Dissertation

Metin Ata

Leibniz-Institut für Astrophysik Potsdam

Kosmologie und großräumige Strukturen

Phase-Space Reconstructions of Cosmic Velocities and the Cosmic Web

Structure Formation Models – Galaxy Bias Models – Galaxy Redshift Surveys – Inference Analysis

Dissertation
zur Erlangung des akademischen Grades
„doctor rerum naturalium“
(Dr. rer. nat.)
in der Wissenschaftsdisziplin „Astrophysik“

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von

Metin Ata

November 2016

Abstract

In the current paradigm of cosmology, the formation of large-scale structures is mainly driven by non-radiating dark matter, making up the dominant part of the matter budget of the Universe. Cosmological observations however, rely on the detection of luminous galaxies, which are biased tracers of the underlying dark matter.

In this thesis I present cosmological reconstructions of both, the dark matter density field that forms the cosmic web, and cosmic velocities, for which both aspects of my work are delved into, the theoretical formalism and the results of its applications to cosmological simulations and also to a galaxy redshift survey.

The foundation of our method is relying on a statistical approach, in which a given galaxy catalogue is interpreted as a biased realization of the underlying dark matter density field. The inference is computationally performed on a mesh grid by sampling from a probability density function, which describes the joint posterior distribution of matter density and the three dimensional velocity field. The statistical background of our method is described in Chapter ”*Implementation of* ARGO”, where the introduction in sampling methods is given, paying special attention to Markov Chain Monte-Carlo techniques.

In Chapter ”*Phase-Space Reconstructions with N-body Simulations*”, I introduce and implement a novel biasing scheme to relate the galaxy number density to the underlying dark matter, which I decompose into a deterministic part, described by a non-linear and scale-dependent analytic expression, and a stochastic part, by presenting a negative binomial (NB) likelihood function that models deviations from Poissonity. Both bias components had already been studied theoretically, but were so far never tested in a reconstruction algorithm. I test these new contributions against N-body simulations to quantify improvements and show that, compared to state-of-the-art methods, the stochastic bias is inevitable at wave numbers of $k \geq 0.15\,h\,\mathrm{Mpc}^{-1}$ in the power spectrum in order to obtain unbiased results from the reconstructions. In the second part of Chapter ”*Phase-Space Reconstructions with N-body Simulations*” I describe and validate our approach to infer the three dimensional cosmic velocity field jointly with the dark matter density. I use linear perturbation theory for the large-scale bulk flows and a dispersion term to model virialized galaxy motions, showing that our method is accurately recovering the real-space positions of the redshift-space distorted galaxies. I analyse the results with the isotropic and also the two-dimensional power spectrum.

Finally, in Chapter ”*Phase-space Reconstructions with Galaxy Redshift Surveys*”, I show how I combine all findings and results and apply the method to the CMASS (for *Constant (stellar) Mass*) galaxy catalogue of the Baryon Oscillation Spectroscopic Survey (BOSS). I describe how our method is accounting for the observational selection effects inside our reconstruction algorithm. Also, I demonstrate that the renormalization of the prior distribution function is mandatory to account for higher order contributions in the structure formation model, and finally a redshift-dependent bias factor is theoretically motivated and implemented into our method. The various refinements yield unbiased results of the dark matter until scales of $k \leq 0.2\,h\,\mathrm{Mpc}^{-1}$ in the power spectrum and isotropize the galaxy catalogue down to distances of $r \sim 20\ h^{-1}\,\mathrm{Mpc}$ in the correlation function.

We further test the results of our cosmic velocity field reconstruction by comparing them to a synthetic *mock* galaxy catalogue, finding a strong correlation between the mock and the reconstructed

velocities.

The applications of both, the density field without redshift-space distortions, and the velocity reconstructions, are very broad and can be used for improved analyses of the baryonic acoustic oscillations, environmental studies of the cosmic web, the kinematic Sunyaev-Zel'dovic or integrated Sachs-Wolfe effect.

Contents

Nomenclature

δ	The density contrast
δ^{D}	Dirac delta function
Λ	Cosmological constant
$\mathscr{H}$	Hamiltonian
$\mathscr{P}, \mathscr{L}$	Probability density, Likelihood
Ω_x	Energy density of x in ratio to critical density
Φ	Peculiar potential
σ	Dispersion
$\mathit{\Phi}$	Newtonian potential
ϱ	Matter density
$\mathit{\Theta}$	Step function in the threshold bias model
ϑ, ϕ	The angular coordinates in RW metric
$\boldsymbol{\mu}$	Renormalized mean field
$\boldsymbol{\Psi}$	Displacement field
$\boldsymbol{k}$	Wave vector
$\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{s}$	(Comoving) Coordinates in real-space, redshift-space
$\boldsymbol{u}$	The total velocity
$\boldsymbol{v}_{\mathrm{p}}$	The peculiar velocity
a, z	scale factor, redshift
b_i	Bias factor at position i
D	Growth factor
f_Ω	Growth rate
H_0	Present Hubble constant
K	Spatial curvature or Kaiser factor

1

Introduction

Cosmology is the science of the origin and the evolution of the Universe, of which the cornerstone was laid with Einstein's theory of General Relativity. Later, in the 1920s, Friedmann and Lemaître derived the laws describing the expanding Universe. When Zwicky in 1933 measured the escape velocities of galaxies in large clusters, he could not explain the gravitational stability of these clusters with the pure observed stellar masses. These observations gave rise to the concept of dark matter. However, due to the lack of observations, research in cosmology was initially on hold as the available data could not validate most of the cosmological models. As an example, Hubble's measurement of the expansion rate of the Universe (Hubble, 1929) was off by an order of magnitude and with large uncertainties. Up to the 1940s-50s observational data in cosmology was rare and inaccurate. This changed dramatically in the 1990s when measurements of the Cosmic Microwave Background (CMB) with COBE (Smoot et al., 1992) detected for the first time the fluctuations of the early Universe, making the most accurate measurement of a black body spectrum in nature. Great progress had already been made in the 1980s, mapping for the first time the distribution of thousands of galaxies with the CfA survey (e.g Davis & Huchra, 1982; Geller & Huchra, 1989). This led to the cold dark matter model (CDM model: Blumenthal et al. (1984) , Davis et al. (1985)) in which structure grows in a hierarchical way, where objects are assumed to collapse under self-gravity and merge to form larger objects. After observations of Type Ia Supernovae (Riess et al. (1998), Perlmutter et al. (1999)) showed the accelerated expansion of the Universe, the cosmological constant was introduced into the CDM model, leading to the ΛCDM model (Λ for dark energy, i.e. the cosmological constant as special case), which has become the ***standard model*** in cosmology. This model can be determined by a set of six parameters and has been proven to explain a large variety of observations to great accuracy.

Modern galaxy surveys have shown that galaxies are spatially not randomly distributed, but form a characteristic pattern, clustering in knots that are connected through a filamentary network, leaving large void areas. This network is commonly called the *cosmic web*, which is believed to have arisen from primordial fluctuations through gravitational evolution.

In the standard cosmological paradigm these primordial fluctuations are assumed to have been formed from quantum fluctuations in an inflationary evolution stage, which were stretched and frozen through an inflationary expansion epoch shortly after the origin of the Universe, also known as the Big Bang.

Ideally, at the largest scales, the evolution of fluctuations is close to a linear growth, allowing a precise analysis of the cosmological parameters. However, gravitational clustering is impeding a robust analysis at smaller scales. Therefore, a detailed understanding and modelling of the transition from the linear to the non-linear range of structure formation is a key to overcome these complications. Systematical cumbrousness, such as the survey geometry or selection effects, caused by the limited sensitivity of the observations, introduce uncertainties on large scales. In addition to systematics, other effects which are due to intrinsic physical phenomena, must be understood and modelled in order to extract maximum information of cosmological parameters and processes. The research of the

Large-Scale Structure (LSS) of the Universe addresses all above mentioned issues. It is the study of the physics of clustering and structure formation at cosmological scales. Two major objectives of the LSS research are to understand how galaxies spatially trace the distribution of dark matter, called *galaxy bias*, and, how cosmic velocities and the cosmic web are connected. The wealth of information, which we can extract from the LSS, is indispensably relying on the amount and the quality of the observational data. Since increasingly larger surveys become possible, cosmology has made a transition to a data driven science. This unprecedented amount of data also come with new challenges and pushes cosmology furthermore to develop methods for *big data* analyses. The large variety of accurate data today paved the way to test models at percentage accuracy, reaching the era of ***precision cosmology***.

Observations play undeniably an significant role in cosmology, nevertheless, according to the current ΛCDM model, the majority of the matter density in the Universe, which is dark, is not directly accessible via measurements. Moreover, ΛCDM model predicts that the dynamics of the galaxies are mainly driven by the underlying dark matter. As a consequence of this inaccessibility, we need a profound understanding of the bias in order to study the full matter density in the Universe.

All these challenges can only be addressed, if adequately precise reconstructions of the dark matter density, the cosmic web, cosmic flows, and other interrelated observational quantities that are not accessible by pure observations, are available. These reconstructions are inferred from tracers, i.e, galaxies or halos (commonly called tracers as they trace the dark matter density) and rely on a statistical Bayesian framework utilizing Markov Chain Monte-Carlo techniques. As reconstructions make (amongst others) dark matter density and velocity accessible for studies, they work as a link between theoretical predictions and galaxy surveys. Reconstruction algorithms demand an accurate handling of the observational data, compensating for systematic effects induced by the survey geometry or the sensitivity of the telescopes. The distance estimation at cosmological scales is done via *redshift* measurements, which, in absence of additional velocities, is determined by the expansion rate of the Universe. *Peculiar velocities* of individual galaxies however, may distort the distance measurement, called *redshift-space distortions* (RSDs). Also, sophisticated theoretical models must be elaborated to adequately describe galaxy bias and structure formation models.

For all the above mentioned reasons the reconstructions of cosmic density and velocity fields are crucially needed to proof or falsify models and to infer physical quantities precisely and study the physics of structure formation.

These are the challenges that precision cosmology addresses and that my thesis is dedicated to.

In my thesis I explored the numerical and theoretical domains of cosmological reconstruction techniques based on synthetic galaxy catalogues from simulations, and of observed galaxies from surveys. In the following I will describe the projects that my thesis comprises. In three separate projects we applied the reconstruction code, called ARGO, to data-sets and included vital refinements, which we found to be inevitable in order to yield robust results.

We rely on a statistical approach and formulate a *probability density function* (PDF), that expresses the joint probability to infer the dark matter density, cosmic velocities, and other related quantities, given a set of galaxy positions. Within this approach we account for the systematic effects due to the observational strategy of the galaxy survey, and also model intrinsic physical processes, including growth of structures, galaxy bias, distortion of galaxy positions due to RSDs, and higher order effects of structure formation beyond linear theory.

We solve the PDF in a Gibbs-sampling scheme by splitting the joint probability into separate conditional probabilities. Within this scheme, the dark matter density is inferred with an (adaptive) Hybrid Monte-Carlo algorithm, linking the dark matter density to the galaxy number counts via a galaxy bias model, whereas the cosmic velocities are obtained from the inferred densities, using linear theory and an additional random dispersion term.

In order to statistically examine density and velocity perturbations, we use the *power spectrum*, which measures the amplitude of fluctuations as a function of distance scales in Fourier space. Cosmological information is encoded in the shape and the amplitude of the power spectrum. With upcoming very large surveys that will provide a dense cosmological cartography of the Universe, (e.g. LSST Dark Energy Science Collaboration, 2012), the precise understanding of the power spectrum at all scales is an urgent challenge. The monopole power spectrum measures the fluctuations isotropically, as the fluctuation modes are computed spherically averaged. However, the quadrupole power spectrum can be used to estimate the power of fluctuations depending on the direction. Thus the quadrupole power spectrum is an important tool to study anisotropies. Current reconstruction models that are based on Poisson statistics and/or unrealistic bias models for galaxy counts are failing to be accurate on a wide range of scales. We implement several crucial improvements to address this problem.

