis of the parent Minerva sample. The contours correspond to the 68% and 95% confidence levels. *Upper panel*: Results from the analysis of the multipoles measured from the dens1 samples. *Lower panel*: Results from the analysis of the clustering wedges measured from the bias2 samples.

3.10 Performance of the covariance matrices

For the final validation of the covariance matrices inferred from the different approximate methods, we analyse their performance on cosmological parameter constraints. We perform fits to the synthetic clustering measurements described in Section 3.7, using the estimates of $\mathbf{C}$ obtained from the different halo samples and approximate methods. We focus on the constraints on the BAO shift parameters $\alpha_\parallel$, $\alpha_\perp$, and the growth rate $f\sigma_8$.

Fig. 3.7 shows the two-dimensional marginalised constraints in the $\alpha_\perp$-$f\sigma_8$ plane for the analysis of our two examples cases, the Legendre multipoles measured from the dens1 samples (upper panels), and the clustering wedges recovered from the bias2 samples (lower panel).

In general, the allowed regions for these parameters obtained using the estimates of $\mathbf{C}$ inferred from the different approximate methods (shown by the solid lines) agree well with those obtained using the covariance matrices from Minerva (indicated by the dotted lines in all panels). However, most cases exhibit small deviations, either slightly under- or over-estimating the statistical uncertainties. We find that, for all samples and clustering statistics, the mean parameter values inferred using approximate methods are in excellent agreement with the ones from the corresponding N-body analysis, showing differences that are much smaller than their associated statistical errors. The parameter uncertainties recovered using covariances from the approximate methods show differences with respect to the N-body constraints ranging between 0.3% and 8% for the low mass samples, while most of the results agree within 5% with the N-body results, and between 0.1% and 20% for the high-mass cut, while most of the results agree within 10% with the N-body results. For the comparison of the obtained parameter uncertainties it is important to point out that in our companion paper Blot et al. (2019) estimate that the statistical limit of our parameter estimation is about 4% to 5%. Fig. 3.8 shows the ratios of the marginalised parameter errors drawn from the analysis with the different approximate methods with respect to the N-body results. We observe that for the samples corresponding to the first mass cut, all methods reproduce the N-body errors within 10% for all parameters, and in most cases within 5% corresponding to the statistical limit of our analysis. For the samples corresponding to the second mass cut also most methods reproduce the N-body errors within 10% with exception of the Peak Patch mass-matched and the Halogen bias-matched samples. This might be due to the fact that these two samples have 15-20% less halos than the corresponding N-body sample.

In order to evaluate the parameter errors, we use the volume of the allowed region in the three-dimensional parameter space of $\alpha_\parallel$, $\alpha_\perp$ and $f\sigma_8$, which can be estimated as

$$V = \sqrt{\det \mathrm{Cov}(\alpha_\parallel, \alpha_\perp, f\sigma_8)} \tag{3.22}$$

where $\det \mathrm{Cov}(\alpha_\parallel, \alpha_\perp, f\sigma_8)$ is the determinant of the parameter covariance matrix. For a Gaussian posterior distribution, the allowed volume is proportional to the volume enclosed by the three-dimensional 68% C.L. contour. This definition is similar to the two-dimensional Dark Energy Task Force figure of merit of the dark-energy equation-of-state

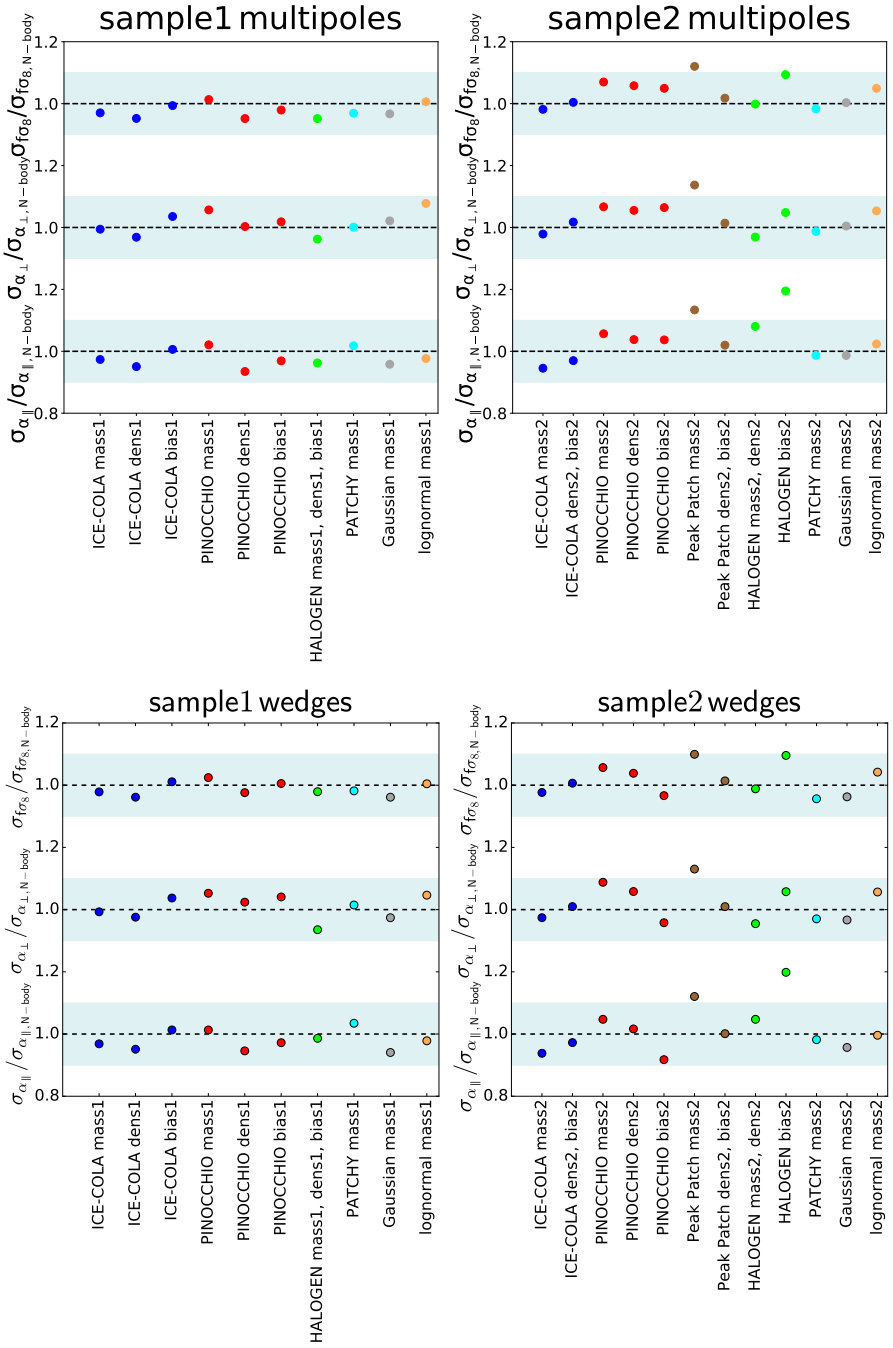

Figure 3.8: Comparison of the marginalised error on the parameters $\alpha_\parallel$, $\alpha_\perp$ and $f\sigma_8$ which are obtained from the analysis using the covariance matrices from the approximate methods to the corresponding ones from the N-body catalogues. The light grey band indicates a range of $\pm 10\%$ deviation from a ratio equal to 1. The different panels show the results obtained from the analysis of *upper, left panel*: the multipoles drawn from the samples corresponding to the first N-body parent sample with the lower mass cut, *upper, right panel*: the multipoles drawn from the samples corresponding to the second N-body parent sample with the higher mass cut, *lower, left panel*: the wedges drawn from the samples corresponding to the first N-body sample, *lower, right panel*: the wedges drawn from the samples corresponding to the second N-body sample.

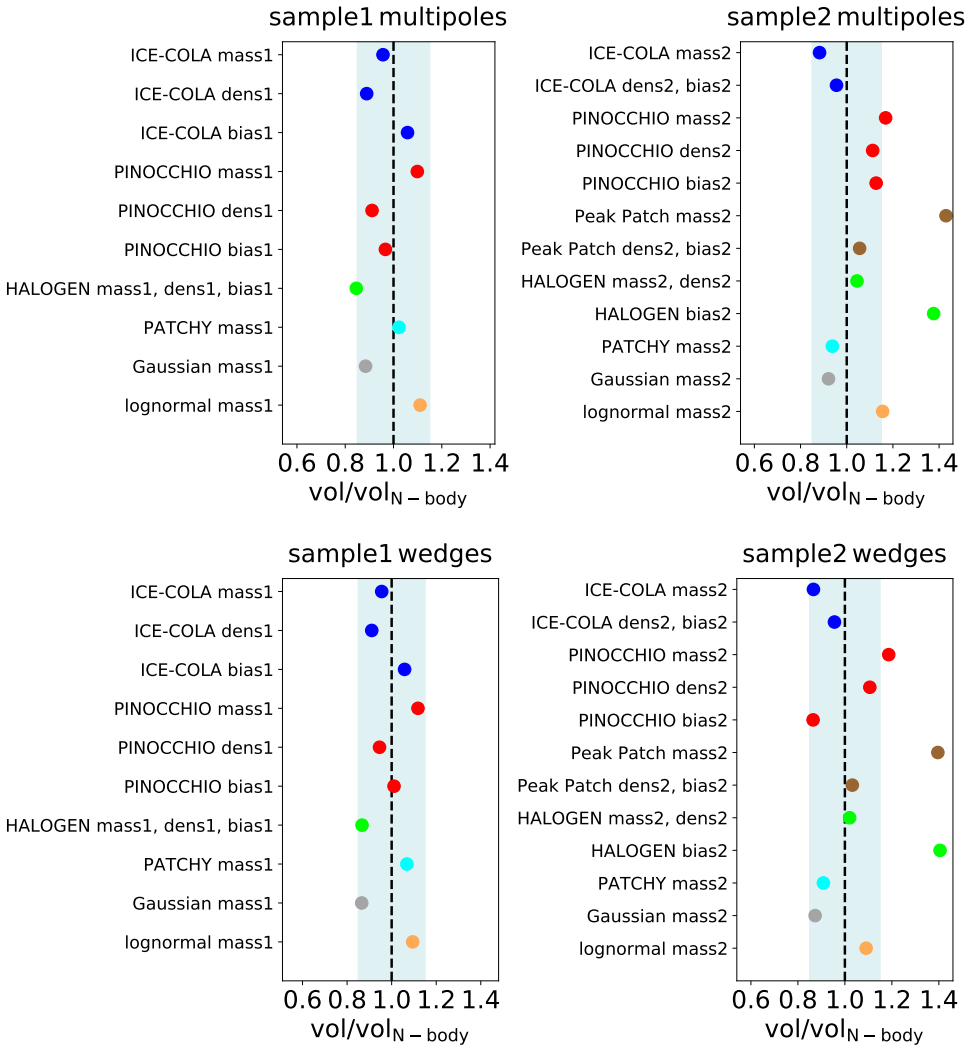

Figure 3.9: Comparison of the volume ratios between the allowed statistical volumes obtained from the analysis using the covariance matrices from the approximate methods to the corresponding ones from the N-body catalogues. The light grey band indicates a range of $\pm 10\%$ deviation from a ratio equal to 1. The different panels show the results obtained from the analysis of *upper, left panel*: the multipoles drawn from the samples corresponding to the first N-body parent sample with the lower mass cut, *upper, right panel*: the multipoles drawn from the samples corresponding to the second N-body parent sample with the higher mass cut, *lower, left panel*: the wedges drawn from the samples corresponding to the first N-body sample, *lower, right panel*: the wedges drawn from the samples corresponding to the second N-body sample.

parameters w_0–w_a (Wang, 2008; Albrecht et al., 2006), but without taking the inverse of the allowed volume. The ratios of the allowed volumes obtained from the analysis with the different approximate methods and the N-body results are shown in Fig. 3.9. Here the differences in the performance of the methods become clearer. For the first mass cut we notice that most approximate methods can reproduce the N-body volume at a 10% level, with the exception of HALOGEN and the Gaussian and log-normal models, which lead to slightly worse results and show 10%–15% agreement. For the second mass cut we find overall larger differences between the samples. The results from the majority of the samples agree within 10% with the N-body results, the rest shows differences of 10%–15%, and for the PEAK PATCH mass2 and HALOGEN bias2 samples differences of up to 40%. For both mass cuts, we find significant differences in the performances of samples drawn from the same approximate method but using different selection criteria.

3.11 Discussion

In this section we discuss our results on the allowed parameter space volumes obtained in Section 3.10. Fig. 3.9 clearly shows that there are significant differences in the volume ratios between samples drawn from the same approximate method when applying different selection criteria to define the halo catalogues. Matching the parent samples from Minerva by mass limit, number density or bias can lead to differences of up to 20% on the obtained results.

For each approximate method, mass limit, and clustering statistic, we identified the best selection criteria for matching to the N-body parent samples. As discussed in Section 3.6, for PATCHY, log-normal and the Gaussian model we only have samples characterized by the same mass cuts as the N-body catalogues. The best cases in decreasing order of the accuracy with which the results of the N-body covariances are reproduced are:

- Lower mass cut, Legendre multipoles: PATCHY ($V/V_{Min} = 1.02$), PINOCCHIO bias matched ($V/V_{Min} = 0.97$), ICE-COLA mass matched ($V/V_{Min} = 0.96$), log-normal ($V/V_{Min} = 1.11$), Gaussian ($V/V_{Min} = 0.88$), HALOGEN mass, density, bias matched ($V/V_{Min} = 0.85$)

- Lower mass cut, clustering wedges: PINOCCHIO bias matched ($V/V_{Min} = 1.01$), ICE-COLA mass matched ($V/V_{Min} = 0.96$), PATCHY ($V/V_{Min} = 1.07$), log-normal ($V/V_{Min} = 1.09$), HALOGEN mass, density, bias matched ($V/V_{Min} = 0.87$), Gaussian ($V/V_{Min} = 0.87$)

- Higher mass cut, Legendre multipoles: ICE-COLA density matched ($V/V_{Min} = 0.96$), HALOGEN mass, density matched ($V/V_{Min} = 1.04$), PEAK PATCH density, biased matched ($V/V_{Min} = 1.06$), PATCHY ($V/V_{Min} = 0.94$), Gaussian ($V/V_{Min} = 0.92$), PINOCCHIO density matched ($V/V_{Min} = 1.11$), log-normal ($V/V_{Min} = 1.16$)

- Higher mass cut, clustering wedges: HALOGEN mass, density matched ($V/V_{Min} = 1.02$), ICE-COLA density matched ($V/V_{Min} = 0.97$), PEAK PATCH density, biased

matched ($V/V_{Min} = 1.03$), PATCHY ($V/V_{Min} = 0.91$), log-normal ($V/V_{Min} = 1.09$), PINOCCHIO density matched ($V/V_{Min} = 1.1$), Gaussian ($V/V_{Min} = 0.87$)

For a better illustration, Fig. 3.10 shows the two-dimensional marginalised constraints on $\alpha_\perp$ and $f\sigma_8$ obtained from the Legendre multipoles for the low (upper panels) and high (lower panels) mass limits. The different panels show the results obtained from the different approximate methods when the best selection criteria for each case is implemented. The overall agreement with the results derived from the N-body covariances is better in this case than when the same definition is applied to all methods.

The best strategy to define the halo samples for a given approximate method is often different for our two mass limits. For example, considering the results from PINOCCHIO, while for our first mass limit the bias-matched halo samples lead to the best agreement with the constraints inferred from the N-body covariances, for the second mass threshold the density-matched samples provide a better performance. Focusing on the results from the multipole analysis, we observe that for the first mass limit PATCHY, ICE-COLA and PINOCCHIO perform slightly better than the other methods. These methods reproduce the statistical volume of the allowed parameter regions obtained using the N-body covariances within 5% while the other methods only reach a 10%-15% agreement. For the second mass limit ICE-COLA, HALOGEN and PEAK PATCH can reproduce the N-body results within 5%, PATCHY and the Gaussian model within 10%, and PINOCCHIO and the log-normal model within 15%. It is also interesting to note that the order of performance of the methods is slightly different for the multipole and the wedges analysis. For example, the multipole analysis using the PATCHY covariance matrix leads to a better than 2% agreement with the N-body results, whereas the wedge analysis only reaches 7%.

Our analysis is part of a general comparison project of approximate methods involving also the covariances of power spectrum and bispectrum measurements (Blot et al., 2019; Colavincenzo et al., 2019). The power spectrum analysis of Blot et al. (2019) is more closely related to the one presented here, as it is based on the same baseline model of the two-dimensional power spectrum and explore constraints on the same nuisance and cosmological parameters. The bispectrum covariance analysis of Colavincenzo et al. (2019) is different in terms of the model and the parameter constraints included in the comparison. Both of our companion papers consider the same approximate methods and mass cuts used here, but focus on the abundance-matched samples. A comparison of the results of the three studies shows that the differences between the predictive, calibrated and PDF-based approximate methods are less evident for the correlation function analysis than for the power spectrum and bispectrum. This can be clearly seen by comparing the variations of the statistically allowed volumes recovered from the different approximate methods when applied to the correlation function, power spectrum and bispectrum covariances. Since our companion papers focus on the density-matched samples, we also show the allowed volumes only for the "dens" samples in Fig. 3.11. The differences between the approximate methods are less evident in configuration space, become more evident for the power spectrum and are strongest for the bispectrum analysis.

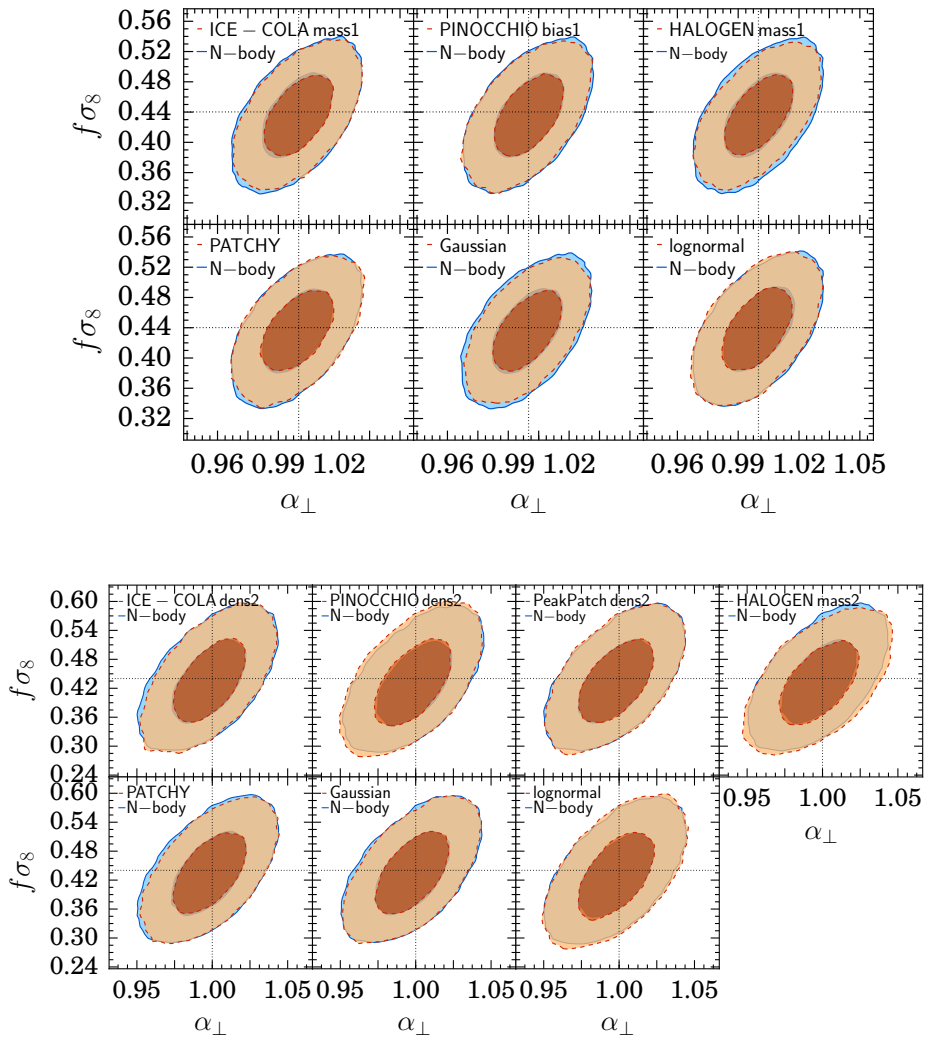

Figure 3.10: Comparison of the marginalised two-dimensional constraints in the $\alpha_\perp$-$f\sigma_8$ plane for the multipole analysis using the best choice of matching for each approximate method individually to the corresponding constraints obtained from the N-body analysis analysis. The contours correspond to the 68% and 95% confidence levels. *Upper panel*: Results for the samples corresponding to the first mass cut. *Lower panel*: Results for the samples corresponding to the second mass cut.