Firstly we applied a deterministic bias model that assumes the form of a power law and takes also a Heaviside step-function into account. Secondly we overcome the problem to model the stochastic component of the bias by introducing a super-Poissonian distribution function, in our case a negative binomial (NB) distribution, for the data, which has a larger variance than a Poisson distribution at the same mean. We will use the term *over-dispersion* throughout this thesis to describe super-Poissonian behaviour. Galaxy surveys are relying on the measurements of redshifts to estimate the distance of galaxies. RSDs spoil the distance measurement to some extend. Thus, it is impossible to bypass the problem of RSDs if we do not precisely reconstruct the galaxies' peculiar velocities. We distinguish between the coherent bulk flow, which is entirely characterized by the gradient of the potential and velocities caused by collapsed, quasi-virialized objects. We can model bulk flows analytically within linear theory, however, the quasi-virialized motions are highly non-linear effects. In our approach we address this problem with a statistical treatment and a classification of the cosmic web.

Structure of the thesis

The structure of my thesis is in the same order as the papers we published. The abstracts are given at the end of this introduction.

In Chapter 2, I revise the necessary theory of the Friedmann-Robertson-Walker cosmology for a isotropic and homogeneous Universe in Section 2.2, and the ΛCDM standard model in Section 2.3. I put special emphasis on the fluctuations that lead to the cosmic structures in the Universe, shown in Section 2.6. The galaxy bias and RSDs, which are the main ingredients of our model, are discussed in Section 2.8 and 2.9 respectively.

In chapter 3, I give an overview of the Sloan Digital Sky Survey (SDSS) and a brief introduction to simulation techniques in the cosmological context.

Further, I present the statistical background and the ARGO framework in Chapter 4.

In Section 4.1 the Bayesian approach up to the Hybrid Monte-Carlo (HMC) technique is described. The ARGO-code is presented in detail in Sections 4.5, 4.7, and 4.8.

Chapter 5 is entirely focusing on reconstructions with synthetic data from simulations with special emphasis on the deviation from Poissonity in Section 5.2. In Sections 5.3 and 5.3.1 the corrections of RSDs are discussed as we focus on the reconstruction of the peculiar velocities. The convergence behavior of our statistical inference model is shown in 5.3.2.

Chapter 6 is showing the results we obtained for the joint density and velocity reconstruction by applying the ARGO code to the SDSS galaxy redshift survey. In particular, the density field reconstruction is shown in Section 6.3, whereas the velocity reconstructions are presented in Section 6.4. This Chapter then is concluded in Section 6.4.4.

I summarize the thesis in Chapter 7 and give a perspective of the future work that can follow up the here presented results and findings.

Bayesian inference of cosmic density fields from non-linear, scale-dependent, and stochastic biased tracers

We present a Bayesian reconstruction algorithm to generate unbiased samples of the underlying dark matter field from halo catalogues. Our new contribution consists of implementing a non-Poisson likelihood including a deterministic non-linear and scale-dependent bias. In particular we present the Hamiltonian equations of motions for the negative binomial (NB) probability distribution function. This permits us to efficiently sample the posterior distribution function of density fields given a sample of galaxies using the Hamiltonian Monte Carlo technique implemented in the Argo code. We have tested our algorithm with the Bolshoi N-body simulation at redshift $z = 0$, inferring the underlying dark matter density field from sub-samples of the halo catalogue with biases smaller and larger than one. Our method shows that we can draw closely unbiased samples (compatible within 1-σ) from the posterior distribution up to scales of about $k \sim 1\,h\,\mathrm{Mpc}^{-1}$ in terms of power-spectra and cell-to-cell correlations. We find that a Poisson likelihood yields reconstructions with power spectra deviating more than 10% at $k = 0.2\,h\,\mathrm{Mpc}^{-1}$. Our reconstruction algorithm is especially suited for emission line galaxy data for which a complex non-linear stochastic biasing treatment beyond Poissonity becomes indispensable.

Bayesian redshift-space distortions correction from galaxy redshift surveys

We present a reconstruction method which maps a galaxy distribution from redshift-space to real-space inferring the distances of the individual galaxies. The method is based on sampling density fields assuming a lognormal prior with a likelihood given by the negative binomial distribution function modelling a stochastic bias. We assume a deterministic bias given by a power law relating the dark matter density field to the expected halo or galaxy field. Coherent redshift-space distortions are corrected in a Gibbs-sampling procedure by moving the galaxies from redshift-space to real-space according to the peculiar motions derived from the recovered density field using linear theory with the option to include tidal field corrections from second order Lagrangian perturbation theory. The virialised distortions are corrected by sampling candidate real-space positions (being in the neighbourhood of the observations along the line of sight), which are compatible with the bulk flow corrected redshift-space position adding a random dispersion term in high density collapsed regions. The latter are defined according to the eigenvalues of the Hessian. This approach presents an alternative method to estimate the distances to galaxies using the three dimensional spatial information, and assuming isotropy. Hence the number of applications is very broad. In this work we show the potential of this method to constrain the growth rate up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$. Furthermore it could be useful to correct for photometric redshift errors, and to obtain improved BAO reconstructions.

The Clustering of Galaxies in the Completed SDSS-III Baryon Oscillation Spectroscopic Survey: Phase-space Reconstructions of Cosmic Flows and Cosmic Web from Luminous Red Galaxies

We present a space reconstruction of the cosmic large-scale matter density and velocity fields from the SDSS-III Baryon Oscillations Spectroscopic Survey Data Release 12 (BOSS DR12) CMASS galaxy clustering catalogue. We rely on a given ΛCDM cosmology, a mesh resolution of 6.25 h^{-1} Mpc, and a lognormal-Negative Binomial model with a redshift dependent non-linear bias. The bias parameters

are derived from the data and a general renormalized perturbation theory approach. We use combined Gibbs and Hamiltonian sampling, implemented in the ARGO code, to iteratively reconstruct the dark matter density field and the coherent peculiar velocities of individual galaxies, correcting hereby for coherent redshift-space distortions (RSD). Our tests relying on accurate N-body based mock galaxy catalogues, show unbiased real-space power spectra of the non-linear density field up to $k = 0.2\,h\,\mathrm{Mpc}^{-1}$, and vanishing quadrupoles down to $\sim 20\,h^{-1}\,\mathrm{Mpc}$. We also demonstrate that the non-linear cosmic web can be obtained from the tidal field tensor based on the Gaussian component of the reconstructed density field. We find that the reconstructed velocities have a statistical correlation coefficient compared to the true velocities of each individual lightcone mock galaxy of $r \sim 0.65$ including about 10% of satellite galaxies with virial motions (about $r = 0.79$ without satellites). The power spectra of the velocity divergence agree well with theoretical predictions up to $k = 0.2\,h\,\mathrm{Mpc}^{-1}$. This work will be especially useful to improve, e.g. BAO reconstructions, kinematic Sunyaev-Zel'dovich (kSZ), warm hot inter-galactic medium (thermal SZ or X-rays), and integrated Sachs-Wolfe (ISW) measurements, or environmental studies.

2 Theoretical Background

This chapter is a brief revision of the theoretical background which is necessary to follow up cosmological relations and quantities we are using, mostly relying on textbooks and recent reviews (Peebles (1980), Peacock (1999), Liddle & Lyth (2000) Coles & Lucchin (2002), Dodelson (2003),Carroll (2004), Mukhanov (2005), Plebanski & Krasinski (2006), Grøn & Hervik (2007), Weinberg (2008)). I emphasize quantities that are a crucial part of our reconstruction algorithm and refer to the corresponding section of the thesis, where these quantities are referred to.

2.1 Einstein's field equations

Cosmology is the study of the Universe, describing the evolution and origin of its structures, based upon Einstein's theory of general relativity (GR). According to GR, cosmic space-time is described by a metric $g_{\mu\nu}$, that defines the coordinate invariant distance defined as

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \,, \tag{2.1}$$

where the indices μ, ν are space-times coordinates, assuming values from 0 (time) till $1, 2, 3$ (space). The metric is obtained by solving Einstein's field equation

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \,, \tag{2.2}$$

where Λ is called cosmological constant or vacuum energy and was introduced by Einstein himself initially to yield a static universe solution, $R_{\mu\nu}$ is the Ricci tensor, R the Ricci scalar, Newton's gravitational constant G, the speed of light c and the energy-stress-momentum tensor $T_{\mu\nu}$.

So generellay, Equation 2.2 represents a system of ten nonlinear differential equations of second order.

2.2 Friedmann-Robertson-Walker Cosmology

Modern cosmology relies on the cosmological principle. This principle states that the distribution of matter in our Universe is homogeneous and isotropic when measured on large enough scales. Thus we can rewrite Equation 2.1 to obtain the *Robertson-Walker metric* in terms of *comoving coordinates* (r, ϑ, φ) and the cosmic time of an comoving observer t

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -c^2 \mathrm{d}t^2 + a^2(t) \left[\frac{\mathrm{d}r^2}{1 - Kr^2} + r^2 \mathrm{d}\vartheta^2 + r^2 \sin^2\vartheta \mathrm{d}\varphi^2 \right] \,, \tag{2.3}$$

where in most general form the scale factor $a(t)$ is a function of time to be determined by Equation 2.2, and K spatial curvature parameter. The spatial part of Equation 2.3 can be interpreted by embedding a three-dimensional sphere (pseudo-sphere, depending on K) in a four-dimensional Euclidean space.

We have the freedom to *rescale* the quantities a, K and r mutually,

$$\begin{aligned} a &\rightarrow \lambda a\,, \\ r &\rightarrow r/\lambda\,, \\ K &\rightarrow \lambda^2 K\,, \end{aligned} \tag{2.4}$$

so that the geometry of space-time is not altered and consequently this transformation leaves Equation 2.3 invariant. In literature, a common choice is a rescaling so that K becomes dimensionless, achieved by setting $\lambda^2 = 1/|K|$ in Equations 2.4. In this way, the coordinate r is dimensionless as well. However, the scale factor $a(t)$ inherits the dimension of length. For $K = 1$ the spatial part of the line element reduces to a three-dimensional sphere with fixed radius $a(t)r$, called *closed* universe. In the case of $K = 0$, we obtain a spatially *flat* Euclidean space, and finally $K = -1$ describing a pseudo-sphere called *open* universe. A particularly convenient definition of the comoving radial coordinate, χ, is related to r by

$$\mathrm{d}\chi = \frac{\mathrm{d}r}{\sqrt{1 - Kr^2}}\,. \tag{2.5}$$

With this coordinate choice, the *Robertson-Walker metric* writes as (setting $c = 1$ in natural units)

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\mathrm{d}\chi^2 + \Sigma^2(\chi)(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2\right]\,, \tag{2.6}$$

where

$$\Sigma(\chi) = \begin{cases} \sin\chi\,, & \text{if } K = +1\,, \\ \chi\,, & \text{if } K = 0\,, \\ \sinh\chi\,, & \text{if } K = -1\,. \end{cases} \tag{2.7}$$

Only for zero curvature $K = 0$, the coordinates $\chi = r$ are identical and thus the normalization of $a(t)$ is arbitrary. Therefore, in flat space, we can set set $a(t_0) = a_0 = 1$ at present time t_0 so that the metric coordinate r has the dimension of length (just as $K^{-1/2}$).

In cosmology, the energy-stress tensor is described by a perfect, frictionless fluid specified by an energy density ϱ and an isotropic pressure p in its rest frame, written as

$$T_{\mu\nu} = (\varrho + p)u_\mu u_\nu + p g_{\mu\nu}\,, \tag{2.8}$$

where u_μ is the four-velocity of the fluid, assuming the form $u_\mu = (1, 0, 0, 0)$ at rest.

Now we can solve Equations 2.2 for the metric given in Equation 2.3 and write the two so-called *Friedmann equations* as

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\varrho + 3p) + \frac{\Lambda}{3}\,, \tag{2.9}$$

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\varrho - \frac{K}{a^2} + \frac{\Lambda}{3}\,, \tag{2.10}$$

that connect the scale factor a to curvature and energy density in the Universe. $H(t) = \dot{a}/a$ is called *Hubble parameter* and gives the expansion rate of the Universe. Its present value H_0 is the *Hubble constant*, given by $H_0 = h\,100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, where h is a parameter. Let us further look at Equation 2.10, defining an energy density of the vacuum, $\varrho_\Lambda = \Lambda/(8\pi G)$, we now identify the different contributions to the *total* energy density as $\varrho_{\mathrm{tot}} = \varrho_{\mathrm{M}} + \varrho_{\mathrm{R}} + \varrho_{\Lambda}$, with the matter energy density ϱ_{M} and radiation density ϱ_{R}. The energy density of neutrinos ϱ_{ν} is neglected in this work (in some textbooks however, the neutrino energy density is absorbed inside the radiation density). The resulting model is called *Friedmann-Robertson-Walker (FRW) cosmology.*

Equation 2.10, describing the time-time component, and Equation 2.9, describing the space-space component, are not independent but connected by the dynamical eqaution $T^{\mu\nu}_{;\nu}$=0. Therefore, the expansion is adiabatic, meaning that the entropy is not changing, $\mathrm{d}S = 0$, so that

$$\begin{aligned} \mathrm{d}U + \mathrm{d}W = T\mathrm{d}S = 0 \quad &\Rightarrow \quad \mathrm{d}(\varrho a^3) = -p\mathrm{d}a^3\,, && (2.11) \\ \frac{\mathrm{d}}{\mathrm{d}t}(\varrho a^3) \quad &= \quad -p\frac{\mathrm{d}}{\mathrm{d}t}a^3\,, && (2.12) \end{aligned}$$

where W is the work and T the temperature. This equation shows that for isotropic and homogeneous models the expansion is adiabatic, since homogeneity and isotropy forbids temperature gradients. Further, we assume that the fluid is described by its *equation of state*, also called barotropic equation, as

$$p = w\varrho\,. \tag{2.13}$$

We find the solution to Equation 2.12 to be

$$\varrho a^{3(w+1)} = \varrho_0\,, \tag{2.14}$$

with ϱ_0 being the present density and a dimensionless parameter w describing the fluid. Equation 2.9 also shows that the total gravitational source term is $\varrho + 3p$, so that pressure itself also contributes to gravity and that gravitation is repulsive if $p < -\varrho/3$. Ordinary matter, often called dust, is characterized by $w_{\mathrm{M}} = 0$. Vacuum energy is described by $w_{\Lambda} = -1$, which means that this energy density is constant. Finally, radiation defined by $w_{\mathrm{R}} = 1/3$.

Equation 2.14 illustrates that the radiation energy density scales like $\varrho_{\mathrm{R}} \propto a^{-4}$. This is due to the scaling with a^{-3} like ordinary matter including an extra a^{-1} term due to *redshift* (defined in Equation 2.20), caused by the expansion rate of the Universe.

We now may rearrange Equation 2.10 and find

$$\varrho_{\mathrm{tot}} = \frac{3H^2}{8\pi G} + \frac{3K}{8\pi G a^2}\,. \tag{2.15}$$

Thus, by defining the *critical energy density*

$$\varrho_{\mathrm{crit}}(t) = \frac{3H^2}{8\pi G}\,, \tag{2.16}$$

we can constrain the curvature term depending on the total energy density as

$$\begin{aligned} \varrho_{\mathrm{tot}} > \varrho_{\mathrm{crit}} \quad &\Rightarrow \quad K = 1\,, \\ \varrho_{\mathrm{tot}} = \varrho_{\mathrm{crit}} \quad &\Rightarrow \quad K = 0\,, \\ \varrho_{\mathrm{tot}} < \varrho_{\mathrm{crit}} \quad &\Rightarrow \quad K = -1\,. && (2.17) \end{aligned}$$

It is convenient to write any energy density of X in terms of the critical energy density as $\Omega_X = \varrho_X/\varrho_{\mathrm{crit}}$. Also, by defining an energy density that corresponds to curvature, $\Omega_{\mathrm{K}} = -K/(H^2a^2)$, we find the Friedmann equation to be

$$H^2(a) \quad = \quad H_0^2(\Omega_{\mathrm{R},0}(a_0/a)^4 + \Omega_{\mathrm{M},0}(a_0/a)^3 + \Omega_{\mathrm{K},0}(a_0/a)^2 + \Omega_{\Lambda,0}) = H_0^2E^2(a)\,, \tag{2.18}$$

where we also defined the function $E(a)$ for convenience.

If the vacuum energy density is not constant, we can replace Ω_Λ with $\Omega_{\rm de}(1+z)^{3(1+w)}$.

The above equation is formulated with all density parameter and the Hubble parameter at present time, denoted by subscript 0. We find that for all times the Friedmann equation is equivalent to

$$1 = \Omega_{\rm R} + \Omega_{\rm M} + \Omega_{\rm K} + \Omega_\Lambda \,. \tag{2.19}$$

Special solutions of the Friedmann equation for flat space are of particular interest:

- **Matter dominated universe:**
  If non-relativistic matter is the only contribution to the energy density, we can write $\varrho_{\rm M} = \varrho_{\rm M,0}(a_0/a)^3$, with $\varrho_{\rm M,0}$ being the present energy density of matter. We find $a(t) \propto (3/2H_0 t)^{2/3}$ is a solution of the Friedmann equation. This model is also referred as the *Einstein-de Sitter* (EdS) universe and states that the Universe is decelerating due to gravitational attraction.

- **Radiation dominated universe:**
  Applying only $\varrho_{\rm R} = \varrho_{\rm R,0}(a_0/a)^4$ in the Friedmann equation leads to the solution $a(t) \propto (2H_0 t)^{1/2}$. We find also a decelerating behavior as in the EdS universe, but with different scaling of $a(t)$.

- **Cosmological constant dominated universe:**
  Here we solve $\dot{a}/a = H_0$ and find $a(t) = a_0 \exp{(H_0 t)}$, which is an exponential expansion without an initial singularity at $a = 0$. Thus, this model has no Big Bang and expands forever.

2.2.1 Cosmic time and distance measures

To describe large-scale structures, we have to introduce time and distance measures in an expanding universe. Also, we introduce *redshift* as a key observable in cosmology that is defined in terms of the scale factor and wavelength as

$$z \equiv \frac{\lambda_{\rm o} - \lambda_{\rm e}}{\lambda_{\rm e}} = \frac{a(t_{\rm o})}{a(t_{\rm e})} - 1 \,, \tag{2.20}$$

between the time of emission $t_{\rm e}$ and observation $t_{\rm o}$.

The concept of distance for an expanding universe is a little more challenging. For light propagation, $\mathrm{d}s = 0$ in Equation 2.3, we can write the radial distance as[1]

$$c\frac{\mathrm{d}t}{a(t)} = \mathrm{d}\chi \,. \tag{2.21}$$

The right-hand site of Equation 2.21 corresponds to *comoving distance* r for zero curvature (which we will assume from now on for the distance measures). For the left-hand side we utilize Equation 2.23 and write

$$\chi = \frac{c}{H_0}\int \frac{\mathrm{d}a}{a^2 E(a)} = \frac{c}{H_0}\int \frac{\mathrm{d}z}{\tilde{E}(z)} \,, \tag{2.22}$$

where we introduced $\tilde{E}(z)$ as the parametrization of $E(a)$ in terms of redshift, according to Equation 2.20. This distance measure is constant for all objects that have no peculiar motion but reside at fixed comoving coordinates in the RW metric. Thus, we can compare distances of objects at different cosmic times.

---

[1] Only the radial distance is crucial, as we can always place the observer in the origin. Also we reintroduced c for clarification of the units.

Let us now define the time scale for the present, thus the *age of the Universe*

$$t_0 = \int_0^{t_0} \mathrm{d}t = H_0^{-1} \int_0^{a_0} \frac{\mathrm{d}a}{aE(a)} = H_0^{-1} \int_{z_0}^{\infty} \frac{\mathrm{d}z}{(1+z)\tilde{E}(z)} \,, \tag{2.23}$$

using $z_0 = 0$ for present (for zero curvature models we use $a_0 = 1$). In this way we can also estimate the time scale corresponding to a particular scale factor or redshift.

Complementary to the comoving distance measure there is the *proper physical distance* x, defined by the distance the light is travelling from one object to the other. It is written as

$$x(t) = a(t)r = a\frac{c}{H_0} \int \frac{\mathrm{d}a}{a^2 E(a)} \,. \tag{2.24}$$

There are two more widely used distance measures in cosmology:

- *Angular diameter distance*
  The angular diameter distance d_A is defined by the size of an object s and the angle ϑ as $d_\mathrm{A} = s/\vartheta$. Integrating the RW metric we obtain $s = a(t)r\vartheta$, and therefore we write

  $$d_\mathrm{A}(z) = \frac{r(z)}{1+z} \,. \tag{2.25}$$

  If the diameter s is scaling according to the expansion of the Universe, we write $s = (1+z)s_0$. This distance scale is interesting when analyzing standard rulers, objects that have the same (comoving) size at different redshifts.

- *Luminosity distance*
  The luminosity distance is defined as the relationship between the absolute *luminosity* L of an object and the *flux* F as

  $$d_\mathrm{L} = \sqrt{\frac{L}{4\pi F}} \,. \tag{2.26}$$

  The relation to the angular diameter distance is given as

  $$d_\mathrm{L} = (1+z)^2 d_\mathrm{A}(z) \,, \tag{2.27}$$

  where the two powers of $(1+z)$ are firstly due to redshift and secondly due to the cosmological time dilation when receiving the emitted radiation.

2.2.2 Horizons

As distances increase following the expansion of the Universe, superluminal recession velocities may occur at large enough distances. This is why determine *horizons*.

- *Particle horizon*
  The particle horizon d_PH is defined as the maximum proper distance over which two objects can be causally connected at a time t

  $$d_\mathrm{PH} = a(t) \int_0^t \frac{c\mathrm{d}t'}{a} \,. \tag{2.28}$$

  Thus, the particle horizon is also defining the size of the *observable universe*.

- *Event horizon*
  A universe that is dark energy dominated undergoes accelerated expansion. This means, that objects that an observer may see today (within the particle horizon) will however never receive light that the observer emits today. If this is the case, the objects is outside the today's event horizon, defined as

$$d_{\mathrm{EH}} = a(t) \int\limits_{t}^{\infty} \frac{c\mathrm{d}t'}{a}\,. \tag{2.29}$$

- *Hubble radius*
  Additionally of interest is the Hubble radius d_{H}, defined as

$$d_{\mathrm{H}}(t) = \frac{c}{H(t)}\,. \tag{2.30}$$

  In an ever accelerating universe, two galaxies separated by a distance greater than the Hubble radius today, stay causally disconnected in future.

In a matter dominated universe the Hubble radius and the particle horizon, although being different concepts, are roughly the same. Thus, to judge horizon problems, it can be more convenient to use the Hubble radius.

2.3 Λ Cold Dark Matter model

We discussed all possible energy sources in the FRW Universe but there is no theory atop which does predict the composition of the density parameters. As a consequence, we need observations to measure them. Also we need to incorporate these findings in a complete picture of the history of the Universe. The current *concordance model* for the dynamics in our Universe is called Λ Cold Dark Matter (ΛCDM). It is based on the FRW cosmology, implying a hot big bang as the origin of space-time. In this section I will discuss the concept of the ΛCDM model to lay the foundation in order to comprise the work done in this thesis. I will further discuss the most important observations in cosmology which led to the present ΛCDM model.

2.3.1 Dark Matter

The earliest hint for dark matter is due to Zwicky (1933). He observed that the relative velocities of galaxies in galaxy clusters were much larger than the velocity to escape the gravitational bound due to the mass of the cluster. Within this assumption, he estimated the total cluster mass from the amount of light emitted by the galaxies. This suggested that there must actually be much more mass in the galaxy clusters than only the luminous galaxies we can see. He was not able to give a solution to this mismatch. However, he mentioned it could be overcome by the existence of (cold) and dark matter that only interacts gravitationally.

Similar discrepancies were measured when rotation curves of galaxies were observed. Kepler's third law states that the velocity of an orbiting body around a central mass should scale with distance like $v \propto 1/\sqrt{r}$. However, the rotation velocities found in Rubin et al. (1980) were deviating significantly from Kepler's predictions, being flat over a large distance range. This can be explained by a dark matter halo in which a galaxy is placed. As dark matter is interacting gravitationally, we also expect lensing effects. These are also found to be in agreement with the dark matter assumptions, shown in Clowe et al. (2007).