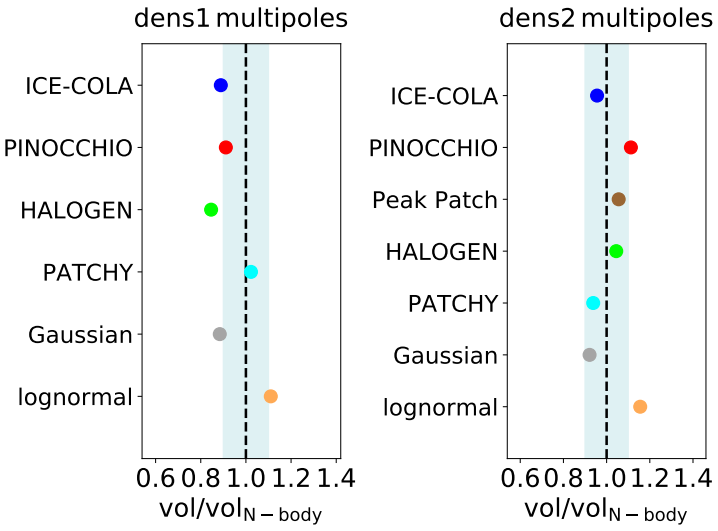

Figure 3.11: Volume ratios between the allowed statistical volumes obtained from the analysis using the covariance matrices from the approximate methods to the corresponding ones from the N-body catalogues for the density matched samples. The light grey band indicates a range of $\pm 10\%$ deviation from a ratio equal to 1.

In summary, our results and those of our companion papers indicate that approximate methods can provide robust covariance matrix estimates for cosmological parameter constraints. However, the differences seen between the various recipes, statistics, and selection criteria considered here highlight the importance of performing detailed tests to find the best strategy to draw halo samples from any given approximate method.

Supplementary note

After the covariance matrix comparison described here, I carried out a similar analysis for a further calibrated method called BAM (Bias Assignment Method; Balaguera-Antolínez et al., 2019). The methodology for the performance tests of the covariance matrices estimated from BAM halo catalogues is analogous to the one described here. The main difference is that it focuses on the real-space power spectrum covariance, since redshift-space distortions were not included yet into BAM. Therefore, only constraints on the nuisance parameters b_1, b_2 and γ_3^- are compared to the corresponding reference constraints from the N-body simulations. The BAM covariance matrices reproduce the N-body errors within 5%-10%, which corresponds to the same level of agreement of the approximate methods considered here. This makes BAM a promising method for the generation of mocks for covariance matrix estimation. Once redshift-space distortions are included, further performance tests would need to follow to compare BAM against N-body simulations and other approximate methods. A detailed description of the BAM analysis in real space can be found in Balaguera-Antolínez et al. (2020).

Chapter 4

Minkowski functionals of the Large-Scale Structure

The previous chapter focused on a key aspect for extracting unbiased information from two-point correlation function measurements. However, we know that the underlying matter density field is not simply Gaussian distributed, and therefore two-point statistics cannot provide a complete description of the large-scale structure of the Universe. The aim of this chapter is to develop a complementary approach to extract the non-Gaussian, or equivalently the higher-order, information from the cosmic density field.

Chapter 2 introduced the isodensity Minkowski functionals (MFs) as geometric and topological descriptors of the cosmic density field that contain compressed higher-order information (see Section 2.7). Most previous analyses are based on two calculation methods for the isodensity MFs, Koenderink invariants from differential geometry and Crofton's intersection formula from integral geometry (Schmalzing & Buchert, 1997; Schmalzing et al., 1999a).

An alternative technique for the estimation of the isodensity MFs is to compute these statistics on triangulated isodensity surfaces constructed from the underlying density field. Although this approach follows very closely the geometry of the isodensity surfaces, there have been very few efforts in this direction. Sheth et al. (2003), who first introduced this idea to large-scale structure analysis, developed a code to construct triangulated surfaces from fixed lattice cubes. The main cosmological application of their code has been the MF measurement from mock catalogues with different cosmologies (Sheth, 2004).

Yaryura et al. (2004) and Aragon-Calvo et al. (2010) proposed a more efficient technique by defining the triangulated surface directly from the Delaunay tessellation of the galaxy distribution instead of using a regular grid. In three dimensions, the Delaunay tessellation divides up the space into tetrahedra, whose vertices are formed by the points of the distribution, here the galaxies. It is defined such that the circumsphere of a Delaunay tetrahedron contains no points from the distribution in its interior. The main advantage of the Delaunay tessellation compared to a regular grid is its adaptive resolution: high density regions are automatically resolved with a large number of small tetrahedra, while low density regions are probed with fewer and larger tetrahedra.

This first part of the chapter presents MEDUSA, a new implementation of an algorithm to estimate MFs based on the Delaunay tessellation of the three-dimensional galaxy distribution, and its first application to synthetic galaxy catalogues. MEDUSA is based on an earlier implementation of the same basic algorithm that was described in Yaryura et al. (2004). A crucial extension is the implementation of periodic boundary conditions, which are required for the analysis of density fields from N-body simulations and the correct comparison of the measurements against theory predictions.

The two main steps of the algorithm, consisting of the construction of the triangulated isodensity surfaces from the Delaunay tessellation and the estimation of the MFs, are outlined in Section 4.1. A thorough validation of MEDUSA using a series of test samples whose MFs can be theoretically predicted follows in Section 4.2. The construction of the isodensity surface requires the values of the density field at the galaxy positions. In galaxy catalogues the underlying density field is not known and needs to be reconstructed from the galaxy distribution itself. Section 4.3 describes the approach implemented for the density estimation. Having all the steps for the MF measurements implemented and validated, we apply MEDUSA to the synthetic galaxy catalogues of the Minerva simulations, which were introduced in Section 3.4. Our main focus lies on three main issues of great importance for the analysis of the MFs inferred from real galaxy surveys: non-Gaussian features due to non-linear gravitational evolution (Section 4.4.1), redshift-space distortions (Section 4.4.2), and Alcock-Paczynski (AP) distortions (Section 4.4.3).

The work presented in these sections has been submitted as "MEDUSA: Minkowski functionals estimated from Delaunay tessellations of the three-dimensional large-scale structure" by M. Lippich and A. G. Sánchez to the Monthly Notices of the Royal Astronomical Society (Lippich & Sánchez, 2020). Sections 4.1 to 4.4 reproduce the corresponding sections of the paper draft, adapted such that the format and references match the thesis format.

The second part of this chapter sets the course to extract further information from the MF measurements in the future. For a standard likelihood analysis (see Section 3.2), we will need a model for the MFs of the non-Gaussian density field and an estimate for their covariance. In this context, Section 4.5 describes the covariance matrix of the MF measurements from the Minerva simulations.

Finally, Section 4.6 presents the novel approach of evolution mapping for MFs, which can represent the basis to build a model for the MFs of the non-Gaussian density field. The research presented in this section is planned for submission to a peer-reviewed scientific journal.

4.1 The MEDUSA code

In real or synthetic galaxy catalogues in general we do not know the underlying continuous density field. Instead, we have to estimate the isodensity MFs from a discrete three-dimensional point distribution. For this, we need three main ingredients

(i) an estimate of the density at each point of the distribution,

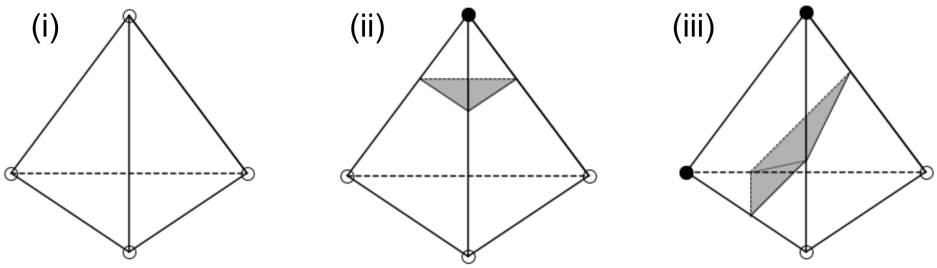

Figure 4.1: The three different tetrahedron configurations considered in MEDUSA: (i) all vertices are the same, either underdense or overdense compared to the density threshold $\rho_{\rm th}$, here shown as empty circles, (ii) one vertex is different to the others, here shown as filled circle, (iii) two vertices are underdense and the other two are overdense. The different fillings of the circle indicate the different densities. The grey area shows the intersection of the triangulated isodensity surface with the tetrahedron.

(ii) a fast and accurate extraction of the isodensity surfaces based on the point distribution for any given density threshold,

(iii) the computation of the MFs of the resulting isodensity surface.

These three steps are implemented into our code MEDUSA (Minkowski functionals Estimated from DelaUnay teSsellAtion). There are several approaches to estimate densities based on a discrete set of points. However, the second and third steps only require a set of points with known densities as an input and are independent of the particular method used to obtain such values. In the following two subsections, we describe each of the two steps (ii) and (iii) in detail and come to point (i) in Section 4.3 after testing MEDUSA extensively on point distributions with known densities.

4.1.1 Extraction of isodensity surfaces

The crucial step for the estimation of the MFs is the extraction of the isodensity surfaces at a desired threshold from a set of points. For MEDUSA we chose a similar approach to Sheth et al. (2003), Yaryura et al. (2004) and Aragon-Calvo et al. (2010), and compute a triangulated isodensity surface directly from a three-dimensional point distribution. We extend these previous approaches by also including a recipe to account for periodic boundary conditions. Analogously to Yaryura et al. (2004) and Aragon-Calvo et al. (2010), we perform a Delaunay tessellation on the three-dimensional point distribution, which we use as the basis for the interpolation of the density field. This approach is simpler than the regular grid used by Sheth et al. (2003) and automatically provides us with higher resolution in the regions where the density is higher. In the case of point distributions from boxes with periodic boundary conditions, we add buffer zones around the box that replicate the particles from the opposite sides. MEDUSA assigns a flag to each tetrahedron resulting from the Delaunay tessellation. This flag depends on how many particles of the

tetrahedron are inside the box and, if the tetrahedron lies (partially) outside the box, on its position. The following cases need to be considered:

(i) Tetrahedra that lie completely inside the box.

(ii) Tetrahedra that are partially outside the box and cross one face of the box far from its edges. Each of these tetrahedra has one copy at the opposite side of the box.

(iii) Tetrahedra that are partially outside and lie close to the edges, but far from the corners of the box. Each of these tetrahedra has three copies at the three opposite edges of the box.

(iv) Tetrahedra that are partially outside and lie close to the corners of the box. Each of these tetrahedra has seven copies at the other seven corners of the box.

(v) For tetrahedra that are located at the corners of the box there is the special case that the vertices of the tetrahedron are all outside, but the tetrahedron is still partially inside the box.

(vi) Tetrahedra close to the edges or corners that lie completely outside the box and are copies of tetrahedra that are completely inside the box, but are neighbours of tetrahedra that are partially inside.

(vii) Tetrahedra that are completely outside the box and do not belong to the previous case (vi). They are also copies of tetrahedra that are completely inside the box.

The tetrahedra that belong to the last category (vii) can be discarded. All other tetrahedra are assigned a flag that takes into account to which category they belong and at which side of the box they are located, in order to prevent double counting. These flags are used in the estimation of the MFs as described in Section 4.1.2.

Additionally, all particles are considered as "overdense" or "underdense" depending on whether their corresponding densities are larger than the density threshold $\rho_{\rm th}$ being considered or not (see Section 2.7.1 for the details on $\rho_{\rm th}$). Given this classification, there are only three different types of tetrahedron configurations:

(i) All vertices of the tetrahedron are either overdense or underdense.

(ii) One vertex is different to the other three vertices.

(iii) Two vertices are overdense and the other two are underdense.

These three cases are illustrated in Fig. 4.1, where vertices with the same density property, i.e. underdense or overdense, are shown with circles with the same filling. The isodensity surface will only intersect tetrahedra of the last two types. This intersection will occur at the edges between particles with different density properties. The intersection points of the surface with the tetrahedron edges correspond to the points where the density matches the threshold $\rho_{\rm th}$, which are obtained by linearly interpolating the densities of the

two corresponding particles. This is equivalent to assuming a constant density gradient within the tetrahedron. For case (ii), where one particle is different to the others, we obtain an intersection triangle. For case (iii), where two particles have the same density property, we obtain four points of intersection on the edges and the resulting surface can be decomposed into two triangles. Following this approach, MEDUSA computes the intersection triangles for all tetrahedra where at least one vertex is different to the others. These are 12 configurations less to take into account than for cubic lattice intersections as in Sheth et al. (2003), which makes this step significantly simpler. Once all tetrahedra of types (ii) and (iii) have been considered, we obtain a triangulated surface representing the isodensity contour corresponding to $\rho_{\rm th}$.

4.1.2 Minkowski Functionals of a triangulated surface

Since the MFs are additive (see Section 2.7.1), the global MFs of the density distribution can be obtained by summing over the MFs of the isodensity surfaces enclosing the individual excursion sets. As described in Sheth et al. (2003), the MFs of a triangulated surface can be computed in a straightforward way:

1. The surface area S of the triangulated surface is given by the sum over the areas of all triangles of the surface,

$$S = \sum_{i=1}^{N_{\rm t}} S_i, \tag{4.1}$$

   where $N_{\rm t}$ is the total number of triangles contributing to the surface.

2. The volume V is the sum over the volumes of all fully enclosed tetrahedra, denoted with T, and the fraction of the volumes of the intersected tetrahedra that lie within the surface, denoted with S,

$$V = \sum_{i=1}^{N_{\rm T}} V_i + \sum_{j=1}^{N_{\rm S}} V_j. \tag{4.2}$$

   If only one vertex is overdense or underdense, corresponding to case (ii) in Fig. 4.1, the volume of the tetrahedron defined by this point and the triangle of the isodensity surface as a base has to be added or subtracted, respectively. If the tetrahedron contains two overdense vertices, as in case (iii) of Fig. 4.1, the contributing volume can be split into three tetrahedra.

3. The integrated mean curvature C is obtained by summing over the edges of all adjacent triangles i and j,

$$C = \frac{1}{2} \sum_{i,j} \ell_{ij} \phi_{ij} \epsilon \tag{4.3}$$

where ℓ_{ij} is the length of the common edge, ϕ_{ij} is the angle between the normals, $\hat{n}_i$ and $\hat{n}_j$, of the two triangles,

$$\cos\phi_{ij} = \hat{n}_i \cdot \hat{n}_j, \tag{4.4}$$

and the value of ϵ distinguishes the cases in which the surface is locally convex, indicated by the value $\epsilon = 1$, or locally concave, in which case $\epsilon = -1$.

4. The Euler characteristic χ of a triangulated surface can be determined by

$$\chi = N_{\rm t} - N_{\rm e} + N_{\rm v}, \tag{4.5}$$

where $N_{\rm t}$, $N_{\rm e}$ and $N_{\rm v}$ are the total number of triangles, triangle edges and triangle vertices contained in the surface.

As mentioned in Section 4.1.1, MEDUSA assigns a flag to every tetrahedron depending on its position in the box and/or the buffer zone. From the tetrahedra that are partially inside the box, and hence are repeated on its other sides, only those that are closer to the origin (0,0,0) are taken into account in the sums of equations (4.1)-(4.2), while all other copies are discarded. The flags that MEDUSA assigns to each tetrahedron ensure that all triangle edges and vertices from tetrahedra that are partially inside the box are taken into account in the sums of equations (4.3)-(4.5), and that their contribution is counted only once.

4.2 Results for test models

In this section we test the performance of MEDUSA by measuring the MFs of point distributions following known density profiles.

4.2.1 Spherical density distribution

As a first test sample we consider a distribution of points following a spherically-symmetric Gaussian density profile given by

$$\rho(\boldsymbol{r}) = \rho_{\rm max} \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{4.6}$$

where $r = |\boldsymbol{r}|$. We generated a set of points following this density profile with 3.5×10^5 particles, $\sigma = 0.6$ and a maximum radius, $r_{\rm max} = 4.0$. The density at each point was obtained by evaluating the true density profile of equation (4.6) at the corresponding location. We used MEDUSA to measure the MFs of 20 equispaced density thresholds from $\rho/\rho_{\rm max} = 0.0$ to 1.0. The analytical MFs for a sphere are:

(i) $S(\rho_{\rm th}) = 4\pi r(\rho_{\rm th})^2$

(ii) $V(\rho_{\rm th}) = \dfrac{4}{3}\pi r(\rho_{\rm th})^3$

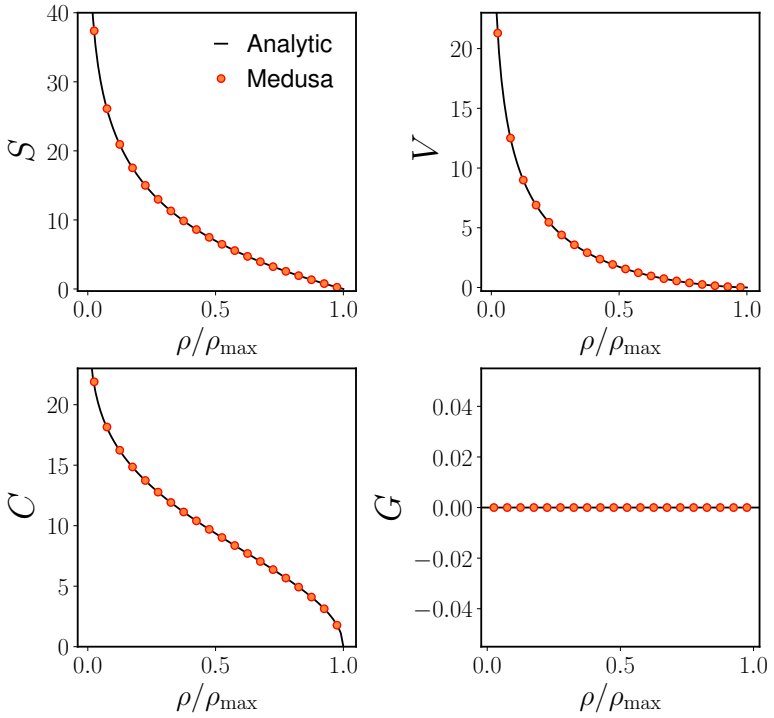

Figure 4.2: Minkowski Functionals inferred from a set of points following a spherically-symmetric Gaussian distribution as a function of the normalized density threshold $\rho/\rho_{\max}$. The red circles correspond to the measurements from MEDUSA using 20 equispaced density thresholds, the black lines show the analytical predictions.

(iii) $C(\rho_{\rm th}) = 4\pi r(\rho_{\rm th})$

(iv) $\chi(\rho_{\rm th}) = 2$ and hence $G = 0$

The radius corresponding to a given density threshold, $r(\rho_{\rm th})$, can be obtained by inverting equation (4.6). A lower density threshold corresponds to a larger radius of the spherical isodensity surface. Fig. 4.2 shows that the measured MFs are in good agreement with the analytical predictions. The overall agreement of the measurements of the first three MFs with the corresponding predictions is significantly better than 1%. The measured genus is always zero.

In order to test the implementation of periodic boundary conditions, we generated sets of points following the same spherically symmetric density profile of equation (4.6) but where the density distributions were cut into two half-spheres located at two opposite faces of a cubic box, four quarter-spheres located at the center of four opposite edges of the box, and eight partial spheres located at each corner of the box. Without the implementation of periodic boundary conditions, the isodensity surface cannot be extracted correctly at the boundaries of the box, since tetrahedra cannot extend outside it.