2.3.2 Supernova Cosmology Project & High-z Supernova Search Team

The Supernova Cosmology Project & High-z Supernova Search Team are two research groups who found that the Universe today is undergoing an accelerated expansion and therefore proved that a positive cosmological constant is dominating the present energy density in the Universe by analyzing redshifts and apparent magnitudes m_B (alternatively the distance modulus $\mu(z) = 5\log_{10}(d_{\mathrm{L}}(z)/10\,\mathrm{pc})$ of Type Ia supernovae (SNIa). Supernovae of type Ia are standard candles, as their intrinsic luminosity is related to the decay rate and can therefore be robustly determined. Within these analyzes, the apparent magnitude of a SNIa is plotted against the redshift. The teams found that the data is supporting a positive cosmological constant as the observed supernovae at a given redshift appeared fainter than a decelerating universe would predict, shown in Figure 2.1. The plot shows that a dominating Ω_Λ is favored by the statistical analysis of the data.

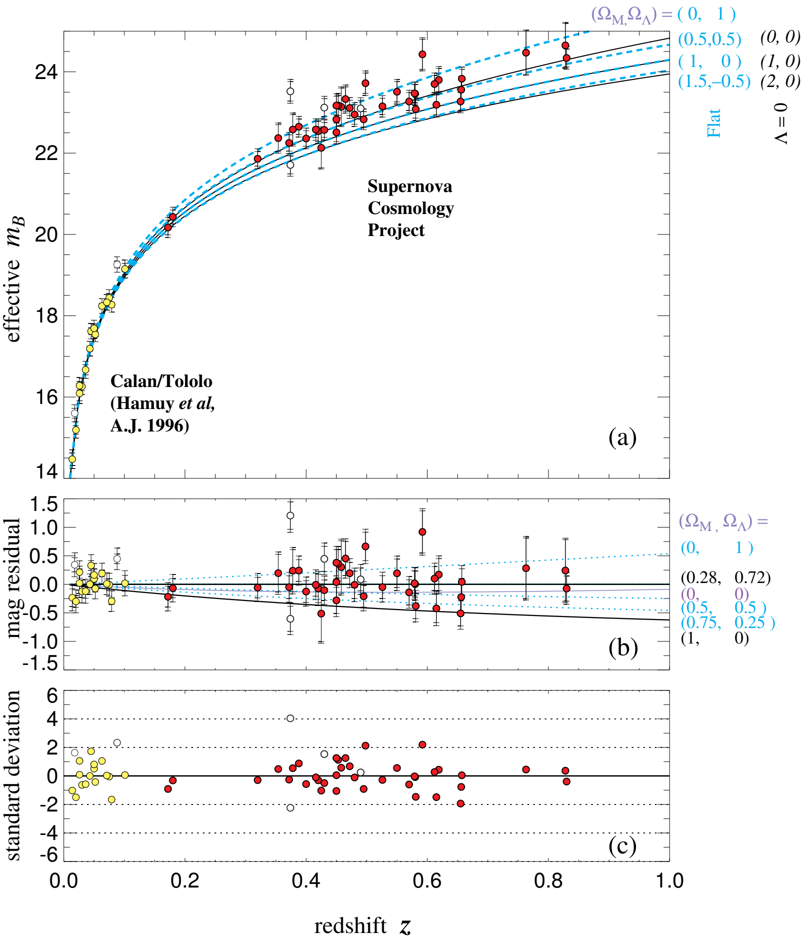


Figure 2.1: Original plot from the Supernova Cosmology project (Perlmutter et al., 1999). The magnitude m_B of the observed Supernova Ia is plotted as function of redshift z. Solid and dashed lines corresponds to the theoretical predictions for different combinations of Ω_{M} and Ω_Λ. The residuals of the fits and also the standard deviations are shown below.

2.3.3 Cosmic Microwave Background & Baryonic Acoustic Oscillations

The cosmic microwave background (CMB) radiation is the best observational confirmation of the Big Bang model. It was predicted by Gamow in the 1940s (Gamow, 1946) to be a relic radiation from the epoch when the Universe became transparent, also known as *decoupling*, as before matter and radiation formed a primordial plasma. The spatial origin of the CMB is also called *last scattering surface*. The CMB radiation possess a nearly perfect black-body spectrum of 2.726±0.005 K. The temperature anisotropies of the CMB contain a wealth of information that are analyzed with a multipole expansion in spherical harmonics as $\Delta T/T(\varphi, \vartheta)$. An application of particular interest for this thesis is the Sachs-Wolfe effect (Sachs & Wolfe, 1967), which causes photons from the cosmic microwave background to be gravitationally redshifted or blueshifted due to a changing gravitational field in between the source of radiation and the observer. A photon coming from a overdense region will have a slightly larger redshift due to the deeper gravitational well at the surface of last scattering. Conversely, a photon coming from an underdense region will have a slightly smaller redshift. Thus, we can calculate the CMB temperature anisotropy due to the varying potential Φ from density fluctuations at the surface of last scattering given by $\Delta T/T \propto \Phi$, shown in Figure 2.2.

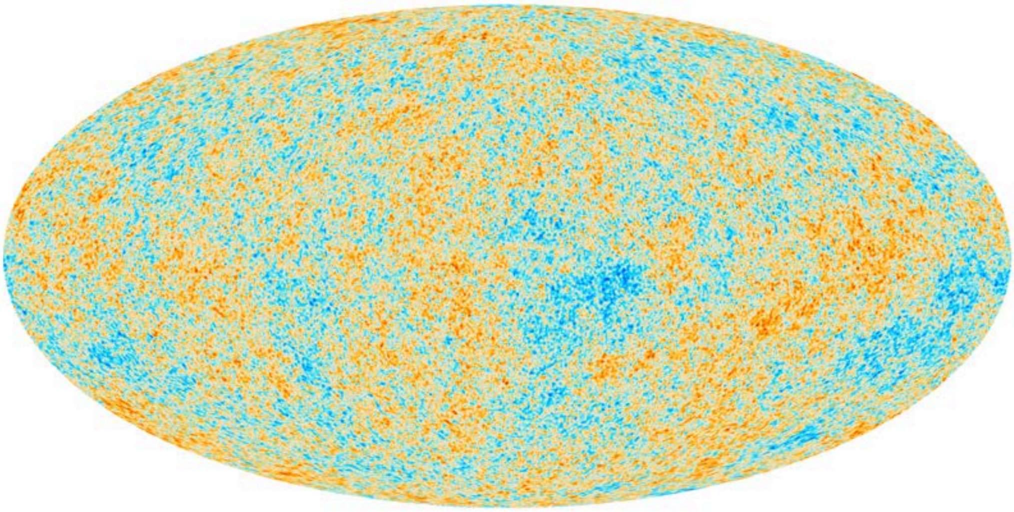


Figure 2.2: Most recent CMB temperature map taken with the Planck satellite (Planck Collaboration, 2015). The color code represents the temperature fluctuations, hotter regions shown in red, colder in blue.

The fluctuations in the CMB are well described by a Gaussian random field, that we will discuss in Section 2.6.5.

Also an inhomogeneous distribution of matter between the observer and the last scattering surface may induce anisotropy by inverse Compton scattering of CMB photons by free electrons in a hot intergalactic plasma (Sunayev-Zel'dovich effect Sunyaev & Zeldovich (1970)). The Baryonic Acoustic Oscillations (BAO) are related to the CMB. Before decoupling, the photons are coupled tightly to the baryons, forming a primordial plasma. The perturbations in the baryon-photon fluid are oscillating, whereas dark matter perturbations grow. Characteristic for these oscillations is the *sound horizon* in comoving coordinates

$$s = \int\limits_0^{t_{\mathrm{dec}}} \frac{c_{\mathrm{S}}\,\mathrm{d}t}{a} \sim 100\,h^{-1}\,\mathrm{Mpc}\,, \tag{2.31}$$

with c_{S} being the speed of the sound waves. These waves freeze when baryons and radiation decouple. The sound horizon is imprinted in the CMB and also in the perturbations of the matter density, although the non-linear evolution under gravitational interaction is diminishing this effect. The galaxy correlation function can be seen in Figure 2.3 on the left, showing an excess at separations $\sim 100\,h^{-1}$ Mpc. The geometric interpretation of the BAO is shown on the right of Figure 2.3. At distances of $100\,h^{-1}$ Mpc the probability to find another galaxy is increased wrt a random distribution.

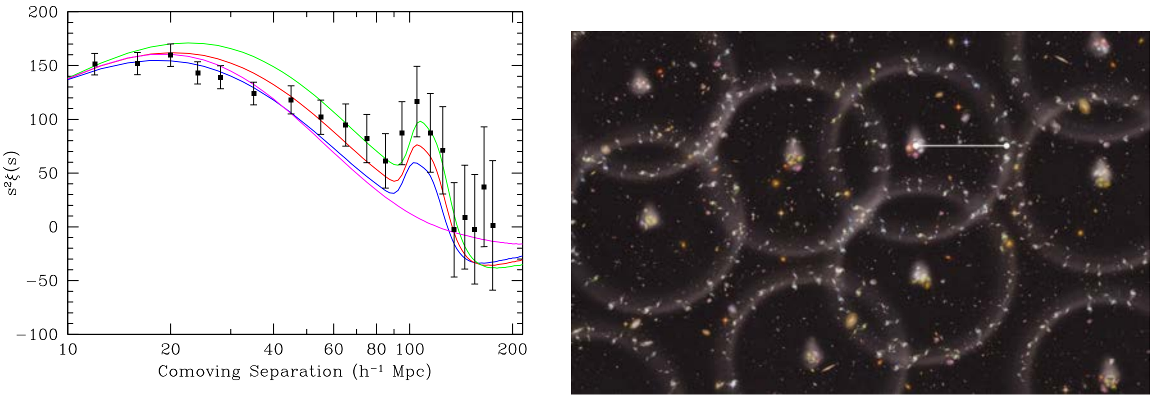


Figure 2.3: Left: original plot from (Eisenstein et al., 2005) showing two-point correlation function. Right: schematic illustration of the BAO, showing a density ring around the central galaxies (courtesy `http://www.astro.ucla.edu`.

The first significant detection of a BAO signal was reported in Eisenstein et al. (2005) using the two-point correlation function with luminous red galaxies observed with the SDSS survey, and simultaneously in Cole et al. (2005) in a power spectrum analysis with the 2dF Galaxy Redshift Survey. Around the characteristic scale of the BAO we expect a higher chance to find a galaxy at a distance of the sound horizon around another galaxy. We will see later that the bump in the correlation function in configuration space will assume the form of wiggles when the density perturbations are analyzed in Fourier space.

2.3.4 Number counts

The evolution of the Universe and properties of homogeneity can also be tested with the cumulative number of objects one sees in samples that probe larger and larger distances. The counts per volume and galaxy type, color or other spectral properties can be a helpful statistical tool. By defining models for the evolution of the galaxy's luminosity (luminosity function), one can predict how many sources one should see above an certain threshold and as a function of redshift. If one accounts for evolution of the intrinsic properties of the sources correctly, then any residual dependence on redshift is due to the volume of space encompassed by a given interval in redshift; this depends quite strongly on $\Omega_{\rm tot}$. Also the matter statistics can be derived from these kind of analyses (Yang & Saslaw, 2011). In Figure 2.4 an illustration of the SDSS galaxies are shown, with the rest-frame $g-r$ color (Zehavi et al., 2011).