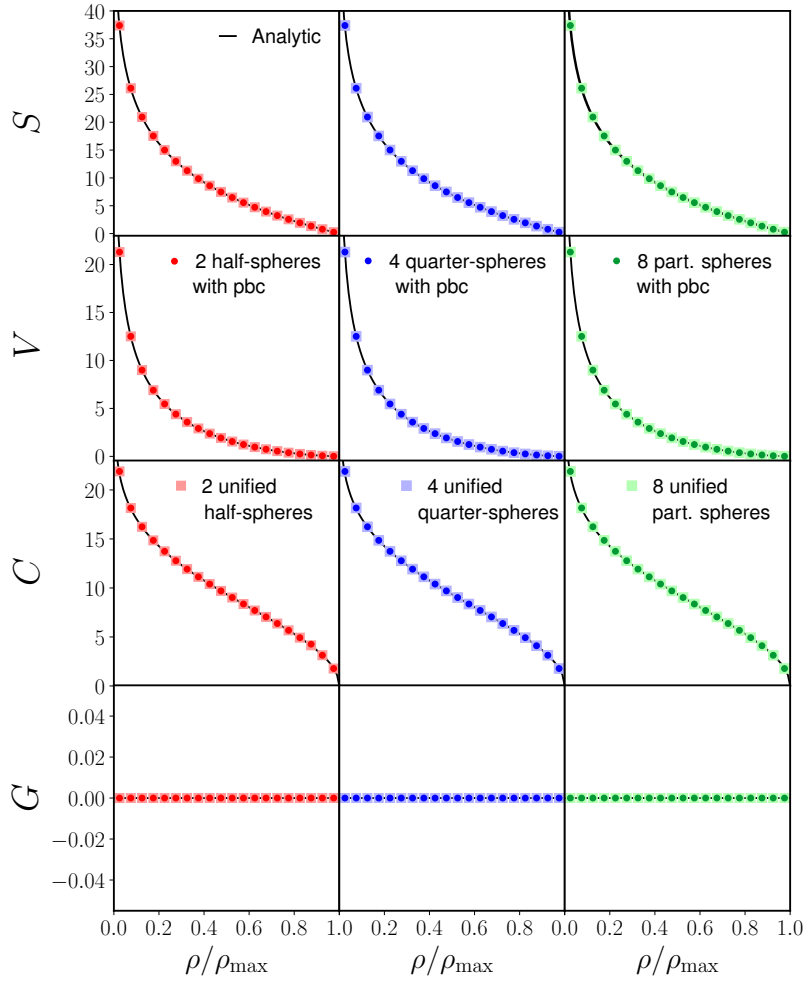

Figure 4.3: Minkowski Functionals of spherical density distributions where the spheres were cut into two half-spheres located at two opposite faces of a cubic box (red), four quarter-spheres located at the centres of four opposite edges of the box (blue), and eight partial spheres located at each corner of the box (green), measured with periodic boundary conditions. The measurements agree perfectly with the ones obtained from the corresponding unified spheres and their analytical predictions. All MFs are measured as a function of threshold $\rho/\rho_{\max}$.

Fig. 4.3 shows the agreement between the MFs measured from these three distributions taking into account periodic boundary conditions and the results obtained from the unified spherical distribution, for which no periodic boundary conditions are required. The agreement with the analytical predictions is also excellent. These results show that MEDUSA can correctly account for distributions with periodic boundary conditions. In particular, an error in the implementation of the periodic boundary conditions leading to a single incorrectly counted triangle, triangle vertex or triangle edge would result in values of genus $G \neq 0$.

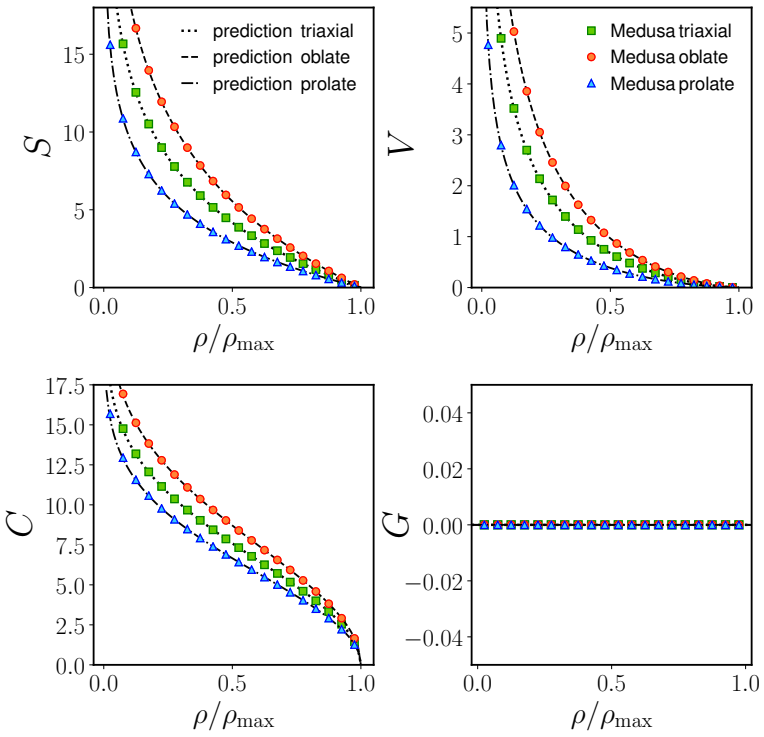

Figure 4.4: Minkowski Functionals for three different ellipsoidal point distributions, oblate, prolate, and triaxial, defined by the density profile of equation (4.7) expressed as a function of the normalized density $\rho/\rho_{\max}$. The lines indicate the corresponding theoretical predictions.

4.2.2 Ellipsoidal density distribution

We now consider as test samples ellipsoidal density distributions given by,

$$\rho(x, y, z) = \rho_{\max} \exp\left[-\left(\frac{x^2}{\sigma_a^2} + \frac{y^2}{\sigma_b^2} + \frac{z^2}{\sigma_c^2}\right)\right]. \tag{4.7}$$

We generated three different point distributions corresponding to oblate ($\sigma_a = \sigma_b = 1.0$, $\sigma_c = 0.4$), prolate ($\sigma_a = \sigma_b = 0.4$, $\sigma_c = 1.0$), and triaxial ($\sigma_a = 0.4$, $\sigma_b = 0.7$, $\sigma_c = 1.0$) ellipsoids using the same number of points and density thresholds as for the spherical case.

For the oblate and the prolate case, we compared the measured MFs against analytical predictions. For the triaxial case no analytical predictions are known for the surface and the curvature and therefore we computed numerical predictions. As for the case of the spherical distributions, a lower density threshold corresponds to larger principal axes a, b, c of the ellipsoid, while conserving the constant axes ratios for all density thresholds, such that $b = a\frac{\sigma_b}{\sigma_a}$ and $c = a\frac{\sigma_c}{\sigma_a}$. The analytical predictions for the MFs are given by:

(i)

$$S_{\rm obl} = 2\pi a\,(\rho_{\rm th})^2 \left[1 + \frac{\sigma_c^2}{\sigma_a\sqrt{\sigma_a^2-\sigma_c^2}}\,\mathrm{arctanh}\left(\sqrt{1-\frac{\sigma_c^2}{\sigma_a^2}}\right)\right] \tag{4.8}$$

$$S_{\rm pro} = 2\pi a\,(\rho_{\rm th})^2 \left[1 + \frac{\sigma_c^2}{\sigma_a\sqrt{\sigma_c^2-\sigma_a^2}}\,\arcsin\left(\sqrt{1-\frac{\sigma_a^2}{\sigma_c^2}}\right)\right] \tag{4.9}$$

(ii)

$$C_{\rm obl} = 2\pi a\,(\rho_{\rm th}) \left[\frac{\sigma_c}{\sigma_a} + \frac{\sigma_a}{\sqrt{\sigma_c^2-\sigma_a^2}}\,\mathrm{arccosh}\left(\frac{\sigma_c}{\sigma_a}\right)\right] \tag{4.10}$$

$$C_{\rm pro} = 2\pi a\,(\rho_{\rm th}) \left[\frac{\sigma_c}{\sigma_a} + \frac{\sigma_a}{\sqrt{\sigma_a^2-\sigma_c^2}}\,\arccos\left(\frac{\sigma_c}{\sigma_a}\right)\right] \tag{4.11}$$

(iii) $V = \frac{4}{3}\pi\, a\,(\rho_{\rm th})^3\,\frac{\sigma_b\sigma_c}{\sigma_a^2}$

(iv) $\chi = 2$ and hence $G = 0$

Fig. 4.4 shows that in all cases the measured MFs agree perfectly with the theoretical predictions. As in the case for the spherical density distribution, the overall agreement of the measurements of the first three MFs with the corresponding predictions is better than 1%. The measured genus is always exactly zero.

4.2.3 Toroidal density profiles

The point distributions considered in the previous sections have isodensity surfaces without holes, and therefore their genus is zero for all density thresholds. In order to test the estimation of the Euler characteristic and the genus, we studied sets of points corresponding to one or more overlapping toroidal distributions. For the case of one torus, the point distribution is generated following a density profile

$$\rho(x,y,z) = \rho_{\rm max}\exp\left[-\frac{(R-\sqrt{x^2+y^2})^2+z^2}{\sigma^2}\right], \tag{4.12}$$

where R and r are the major and minor radii of the torus, respectively, and $r^2 = \left(R-\sqrt{x^2+y^2}\right)^2 + z^2$. The upper panel of Fig. 4.5 shows the measured genus of such a density distribution, generated with $R = 1.1$ and $\sigma = 0.9$. The upper panels of Fig. 4.6 show the corresponding triangulated isodensity surfaces obtained by MEDUSA for $\rho/\rho_{\rm max} = 0.225$, $\rho/\rho_{\rm max} = 0.475$ and $\rho/\rho_{\rm max} = 0.725$. For low density thresholds, no hole is visible in the isodensity surface and MEDUSA measures $G = 0$. For density

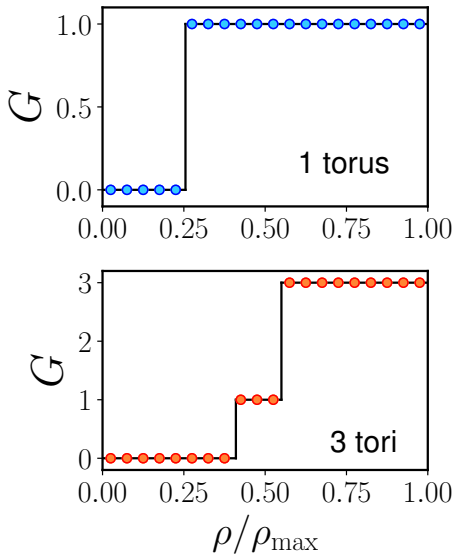

Figure 4.5: Genus measurements as a function of the normalized density $\rho/\rho_{\rm max}$ of point distributions following profiles of one torus (upper panel) and three overlapping tori (lower panel). The solid lines show the theoretical predictions for each case.

thresholds $\rho/\rho_{\rm max} > 0.25$ a hole in the center of the isodensity surface becomes visible and the code correctly recovers $G = 1$.

We also considered a set of points corresponding to three overlapping tori with $R = 1.3$ and $\sigma = 0.8$, and centred at $(1, 0, 0)$, $(-1, 0, 0)$ and $(0, 2, 0)$, respectively. The lower panel of Fig. 4.5 shows the genus measured from this density distribution, and the lower panels of Fig. 4.6 show three characteristic isodensity surfaces at the same thresholds as before. As in the case of a single torus, for low density thresholds the isodensity surface contains no holes and the measurement of the genus is $G = 0$. For a density threshold $0.4 < \rho/\rho_{\rm max} < 0.55$ the corresponding isodensity surface shows the hole of the torus whose center is furthest from the other two and hence the measured genus is $G = 1$. For a density threshold $\rho/\rho_{\rm max} > 0.55$, the isodensity surface contains three holes and we recovered the correct value $G = 3$.

4.2.4 Effect of using particles tracing the density field

When estimating MFs, MEDUSA uses the values of the density field directly at the positions of the points in the sample being analysed. In the test samples considered in the previous sections, as in numerical simulations or real galaxy surveys, the points trace the underlying density field. Using their positions as the nodes to interpolate the density field, as opposed to, e.g. the vertices of a regular grid, has the advantage of automatically providing a higher resolution in high-density regions. Note however, that the procedure described in Section 4.1 does not require the points used as the basis of the Delaunay

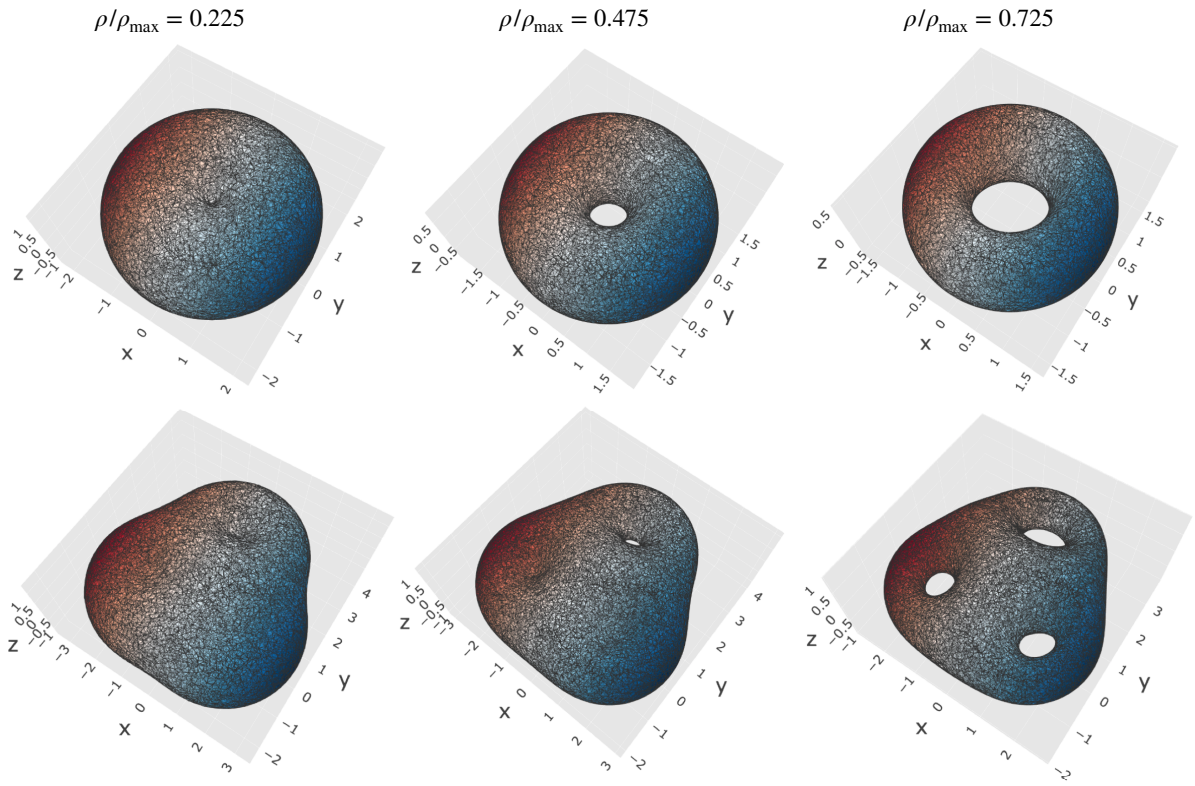

Figure 4.6: Isodensity surfaces for three different density thresholds $\rho/\rho_{\max} = 0.225$, $\rho/\rho_{\max} = 0.475$ and $\rho/\rho_{\max} = 0.725$ for density distributions following profiles of one torus (upper panel) and three overlapping tori (lower panel).

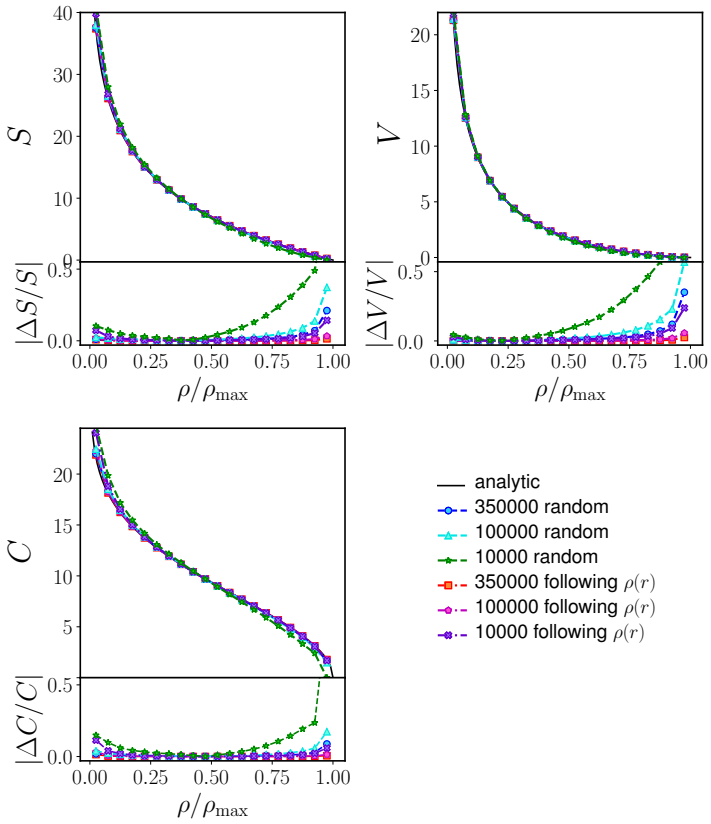

Figure 4.7: Minkowski Functionals for the spherical density distribution of Section 4.2.1 obtained using samples with 350 000, 100 000 and 10 000 points that are placed following the density profile or randomly within the same volume. The genus is not shown as it is consistently zero in all cases.

tessellation to follow the density field.

As a test, Fig. 4.7 shows the MFs for the same spherical density distribution as in Section 4.2.1 estimated using sets of particles of different size that are placed following the density distribution or randomly within the same volume. The computation of the genus is consistently zero for all considered cases and density thresholds. The remaining three MFs computed from the 100 000 and 350 000 particles tracing the density field agree with the analytical predictions at better than 1% level. Even for the case of 10 000 points the agreement between measurements and analytical predictions is better than 2% on densities $0.1 < \rho/\rho_{\rm max} < 0.8$.

For the case of the randomly distributed particles, we obtain a comparable precision only when using 350 000 particles. For smaller samples, the deviations from the analytical predictions become significantly larger, particularly for high density thresholds. This comparison illustrates the advantage of using particles tracing the density field as the nodes of the Delaunay tessellation, which provides a better resolution on high-density regions and allows for a robust determination of all MFs even for sparse samples.

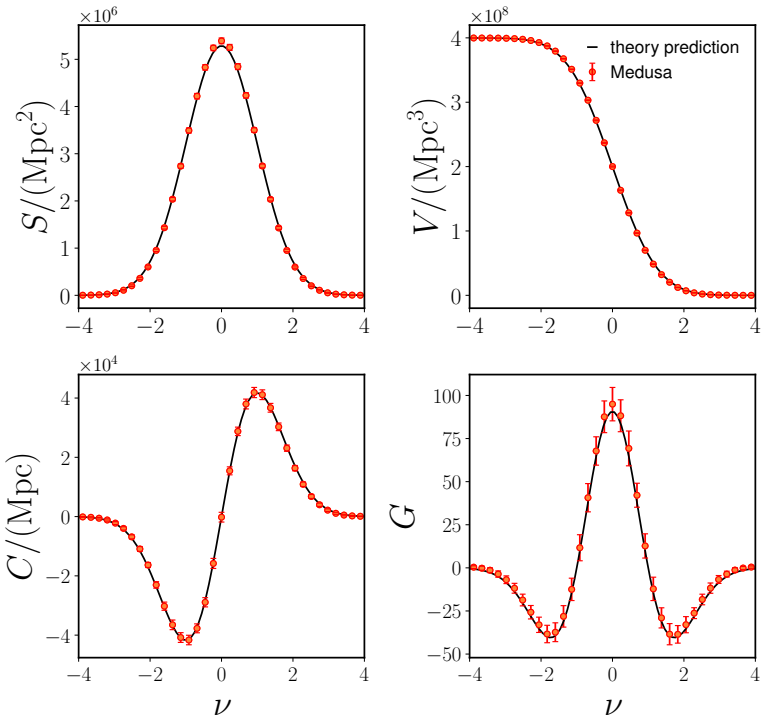

Figure 4.8: Mean MFs of 100 GRFs, generated with the same linear ΛCDM power spectrum as the Minerva simulations in a cubic box with length 737 Mpc and smoothed with a Gaussian kernel with $\lambda = 20$ grid units ($\lambda = 28.8$ Mpc). We used the densities at 200 000 random points within the box. The red points show the mean values determined using MEDUSA together with the standard deviation from 100 realizations. The theoretical predictions are shown as black solid lines.