2.4 The horizon & the flatness problem

Gravity has a tendency to enhance inhomogeneities. This irreversible process causes structures like galaxies to form. Hence, if gravity steadily is clumping matter together and forms inhomogeneities, then the Universe must have been in an extreme state of homogeneity initially. This seems very unlikely because one expects that the Universe was formed from tiny fluctuations in the *initial conditions*. A

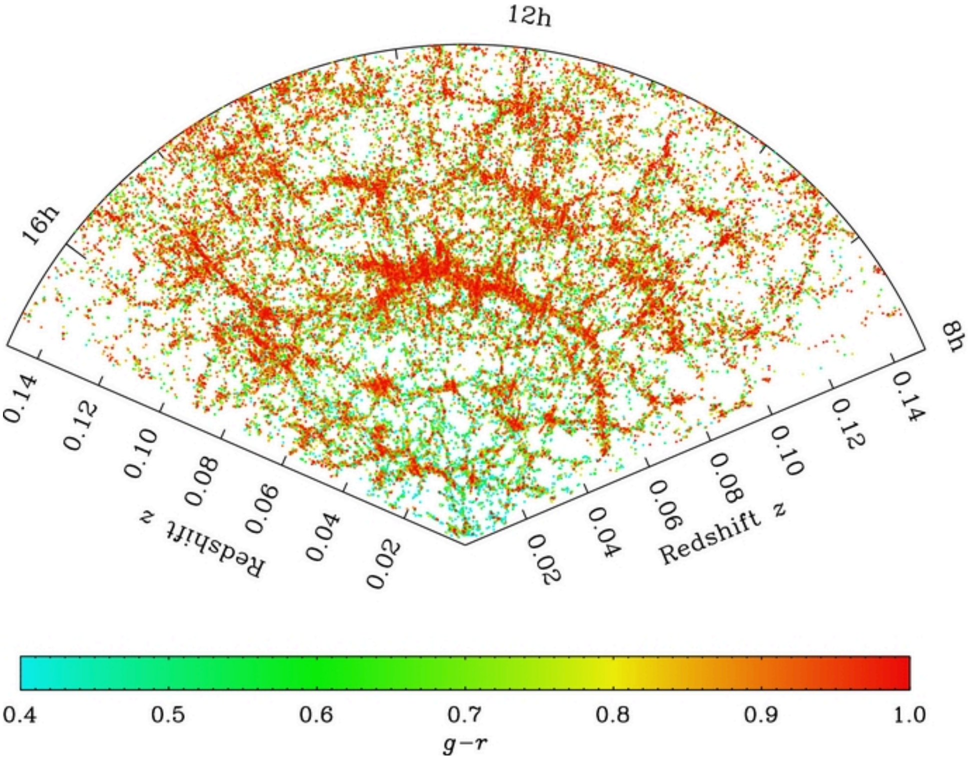


Figure 2.4: Slice through the SDSS main galaxy sample, with galaxies color-coded based on rest-frame $g - r$ color. The slice shows galaxies within ±4 deg of the celestial equator, in the north Galactic cap. (Zehavi et al., 2011)

homogeneous and isotropic state is quite special; an inhomogeneous state is by far a more general state than a homogeneous one.

The related *horizon problem* disturbed the cosmologists for a long time. The cosmic microwave background was seen to be very homogeneous and isotropic. Actually, it is the most perfect black-body radiation known, better than any spectrum that can be produced in the laboratory. The isotropy in the radiation indicates that the radiation had thermal contact once in the past, before it was emitted. Photons of the CMB must have been in thermal contact even if coming from different directions. This is what is called the horizon problem because the particle horizon to each photon in the last scattering surface only covers a small patch of the sky of about two degrees.

It can be solved within the theory of *inflation*. This theory predicts an era with exponential growth in the very early Universe shortly after the Big Bang. Matter and radiation domination predict a decelerating expansion $\ddot{a} < 0$, so two objects within the particle horizon will stay connected. But for models with $\ddot{a} > 0$, like in the cosmological constant dominated universe, two objects that were within their particle horizon are able to leave it and being outside the horizon at a later time. So, this scenario explains how photons in the CMB may have had physical interactions when they were within the horizon of each other and later left the horizon due to exponential expansion.
Another crucial issue inflation is potentially solving is the *flatness* problem. With the Friedmann Equations we derived that $\Omega_{\rm tot} = 1$ at any time. Models with zero spatial curvature demand

$$\Omega_{\rm tot} - 1 = \Omega_{\rm K} = 0\,. \tag{2.32}$$

We showed that the energy density of curvature can be expressed as $|\Omega_{\rm K}| = |K|/(H^2a^2)$. For a radiation dominated universe the scale factor evolves as $a(t) \propto t^{\frac{1}{2}}$, so that $\Omega_{\rm K} \propto t$, while a matter dominated universe implies that $a(t) \propto t^{\frac{2}{3}}$ so that $\Omega_{\mathrm{K}} \propto t^{\frac{2}{3}}$. This means that Ω_{K} must be *fine-tuned* to zero in the early Universe as Ω_{K} grows with time. Tiny deviations from the curvature energy density from zero would have grown largely until today.

The concept of *inflation* is solving this problem. If the Universe would have gone through an exponential expansion like the cosmological constant dominated universe with $a(t) \propto \exp(Ht)$ from just $t_{\mathrm{i}} = 10^{-36}\,\mathrm{s}$ to $t_{\mathrm{f}} = 10^{-34}\,\mathrm{s}$, this would flatten the space like $\Omega_{\mathrm{K}} \propto \exp(-2Ht)$ by 60 e-holds. The inflation could have been driven by a scalar field undergoing a phase-transition (summarized in Guth (1984)). The dynamics of such a field is described by a Lagrangian of a spin-0 particle $\mathscr{L} = \frac{1}{2}\partial^{\mu}\phi\partial_{\mu}\phi - V(\phi)$, where the first term described the kinetic energy and the second is the potential function.

2.5 Reference model for this thesis

We adopt the most recent set of cosmological parameters in the findings of the CMB analysis (Planck Collaboration, 2015), called *Planck Cosmology*, if nothing else is stated. Besides the cosmological parameters we also assume the initial conditions of the early Universe to be Gaussian from which today's structures grew. The cosmological parameters assert a spatial flat space and that the Universe is undergoing an accelerated expansion due to the density parameter of the dark energy Ω_{Λ}. The radiation density is negligible today and also the major part of the matter density in the Universe is dark.

2.6 Perturbed Universe

Up to this point I have described the Universe with a homogeneous and isotropic model. I will keep this FRW model to describe the background. In this Section I will discuss a formalism to analyze the generation and evolution of inhomogeneities. I therefore solve the equations of a fluid in presence of gravity and an expanding universe. In order to keep the handling of the perturbations feasible I will focus on the linear regime of the solutions. Perturbation theory in cosmology (Peebles (1980), Fry (1984), Bernardeau et al. (2002)) is further developed to higher orders (McDonald & Roy, 2009). Also renormalized perturbation formalism up to a certain scale have been developed (e.g. Crocce & Scoccimarro, 2006).

2.6.1 Eulerian description of non-relativistic density perturbations

A Newtonian non-relativistic fluid with mass density ϱ, pressure p and potential Φ is described in Eulerian space by

- the *continuity equation*, which results from mass conservation, revealing change of a local density in time is equal to the flux through the surface,
- the *Euler equation*, which describes the force acting on a fluid due to the gradient of the pressure and a gravitational potential,
- the *Poisson equation*, describing the potential induced by the mass inhomogenity.

We write the continuity equation for the mass density $\varrho(\boldsymbol{x}, t)$ and the flux $\varrho\boldsymbol{u}(\boldsymbol{x}, t)$, where $\boldsymbol{x}$ is the proper physical distance, as

$$\frac{\partial\varrho}{\partial t} = -\boldsymbol{\nabla}_{\boldsymbol{x}}(\varrho\boldsymbol{u})\,. \tag{2.33}$$

The Euler equation with pressure p writes as

$$\left(\frac{\partial}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}\right) \boldsymbol{u} = -\frac{\boldsymbol{\nabla}_{\boldsymbol{x}} p}{\varrho} - \boldsymbol{\nabla}_{\boldsymbol{x}} \mathit{\Phi} \,. \tag{2.34}$$

The potential $\mathit{\Phi}$ is derived from the Poisson equation

$$\nabla^2_{\boldsymbol{x}} \mathit{\Phi} = 4\pi G \varrho \,. \tag{2.35}$$

Perturbation analyses assume that the equations can be expanded in series, e.g. expanding the mass density like $\varrho(\boldsymbol{x}, t) = \bar{\varrho}(t) + \delta\varrho(\boldsymbol{x}, t)$, with a time dependent background density, but *independent* of spatial position $\bar{\varrho}(t)$, and a small perturbation at position $\boldsymbol{x}$, written as $\delta\varrho(\boldsymbol{x}, t) << 1$. We can expand all other quantities in our equations according to this scheme, namely, pressure, velocity, and finally potential. Also we assume a curl-free velocity field $\nabla \times \boldsymbol{u} = 0$, as any initial curl would have asymptotically vanished with the expansion of the Universe.

Linearization implies now that only linear terms are considered and all higher orders can be neglected, i.e. the product of two and more perturbations, e.g. $\delta\varrho^2$ or $\delta\boldsymbol{u}\delta\varrho$. Assuming no expansion and no gravity, we obtain the linear expressions

$$\frac{\partial \delta\varrho}{\partial t} = -\boldsymbol{\nabla}_{\boldsymbol{x}} \cdot (\bar{\varrho}\boldsymbol{u}) \tag{2.36}$$

$$\bar{\varrho}\frac{\partial \boldsymbol{u}}{\partial t} = -\boldsymbol{\nabla}_{\boldsymbol{x}} \delta p \,. \tag{2.37}$$

We can connect pressure and density perturbations for adiabatic processes by introducing the adiabatic sound speed $c_{\mathrm{S}}^2 = \left(\frac{\delta p}{\delta \varrho}\right)_s$, where subscript s denotes that the entropy does not change. Combining these equations we find the differential equation for the density fluctuation

$$\frac{\partial^2 \delta\varrho}{\partial t^2} - c_{\mathrm{S}}^2 \nabla^2_{\boldsymbol{x}} \delta\varrho = 0 \,. \tag{2.38}$$

This differential equations for a non-expanding universe is solved by a plane wave $\delta\varrho = A \exp\left[i(\omega t - \boldsymbol{k} \cdot \boldsymbol{x})\right]$ with constant amplitude and frequency $\omega = c_{\mathrm{S}} k$. It is customary to define the *density contrast* $\delta = \delta\varrho/\bar{\varrho}$ and also to consider the *peculiar potential* Φ, that is a perturbation of the mean background potential as

$$\mathit{\Phi} = \frac{2\pi G}{3} \bar{\varrho} r^2 + \Phi \,. \tag{2.39}$$

The gravitational interaction is described by the perturbed Poisson equation

$$\nabla^2_{\boldsymbol{x}} \Phi = 4\pi G \bar{\varrho} \delta \,, \tag{2.40}$$

and consequently we obtain

$$\frac{\partial^2 \delta}{\partial t^2} - c_{\mathrm{S}}^2 \nabla^2_{\boldsymbol{x}} \delta = 4\pi G \bar{\varrho} \delta \,. \tag{2.41}$$

The solutions to this equation are oscillation in which frequency $\omega^2 = c_{\mathrm{S}}^2 k^2 - 4\pi G \bar{\varrho}$. This shows that the fluctuations with wavelength larger that $\lambda_{\mathrm{J}} = \frac{2\pi c_{\mathrm{S}}}{\sqrt{4\pi G \bar{\varrho}}}$ grow exponentially while at smaller scales the fluctuations oscillate. This wavelength is called *Jeans' length.*

We switch from physical to comoving coordinates, using the convention we introduce above as $\boldsymbol{x} = a(t)\boldsymbol{r}$, and decompose the total velocity as

$$\boldsymbol{u} = \dot{\boldsymbol{x}} = \underbrace{\dot{a}\boldsymbol{r}}_{H\boldsymbol{x}} + \underbrace{a\dot{\boldsymbol{r}}}_{\boldsymbol{v}_{\mathrm{P}}} \,, \tag{2.42}$$

which is a composition of the Hubble recession velocity and the peculiar velocity. We also need to refine the time derivative as

$$\left(\frac{\partial}{\partial t}\right)_{\boldsymbol{r}} = \left(\frac{\partial}{\partial t}\right)_{\boldsymbol{x}} - H\boldsymbol{x}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}\,, \tag{2.43}$$

where we defined $\boldsymbol{\nabla}_{\boldsymbol{r}} = \frac{1}{a}\boldsymbol{\nabla}_{\boldsymbol{x}}$. We drop the subscripts in the following. Applying all these relations into the continuity equation we obtain

$$\dot{\bar{\varrho}} + 3H\bar{\varrho} = 0\,, \tag{2.44}$$

which is simply the evolution of the background density and independent of perturbations, and on the other hand the perturbative part

$$\dot{\delta} = -\frac{1}{a}\boldsymbol{\nabla}\cdot\boldsymbol{v}_{\mathrm{p}}\,. \tag{2.45}$$

The Euler equation is now written as

$$\dot{\boldsymbol{v}}_{\mathrm{p}} + H\boldsymbol{v}_{\mathrm{p}} = -\frac{1}{a\bar{\varrho}}\boldsymbol{\nabla}\delta p = \frac{1}{a}\boldsymbol{\nabla}\Phi\,. \tag{2.46}$$

The Poisson equation in Equation 2.40 is already linear. We combine the equations that describe a fluid in an expanding coordinate system and obtain the so-called *perturbation equation*

$$\ddot{\delta} + 2H\dot{\delta} = 4\pi G\bar{\varrho}\delta\,, \tag{2.47}$$

where we dropped the pressure term as we assume $c_{\mathrm{S}} = 0$ for cold dark matter (CDM). Equation 2.47 is generally solved by the ansatz

$$\delta(\boldsymbol{r}, t) = A(\boldsymbol{r})D(t) + B(\boldsymbol{r})C(t)\,. \tag{2.48}$$

$D(t)$ is the growing solution (called growth factor) and $C(t)$ the decaying one. We will only consider the growing solution in the following to be physically relevant. Thus we see that in the linear solutions of perturbations the spatial and time depending parts factorize. This means that as long as this approximation is valid, the fluctuations grow in space in comoving coordinates. A solution for $D(t)$ can be given up to a normalization. A common choice is to normalize $D(z) = 1$ at $z = 0$. So we write the solution for flat space as

$$D(z) = \frac{H(z)}{H_0}\int_z^\infty \frac{\mathrm{d}z'}{H^3(z')}\left[\int_0^\infty \frac{\mathrm{d}z'}{H^3(z')}\right]^{-1}\,. \tag{2.49}$$

The density contrast of point $\boldsymbol{r}$ at time t can be calculated from the initial density contrast as $\delta(\boldsymbol{r}, t) = \delta(\boldsymbol{r}, t_i)\frac{D(t)}{D(t_i)}$.