4.2.5 Gaussian density field

For a final test of MEDUSA, we computed the MFs of a smoothed Gaussian random field (GRF), which have known analytical expressions that are sensitive to the power spectrum of the field, $P(k)$, as described in Section 2.7.2. This case also serves as an additional validation of the implementation of periodic boundary conditions in our code, as an incorrect treatment would lead to deviations from the analytical predictions. We generated 100 realizations of a GRF with the same linear power spectrum as our Minerva simulations, which were introduced in Section 3.4, at redshift $z = 0.57$ on a cubic grid with side length $L = 737$ Mpc and periodic boundary conditions. The field, f, was smoothed with a Gaussian kernel

$$W(x) = \frac{1}{(2\pi)^{3/2} R^3} \exp\left(-\frac{x^2}{2R^2}\right), \tag{4.13}$$

with a smoothing scale $R = 20$ grid units ($= 28.8$ Mpc) and normalized by its standard deviation, $\nu = f/\sigma_0$, such that it has zero mean, $\langle\nu\rangle = 0$, and unit variance, $\langle\nu^2\rangle = 1$.

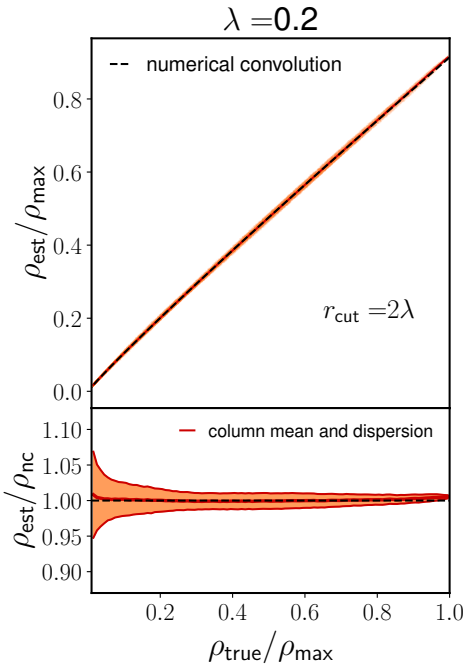

Figure 4.9: Upper panel: estimated densities for the same spherical point distribution of Section 4.2.1 plotted against their true values. The densities are estimated using the Gaussian kernel of equation (4.14) with a smoothing length $\lambda = 0.2$ and truncated at a radius $r_{\mathrm{cut}} = 2\lambda$ and closely follow the convolution of the true density profile with the same kernel, indicated by a black dashed line. Lower panel: the ratios of the estimated densities and their expected values. The red lines indicate the column mean and corresponding dispersion.

Since there are grid cells with negative values of ν, it cannot be treated as a density field and sampled with points. Instead, we follow the approach tested in Section 4.2.4 and use the values of ν at 200 000 randomly placed points in each box. The resulting mean interparticle separation is approximately half of the smoothing length, and thus it should be possible to resolve the full structure of the smoothed density field.

As introduced in Section 2.7.2, the theoretical predictions for the MFs of a smoothed GRF only depend on the parameter λ_{c} which is sensitive to the underlying power spectrum. We compute the theoretical predictions according to equations (2.68) to (2.77) using the linear power spectrum convolved with the kernel of equation 4.13 as input. Fig. 4.8 shows the mean MFs computed with MEDUSA from the 100 realizations of the GRF and their corresponding theoretical predictions, which are in good agreement. This shows that MEDUSA can accurately determine MFs of cosmological density fields and that the periodic boundary conditions are correctly implemented.

4.3 Density estimation

The procedure to extract isodensity surfaces and estimate MFs described in Section 4.1 requires as input the values of the density at each point of our discrete distribution. In the test cases of Section 4.2, we used the true values of the underlying density field evaluated at the position of the points. When analysing N-body simulations or galaxy surveys, these densities need to be estimated from the point distribution itself. Here, we estimate densities by applying a Gaussian kernel with a fixed smoothing scale λ,

$$W(r) = \frac{1}{A} \exp\left(-\frac{r^2}{2\lambda^2}\right), \tag{4.14}$$

where r represents the distance between the points. This kernel is truncated at a scale $r_{\rm cut}$ and the normalization A is defined such that the volume integral of $W(r)$ up to this maximum scale is equal to one and hence the total mass is conserved. We tested this approach by applying it to the spherical Gaussian density distribution described in Section 4.2.1 for which the true underlying density distribution is known. We examined the impact of using different kernel smoothing lengths and truncation radii. The true underlying density profile corresponding to each case can be obtained by convolving the Gaussian density field of equation (4.6), which is truncated at $r_{\rm max} = 4$, with the kernel of equation (4.14).

Fig. 4.9 shows the densities estimated at each point of the spherical Gaussian distribution of Section 4.2.1 by applying a Gaussian kernel with a smoothing length $\lambda = 0.2$ and a truncation radius of $r_{\rm cut} = 2\lambda$, which follow closely the true profile, which is indicated by a black dashed line. Fig. 4.10 shows the MFs measured using these density estimates and the corresponding theory predictions computed using the convolved density profile. The measurements match the theory predictions remarkably well, with a similar level of agreement as for the case in which the true densities were used, which was discussed in Section 4.2.1. Note that, as the convolution with the Gaussian kernel reduces the maximum densities in the profile, the highest density threshold considered in this case is $\rho/\rho_{\rm max} = 0.875$. We tested the impact of using different values of λ and $r_{\rm cut}$ and found similar results but with a larger variance.

When this method is applied to realistic point distributions such as galaxy catalogues, the smoothing length and truncation radius of the kernel need to be adjusted to provide the necessary smoothing to avoid discreteness effects without erasing too much information. In principle, steps (ii) and (iii) of the MEDUSA code described in Section 4.1 could be applied to density estimates obtained using a different approach. Other possibilities include non-parametric methods in which the densities are derived from the size of the Voronoi or Delaunay cells (e.g., Schaap & van de Weygaert, 2000). Although these approaches can better resolve high-density regions due to their varying resolution, we have found that these density estimates are highly affected by Poisson noise in the low density regions of sparse samples, and are therefore not optimal for the analysis of real galaxy catalogues. An additional advantage of using an isotropic Gaussian kernel with a fixed smoothing length is that it is more convenient to compute theory predictions.

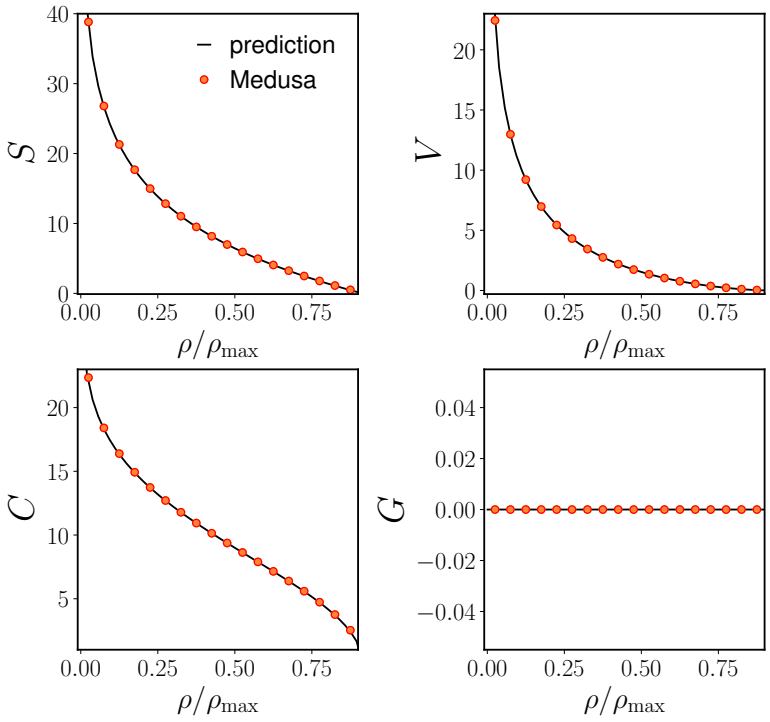

Figure 4.10: MFs of the same spherical point distribution of Section 4.2.1 but inferred from density estimates based on a Gaussian kernel with a smoothing length $\lambda = 0.2$ and truncated at a radius $r_{\rm cut} = 2\lambda$, expressed as a function of the normalized density $\rho/\rho_{\max}$. The black lines shows the theoretical predictions corresponding to the convolved density profile.

4.4 Minkowski functionals of the Minerva HOD galaxy catalogues

4.4.1 Real-space measurements

After validating the performance of MEDUSA for several test cases with different geometries and topologies, we now show the results obtained by applying the code to synthetic cosmological galaxy samples. We use catalogues derived from the set of 300 N-body simulations Minerva, which were described in Section 3.4.

To create a synthetic galaxy catalogue, the halos of the snapshot at $z = 0.57$ were populated using the halo occupation distribution (HOD) parametrization of Zheng et al. (2007). The HOD gives the average number N of galaxies in a halo as a function of its mass M by decomposing it into contributions from central and satellite galaxies, $\langle N(M)\rangle = \langle N_{\rm cen}(M)\rangle + \langle N_{\rm sat}(M)\rangle$. This derives from the idea that only halos that already host a central galaxy can host a further satellite galaxy and that the probability that a halo hosts a central galaxy can be characterized by a minimum mass.

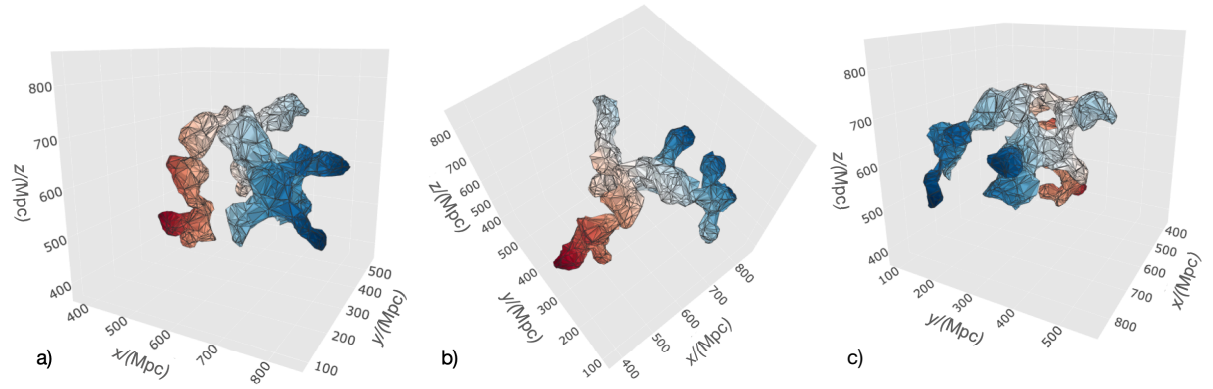

Figure 4.11: Isodensity surface of one structure found in the first Minerva HOD galaxy catalogue at a density threshold of $\delta = 0.584$.

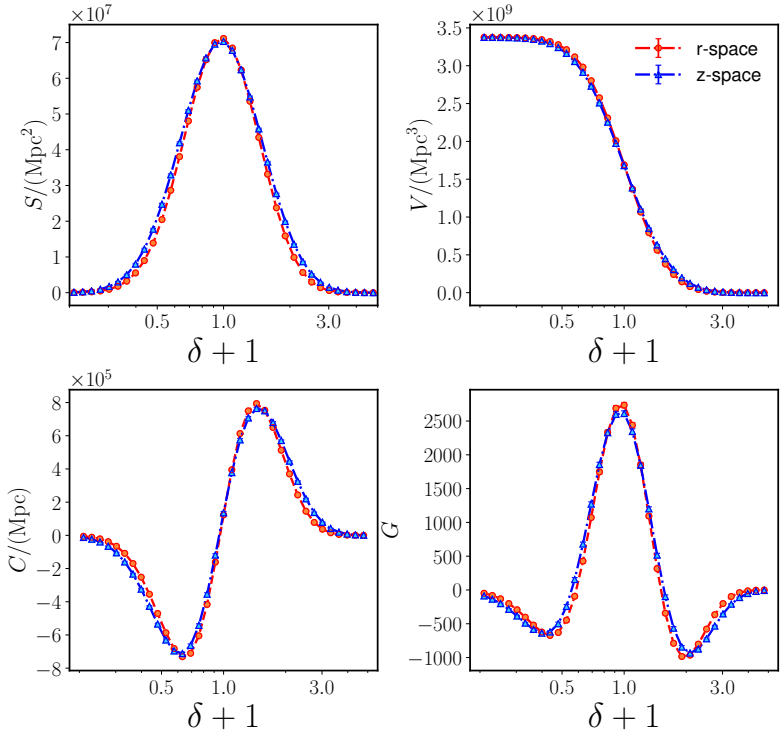

Figure 4.12: Mean MFs of the 300 Minerva HOD catalogues as a function of the density contrast δ for the real- and redshift-space galaxy density fields (orange and blue, respectively). The error bars corresponding to the standard deviation from the 300 realizations are of the size of the points or smaller and therefore not visible. The densities were estimated with a smoothing length λ corresponding to the mean interparticle separation and a truncation radius $r_{\mathrm{cut}} = 3\lambda$.

In total, the mean occupation function $\langle N(M)\rangle$ by Zheng et al. (2007) has five free parameters, which are fitted such that they best match a specific observation. The position and velocity of a central galaxy correspond to those of the most-bounded DM particle of the halo. The position and velocity of a satellite galaxy is randomly assigned from the other DM particles of the halo. The redshift $z = 0.57$ corresponds to the mean redshift of the CMASS sample of the BOSS survey (Alam et al., 2017). Grieb et al. (2016) produced 100 HOD catalogues for the first set of Minerva simulations choosing the HOD parameters such that the monopole of the mean correlation function from the resulting sample matches the one measured from the CMASS galaxies. For the remaining 200 Minerva realizations we generated HOD catalogues with the same HOD parameters (for more details see Grieb et al. (2016)).[1]

As a first step for the MF measurements, we estimated the number density, $n_{\rm est}$, at the position of each galaxy by smoothing the distribution with a Gaussian kernel as described in Section 4.3. We used a smoothing length corresponding to the mean interparticle separation, $\lambda = 19.7\,\mathrm{Mpc}$, close to what was found to be the optimal smoothing length for BAO reconstruction in the final BOSS analyses (Alam et al., 2017). This smoothing length is sufficiently large to avoid discreteness effects, but without erasing too much information on small scales. The truncation radius of the kernel was set to $r_{\rm cut} = 3\lambda$, which gives the highest signal-to-noise. The density contrast at the position of each galaxy was obtained as $\delta = n_{\rm est}/\bar{n} - 1$, where $\bar{n}$ is the mean number density.

We computed the MFs on 35 density thresholds equispaced in logarithmic scale around the mean density contrast $\delta = 0$. Fig. 4.11 shows a section of the isodensity surface corresponding to the threshold $\delta_{\rm th} = 0.584$ viewed from three different angles. This sample is sparser than the test samples of Section 4.2, which makes the triangles contributing to the surface more visible than for the toroidal profiles of Fig. 4.6. This structure has a hole in the center that is visible in panel c), and can then be described by a local genus of 1.

Fig. 4.12 shows the mean global MFs from the 300 Minerva realizations as a function of $\delta_{\rm th}$. In logarithmic scale, the shape of the MFs resembles that of the Gaussian predictions from Fig. 4.8, but the genus is clearly not symmetric and exhibits different depths for the two minima.

In order to compare MF measurements to theory predictions, the MF densities are typically expressed as functions of the volume-filling fraction f_V. The advantage of this is that the MF densities are expected to be invariant under any local monotonic transformation, if the threshold is adjusted such that that it gives the same volume-filling fraction (Codis et al., 2013). Fig. 4.13 shows the measured mean MF densities and the corresponding Gaussian predictions as a function of f_V. The Gaussian predictions are obtained from equations (2.68)-(2.72) using the measured mean galaxy power spectrum multiplied by the Fourier transform of the smoothing kernel. There are clear differences between the measurements and the Gaussian predictions. In particular, the asymmetry of the genus is also obvious here, with the Gaussian prediction providing a better match to the measurement

[1]This paragraph is a slightly extended and modified version of the corresponding one in Lippich & Sánchez (2020).

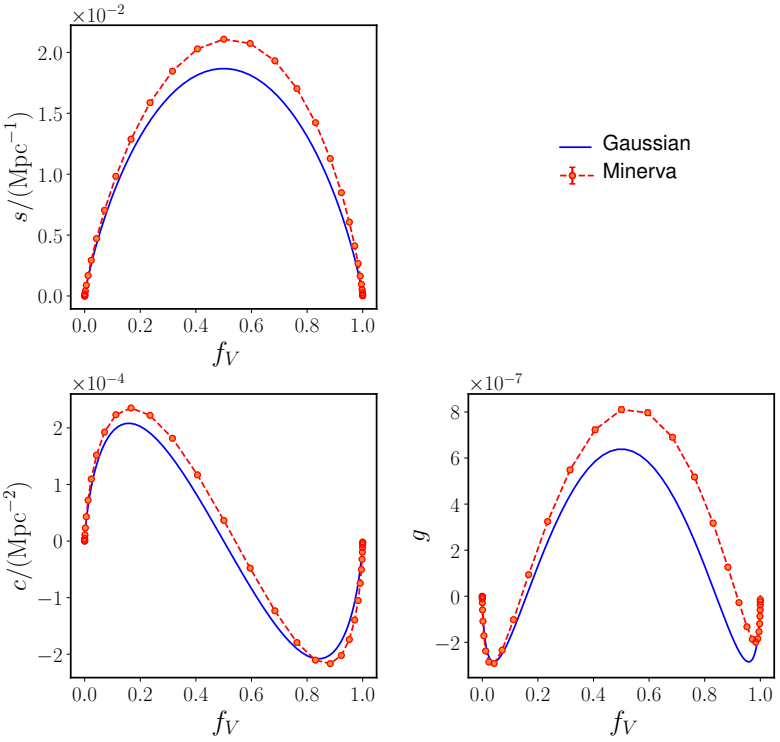

Figure 4.13: MF densities for the mean of the 300 Minerva HOD catalogues measured from the same smoothed galaxy density field in real space as in Fig. 4.12, but plotted as a function of the volume-filling fraction f_V.

at low f_V values. The measured power spectrum, which is well in the non-linear regime, is dominated by the high-density regions. Hence, it is to be expected that the Gaussian prediction derived from it is in better agreement with the measurements at the high-density end, which corresponds to low f_V values.

It is clear that the Gaussian model cannot be used to analyse the MFs of galaxy catalogues with comparable number density and redshift as our HOD sample. Since the measurements of the surface area, curvature and, in particular, the genus are sensitive to the non-Gaussian features of the density field, they contain complementary information to that of the galaxy power spectrum. We will explore the cosmological information content of these measurements in detail in upcoming work. In the next sections, we will focus on two important observational effects that must be taken into account before measuring the MFs of a real galaxy survey, namely RSD and AP distortions.

4.4.2 Effect of redshift-space distortions

To study the effect of RSD on the MFs, we distorted the positions of the HOD galaxies by taking into account the component of their peculiar velocities along one Cartesian axis of the box, which was treated as the line-of-sight direction. Since the total volume

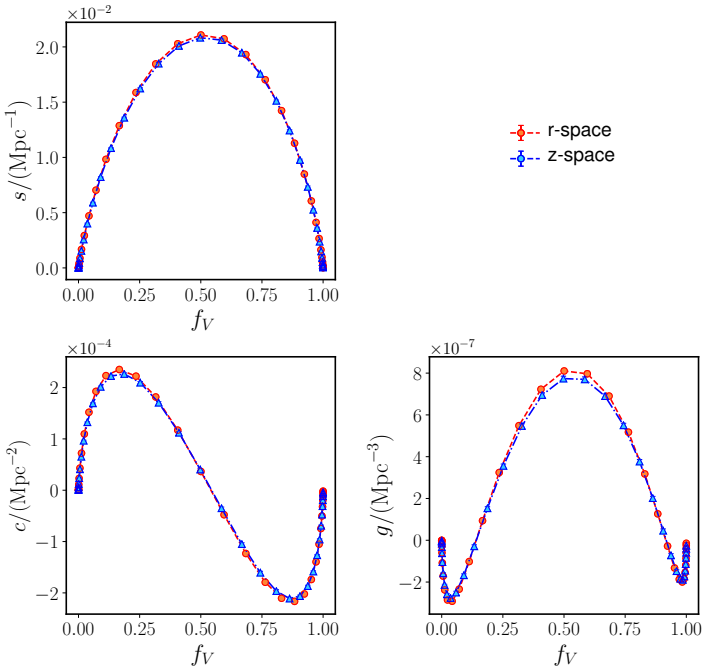

Figure 4.14: MF densities for the mean of the 300 Minerva HOD catalogues measured from the same smoothed galaxy density field in real and redshift space as in Fig. 4.12, but plotted as a function of the volume-filling fraction f_V.

and number density are not altered by RSD, we used the same smoothing length as in Section 4.4.1 to estimate the densities at the distorted galaxy positions. Fig. 4.12 compares the measurements of the MFs in real and redshift space. The amplitudes of both sets of measurements are very similar, but the redshift-space MFs appear to be stretched towards lower and higher densities than $\delta = 0$ compared to the corresponding ones in real-space.

Fig. 4.14 shows the same measurements from Fig. 4.12, but plotted as functions of the volume-filling fraction, f_V. Expressed in this way, the agreement between the MFs in real and redshift space is significantly improved, with only small deviations in their amplitude. The surface measurements agree at a 2% level, while the deviations in the curvature and genus are smaller than 5% (except for the density thresholds where these MFs are close to zero). RSD do not correspond to a monotonic transformation of the density field. Nonetheless, on average, the mapping from the real-space density threshold, δ_{rs}, to the corresponding value in redshift space, $\delta_{\mathrm{zs}}(\delta_{\mathrm{rs}})$, can be well described by matching the values of f_V in the two spaces (although the scatter for the individual densities is large). For this reason, the global effect of RSD on the MFs of the Minerva HOD galaxy catalogues is small when these are expressed as functions of f_V. However, this result cannot be generalised to other samples with different number densities or mean redshifts without careful study.

The fact that RSD have only a small effect on the MF densities when expressed in terms of f_V implies that it should be possible to probe the impact of deviations from Gaussianity or the sensitivity to the underlying cosmology without a detailed characterization of the

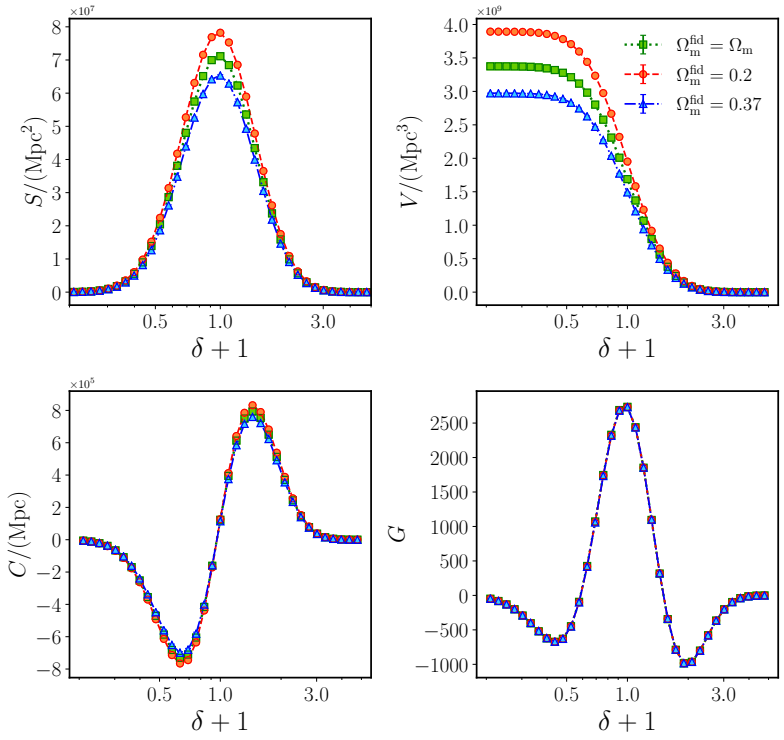

Figure 4.15: MFs for the mean of the 300 Minerva HOD catalogues as a function of the galaxy density contrast δ for three AP distorted boxes with different fiducial Ω'_m : a) the undistorted box with $\Omega'_\mathrm{m} = \Omega_\mathrm{m}$, b) $\Omega'_\mathrm{m} = 0.20$, c) $\Omega'_\mathrm{m} = 0.37$ .The densities are estimated with a kernel of a smoothing length λ corresponding to specific the interparticle and a truncation radius $r_\mathrm{cut} = 3\lambda$.

mapping between real and redshift space. However, an accurate modelling of such mapping would open up the possibility to use the measurements of all four MFs and to extract constraints on the growth-rate of cosmic structure. We leave such analysis for a future study.

4.4.3 Effect of Alcock-Paczynski distortions

The MFs measured from a real galaxy survey will depend on the fiducial cosmology assumed to transform the observed redshifts into comoving distances. Any difference between this cosmology and the true underlying one gives rise to AP distortions (Alcock & Paczynski (1979), see Section 2.8.3). The modelling of AP distortions is standard in the analysis of two-point statistics, but has mostly been ignored for MFs. We mimic the effect of AP distortions on our Minerva HOD samples by distorting the galaxy positions according to equations (2.99) to (2.101) by

$$x'_\perp = q_\perp^{-1} x_\perp, \tag{4.15}$$

for the two Cartesian axes perpendicular to the line of sight and

$$x'_\parallel = q_\parallel^{-1} x_\parallel, \tag{4.16}$$

for the line-of-sight coordinate. To obtain two different pairs of $q_\perp$ and $q_\parallel$, we compute the values of the comoving angular-diameter distance, $D_\mathrm{M}(z)$, and the Hubble parameter $H(z)$ using the true underlying matter density of the Minerva simulations and the fiducial values, $D'_\mathrm{M}(z)$ and $H(z)'$, using two different fiducial matter densities $\Omega'_\mathrm{m} = 0.20$ and $\Omega'_\mathrm{m} = 0.37$. We apply the same smoothing procedure as in Section 4.4.1 to the AP distorted HOD galaxy samples, where we again set the smoothing scale λ as the mean interparticle separation and $r_\mathrm{cut} = 3\lambda$. As the volumes of the AP distorted boxes change with respect to the undistorted reference one, also the mean interparticle separations, and therefore the corresponding values of λ and r_cut, are adjusted accordingly.

We used MEDUSA to measure the MFs of the resulting density fields using the same density thresholds as in Section 4.4.1. Fig. 4.15 shows the mean global MFs as function of the density contrast δ of the original boxes (green points) and the two distorted cases (orange and blue points). There are obvious differences in the amplitudes of S, V, and C for the three different choices of fiducial matter densities. As the topology of the galaxy density field is not changed by the coordinate transformations of equations (4.15) and (4.16), the genus is the same in all cases.

As MFs are angle-averaged measurements, they are sensitive to the isotropic AP parameter q from equation (2.104), which depends on the volume-averaged distance $D_\mathrm{V}(z)$ given by the combination of $D_\mathrm{M}(z)$ and $H(z)$ of equation (2.19). The coordinate transformation associated with AP distortions is described by the Jacobian of the volume and surface integrals of the MFs. Following from this, the global MFs transform under AP distortions as

$$S = q^2\, S', \tag{4.17}$$
$$V = q^3\, V', \tag{4.18}$$
$$C = q\, C', \tag{4.19}$$

while the genus remains unaffected. Equivalently, the AP distorted MF densities can be rescaled by the factors q^α, with $\alpha = 0$, -1, -2, -3 for f'_V, s', c', and g' to obtain the undistorted MF densities. Fig. 4.16 shows the global MFs rescaled by the corresponding powers of q, which are in excellent agreement with the undistorted reference measurements.

The correction factors of equations (4.17) – (4.19) must be taken into account before any model of the MFs can be compared against measurements inferred from galaxy redshift surveys. They also show that these measurements can be used to constrain q, and hence the volume-averaged distance $D_\mathrm{V}(z)$. This was the approach followed by Blake et al. (2014), who used Gaussian theory predictions to derive constraints on $D_\mathrm{V}(z)$ from the differential MFs of WiggleZ. As we discussed in Section 4.4.1, the Gaussian predictions do not give a correct description of the MFs of our HOD catalogues, indicating that the derivation of unbiased constraints on $D_\mathrm{V}(z)$ from real galaxy samples with similar clustering properties (such as the BOSS CMASS sample) would require a more accurate treatment of the impact of non-linearities on the MFs.

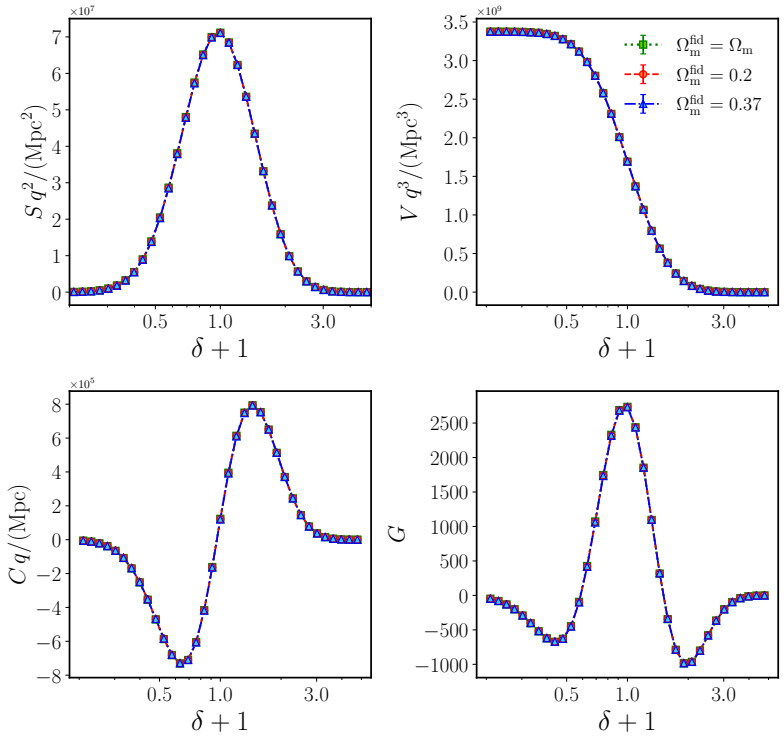

Figure 4.16: The same as Fig. 4.15, but now the MFs are rescaled by the corresponding powers of the isotropic AP parameter q.

4.5 Covariance matrix for Minkowski functionals

One of the main prerequisites for performing a likelihood analysis of MF measurements will be the estimation of the covariance matrix (see Section 3.2). In this context, this section shows the measurements of the covariance matrices from the MFs of the Minerva HOD catalogues.

We compute the covariance matrix by using equation (3.8), where we insert the MF measurements from the 300 Minerva catalogues in place of the two-point statistics. We measure the MFs for 10 density thresholds equispaced in logarithmic scale around the mean density contrast $\delta = 0$, and hence the total size of the resulting covariance matrix is 40x40. Fig. 4.17 shows the resulting correlation matrices, computed from equation 3.15, for real and redshift space. We note the rich structure of the correlation matrices. The MF measurements of the high and the low density thresholds are correlated with each other and there is pronounced cross-correlation between the different MFs, in particular for the surface, volume and curvature MF.

The real and redshift-space correlation matrices exhibit a very similar structure. Small differences in the variances $\sigma = \sqrt{C_{ii}}$ of the MF measurements in real and redshift space are displayed in Fig. 4.18. The redshift-space variances are slightly increased at the low- and at the high-density ends.

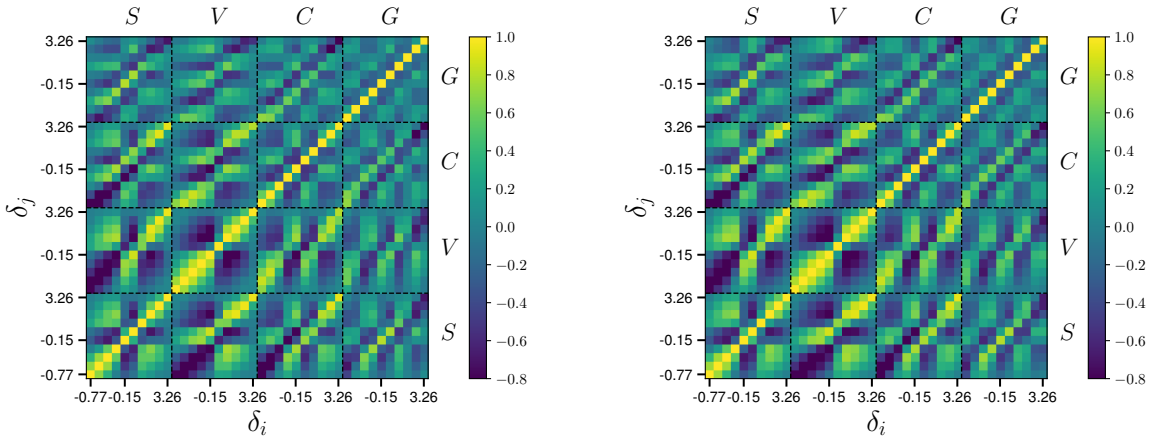

Figure 4.17: The full correlation matrix inferred from the Minkowski functionals of the 300 Minerva HOD catalogues in real space (left panel) and redshift space (right panel).

In their analysis of the MFs measured from the WiggleZ survey, Blake et al. (2014) propose to use differential MFs, in order to reduce the covariance between the different density thresholds and obtain a covariance matrix that is more closely diagonal. Due to the additivity of the MFs, the differential MFs can be defined as

$$M'(\delta) = \frac{\Delta M(\delta)}{\Delta \delta}, \tag{4.20}$$

where M is the surface, volume, curvature or genus MF, $M = S, V, C, G$, and $\Delta M(\delta)$ is the difference between the MF measurements of two adjacent density thresholds. Fig. 4.19 shows the correlation matrices computed from the measurements of the differential MFs of Minerva HOD catalogues in real and redshift space for the same density thresholds as in the previous Fig. 4.17. The correlations between the different density thresholds appears to be alleviated compared to the Fig. 4.17, but the correlation matrix still exhibits notable off-diagonal structure.

The measurements in this section show that we have to take the covariance between the different density thresholds and the cross-covariance between the surface, volume, curvature and genus MFs into account for future likelihood analyses. As in the case of the two-point clustering measurements, also for MF analyses we might rely on the approximate methods described in Section 3.5, in order to obtain a large enough number of mocks for robust covariance matrix estimates. We expect that the approximate methods that reproduce well the bispectrum covariances in Colavincenzo et al. (2019), in particular the predictive methods, will also be a suitable choice for the estimation of the MF covariance matrix, since the bispectrum already includes the 3-point order correlations. A thorough validation of the approximate methods for MF covariances is left for future study.

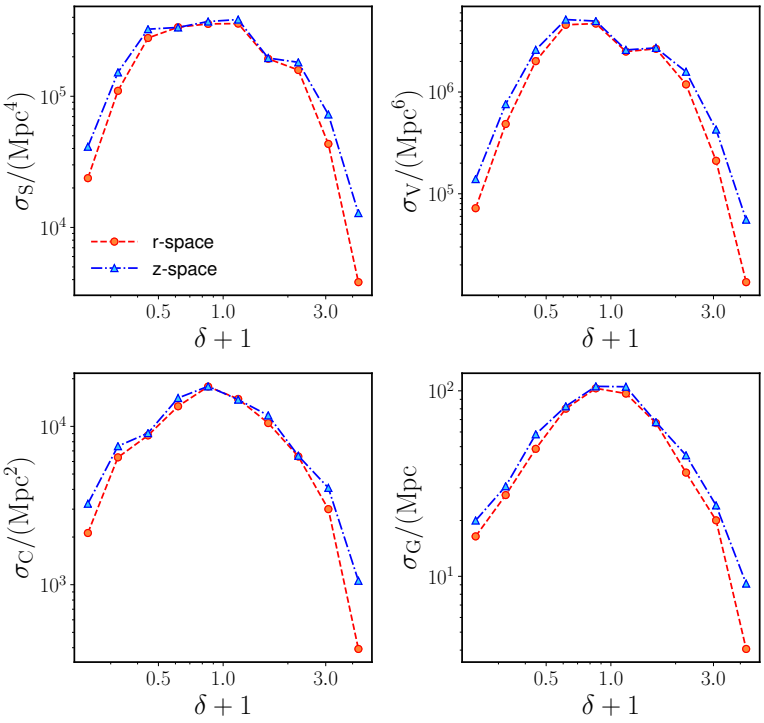

Figure 4.18: Comparison of the variance of the Minkowski functionals measurements from the Minerva HOD catalogues in real and redhsift space

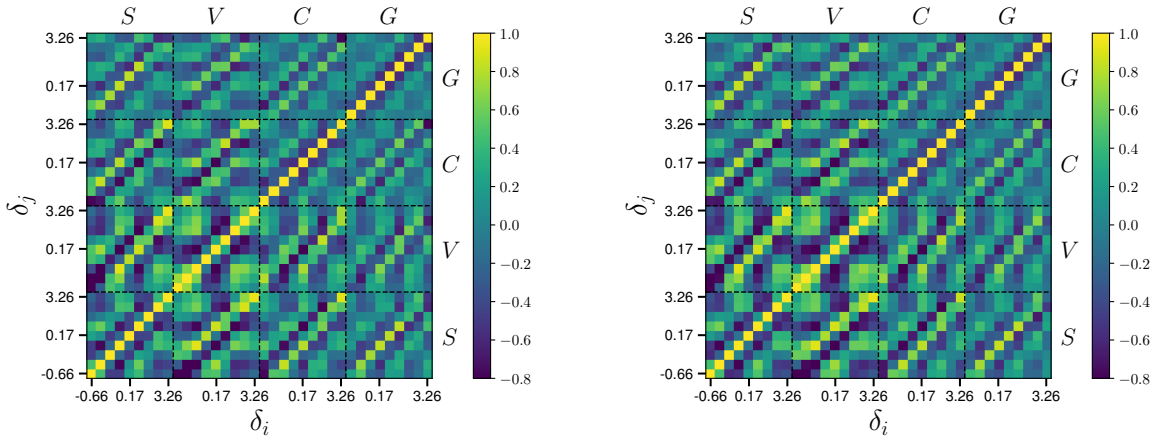

Figure 4.19: The same as Fig. 4.17, but now for the differential Minkowski functionals.

4.6 Evolution mapping of Minkowski Functionals

4.6.1 Evolution mapping of two-point statistics and beyond

The theoretical modelling of previous isodensity MF analyses, such as the one by Blake et al. (2014), is commonly based on the predictions for a Gaussian density field of equations (2.68) to (2.77). However, in the MF measurements from the Minerva simulations we found clear deviations from the Gaussian case. Hence, for a likelihood analysis of MF measurements from real galaxy catalogues of comparable or better quality, we need a model for the MFs valid for the non-linear and thus non-Gaussian density field.

A possible strategy is to implement the analytical models for the MFs of weakly non-Gaussian fields that have been derived in the recent years (Pogosyan et al., 2009; Matsubara, 2010; Gay et al., 2012; Codis et al., 2013; Matsubara & Kuriki, 2020; Matsubara et al., 2020). Their application to large-scale structure analysis has practically not been explored yet, and it will be interesting to test these predictions for density fields in the highly non-linear regime and to study their dependence on specific cosmological parameters. We leave this approach for future research.

An alternative for predicting the MFs of non-linear density fields is to design a suitable emulator based on simulations. Several emulators for the non-linear power spectrum have been built, such as the CosmicEmu (Lawrence et al., 2010; Heitmann et al., 2014; Casarini et al., 2016) and the EuclidEmulator (Euclid Collaboration et al., 2019). They were constructed by first performing a set of N-body simulations that sample a specific cosmological parameter space in a given redshift range, and by finally incorporating an interpolation scheme to obtain predictions for the non-linear power spectrum for any parameter combination in the sampled space. The limitations of such emulators are that their predictions might not be valid for different cosmological parameter spaces and arbitrary redshifts.

To alleviate the limitations, Sánchez et al. (in prep.) recently developed a novel approach to construct emulators for the non-linear power spectrum for general cosmological parameter spaces and redshift ranges. The underlying idea is based on the fact that the cosmological parameters can be classified according to their impact on the linear matter power spectrum into *shape* and *evolution* parameters. The shape parameters, $\boldsymbol{\Theta}_\mathrm{s}$, define the shape of the linear theory power spectrum. These are, for example, the physical densities (see Section 2.1.1) of baryons ω_b, cold dark matter ω_c, or the spectral index n_s,

$$\boldsymbol{\Theta}_\mathrm{s} = (\omega_\mathrm{b}, \omega_\mathrm{c}, n_\mathrm{s}, \dots)\,. \tag{4.21}$$

The evolution parameters, $\boldsymbol{\Theta}_\mathrm{e}$, determine the amplitude of the linear power spectrum at a given redshift z. This group of parameters consists of, for example, the scalar spectral amplitude A_s, the curvature density ω_K, the dark energy density ω_DE, and dark energy equation-of-state parameter w_DE, which can either be constant or evolving with the scale factor a (c.f. equation (2.14)),

$$\boldsymbol{\Theta}_\mathrm{e} = (A_\mathrm{s}, \omega_K, \omega_\mathrm{DE}, w_\mathrm{DE}(a), \dots)\,. \tag{4.22}$$

The impact of the evolution parameters can be completely characterized by $\sigma_{12}(z, \mathbf{\Theta}_\mathrm{e})$, the rms linear perturbation theory variance in spheres of radius $r = 12\,\mathrm{Mpc}$ (see equation (2.42), and also Sánchez, 2020). The linear power spectrum for a given set of shape parameters $\mathbf{\Theta}_\mathrm{s}$ and evolution parameters $\mathbf{\Theta}_\mathrm{e}$ at a redshift z can then be written as

$$P_\mathrm{L}(k|z, \mathbf{\Theta}_\mathrm{s}, \mathbf{\Theta}_\mathrm{e}) = P_\mathrm{L}\left(k|\mathbf{\Theta}_\mathrm{s}, \sigma_{12}\left(z, \mathbf{\Theta}_\mathrm{e}\right)\right). \tag{4.23}$$

Note that the traditional σ_8 is not a suitable choice, since it depends on h, which is a combination of shape and evolution parameters.