This solution predicts:

- $D(t) \propto a(t)$ in a matter dominated epoch (also called Einstein–de Sitter universe, EdS),
- $D(t) \propto a^2(t)$ in the radiation dominated universe,
- and $D(t) \propto a^{-2}(t)$ if the universe is Λ dominated.

Applying these relations to Equation 2.45 we find

$$\nabla \cdot \boldsymbol{v}_{\mathrm{p}} = -\dot{\delta} = -\dot{a}\frac{\partial \delta}{\partial a} = -H\frac{\partial \delta}{\partial \ln a}\,. \tag{2.50}$$

This can be more elegantly written defining the *growth rate*

$$f_{\Omega} = \frac{1}{H}\frac{\dot{D}}{D} = \frac{\mathrm{d}\ln D}{\mathrm{d}\ln a}\,, \tag{2.51}$$

as

$$\boldsymbol{\nabla} \cdot \boldsymbol{v}_{\mathrm{p}} = -aHf_{\Omega}\delta\,. \tag{2.52}$$

In linear theory the divergence of the peculiar velocity field is proportional to the density contrast in comoving coordinates.

A non-linear treatment of perturbations has also been studied in literature (Goroff et al., 1986; Jain & Bertschinger, 1994). They rely on a Fourier decomposition of the terms that are neglected in the linear approach. These functions are described as convolutions in Fourier space, thus introducing terms that are known as *mode couplings*.

2.6.2 Statistical description

Pioneering work in statistical descriptions of the density fluctuations stems from Peebles (1973), Peacock & Heavens (1985). We will stick to the common notations and discuss the most important findings for large-scale structure analyses. To understand better the validity of perturbation theory and also being able to statistically describe the density fluctuations we discuss now the Fourier decomposition of density and velocity perturbations. We decompose the density contrast and the divergence of the velocity field $\theta = -\boldsymbol{\nabla} \cdot \boldsymbol{v}$ in terms of plane waves as

$$\delta(\boldsymbol{k}, t) = \frac{1}{(2\pi)^3}\int \mathrm{d}^3x\, \delta(\boldsymbol{r}, t)\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}} \tag{2.53}$$

$$\theta(\boldsymbol{k}, t) = \frac{1}{(2\pi)^3}\int \mathrm{d}^3x\, \theta(\boldsymbol{r}, t)\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}}\,, \tag{2.54}$$

where $\boldsymbol{k}$ is the wave vector of a particular plane wave, also called *mode*. The magnitude of $|\boldsymbol{k}| \equiv k$ is connected to the wavelength via $k = \frac{2\pi}{\lambda}$.

By definition the mean of the density contrast $\delta = \frac{\varrho}{\bar{\varrho}} - 1$ is $\langle\delta\rangle = 0$. However, the variance of the density contrast does not vanish and is one of the most important observables in cosmology, given as

$$\sigma^2 = \left\langle \delta^2 \right\rangle - \langle\delta\rangle^2 = \left\langle \delta^2 \right\rangle\,. \tag{2.55}$$

As we are interested in the variance as a function of different modes, it is convenient to use the *power spectrum* defined as

$$\langle \delta(\boldsymbol{k})\delta(\boldsymbol{k}')\rangle = P(\boldsymbol{k})\delta^{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}')\,, \tag{2.56}$$

where the Dirac delta function states that all modes of $\delta(\boldsymbol{k})$ are uncorrelated. Isotropy implies that $P(\boldsymbol{k}) = P(k)$, i.e. the power spectrum is only a function of $|\boldsymbol{k}| \equiv k$. For large enough volumes we can write

$$\sigma^2 = \frac{1}{2\pi^2}\int \mathrm{d}k\, k^2\, P(k)\,. \tag{2.57}$$

The spatial (auto-)correlation function ξ (also called two-point correlation function ξ_2, see below) is defined as a excess probability $\mathrm{d}P$, relative to a random distribution, of finding a pair of galaxies at a distance $\boldsymbol{r}_{12}$ in the volumes $\mathrm{d}V_1$ and $\mathrm{d}V_2$

$$\mathrm{d}P = \bar{n}^2(1+\xi(\boldsymbol{r}_{12}))\mathrm{d}V_1\mathrm{d}V_2\,, \tag{2.58}$$

where $\bar{n}$ is the mean number density. The correlation function can be expressed for a spatial separation of $\boldsymbol{\Delta}$ as $\xi(\boldsymbol{\Delta}) = \langle\delta(\boldsymbol{r})\delta(\boldsymbol{r}+\boldsymbol{\Delta})\rangle$.

It is straight forward to generalize the correlation function to higher degrees (called two-point, three-point, ..., N-point correlation function) as

$$\begin{aligned}
\langle\delta(\boldsymbol{r}_1)\delta(\boldsymbol{r}_2)\rangle &= \xi_2\,, && (2.59)\\
\langle\delta(\boldsymbol{r}_1)\delta(\boldsymbol{r}_2)\delta(\boldsymbol{r}_3)\rangle &= \xi_3\,, && (2.60)\\
&\vdots \\
\langle\delta(\boldsymbol{r}_1)\delta(\boldsymbol{r}_2)\delta(\boldsymbol{r}_3)...\delta(\boldsymbol{r}_N)\rangle &= \xi_N\,. && (2.61)
\end{aligned}$$

The correlation function is the counterpart to the power spectrum in configuration space.

$$\begin{aligned}
\langle\delta(\boldsymbol{r})\delta(\boldsymbol{r}+\boldsymbol{\Delta})\rangle &= \left\langle\int\frac{\mathrm{d}^3k'}{2\pi^3}\int\frac{\mathrm{d}^3k}{2\pi^3}\delta(\boldsymbol{k}')\delta(\boldsymbol{k})\mathrm{e}^{i\boldsymbol{k}'\cdot\boldsymbol{r}}\mathrm{e}^{-i\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{\Delta})}\right\rangle\,, && (2.62)\\
&= 2\pi\int\frac{\mathrm{d}^3k'}{2\pi^3}\int\frac{\mathrm{d}^3k}{2\pi^3}\delta^{\mathrm{D}}(\boldsymbol{k}-\boldsymbol{k}')P(\boldsymbol{k})\mathrm{e}^{i\boldsymbol{r}\cdot(\boldsymbol{k}'-\boldsymbol{k})-i\boldsymbol{k}\boldsymbol{\Delta}}\,, && (2.63)\\
&= \int\frac{\mathrm{d}^3k}{2\pi^3}P(\boldsymbol{k})\mathrm{e}^{-i\boldsymbol{k}\boldsymbol{\Delta}}\,. && (2.64)
\end{aligned}$$

This is also known as the Wiener-Khinchin theorem: The power spectrum is the Fourier transform of the (auto-)correlation function.

We use the condition of isotropy to simplify the argument of the correlation function as $\xi(|\boldsymbol{r}|) = \xi(r)$. In this case we can simplify Equation 2.64 by directly integrating the angular dependencies of the wave vector $\boldsymbol{k}$ and obtain

$$\xi(r) = \frac{1}{2\pi^2}\int dkP(k)k^2\frac{\sin(kr)}{kr}\,. \tag{2.65}$$

Obviously, we find that $\xi(r=0) = \langle\delta^2\rangle = \sigma^2$, thus the correlation function at zero separation corresponds to the variance of the mean density squared. Density fluctuations appear on all scales of k. It is often useful to apply a *window function* $W_{\mathrm{r_s}}(k)$ on the density contrast, which is either convoluted in configuration space or multiplied in Fourier space. The resulting density will be smoothed according to the smoothing length of the window $\delta_{\mathrm{smooth}}(k) = \delta(k)W_{\mathrm{r_s}}(k)$ and thus the power spectrum becomes $P_{\mathrm{smooth}}(k) = P(k)W^2_{\mathrm{r_s}}(k)$. Of particular interest is the fluctuation in a spherical volume of radius $r_s = 8\,h^{-1}\,\mathrm{Mpc}$, filtered with a *top-hat* window function in Fourier space

$$W_{\mathrm{r_s}}(k) = \frac{3(\sin kr_s - kr_s\cos kr_s)}{(kr_s)^2}\,, \tag{2.66}$$

assuming the form

$$\sigma_8^2 = \frac{1}{2\pi^2}\int\limits_0^\infty \mathrm{d}kk^2P(k)|W_{\mathrm{r_s}}(k)|^2\,. \tag{2.67}$$

2.6.3 Discretization effects

The power spectrum supposes continuous density fields. However, galaxies are discrete objects. Here I will discuss the numerical effects, when discrete objects' positions are interpolated to a mesh grid. Due to the discrete number counts the total power spectrum will have another constant additive term, called shot noise

$$P(k)_{\mathrm{tot}} = P(k)_{\mathrm{shot}} + P(k)\,. \tag{2.68}$$

$P(k)_{\mathrm{shot}}$ is also also called the Poissonian white noise and is written as $P(k)_{\mathrm{shot}} = \frac{1}{\bar{n}}$, where $\bar{n}$ is the mean number density of galaxies in the considered volume.

The interpolation of the galaxy position on a mesh grid is done with a *mass assignment kernel* (MAK), which can be written as a polynomial function. In this thesis I use only the *nearest grid point* (NGP) and the *cloud in cell* (CIC) scheme. The interpolated density fields are a convolution of the original field, e.g. $\delta(\boldsymbol{k})$ and the kernels $W(\boldsymbol{k})$. The kernels can be written as a product over the spatial dimensions i

$$W(\boldsymbol{k}) = \left[\prod_i \left(\frac{\sin(\pi k_i/2k_{\mathrm{Ny}})}{\pi k_i/2k_{\mathrm{Ny}}}\right)\right]^{\zeta}, \tag{2.69}$$

where ζ gives the order of the MAK, $\zeta = 1, 2$ corresponds to NGP or CIC respectively (Jing, 2005). $k_{\mathrm{Ny}} = \pi/\Delta$ is the one-dimensional Nyquist frequency and Δ the mean spacing on the mesh grid. Mistreatment of the Nyquist frequency would lead to so-called *aliasing* effects.