Based on equation 4.23, power spectra defined by the same set of shape parameters and different evolution parameters are the same if the corresponding values of σ_{12} are identical. At the linear level, the time evolution of these power spectra can then be mapped from one to the others by matching the redshifts that correspond to the same values of σ_{12}.

To validate this approach Sánchez et al. (in prep.) constructed a set of simulations with eight different input cosmologies, which are characterized by the same shape parameters and different evolution parameters that lead to the same value of σ_{12} at redshift $z = 0$. These simulations will be described in more detail in the following Section 4.6.2. The left panel of Fig. 4.20 shows the linear matter power spectra for these eight different cosmologies at five redshifts that are chosen such that the $\sigma_{12}(z)$ values of each case match the $\sigma_{12}(z)$ values of the reference cosmology at $z \in \{2.0, 1.0, 0.57, 0.3, 0.0\}$. As expected, the linear power spectra all perfectly agree. Note that they use Mpc units and not the traditional units h^{-1}Mpc, which would lead to changes in the power spectrum shape for cosmologies with different h values.

Furthermore, Sánchez et al. (in prep.) show that this approach can be extended to the non-linear power spectrum with high accuracy. The right panel of Fig. 4.20 displays the non-linear matter power spectra measured from the simulations with these eight different input cosmologies (models) at the same redshifts as the left panel. The agreement of these measurements is remarkable. For the three lower σ_{12} values corresponding to higher redshifts the deviations compared to the measurements from the reference case are less than 1%. For the second-highest and highest σ_{12} values slightly larger deviations of up to 4% arise for some of the models at scales much larger than those currently used by power spectrum analyses.

This means that from a single set of simulations with one input cosmology, the redshift evolution for various non-linear power spectra with different evolution parameters can be predicted, also for non-ΛCDM models. Therefore, the evolution mapping is a promising approach for the construction of emulators that are valid for a wide range of cosmologies and redshifts.

The aim of this section is to test if this approach can be extended to the MFs. In particular, we test if the impact of the evolution parameters on the MFs can also be completely characterized by $\sigma_{12}(z, \mathbf{\Theta}_\mathrm{e})$. This is related to the question whether the evolution mapping approach is also valid for higher-order correlations.

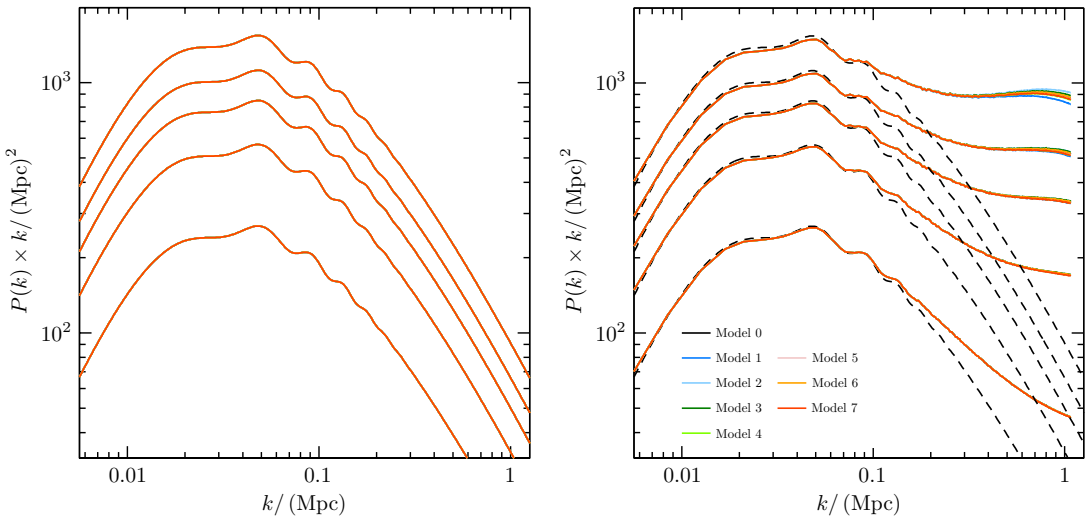

Figure 4.20: Redshift evolution of the linear matter power spectra (left panel) and the non-linear matter power spectra (right-panel) from the Columbus simulations with eight different input cosmologies characterized by equal shape parameters, but different evolution parameters (more details in Section 4.6.2). Each of the five redshifts for each cosmology is chosen such that it gives the same value of σ_{12} as the reference cosmology at $z \in \{2.0, 1.0, 0.57, 0.3, 0.0\}$. Image credit: Sánchez et al. (in prep.)

4.6.2 Columbus simulations and halo catalogues

The setup for our analysis is the same as in Sánchez et al. (in prep.). It is based on the Columbus simulations, a set of "fixed-paired" N-body simulations with eight different cosmologies. The name "fixed-paired" derives from the fact that the initial realizations were generated in pairs by fixing the amplitudes of the initial Fourier modes to a specific power spectrum, and with the initial modes exactly out of phase. Angulo & Pontzen (2016) proposed this "fixed-paired" technique to reduce the variance caused by the sparse sampling of modes in simulations. They showed that one pair of fixed simulations can correctly reproduce the average dark matter power spectrum from an ensemble of 300 simulations that were generated with the standard Gaussian random initial conditions. Therefore, this technique is ideal for exploring the cosmological parameter space with simulations.

Table 4.1: The cosmological parameters for the reference Columbus simulations.

ω_b	ω_c	h	n_s	ω_K	ω_Λ	w_DE
0.022445	0.120567	0.67	0.96	0	0.305887	-1

The eight different input cosmologies are chosen such that they have the same shape parameters and the same value of σ_{12} at redshift $z = 0$, $\sigma_{12}(z = 0) = 0.825$. The first cosmology corresponds to a standard flat ΛCDM model close to the Planck cosmology

Table 4.2: The cosmological parameters that define the different eight input cosmologies of the Columbus simulations and the redshifts of the five snapshots for each cosmology that yield the same $\sigma_{12}(z)$ as the reference simulation (model 0).

Model	Cosmology	z_0 : $\sigma_{12}(z) = 0.343$	z_1 : $\sigma_{12}(z) = 0.499$	z_2 : $\sigma_{12}(z) = 0.611$	z_3 : $\sigma_{12}(z) = 0.703$	z_4 : $\sigma_{12}(z) = 0.825$
0	Reference ΛCDM from table 4.1.	2.000	1.000	0.570	0.300	0.00
1	ΛCDM, $h = 0.55$.	1.761	0.859	0.480	0.248	0.00
2	ΛCDM, $h = 0.79$.	2.231	1.137	0.659	0.352	0.00
3	wCDM, $w_{\rm DE} = -0.85$.	2.100	1.044	0.590	0.307	0.00
4	wCDM, $w_{\rm DE} = -1.15$.	1.923	0.964	0.553	0.293	0.00
5	Dynamic DE, $w_0 = -1$,$w_a = -0.2$.	1.973	0.990	0.566	0.299	0.00
6	Dynamic DE, $w_0 = -1$,$w_a = 0.2$.	2.031	1.011	0.574	0.301	0.00
7	curved ΛCDM, $\Omega_K = -0.05$.	1.938	0.978	0.561	0.297	0.00

from the *Planck* Collaboration et al. (2020, c.f. Section 2.1.2) and is denoted by "model 0" as in Fig. 4.20. The parameter values for this reference cosmology are listed in table 4.1. Note that we express the energy densities as the physical densities (see Section 2.1.1), which do not depend on the value of h, such that we can classify shape and evolution parameters as described in the previous section. For each case of the remaining seven cosmologies (model 1-7), the value of one evolution parameter is changed compared to the reference cosmology. There are two cosmologies with different h values (model 1 & 2), two cosmologies with a different dark energy equation-of-state parameter $w_{\rm DE}$ (model 3 & 4), two dynamic dark energy models with the parametrization of equation (2.14) with different values for w_a (model 5 & 6), and finally one cosmology with non-zero curvature (model 7). These models do not correspond to cosmologies allowed by present-day observations, but are suitable models for our tests. The parameters that are different for the considered cases compared to the reference cosmology can be found in the column "Cosmology" in table 4.2.

For each of these cosmologies, two realizations were generated according to the "fixed-paired" technique. The initial density fields were generated with 2LPT at redshift $z_{\rm ini} = 99$ using the linear power spectra of the different cosmologies computed by CAMB (Lewis et al., 2000, c.f. Section 2.4) as input. Each realization simulates 1500^3 dark-matter particles in a cubic box of side length $L = 1492.5$ Mpc, corresponding to $1000\,h^{-1}$ Mpc for the reference cosmology, with periodic boundary conditions. The simulations were performed using a modified version of GADGET-2 (Springel, 2005, c.f. Section 2.6) that includes the background evolution for dynamic dark energy models. For the simulations of the reference cosmology, the positions and velocities of the evolved DM particles were stored in five snapshots corresponding to $z_0 = 2.0$, $z_1 = 1.0$, $z_2 = 0.57$, $z_3 = 0.3$ $z_4 = 0.0$.

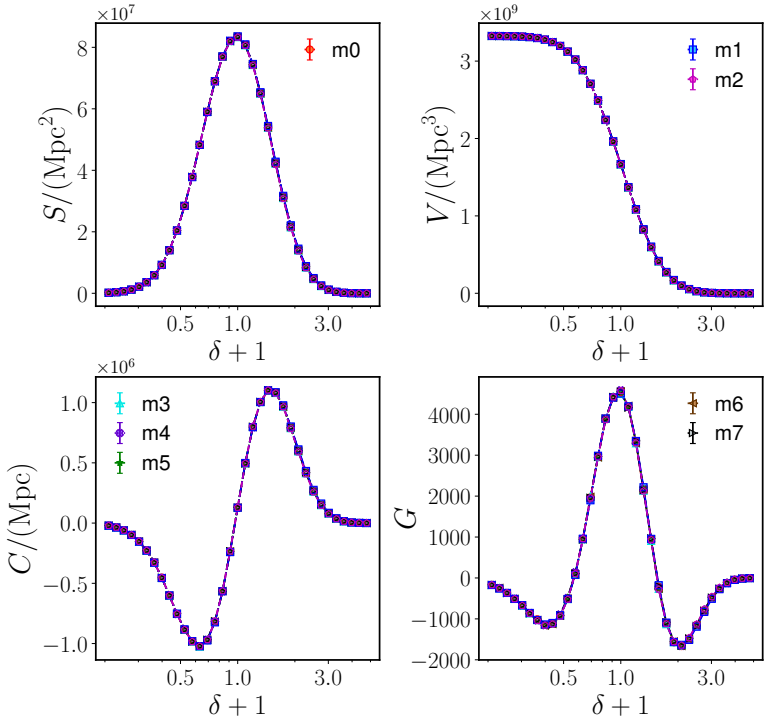

Figure 4.21: Mean MFs measured for the halo samples from the eight pairs of the Columbus simulations with different cosmologies (model 0–7) at $z = 0$ as function of the density contrast δ. The error bars corresponding for each case to the standard deviation from the pair of realizations are of the size of the points or smaller and therefore not visible. The densities were estimated with a smoothing length λ corresponding to the mean interparticle separation of the model 0 halo samples.

For the simulations of the other cosmologies, the redshifts of the snapshots were chosen such that they have the same value of $\sigma_{12}(z)$ as the one of the reference cosmology. The values of $\sigma_{12}(z)$ and the corresponding redshifts of the snapshots are listed in table 4.2.

Halos were identified with the ROCKSTAR halo finder (Behroozi et al., 2013), which is an improved variant of a FoF halo finder. ROCKSTAR builds a hierarchy of FoF subgroups by adaptively reducing the six-dimensional linking length, which also takes the velocities of the particles into account. In that way it can better resolve the substructure compared to other halo finders and identify the halos consistently across the time-steps of the simulations. Here, we consider all halos with masses above $10^{13}\,\mathrm{M}_{\odot}$. More details on the Columbus simulations will be published soon by Sánchez et al. (in prep.).

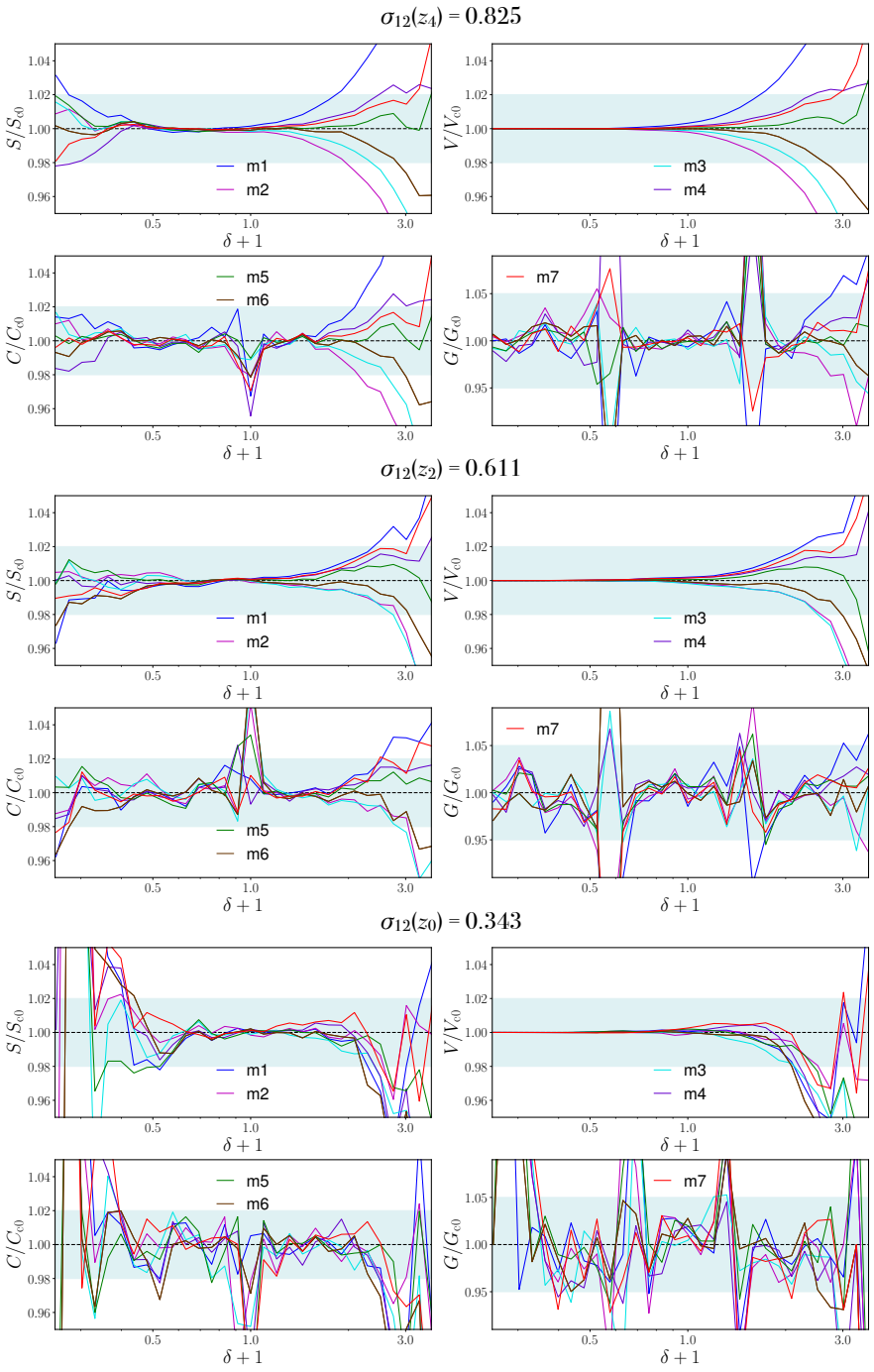

Figure 4.22: Ratios of the mean MF measurements from the halo samples of the cosmologies 1–7 defined in 4.2 with respect to the ones from the reference cosmology 0 for the snapshots corresponding to low, intermediate and high σ_{12} values as defined in Section 4.6.2. The light grey bands indicate a range of $\pm 2\%$ deviation for the volume, surface and curvature, and of $\pm 5\%$ for the genus.

4.6.3 The dependence of the Minkowski functionals on the cosmological evolution parameters

To test the evolution mapping approach for the MFs, we apply MEDUSA to the halo samples from the five snapshots of the Columbus simulations. We use a halo mass cut of $10^{13}\,\mathrm{M}_{\odot}$, which yields roughly the same halo number density for all catalogues inferred from one snapshot defined by $z_{0,\dots,4}$ in table 4.2. For the density estimation at the halo positions, we smooth the distribution with a Gaussian kernel as described in Section 4.3. We apply a smoothing length λ corresponding to the mean interparticle separation in the pair of reference halo samples (model 0) for all samples of one snapshot $z_{0,\dots,4}$, and a truncation radius $r_{\mathrm{cut}} = 3\lambda$.

Analogous to Section 4.4.1, we compute the MFs on 35 density thresholds equispaced in logarithmic scale around the mean density contrast $\delta = 0$. Fig. 4.21 shows the mean MF measurements from the halo samples of the eight pairs of the Columbus simulations (model 0–7) for the snapshot corresponding to $z_4 = 0$. No differences between the measurements from the different cases are visible. We note, that we again find the clear non-Gaussian signature in the asymmetry of the genus. For the other snapshots, we find very similar results and therefore do not show them here.

The differences between the models can be seen in Fig. 4.22, which displays the ratios of the measurements for models 1–7 with respect to the reference ones for model 0 in three panels corresponding to $\sigma_{12}(z) = 0.323$, $\sigma_{12}(z) = 0.611$ and $\sigma_{12}(z) = 0.825$. For all snapshots $z_{0,\dots,4}$, we find that in the density range of $-0.6 \leq \delta \leq 3.0$ the overall agreement with the reference measurement is better than 2% for the surface, volume and curvature. For the genus, the differences are in the 5% range, since its measurements are noisier than the ones of the other MFs. The noise in all measurements increases for lower σ_{12} values due to the lower number density of the corresponding halo samples. We note that the measurements from the different models are more similar for the lower σ_{12} values. The largest differences appear for the cosmologies with different h values (model 1 & 2) at $\sigma_{12} = 0.825$. This behaviour is analogous to the measurements of the non-linear matter power spectrum by Sánchez et al. (in prep.).

Our results show that the impact of the evolution parameters on the MFs can be characterized with high accuracy by σ_{12} and hence imply that the evolution mapping equation (4.23) can be transferred to the MFs as

$$M(\delta|z, \mathbf{\Theta}_{\mathrm{s}}, \mathbf{\Theta}_{\mathrm{e}}) \approx M\left(\delta|\mathbf{\Theta}_{\mathrm{s}}, \sigma_{12}\left(z, \mathbf{\Theta}_{\mathrm{e}}\right)\right). \tag{4.24}$$

This relation can be of great use for building a simulation-based model for the MFs of non-linear density fields. It allows us to describe the dependence of the MFs on the evolution parameters based on σ_{12} and to explore large parameter spaces starting from a single set of shape parameters.

In upcoming studies, we intend to analyse the mapping of the MFs between different redshifts in more detail and to examine the impact of bias, since here we have focused on halo samples defined with the same mass threshold. Finally, our results suggest that the

evolution mapping approach is valid for higher N-point statistics. It would be interesting to further confirm this approach in a similar analysis of the bispectrum.

Chapter 5

Summary and Outlook

The large-scale structure traced by galaxies carries a wealth of cosmological information. Extracting this information is a challenging endeavour involving many different components. This thesis examined two useful tools for the statistical analysis of the large-scale galaxy clustering: covariance matrices and Minkowski functionals.

First, we considered the covariance matrices of anisotropic two-point clustering measurements in configuration space. Chapter 3 presented an extensive comparison of the covariance matrix estimates from several approximate methods. We included seven approximate methods from three broad categories: predictive methods (ICE-COLA, PEAK PATCH, and PINOCCHIO), calibrated methods (HALOGEN and PATCHY), and recipes assuming specific shapes of the density probability distribution function (log-normal and Gaussian density fields).

For the predictive and calibrated methods, we generated sets of 300 halo catalogues, matching the initial conditions of Minerva, our reference N-body simulations. Furthermore, we produced a set of 1000 log-normal catalogues with the same number density and mean correlation function as the N-body simulations, and computed the theoretical prediction for the Gaussian covariance matrix. We defined two halo samples from the Minerva simulations with the lowest halo mass corresponding to 42 and 100 dark matter particles, respectively. We constructed equivalent samples from the approximate mocks by matching the mass threshold, number density and clustering amplitude of the parent samples from the N-body simulations.

Our comparison focused on the performance of the covariance matrices estimated from the approximate halo samples at inferring cosmological parameters. We constructed synthetic clustering measurements based on the theoretical model for the non-linear power spectrum that was used in recent LSS analyses (Sánchez et al., 2017; Grieb et al., 2017; Salazar-Albornoz et al., 2017; Hou et al., 2018). Then, we fitted these synthetic data with the same baseline model, using the covariances from the different methods, and analysed the obtained parameter constraints on the AP parameters $\alpha_{\parallel}$, $\alpha_{\perp}$ and the growth rate $f\sigma_8$.

The mean parameter values obtained from all these fits agree all perfectly with the N-body results. The marginalised parameter errors reproduce those from the N-body analysis within the expected statistical uncertainty of 5% for the lower mass cut and 10% for the

higher mass cut. The allowed statistical volumes in the three-dimensional parameter space of $\alpha_\parallel$, $\alpha_\perp$, and $f\sigma_8$ showed differences of up to 20% for the halo samples based on the same approximate method but different selection criteria, i.e. by mass, number density, or bias matching. Therefore, we selected for each case the matching scheme that yielded the closest agreement with the N-body results. Finally, we found that the allowed statistical volumes from the best-matched halo samples agree with the N-body results mostly within 10%, with no method performing significantly better than the others.

Our main conclusion from the covariance matrix comparison are:

(i) Due to their similar performance, there is no preference for a specific approximate method for the covariance estimates of two-point correlation function measurements.

(ii) The decisive criterion for choosing a particular approximate method might be the computational cost. Although the calibrated methods are computationally less expensive than the predictive methods, it has to be taken into account that also the calibration to N-body simulations can be challenging and time-consuming. In this context, it is noteworthy that the simple Gaussian prediction performs similar to the other approximate methods.

Chapter 4 was devoted to the Minkowski functionals as geometrical and topological descriptors of the galaxy density field. We developed MEDUSA, a code for the accurate estimation of isodensity Minkowski functionals from three-dimensional point distributions. MEDUSA performs three main steps: First, the density values at every point in the input sample are estimated using a Gaussian kernel with a fixed smoothing length. This step can also be skipped if the densities are already known, or adapted to a different recipe for density estimation. Secondly, triangulated isodensity surfaces are constructed from the Delaunay tessellation of the input points. MEDUSA selects the tetrahedra from the Delaunay tessellation that contain vertices with densities above and below a chosen threshold. Then it finds the intersection of the isodensity surface by linearly interpolating the density between those vertices. The third step is the computation of the four Minkowski functionals, volume, surface area, integrated mean curvature and Euler characteristic, on the triangulated surface by summing over all contributions from the tetrahedra and triangles. MEDUSA is a refined version of the algorithm by Yaryura et al. (2004), which can also account for periodic boundary conditions.

We tested MEDUSA for several point samples with different geometrical and topological properties and known densities. We applied MEDUSA to spherical, ellipsoidal and toroidal density distributions and computed the corresponding theoretical predictions for the Minkowski functionals. We also included different spherical distributions intersecting the edges of a box with periodic boundary conditions to our tests. The estimated volume, surface area and extrinsic curvature agreed significantly better than one per cent with the theoretical predictions for all cases. The Euler characteristic, and equivalently the genus, were always computed exactly. Also, we generated 100 Gaussian random fields with the linear power spectrum from the Minerva simulations as input and periodic boundary conditions, since they have known analytical predictions for the Minkowski functionals, which

are sensitive to the power spectrum. We found that the Minkowski functionals estimated by MEDUSA agree well with the theoretical predictions.

After the validation tests, we applied MEDUSA to the 300 HOD galaxy catalogues from the Minerva simulations at $z = 0.57$. For the density estimation, we used a smoothing length matching the mean inter-particle separation of the sample. We computed the Minkowski functionals of the HOD catalogues as functions of the density contrast δ using MEDUSA. We also expressed the resulting measurements as functions of the volume-filling fraction f_V.

Our analysis focused on three aspects crucial for the Minkowski functional measurements of real galaxy redshift surveys: non-Gaussian signatures due to non-linear gravitational evolution, redshift-space distortions (RSD), and Alcock-Paczynski (AP) distortions. We found non-Gaussian signatures in the measured Minkowski functionals, in particular, the asymmetric genus. The measurements of the Minkoswki functionals in redshift space were affected by RSD. However, when expressed as a function of f_V, the impact of RSD was significantly reduced. AP distortions only changed the measured volume, surface area, and curvature. The topology of the density field should not be affected by AP distortions. In agreement with this expectation, we found that the Euler characteristic and hence the genus remained unchanged. We could account for the AP distortions by rescaling the Minkowski functionals by the corresponding powers of the isotropic AP parameter q.

As the next point, we computed the covariance matrices for the Minkowski functional measurements from the Minerva HOD catalogues. We found significant correlations between different density thresholds and between the different Minkowski functionals. The covariance between the density thresholds could be reduced by using the differential Minkowski functionals introduced by Blake et al. (2014).

Finally, we tested the novel approach dubbed evolution mapping by Sánchez et al. (in prep.) to study the cosmology dependence of the Minkowski functionals. The underlying idea is to classify the cosmological parameters into shape and evolution parameters according to their effect on the linear power spectrum, and to characterize the impact of the evolution parameters by σ_{12}, the linear-theory rms mass fluctuation in spheres of radius 12 Mpc. We applied MEDUSA to the halo catalogues from five snapshots of the Columbus simulations, a set of "paired-fixed" simulations with eight different input cosmologies. They were constructed such that they have the same shape parameters and the same value of σ_{12} irrespective of their redshifts, but different evolution parameters. We found that the Minkowski functionals from the different cosmologies overall agree within 5% for all snapshots corresponding to different redshifts.

Our main conclusion from the analysis of the Minkowski functionals are:

(i) Since the measured Minkowski functionals are sensitive to deviations from Gaussianity, and therefore encode information on higher-order correlations, they are promising tools to study the non-linear galaxy density field.

(ii) Expressing the Minkowski functionals as a function of f_V presents an opportunity to probe the deviations from Gaussianity in redshift space even without a detailed

model of the RSD. However, modelling the mapping between the real- and redshift-space densities would allow us to extract information on the growth rate of cosmic structure from Minkowski functionals.

(iii) As the impact of AP distortions on the Minkowski functionals can be described by the isotropic AP parameter q, the directly related volume-averaged distance $D_{\mathrm{V}}(z)$ can be constrained.

(iv) The covariance matrix of the Minkowski functionals is essential for future likelihood analyses due to its significant off-diagonal structure. Using the differential Minkowski functionals can be beneficial because the covariance between the density thresholds is alleviated.

(v) The impact of the evolution parameters on the Minkowski functionals can be characterized to high accuracy by σ_{12}. Therefore, the evolution mapping approach can be a good starting point for building a simulation-based model for the Minkowski functionals of non-linear density fields.

The findings in this thesis lead to several interesting questions for future research. For the covariance matrices, it will be useful to include the effect of the survey geometry in further studies, to assess the impact of applying approximate methods to the analysis of real galaxy surveys. So far, the considered approximate methods have neglected neutrino effects. For the analysis of future clustering measurements, it might be necessary to incorporate massive neutrinos. The likelihood-based analysis of Minkowski functional measurements will also require robust covariance matrix estimates. We expect that the predictive methods, which accurately reproduce the bispectrum covariances in Colavincenzo et al. (2019), are also suitable for the Minkowski functionals. However, this has to be confirmed through careful examination.

In upcoming studies, we can explore different avenues for the Minkowski functionals. An interesting topic is the modelling of the effect of redshift-space distortions since it could allow us to constrain the growth rate from Minkowski functional measurements. Further, accurate predictions for the Minkowski functionals of the non-linear galaxy density field will be crucial for future analysis. We could address this issue by implementing and testing the analytical models for the Minkowski functionals of weakly non-Gaussian density fields that were developed in the last years (e.g., Codis et al., 2013; Matsubara & Kuriki, 2020).

In this thesis, we have laid the foundation for an alternative simulation-based approach. We might exploit the evolution mapping idea and design an emulator for the non-linear Minkowski functionals in forthcoming studies. This would also open the path to assessing the sensitivity of the Minkowski functionals to specific cosmological parameters. In this context, it could also be possible to study the effect from massive neutrinos on the Minkowski functionals. To explore the cosmological implications of Minkowski functional measurements inferred from real galaxy surveys, will further require to examine and account for the survey geometry. The final goal of the described future avenues is to advance the analysis of Minkowski functionals as a powerful complement of the standard two-point clustering statistics.

Bibliography

Agrawal, A., Makiya, R., Chiang, C.-T., Jeong, D., Saito, S. & Komatsu, E., "Generating log-normal mock catalog of galaxies in redshift space", 2017, *J. Cosmol. Astropart. Phys.*, 10, 003, arXiv: 1706.09195.

Alam, S., Albareti, F. D., Allende Prieto, C. et al., "The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III", 2015, *ApJS*, 219, 12, arXiv: 1501.00963.

Alam, S., Ata, M., Bailey, S. et al., "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample", 2017, *MNRAS*, 470, 2617-2652, arXiv: 1607.03155.

Albrecht, A., Bernstein, G., Cahn, R. et al., "Report of the Dark Energy Task Force", 2006, *ArXiv e-prints*, arXiv: astro-ph/0609591.

Alcock, C. & Paczynski, B., "An evolution free test for non-zero cosmological constant", 1979, *Nature*, 281, 358.

Angulo, R. E. & Pontzen, A., "Cosmological N-body simulations with suppressed variance", 2016, *MNRAS*, 462, L1-L5, arXiv: 1603.05253.

Aragon-Calvo, M. A., Shandarin, S. F. & Szalay, A. "Geometry of the cosmic web: Minkowski functionals from the delaunay tessellation", in *2010 International Symposium on Voronoi Diagrams in Science and Engineering*, 2010, pp. 235–243.

Avila, S., Murray, S. G., Knebe, A., Power, C., Robotham, A. S. G. & Garcia-Bellido, J., "HALOGEN: a tool for fast generation of mock halo catalogues", 2015, *MNRAS*, 450, 1856-1867, arXiv: 1412.5228.

Balaguera-Antolínez, A., Kitaura, F.-S., Pellejero-Ibáñez, M., Zhao, C. & Abel, T., "BAM: bias assignment method to generate mock catalogues", 2019, *MNRAS*, 483, L58-L63, arXiv: 1806.05870.

Balaguera-Antolínez, A., Kitaura, F.-S., Pellejero-Ibáñez, M. et al., "One simulation to have them all: performance of the Bias Assignment Method against N-body simulations", 2020, *MNRAS*, 491, 2565-2575, arXiv: 1906.06109.

Behroozi, P. S., Wechsler, R. H. & Wu, H.-Y., "The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores", 2013, *ApJ*, 762, 109, arXiv: 1110.4372.

Bernardeau, F., Colombi, S., Gaztañaga, E. & Scoccimarro, R., "Large-scale structure of the Universe and cosmological perturbation theory", 2002, *Phys. Rep.*, 367, 1-248, arXiv: astro-ph/0112551.

Bertschinger, E., "Self-similar secondary infall and accretion in an Einstein-de Sitter universe", 1985, *ApJS*, 58, 39-65.

Bertschinger, E., "Simulations of Structure Formation in the Universe", 1998, *ARA&A*, 36, 599-654.

Blake, C. & Glazebrook, K., "Probing Dark Energy Using Baryonic Oscillations in the Galaxy Power Spectrum as a Cosmological Ruler", 2003, *ApJ*, 594, 665-673, arXiv: astro-ph/0301632.

Blake, C., James, J. B. & Poole, G. B., "Using the topology of large-scale structure in the WiggleZ Dark Energy Survey as a cosmological standard ruler", 2014, *MNRAS*, 437, 2488-2506, arXiv: 1310.6810.

Blas, D., Lesgourgues, J. & Tram, T., "The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes", 2011, *J. Cosmol. Astropart. Phys.*, 7, 034, arXiv: 1104.2933.

Blot, L., "MPA Lecture on Numerical Methods for Cosmology", 2020, `https://wwwmpa.mpa-garching.mpg.de/~komatsu/lecturenotes/Linda_Blot_on_NumericalMethods.pdf`.

Blot, L., Crocce, M., Sefusatti, E. et al., "Comparing approximate methods for mock catalogues and covariance matrices II: power spectrum multipoles", 2019, *MNRAS*, 485, 2806-2824, arXiv: 1806.09497.

Bond, J. R. & Myers, S. T., "The Peak-Patch Picture of Cosmic Catalogs. I. Algorithms", 1996, *ApJS*, 103, 1.

Casarini, L., Bonometto, S. A., Tessarotto, E. & Corasaniti, P. S., "Extending the Coyote emulator to dark energy models with standard w_0-w_a parametrization of the equation of state", 2016, *J. Cosmol. Astropart. Phys.*, 8, 008, arXiv: 1601.07230.

Chan, K. C. & Scoccimarro, R., "Halo sampling, local bias, and loop corrections", 2012, *Phys. Rev. D*, 86, 10, 103519, arXiv: 1204.5770.

Choi, Y.-Y., Park, C., Kim, J., Gott, I., J. Richard, Weinberg, D. H., Vogeley, M. S., Kim, S. S. & SDSS Collaboration, "Galaxy Clustering Topology in the Sloan Digital Sky Survey Main Galaxy Sample: A Test for Galaxy Formation Models", 2010, *ApJS*, 190, 181-202, arXiv: 1005.0256.

Chuang, C.-H. & Wang, Y., "Measurements of H(z) and D_A(z) from the two-dimensional two-point correlation function of Sloan Digital Sky Survey luminous red galaxies", 2012, *MNRAS*, 426, 226-236, arXiv: 1102.2251.

Chuang, C.-H., Zhao, C., Prada, F. et al., "nIFTy cosmology: Galaxy/halo mock catalogue comparison project on clustering statistics", 2015, *MNRAS*, 452, 686-700, arXiv: 1412.7729.

Codis, S., Pichon, C., Pogosyan, D., Bernardeau, F. & Matsubara, T., "Non-Gaussian Minkowski functionals and extrema counts in redshift space", 2013, *MNRAS*, 435, 531-564, arXiv: 1305.7402.

Colavincenzo, M., Sefusatti, E., Monaco, P. et al., "Comparing approximate methods for mock catalogues and covariance matrices - III: bispectrum", 2019, *MNRAS*, 482, 4883-4905, arXiv: 1806.09499.

Cole, S., Percival, W. J., Peacock, J. A. et al., "The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications", 2005, *MNRAS*, 362, 505-534, arXiv: astro-ph/0501174.

Coles, P. & Jones, B., "A lognormal model for the cosmological mass distribution", 1991, *MNRAS*, 248, 1-13.

Colless, M., Dalton, G., Maddox, S. et al., "The 2dF Galaxy Redshift Survey: spectra and redshifts", 2001, *MNRAS*, 328, 1039-1063, arXiv: astro-ph/0106498.

Crocce, M. & Scoccimarro, R., "Renormalized cosmological perturbation theory", 2006, *Phys. Rev. D*, 73, 063519, arXiv: astro-ph/0509418.

Crocce, M., Pueblas, S. & Scoccimarro, R., "Transients from initial conditions in cosmological simulations", 2006, *MNRAS*, 373, 369-381, arXiv: astro-ph/0606505.

Crocce, M., Scoccimarro, R. & Blas, D., "Galilean-invariant renormalized perturbation theory", in prep.

Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M., "The evolution of large-scale structure in a universe dominated by cold dark matter", 1985, *ApJ*, 292, 371-394.

Dawson, K. S., Schlegel, D. J., Ahn, C. P. et al., "The Baryon Oscillation Spectroscopic Survey of SDSS-III", 2013, *AJ*, 145, 10, arXiv: 1208.0022.

Dawson, K. S., Kneib, J.-P., Percival, W. J. et al., "The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data", 2016, *AJ*, 151, 44, arXiv: 1508.04473.

DESI Collaboration, Aghamousa, A., Aguilar, J. et al., "The DESI Experiment Part I: Science,Targeting, and Survey Design", 2016, *arXiv e-prints*, arXiv: 1611.00036.

Dodelson, S. & Schneider, M. D., "The effect of covariance estimator error on cosmological parameter constraints", 2013, *Phys. Rev. D*, 88, 063537, arXiv: 1304.2593.

Dodelson, S. & Schmidt, F. *Modern Cosmology*. Academic Press, Cambridge, USA, 2nd edition, 2020.

Drinkwater, M. J., Jurek, R. J., Blake, C. et al., "The WiggleZ Dark Energy Survey: survey design and first data release", 2010, *MNRAS*, 401, 1429-1452, arXiv: 0911.4246.

eBOSS Collaboration, Alam, S., Aubert, M. et al., "The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological Implications from two Decades of Spectroscopic Surveys at the Apache Point observatory", 2020, *arXiv e-prints*, arXiv: 2007.08991.

Eisenstein, D. J., Zehavi, I., Hogg, D. W. et al., "Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies", 2005, *ApJ*, 633, 560-574, astro-ph/0501171.

Euclid Collaboration, Knabenhans, M., Stadel, J. et al., "Euclid preparation: II. The EUCLIDEMULATOR - a tool to compute the cosmology dependence of the nonlinear matter power spectrum", 2019, *MNRAS*, 484, 5509-5529, arXiv: 1809.04695.

Fixsen, D. J., "The Temperature of the Cosmic Microwave Background", 2009, *ApJ*, 707, 916-920, arXiv: 0911.1955.

Gay, C., Pichon, C. & Pogosyan, D., "Non-Gaussian statistics of critical sets in 2D and 3D: Peaks, voids, saddles, genus, and skeleton", 2012, *Phys. Rev. D*, 85, 023011, arXiv: 1110.0261.

Gil-Marín, H., Verde, L., Noreña, J. et al., "The power spectrum and bispectrum of SDSS DR11 BOSS galaxies - II. Cosmological interpretation", 2015, *MNRAS*, 452, 1914-1921, arXiv: 1408.0027.

Gil-Marín, H., Percival, W. J., Verde, L., Brownstein, J. R., Chuang, C.-H., Kitaura, F.-S., Rodríguez-Torres, S. A. & Olmstead, M. D., "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", 2017, *MNRAS*, 465, 1757-1788, arXiv: 1606.00439.

Gott, J. R., Choi, Y.-Y., Park, C. & Kim, J., "Three-Dimensional Genus Topology of Luminous Red Galaxies", 2009, *ApJ*, 695, L45-L48, arXiv: 0812.1406.

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S. & Dalla Vecchia, C., "Gaussian covariance matrices for anisotropic galaxy clustering measurements", 2016, *MNRAS*, 457, 1577-1592, arXiv: 1509.04293.

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S. et al., "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the Fourier space wedges of the final sample", 2017, *MNRAS*, 467, 2085-2112, arXiv: 1607.03143.

Guzzo, L., Pierleoni, M., Meneux, B. et al., "A test of the nature of cosmic acceleration using galaxy redshift distortions", 2008, *Nature*, 451, 541-544, arXiv: 0802.1944.

Hadwiger, H. *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie.* Springer-Verlag, Berlin Heidelberg, 1st edition, 1957.

Hamilton, A. J. S., "Measuring Omega and the real correlation function from the redshift correlation function", 1992, *ApJ*, 385, L5-L8.

Hamilton, A. J. S., Rimes, C. D. & Scoccimarro, R., "On measuring the covariance matrix of the non-linear power spectrum from simulations", 2006, *MNRAS*, 371, 1188-1204, arXiv: astro-ph/0511416.

Hand, N., Feng, Y., Beutler, F., Li, Y., Modi, C., Seljak, U. & Slepian, Z., "nbodykit: An Open-source, Massively Parallel Toolkit for Large-scale Structure", 2018, *AJ*, 156, 160, arXiv: 1712.05834.