Finally there is the *variance of the power spectrum* σ_{P}^2. We define the *effective volume* of a galaxy catalogue with real volume V_{survey} (I will discuss surveys and simulations in Chapter 3) as

$$V_{\mathrm{eff}}(k) = \left[\frac{\bar{n}P(k)}{1+\bar{n}P(k)}\right]^2 V_{\mathrm{survey}}\,, \tag{2.70}$$

which is the volume that can be used in clustering analyses in Fourier space at a certain mode k (Feldman, Kaiser & Peacock, 1994). Then the square root of the variance of the power spectrum is defined as

$$\sigma_{\mathrm{P}} = \sqrt{\frac{2\pi^3}{V_{\mathrm{k}} V_{\mathrm{survey}}}} \left(P(k) - \frac{1}{\bar{n}}\right)\,, \tag{2.71}$$

where V_k is the volume of the shell in Fourier space, belonging to mode k. The cosmic variance is the main contribution of the uncertainty of the power spectrum on large scale, described by the first term of Equation 2.71

2.6.4 Power spectrum entering the horizon

I described above how scales get causally disconnected while exponential expansion during inflation. However, as the particle horizon r_{PH} is growing again after inflation, i.e. in the radiation and matter dominated universe, the different scales of the density fluctuations are going to enter the horizon again. We can estimate when a mode enters the horizon and thus predict the rough shape of the power spectrum. A perturbation of scale $\lambda = 2\pi/k$ enters the horizon when its wavelength is equal to the particle horizon

$$\lambda \simeq d_{\mathrm{PH}} = a\frac{c}{H_0}\int_0^{a_{\mathrm{Enter}}} \frac{\mathrm{d}a}{a^2 E(a)}\,. \tag{2.72}$$

In linear theory, we can also predict the evolution of the power spectrum after inflation with

$$P(k,t) = D^2(t) P_0(k)\,, \tag{2.73}$$

with the initial power spectrum $P_0(k)$ and linear growth function $D(t)$. The question arises how $P_0(k)$ may look like. An interesting concept for $P_0(k)$ is the *Zel'dovich-Harrison-Peebles* (Harrison (1970), Zeldovich (1972)) power spectrum, which is called *scale-invariant*. It is written as a power law

$$P_0(k) = Ak^n \,. \tag{2.74}$$

The normalization of the power spectrum A is not predicted by ΛCDM and thus is a free parameter. The power law index n describes how the density fluctuations vary with scale k. Harrison, Zel'dovich, and Peebles argued, that the *spectral index* $n = 1$ would ensure that the peculiar potential fluctuations entering the horizon would be constant. This can be seen when defining the *dimensionless* power spectrum

$$\Delta^2(k) \equiv \frac{1}{2\pi^2} k^3 P(k) \,, \tag{2.75}$$

which is the power in a logarithmic unit length of k. Thus, we write $\Delta_0 \propto k^{3+n}$. Further we define the power spectrum for the potential fluctuations and yield

$$\Delta^2_\Phi(k) \equiv \frac{1}{2\pi^2} k^3 P_\Phi(k) \propto k^{-4} \Delta^2(k) \propto k^{n-1} \,, \tag{2.76}$$

where we used $P_\Phi(k) = k^{-4} P(k)$ (see Equation 2.40). For $n = 1$ we have no preferred scale of the initial power spectrum, this is equivalent with $P(k)k^3 = \mathit{const}$. Impressively, also inflation predicts a spectral index close to unity. The initial power spectrum is modified due to the radiation dominated phase of the Universe. These modifications are described by the *transfer function* $T(k)$. The power spectrum assumes the form

$$P(k,t) \propto k^n T^2(k) D^2(t) \,; \tag{2.77}$$

Eisenstein & Hu (1998) developed a powerful fitting formula to compute the transfer function.

Obviously small scales/wavelengths will enter the horizon first in the radiation dominated era and will grow accordingly with $D(t) \propto a^2(t)$ while larger scales will enter the horizon later while the energy density of matter dominate and therefore grow with $D(t) \propto a(t)$. The growth in the radiation dominated universe is smaller w.r.t. time, as free-streaming and pressure of radiation is impeding structures to collapse as long as radiation is dominating. We define the horizon L_0 at the time t_{eq}, when radiation and matter was equal. Thus, the transfer function is $T \sim (kL_0)^{-2}$ when radiation dominated the Universe and $T \sim 1$ for matter domination. Putting all together, we have an approximate estimate of the linear power spectrum. We expect:

- $P(k) \propto k$ at scales $k \ll \dfrac{1}{L_0}$;
- $P(k) \propto k^{-3}$ at scales $k \gg \dfrac{1}{L_0}$;
- $P(k)$ to have wiggles due to BAO, at the sound horizon s.

The scale-invariant power spectrum is also predicted by inflation as the spectral index n was measured with the Planck satellite to be $n \simeq 0.968 \pm 0.006$.

2.6.5 Random Fields

Random fields are widely used in cosmology to describe statistical properties that we discussed in Section 2.6.2. Particularly important is the *Gaussian random field* (GRF), in which $\delta(\boldsymbol{x})$ or $\delta(\boldsymbol{k})$ is a random variable and $\mathbf{C} = \langle \boldsymbol{\delta}^\dagger \boldsymbol{\delta} \rangle$ the covariance matrix. The probability distribution of a n-dimensional GRF is written as

$$P_n(\delta_1, ..., \delta_n) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \exp\left[-\frac{1}{2}\boldsymbol{\delta}^\dagger \mathbf{C}^{-1} \boldsymbol{\delta}\right] \tag{2.78}$$

A GRF comes with many useful properties as:

- a Fourier transformed GRF remains a GRF;
- a GRF is entirely described by its first two moments, the mean $\boldsymbol{\mu}$ and the covariance $\mathbf{C}$;
- real and imaginary parts of the coefficients of the variables (the density in Fourier space) are independent Gaussian fields;
- the phases of the Fourier modes are randomly distributed.

Although today's density fluctuations are highly non-Gaussian, the CMB and also the LSS at early epochs can be reasonably described with a GRF. We will also discuss density transformations in Section 4.4, in order to gaussianize density fields to extract the information of the linear density field.

For this work also the *lognormal random field* (LNRF) is vital. If a quantity x is distributed lognormally, then $y = \log x$ is distributed according to a Gaussian. Thus we write

$$P_n(y_1, ..., y_n) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \exp\left[-\frac{1}{2}\log(\boldsymbol{y})^\dagger \mathbf{C}^{-1} \log(\boldsymbol{y})\right] \prod \boldsymbol{y}^{-1}\,. \tag{2.79}$$

In Section 2.7.1 we discuss the importance of lognormal solutions for density fields in more detail.

2.6.6 Collapsed objects & density peaks

After scales entered the horizon, an independent growth of scales is not given anymore, hence the scales will be affected by non-linear growth in amplitude and mode coupling, meaning that fluctuations can transfer power from one scale to another. Large-scale fluctuations enter the horizon later than small scales, so we expect non-linearities to have a stronger impact on small-scale fluctuations.

Also, a linear perturbative approach is not able to describe any gravitational collapse nor virialization, two processes that are crucial to describe halo formations. Therefore, it is crucial to know the limit for which perturbative calculations are valid.

One rather intuitive model is to assume a spherical over-density δ which grows independently of the background density (called Birkhoff's theorem in general relativity). Assuming a mass M, the gravitational constant G, and an initial mass perturbation δ_0, we can parametrize the cyclic behaviour of the mass shell as

$$\begin{aligned} r(\theta) &= A(1-\cos\theta)\,, && (2.80)\\ t(\theta) &= B(\theta - \sin\theta)\,, && (2.81)\\ A^3 &= GMB^2\,, && (2.82) \end{aligned}$$

where θ, A and B are parameters of the collapse model. The shell expands from $r = 0$ at $\theta = 0$, reaches its maximum r_max ($t = t_\mathrm{max} = \pi B$) at $\theta = \pi$ and collapses back to $r = 0$ at $\theta = 2\pi$ ($t = 2t_\mathrm{max}$). We Taylor expand these expressions for small θ and combine the equations, finding

$$r \cong \frac{A}{2}\left(\frac{6t}{B}\right)^{\frac{2}{3}}\left[1 - \frac{1}{20}\left(\frac{6t}{B}\right)^{\frac{2}{3}}\right]\,. \tag{2.83}$$

Also, we showed that in linear theory and a matter dominated universe, the scale factor was written as $a(t) = (3/2H_0 t)^{\frac{2}{3}}$. We can match the above equation with linear theory and write for the density contrast in the mass shell

$$\delta \cong \frac{3}{20}\left(\frac{6t}{B}\right)^{\frac{2}{3}} . \tag{2.84}$$

We have assumed CDM cosmology only, neglecting the contribution from Λ. Fortunately, it can be shown that the dependence on Ω_Λ is negligible.

Let us consider a complete collapse, which happens at $\theta = 2\pi$, or $t = t_{\mathrm{tax}} = 2\pi B$. Equation 2.84 then describes the *critical density contrast* of

$$\delta_{\mathrm{C}} = \frac{3}{5}\left(\frac{3\pi}{2}\right)^{\frac{2}{3}} \approx 1.686 . \tag{2.85}$$

Rather a total collapse to $r = 0$, a gravitational system would evolve into a viral equilibrium with the condition $2K = -U$ for the kinetic energy K and potential U. Utilizing this statement we find that $r_{\mathrm{vir}} = 1/2 r_{\mathrm{max}}$ (corresponds to $\theta = 3/2\pi$). Inside this virialized system we compute the density contrast to be $\varrho_{\mathrm{vir}}/\bar{\varrho} \approx 178$ above the mean background density. In Section 3.3.4 we will discuss how halos are identified using these assumptions.

The statistical distribution of massive objects given an underlying dark matter density has been studied in Press & Schechter (1974). Later, also peak background split models have been introduced, aiming to describe statistical properties of density peaks (Bardeen et al. (1986), Mo et al. (1996), Sheth et al. (2001)).

The Press-Schechter formalism postulates that the fraction of mass in objects, that is larger than a certain mass M, is given by the probability that a density exceeds the critical linear density contrast δ_{C}

$$P(>M) = \int_{\delta_{\mathrm{C}}}^{\infty} \mathrm{d}\delta \frac{1}{\sqrt{2\pi\sigma^2(M)}} \exp\left(-\frac{\delta^2}{2\sigma^2(M)}\right) , \tag{2.86}$$

where $\sigma^2(M)$ is the variance in terms of the mass. This formula can be rewritten substituting $\nu = \delta_{\mathrm{C}}/\sigma(M)$, Thus we obtain

$$P(>M) = 1 - \operatorname{erfc}\left(\frac{\nu}{\sqrt{2}}\right) , \tag{2.87}$$

saying that the fraction of mass larger than M is determined by $\sigma(M)$ only. For a given mean density $\bar{\varrho}$, the Press-Schechter formula gives us the number density of objects of a certain mass

$$n(M) = \sqrt{\frac{\pi}{2}} \frac{\bar{\varrho}}{M^2} \left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right| \nu \exp\left(-\frac{\nu^2}{2}\right) , \tag{2.88}$$

also denoted as *mass function*. The pristine Press-Schechter model has been further developed, e.g. by the *peak-background split* (PBS) formalism. In the PBS the density contrast is split into a long-wavelength perturbation for the background, denoted as δ_{b}, and a short-wavelength part δ_{p} for the peaks. The perception of this model is now that the short wavelength perturbations are located on top of the background and describe the density peaks. Therefore, the threshold of collapse is not a fixed δ_{C} anymore, but $\delta_{\mathrm{C}} - \delta_{\mathrm{b}}$. We can now expand Equation 2.88, assuming $n(\nu) \propto \nu \exp(-\nu/2)$:

$$\begin{aligned} n(\nu + \delta\nu) = n\left(\frac{\delta_{\mathrm{C}} - \delta_{\mathrm{b}}}{\sigma}\right) &\approx n(\nu) + \frac{\mathrm{d}n}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\delta}(-\delta_{\mathrm{b}}) , && (2.89) \\ &= n(\nu)\left[1 + \frac{\nu^2 - 1}{\nu\sigma}\delta_{\mathrm{b}}\right] . && (2.90) \end{aligned}$$

We find that

$$\frac{\delta n}{n} = \frac{\nu^2 - 1}{\nu\sigma}\delta_{\mathrm{b}} \equiv b_{\mathrm{L}}\delta_{\mathrm{b}} \,. \tag{2.91}$$

Here, b_{L} is the *Lagrangian bias*. We will discuss this quantity in Section 2.8 in detail. Also, we use a threshold term in our bias model, described in Chapter 4, that is based on the density peak formalism. The PBS formalism has been further developed to yield more precise mass functions (e.g. in Sheth & Tormen (1999) and Tinker et al. (2010)).