Heitmann, K., Lawrence, E., Kwan, J., Habib, S. & Higdon, D., "The Coyote Universe Extended: Precision Emulation of the Matter Power Spectrum", 2014, *ApJ*, 780, 1, 111, arXiv: 1304.7849.

Hikage, C., Schmalzing, J., Buchert, T. et al., "Minkowski Functionals of SDSS Galaxies I : Analysis of Excursion Sets", 2003, *PASJ*, 55, 911-931, arXiv: astro-ph/0304455.

Hill, G. J., Gebhardt, K., Komatsu, E. et al. "The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX): Description and Early Pilot Survey Results", in T. Kodama, T. Yamada & K. Aoki, editors, *Panoramic Views of Galaxy Formation and Evolution*, volume 399 of *Astronomical Society of the Pacific Conference Series*, 2008, p. 115. arXiv: 0806.0183.

Hou, J., Sánchez, A. G., Scoccimarro, R. et al., "The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: anisotropic clustering analysis in configuration space", 2018, *MNRAS*, 480, 2521-2534, arXiv: 1801.02656.

Hou, J., Sánchez, A. G., Ross, A. J. et al., "The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2", 2021, *MNRAS*, 500, 1201-1221, arXiv: 2007.08998.

Huchra, J., Davis, M., Latham, D. & Tonry, J., "A survey of galaxy redshifts. IV - The data", 1983, *ApJS*, 52, 89-119.

Izard, A., Crocce, M. & Fosalba, P., "ICE-COLA: towards fast and accurate synthetic galaxy catalogues optimizing a quasi-N-body method", 2016, *MNRAS*, 459, 2327-2341, arXiv: 1509.04685.

Izard, A., Fosalba, P. & Crocce, M., "ICE-COLA: fast simulations for weak lensing observables", 2018, *MNRAS*, 473, 3051-3061, arXiv: 1707.06312.

James, J. B., Colless, M., Lewis, G. F. & Peacock, J. A., "Topology of non-linear structure in the 2dF Galaxy Redshift Survey", 2009, *MNRAS*, 394, 454-466, arXiv: 0810.2115.

Jones, D. H., Saunders, W., Colless, M. et al., "The 6dF Galaxy Survey: samples, observational techniques and the first data release", 2004, *MNRAS*, 355, 747-763, arXiv: astro-ph/0403501.

Kaiser, N., "Clustering in real space and in redshift space", 1987, *MNRAS*, 227, 1-21.

Kazin, E. A., Sánchez, A. G. & Blanton, M. R., "Improving measurements of H(z) and D_A (z) by analysing clustering anisotropies", 2012, *MNRAS*, 419, 3223-3243, arXiv: 1105.2037.

Kerscher, M. in K. R. Mecke & D. Stoyan, editors, *Statistical physics and spatial statistics: The art of analyzing and modeling spatial structures and pattern formation*, volume 554 of *Lecture notes in physics*, 2000, Springer Verlag, Berlin. arXiv: astro-ph/9912329.

Kerscher, M., Schmalzing, J., Retzlaff, J. et al., "Minkowski functionals of Abell/ACO clusters", 1997, *MNRAS*, 284, 73-84, arXiv: astro-ph/9606133.

Kerscher, M., Schmalzing, J., Buchert, T. & Wagner, H., "Fluctuations in the IRAS 1.2 Jy catalogue", 1998, *A&A*, 333, 1-12, arXiv: astro-ph/9704028.

Kerscher, M., Mecke, K., Schuecker, P. et al., "Non-Gaussian morphology on large scales: Minkowski functionals of the REFLEX cluster catalogue", 2001, *A&A*, 377, 1-16, arXiv: astro-ph/0105150.

Kitaura, F.-S. & Heß, S., "Cosmological structure formation with augmented Lagrangian perturbation theory", 2013, *MNRAS*, 435, L78-L82, arXiv: 1212.3514.

Kitaura, F.-S., Yepes, G. & Prada, F., "Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing", 2014, *MNRAS*, 439, L21-L25, arXiv: 1307.3285.

Kitaura, F.-S., Gil-Marín, H., Scóccola, C. G., Chuang, C.-H., Müller, V., Yepes, G. & Prada, F., "Constraining the halo bispectrum in real and redshift space from perturbation theory and non-linear stochastic bias", 2015, *MNRAS*, 450, 1836-1845, arXiv: 1407.1236.

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H. et al., "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the BOSS Final Data Release", 2016, *MNRAS*, 456, 4156-4173, arXiv: 1509.06400.

Koda, J., Blake, C., Beutler, F., Kazin, E. & Marin, F., "Fast and accurate mock catalogue generation for low-mass galaxies", 2016, *MNRAS*, 459, 2118-2129, arXiv: 1507.05329.

Kullback, S. & Leibler, R. A., "On information and sufficiency", 1951, *Ann. Math. Statist.*, 22, 79–86.

Laureijs, R., Amiaux, J., Arduini, S. et al., "Euclid Definition Study Report", 2011, *ArXiv e-prints*, arXiv: 1110.3193.

Lawrence, E., Heitmann, K., White, M., Higdon, D., Wagner, C., Habib, S. & Williams, B., "The Coyote Universe. III. Simulation Suite and Precision Emulator for the Nonlinear Matter Power Spectrum", 2010, *ApJ*, 713, 1322-1331, arXiv: 0912.4490.

Lewis, A., Challinor, A. & Lasenby, A., "Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models", 2000, *ApJ*, 538, 2, 473-476, arXiv: astro-ph/9911177.

Linder, E. V., "Baryon oscillations as a cosmological probe", 2003, *Phys. Rev. D*, arXiv: astro-ph/0304001.

Linder, E. V. & Cahn, R. N., "Parameterized beyond-Einstein growth", 2007, *Astroparticle Physics*, 28, 481-488, arXiv: astro-ph/0701317.

Lippich, M. & Sánchez, A. G., "MEDUSA: Minkowski functionals estimated from Delaunay tessellations of the three-dimensional large-scale structure", 2020, *arXiv e-prints*, arXiv: 2012.08529.

Lippich, M., Sánchez, A. G., Colavincenzo, M. et al., "Comparing approximate methods for mock catalogues and covariance matrices - I. Correlation function", 2019, *MNRAS*, 482, 1786-1806, arXiv: 1806.09477.

Marín, F. A., Blake, C., Poole, G. B. et al., "The WiggleZ Dark Energy Survey: constraining galaxy bias and cosmic growth with three-point correlation functions", 2013, *MNRAS*, 432, 2654-2668, arXiv: 1303.6644.

Matsubara, T., "Analytic Expression of the Genus in a Weakly Non-Gaussian Field Induced by Gravity", 1994, *ApJ*, 434, L43, arXiv: astro-ph/9405037.

Matsubara, T., "Statistics of Smoothed Cosmic Fields in Perturbation Theory. I. Formulation and Useful Formulae in Second-Order Perturbation Theory", 2003, *ApJ*, 584, 1-33.

Matsubara, T., "Analytic Minkowski functionals of the cosmic microwave background: Second-order non-Gaussianity with bispectrum and trispectrum", 2010, *Phys. Rev. D*, 81, 083505, arXiv: 1001.2321.

Matsubara, T. & Kuriki, S., "Weakly non-Gaussian formula for the Minkowski functionals in general dimensions", 2020, *arXiv e-prints*, arXiv: 2011.04954.

Matsubara, T., Hikage, C. & Kuriki, S., "Minkowski functionals and the nonlinear perturbation theory in the large-scale structure: second-order effects", 2020, *arXiv e-prints*, arXiv: 2012.00203.

Mecke, K. R., Buchert, T. & Wagner, H., "Robust morphological measures for large-scale structure in the Universe", 1994, *A&A*, 288, 697-704, arXiv: astro-ph/9312028.

Monaco, P. *The Cosmological Mass Function*. Ph.D. thesis, Universitá degli studi di Trieste, 1997.

Monaco, P., "Approximate Methods for the Generation of Dark Matter Halo Catalogs in the Age of Precision Cosmology", 2016, *Galaxies*, 4, 53, arXiv: 1605.07752.

Monaco, P., Theuns, T. & Taffoni, G., "The pinocchio algorithm: pinpointing orbit-crossing collapsed hierarchical objects in a linear density field", 2002, *MNRAS*, 331, 587-608, arXiv: astro-ph/0109323.

Munari, E., Monaco, P., Koda, J., Kitaura, F.-S., Sefusatti, E. & Borgani, S., "Testing approximate predictions of displacements of cosmological dark matter halos", 2017, *J. Cosmol. Astropart. Phys.*, 7, 050, arXiv: 1704.00920.

O'Connell, R., Eisenstein, D., Vargas, M., Ho, S. & Padmanabhan, N., "Large covariance matrices: smooth models from the two-point correlation function", 2016, *MNRAS*, 462, 2681-2694, arXiv: 1510.01740.

Park, C., Choi, Y.-Y., Vogeley, M. S. et al., "Topology Analysis of the Sloan Digital Sky Survey. I. Scale and Luminosity Dependence", 2005, *ApJ*, 633, 11-22, arXiv: astro-ph/0507059.

Paz, D. J. & Sánchez, A. G., "Improving the precision matrix for precision cosmology", 2015, *MNRAS*, 454, 4326-4334, arXiv: 1508.03162.

Pearson, D. W. & Samushia, L., "A Detection of the Baryon Acoustic Oscillation features in the SDSS BOSS DR12 Galaxy Bispectrum", 2018, *MNRAS*, 478, 4500-4512, arXiv: 1712.04970.

Percival, W. J., Baugh, C. M., Bland-Hawthorn, J. et al., "The 2dF Galaxy Redshift Survey: the power spectrum and the matter content of the Universe", 2001, *MNRAS*, 327, 1297-1306, arXiv: astro-ph/0105252.

Percival, W. J., Ross, A. J., Sánchez, A. G. et al., "The clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: including covariance matrix errors", 2014, *MNRAS*, 439, 2531-2541, arXiv: 1312.4841.

Planck Collaboration, Aghanim, N., Akrami, Y. et al., "Planck 2018 results. VI. Cosmological parameters", 2020, *A&A*, 641, A6, arXiv: 1807.06209.

Pogosyan, D., Gay, C. & Pichon, C., "Invariant joint distribution of a stationary random field and its derivatives: Euler characteristic and critical point counts in 2 and 3D", 2009, *Phys. Rev. D*, 80, 081301, arXiv: 0907.1437.

Pope, A. C. & Szapudi, I., "Shrinkage estimation of the power spectrum covariance matrix", 2008, *MNRAS*, 389, 766-774, arXiv: 0711.2509.

Ross, A. J., Bautista, J., Tojeiro, R. et al., "The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Large-scale structure catalogues for cosmological analysis", 2020, *MNRAS*, 498, 2354-2371, arXiv: 2007.09000.

Salazar-Albornoz, S., Sánchez, A. G., Grieb, J. N. et al., "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: angular clustering tomography and its cosmological implications", 2017, *MNRAS*, 468, 2938-2956, arXiv: 1607.03144.

Sánchez, A. G., "Arguments against using h^{-1} Mpc units in observational cosmology", 2020, *Phys. Rev. D*, 102, 123511, arXiv: 2002.07829.

Sánchez, A. G. *The formation and evolution of cosmic structures.* in prep.

Sánchez, A. G., Baugh, C. M., Percival, W. J., Peacock, J. A., Padilla, N. D., Cole, S., Frenk, C. S. & Norberg, P., "Cosmological parameters from cosmic microwave background measurements and the final 2dF Galaxy Redshift Survey power spectrum", 2006, *MNRAS*, 366, 189–207, arXiv: astro-ph/0507583.

Sánchez, A. G., Kazin, E. A., Beutler, F. et al., "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological constraints from the full shape of the clustering wedges", 2013, *MNRAS*, 433, 1202-1222, arXiv: 1303.4396.

Sánchez, A. G., Scoccimarro, R., Crocce, M. et al., "The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the configuration-space clustering wedges", 2017, *MNRAS*, 464, 1640-1658, arXiv: 1607.03147.

Sánchez, A. G., Gonzalez Jara, J., Ruiz, A. N. & Padilla, N. D., "Evolution mapping: a new approach to describe the non-linear matter power spectrum", in prep.

Schaap, W. E. & van de Weygaert, R., "Continuous fields and discrete samples: reconstruction through Delaunay tessellations", 2000, *A&A*, 363, L29-L32, arXiv: astro-ph/0011007.

Schmalzing, J. & Buchert, T., "Beyond Genus Statistics: A Unifying Approach to the Morphology of Cosmic Structure", 1997, *ApJ*, 482, L1-L4, arXiv: astro-ph/9702130.

Schmalzing, J., Kerscher, M. & Buchert, T. "Minkowski Functionals in Cosmology", in S. Bonometto, J. R. Primack & A. Provenzale, editors, *Dark Matter in the Universe*, p. 281. arXiv: astro-ph/9508154.

Schmalzing, J., Buchert, T., Melott, A. L., Sahni, V., Sathyaprakash, B. S. & Shandarin, S. F., "Disentangling the Cosmic Web. I. Morphology of Isodensity Contours", 1999a, *ApJ*, 526, 568-578, arXiv: astro-ph/9904384.

Schmalzing, J., Gottlöber, S., Klypin, A. A. & Kravtsov, A. V., "Quantifying the evolution of higher order clustering", 1999b, *MNRAS*, 309, 1007-1016, arXiv: astro-ph/9906475.

Schneider, M. D., Cole, S., Frenk, C. S. & Szapudi, I., "Fast Generation of Ensembles of Cosmological N-body Simulations Via Mode Resampling", 2011, *ApJ*, 737, 11, arXiv: 1103.2767.

Scoccimarro, R., "Redshift-space distortions, pairwise velocities, and nonlinearities", 2004, *Phys. Rev. D*, 70, 083007, arXiv: astro-ph/0407214.

Sellentin, E. & Heavens, A. F., "Parameter inference with estimated covariance matrices", 2016, *MNRAS*, 456, L132-L136, arXiv: 1511.05969.

Sheth, J. V., "Morphology of mock SDSS catalogues", 2004, *MNRAS*, 354, 332-342, arXiv: astro-ph/0310755.

Sheth, J. V., Sahni, V., Shandarin, S. F. & Sathyaprakash, B. S., "Measuring the geometry and topology of large-scale structure using SURFGEN: methodology and preliminary results", 2003, *MNRAS*, 343, 22-46, arXiv: astro-ph/0210136.

Slepian, Z., Eisenstein, D. J., Brownstein, J. R. et al., "Detection of baryon acoustic oscillation features in the large-scale three-point correlation function of SDSS BOSS DR12 CMASS galaxies", 2017a, *MNRAS*, 469, 1738-1751, arXiv: 1607.06097.

Slepian, Z., Eisenstein, D. J., Beutler, F. et al., "The large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies", 2017b, *MNRAS*, 468, 1070-1083.

Springel, V., "The cosmological simulation code GADGET-2", 2005, *MNRAS*, 364, 1105-1134, arXiv: astro-ph/0505010.

Springel, V., White, S. D. M., Tormen, G. & Kauffmann, G., "Populating a cluster of galaxies - I. Results at z=0", 2001, *MNRAS*, 328, 726-750, arXiv: astro-ph/0012055.

Stein, G., Alvarez, M. A. & Bond, J. R., "The mass-Peak Patch algorithm for fast generation of deep all-sky dark matter halo catalogues and its N-body validation", 2019, *MNRAS*, 483, 2236-2250, arXiv: 1810.07727.

Taruya, A., Nishimichi, T. & Saito, S., "Baryon acoustic oscillations in 2D: Modeling redshift-space power spectrum from perturbation theory", 2010, *Phys. Rev. D*, 82, 063522, arXiv: 1006.0699.

Tassev, S., Zaldarriaga, M. & Eisenstein, D. J., "Solving large scale structure in ten easy steps with COLA", 2013, *J. Cosmol. Astropart. Phys.*, 6, 036, arXiv: 1301.0322.

Taylor, A., Joachimi, B. & Kitching, T., "Putting the precision in precision cosmology: How accurate should your data covariance matrix be?", 2013, *MNRAS*, 432, 1928-1946, arXiv: 1212.4359.

The Dark Energy Survey Collaboration, "The Dark Energy Survey", 2005, *arXiv e-prints*, arXiv: astro-ph/0510346.

Tomita, H. "Statistics and geometry of random interface systems,", in K. Kawasaki, M. Suzuki & A. Onuki, editors, *Formation, Dynamics,and Statistics of Patterns*, volume 1, 1990, pp. 113–157.

Vakili, M., Kitaura, F.-S., Feng, Y., Yepes, G., Zhao, C., Chuang, C.-H. & Hahn, C., "Accurate halo-galaxy mocks from automatic bias estimation and particle mesh gravity solvers", 2017, *MNRAS*, 472, 4144-4154, arXiv: 1701.03765.

Verde, L., Treu, T. & Riess, A. G., "Tensions between the early and late Universe", 2019, *Nature Astronomy*, 3, 891-895, arXiv: 1907.10625.

Wang, Y., "Figure of merit for dark energy constraints from current observational data", 2008, *Phys. Rev. D*, 77, 123525, arXiv: 0803.4295.

Wiegand, A. & Eisenstein, D. J., "The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: higher order correlations revealed by germ-grain Minkowski functionals", 2017, *MNRAS*, 467, 3361-3378, arXiv: 1609.08613.

Wiegand, A., Buchert, T. & Ostermann, M., "Direct Minkowski Functional analysis of large redshift surveys: a new high-speed code tested on the luminous red galaxy Sloan Digital Sky Survey-DR7 catalogue", 2014, *MNRAS*, 443, 241-259, arXiv: 1311.3661.

Yaryura, C. Y., Sánchez, A. G. & García Lambas, D., "Geometrical analysis of the large scale structure in the Universe", 2004, *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 47, 377-381.

York, D. G., Adelman, J., Anderson, J., John E. et al., "The Sloan Digital Sky Survey: Technical Summary", 2000, *AJ*, 120, 1579-1587, arXiv: astro-ph/0006396.

Zhang, Y., Springel, V. & Yang, X., "Genus Statistics Using the Delaunay Tessellation Field Estimation Method. I. Tests with the Millennium Simulation and the SDSS DR7", 2010, *ApJ*, 722, 812-824, arXiv: 1006.3768.

Zhao, C., Kitaura, F.-S., Chuang, C.-H., Prada, F., Yepes, G. & Tao, C., "Halo mass distribution reconstruction across the cosmic web", 2015, *MNRAS*, 451, 4266-4276, arXiv: 1501.05520.

Zheng, Z., Coil, A. L. & Zehavi, I., "Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering", 2007, *ApJ*, 667, 760-779, arXiv: astro-ph/0703457.

Acknowledgements

I would like to thank Dr. Ariel G. Sánchez for the scientific supervision of my PhD. Thank you very much for your guidance, your always good advice even when it seems that there is no way out, and for trusting me with one of your dearest projects.

I would also like to express my gratitude to Prof. Dr. Ralf Bender for hosting me in the OPINAS group and for supporting my PhD.

I am grateful to Dr. Daniel Farrow, Dr. Jiamin Hou, Dr. Andrea Pezzotta and Agne Semenaite for the fruitful discussions during our meetings (powered by lots of good chocolate cake). Thanks also for the useful comments on the MEDUSA text to Danny, on the theory chapter to Agne and on the entire thesis to Ariel.

Part of this work has been supported by the Transregional Collaborative Research Centre TR33 'The Dark Universe' of the German Research Foundation (DFG).

The computationally expensive parts of the analysis have been performed on the high-performance computing resources of the Max Planck Computing and Data Facility (MPCDF) in Garching.

During my PhD, I had the great opportunity to visit several inspiring conferences, meetings and research schools. I would like to thank my hosts Prof. Dr. Dante Paz and Prof. Dr. Nelson Padilla, the LACEGAL project and my supervisor Dr. Ariel G. Sánchez for the wonderful stay at IATE in Còrdoba, Argentina, and at PUC in Santiago de Chile.

In the first years of my PhD, there was a lot of offline scientific exchange. This changed drastically and the last year of my PhD was marked by the worldwide Covid-19 pandemic. I am grateful to all the health workers and essential workers who risked their health to be there for us.

This thesis would not have been possible without my family and friends who filled my life in so many facets. Thanks to my colleagues who became friends, my old friends from school, my friends from the "wild" university times and from all the travels, my ballet friends, mi prima de Bochum, the Temples, and the M&Ms with Luis. I will thank you all in person once the pandemic is over.

My warmest thanks go to Till Siebenmorgen. You bring harmony, light and the right amount of inexplicable craziness to my Universe.

I am deeply grateful to my parents. Thank you for your unconditional love and support and for always believing in me.