2.6.7 Hierarchical clustering

The formation of structures is a result of the gravitational growth of the primordial density- and velocity perturbations. Gravitational attraction in slightly denser regions will be stronger than the average gravitational potential. In these regions the slowdown of the initial cosmic expansion is correspondingly stronger, and when the region is sufficiently overdense, it even comes to a halt, turns around and starts to contract again. If or as long as pressure forces are not sufficient to counteract the infall, the structures will grow without bound, assemble more and more matter by accretion from its surroundings and fully collapse to form a gravitationally bound and virialized object, such as dark matter halos. Due to the lack of pressure, dark matter decouples before radiation and barynonic matter do. Thus, dark matter will likely collapse earlier into virialized structures, providing a potential well in which baryonic matter can stream inside.

Hierarchical clustering implies that the first objects to condense are small and that larger structures form through the merging of smaller structures (Bernardeau & Schaeffer, 1992, 1999; Martínez et al., 2009). Usually an object forms through the accretion of all matter and the fusion of all substructures within its surroundings, including the small-scale objects which had condensed out at an earlier stage. The second fundamental aspect is the anisotropic gravitational collapse, which we will discuss in Section 2.10.

Following Mo et al. (2010), we can link the hierarchical growth of structures in the Einstein-de Sitter universe to the initial power spectrum with a few simple assumptions. Let r be the length scale of a spherical fluctuation in which the mass M is enclosed. We defined the variance of density fluctuations in Equation 2.57. Therefore, the variance grows in time like $\sigma^2(r,t) \propto D^2(t)$. For a fixed time we can then write $\sigma^2(r) \propto r^{-(n+3)}$, using the power law assumption of the power spectrum $P(k) \propto k^n$. Thus, these relations yield

$$\sigma^2(r,t) = \left[\frac{r}{r^*(t)}\right]^{-(n+3)} = \left[\frac{M}{M^*(t)}\right]^{-(n+3)/3} \,, \tag{2.92}$$

where

$$r^*(t) \propto [D(t)]^{2/(n+3)} \quad \text{and} \quad M^*(t) \propto [D(t)]^{6/(n+3)} \,, \tag{2.93}$$

are the mass and length scales for which $\sigma^2 = 1$ at time t. As the spherical collapse model predicts a critical density fluctuations of $\delta_{\mathrm{C}} \approx 1.686$, we can assume that for $\sigma(M,t) \sim \delta_{\mathrm{C}}$ non-linear structures of mass M form. As the growth function $D(t)$ grows with time, so does $M^*(t)$ if $n > -3$, which is the case for a scale-invariant initial power spectrum with $n = 1$. In this case, structure formation develops in a ‘bottom-up’ behaviour, meaning that smaller structures form prior to larger ones, as we discussed above.

2.7 Structure Formation Models

In this Section I discuss the different structure formation models beyond the linear theory. Important models for this work are the lognormal model, derived from the continuity equation (Section 2.7.1), and the Zel'dovich approximation (Section 2.7.2), which is the linear order Lagrangian model.

2.7.1 Lognormal model

Already Hubble found that when galaxies are counted in cells, the resulting galaxy distribution function follows closely a lognormal distribution (Hubble, 1934). In Coles & Jones (1991) the authors found that a lognormal solution also applies to describe structure formation. The continuity equation can be written in terms of conformal time $\mathrm{d}\tau = \frac{\mathrm{d}t}{a}$ and the convective derivative $\frac{\mathrm{d}}{\mathrm{d}\tau} = \frac{\partial}{\partial\tau} + \boldsymbol{u}\cdot\boldsymbol{\nabla}$, as

$$\begin{aligned} \frac{1}{\varrho}\frac{\mathrm{d}\varrho}{\mathrm{d}\tau} &= -\boldsymbol{\nabla}\cdot\boldsymbol{u}\,, && (2.94)\\ \int\frac{1}{\varrho}\mathrm{d}\varrho &= -\int\mathrm{d}\tau\boldsymbol{\nabla}\cdot\boldsymbol{u}\,. && (2.95)\end{aligned}$$

This integral gives a lognormal solution for the density contrast if the divergence of the velocity field is Gaussian distributed, as

$$\varrho(\boldsymbol{x}) = \varrho_0\exp\left(\epsilon(\boldsymbol{x})\tau^2\right), \tag{2.96}$$

where $\epsilon(\boldsymbol{x}) = 1/(2\tau_0)(\boldsymbol{\nabla}\cdot\boldsymbol{u})$ is a Gaussian random field. We will discuss the lognormal model for structure formation in detail in Sections 4.4 and 4.6.3, where also the constant of integration is taken in account, that has been neglected in the original work of Coles & Jones (1991).

2.7.2 Zel'dovich Approximation

The Zel'dovich approximation (Zel'dovich, 1970) is describing a fluid element's trajectory by the initial Lagrangian position $\boldsymbol{q}$, the comoving Eulerian position $\boldsymbol{x}(\boldsymbol{q},t)$ and the displacement field $\boldsymbol{\Psi}(\boldsymbol{q},t)$. At the initial time $t=t_0$ the displacement writes $\boldsymbol{\Psi}(\boldsymbol{q},t_0)=0$. Every element of the fluid is uniquely labeled by its Lagrangian coordinate $\boldsymbol{q}$ and mapped to the final position in Eulerian coordinates $\boldsymbol{x}$ according to the displacement field $\boldsymbol{\Psi}$.

The mapping is written as

$$\boldsymbol{x}(\boldsymbol{q},t) = \boldsymbol{q}(t_0) + \boldsymbol{\Psi}(\boldsymbol{q},t)\,. \tag{2.97}$$

The displacement field now can be expanded in a perturbation series

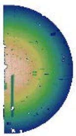
$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^1 + \boldsymbol{\Psi}^2 + \ldots + \boldsymbol{\Psi}^N\,. \tag{2.98}$$

The Zel'dovich approximation further assumes that the series can be truncated at linear order, $\boldsymbol{\Psi}^1 \simeq D(t)\boldsymbol{\nabla}\Phi(\boldsymbol{q})$. The conservation of mass allows us to write

$$\varrho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x} = \varrho(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\,. \tag{2.99}$$

In a good approximation we can neglect the density fluctuations in Lagrangian space, thus obtaining

$$\begin{aligned} 1+\delta(\boldsymbol{x},t) &= \mathbf{J}^{-1}(\boldsymbol{q},t)\,, && (2.100)\\ \left|\frac{\partial\boldsymbol{x}}{\partial\boldsymbol{q}}\right| &= \mathbf{J}(\boldsymbol{q},t)\,, && (2.101)\end{aligned}$$



where $\mathbf{J}$ is the Jacobian of the coordinate transformation, also called *deformation tensor* $\mathscr{D}_{ij}$. The solution of Equation 2.100 can be written in terms of eigenvalues $\lambda_{1,2,3}$ of the Jacobian as

$$1+\delta(\boldsymbol{q},t)=\frac{1}{(1-D(t)\lambda_1(\boldsymbol{q}))(1-D(t)\lambda_2(\boldsymbol{q}))(1-D(t)\lambda_3(\boldsymbol{q}))}\,, \tag{2.102}$$

where the eigenvalues are sorted as $\lambda_1 > \lambda_2 > \lambda_3$. Although intrinsically non-linear, the Zel'dovich approximation is not able to describe virialization or shell-crossing. The displacement of the fluid elements is just determined by the velocity $\boldsymbol{v} \propto \boldsymbol{\nabla}\Phi$ in Lagrangian coordinates. A term which is commonly used in literature is the so-called *Zel'dovich pancake*. This term is related to the model how cosmic structures are forming in the Zel'dovich approximation. According to Equation 2.102, a spherical overdense region will likely start collapsing in direction of the first principle axis (corresponding to the λ_1). Thus, after the first collapse we will obtain a pancake-like structure. I will discuss this formalism in more detail in Section 2.10.

2.8 Galaxy Bias

The term *galaxy bias*, introduced by Kaiser (1984), comprises the fact, that observations target the spatial distribution of luminous galaxies and not directly the total masses. Galaxies (and halos) are tracers of the underlying dark matter distribution and therefore connected by the bias function. Ideally, this relation must be modelled by the gravitational collapse of dark matter into a halo, the accumulation of baryonic gas inside the potential well of the halo, and finally the cooling of the gas and fragmentation to a galactic structure. Rather than modelling all these stages of galaxy formation, the concept of bias is a method to phenomenologically describe the statistical influence of the above mentioned processes (Bardeen et al., 1986; Mo & White, 1996; Mo et al., 1997). So in general, the bias is an effective theory to relate the distribution of galaxies to the underlying dark matter.

In most general form we can write the relation of the density field for galaxies ϱ_{G} (alternatively the density contrast of galaxies δ_{G}) and dark matter fluctuations δ_{DM} as

$$\varrho_{\mathrm{G}}=\mathscr{B}(\delta_{\mathrm{DM}})\,, \tag{2.103}$$

where $\mathscr{B}(\delta_{\mathrm{DM}})$ is a non-linear, non-local and stochastic function. Modelling the galaxy bias is not only necessary to obtain unbiased cosmological information about the underlying dark matter, but it is also useful to understand the physics of galaxy formation better.

Over the time many bias models have been formulated in the cosmological context. I will discuss a few different concepts in the following, which are part of our reconstruction algorithm.

2.8.1 Deterministic Bias

On large scales, much larger than the typical scales of galaxy formation, a linear local bias model is a good approximation (Matarrese et al., 1997). So, we write

$$\delta_{\mathrm{G}} = b\delta_{\mathrm{DM}}\,, \tag{2.104}$$

$$P_{\mathrm{G}}(k) = b^2 P_{\mathrm{DM}}(k)\,. \tag{2.105}$$

This linear model was initially proposed by Kaiser (1984, 1987). Although linear bias is valid for large scales, the model breaks down at scales that are not described by linear theory. Also, the linear model is scale-independent. However, we expect different scales to undergo different cosmic growth. This is why a scale-dependent bias model is desirable. Particularly in Fry & Gaztanaga (1993), the authors suggest a series expansion of the galaxy density field:

$$\delta_{\mathrm{G}} = \mathscr{B}(\delta_{\mathrm{DM}}) = \sum_{k=0}^{\infty} \frac{b_k}{k!} \delta_{\mathrm{DM}}^k \,, \tag{2.106}$$

and linked each order of the bias factor b_k (but b_0) to an order of the density contrast. However this model may lead to unphysical densities if the first order b_1 is negative. Cen & Ostriker (1993) on the other hand suggest to expand the logarithm of the density.

Based on this work, Mann et al. (1998) analyzed different bias relations in Eulerian space, showing that the linear bias relation for the initial Gaussian density field under a lognormal transformation leads to a deterministic, scale-dependent power law bias relation. Also Frusciante & Sheth (2012) studied a formulation of a linear bias in Lagrangian space undergoing lognormal evolution. They found that this bias description needs an additional normalization to ensure that $\langle \delta_{\mathrm{DM}} \rangle = 0$, and therefore suggested a renormalized model. In the following we will use a scale-dependent bias as

$$\varrho_{\mathrm{G}} \approx \exp\left(\sum_{k=0}^{\infty} b_k \cdot \log(1+\delta_k) \right) \,, \tag{2.107}$$

$$\varrho_{\mathrm{G}} \approx (1+\delta_{\mathrm{DM}})^b \,, \tag{2.108}$$

for the deterministic part of our bias model for the rest of this work.

2.8.2 Stochastic Bias

Basically, the assumption of a stochastic bias is based on the interpretation that the underlying dark matter density field is a multivariate probability density of which the galaxy density fields are statistically random realizations. Already in Peebles (1993, p.500-512), the author found that the variance of the density contrast δ is different from a Poisson process when discussing the Layzer-Irvine Equation (Irvine, 1961; Layzer, 1963). Because of this finding the stochasticity of the galaxy bias shifted into focus.

The stochastic relation of the galaxy and dark matter density was firstly studied in Dekel & Lahav (1999). In this work the authors state that the biasing relation can be interpreted as a random process, specified by the *conditional probability distribution* $\mathscr{P}(\delta_{\mathrm{G}}|\delta_{\mathrm{DM}})$ of a galaxy density, given the dark matter density. Thus, the joint probability of δ_{G} and δ_{DM} can be expressed with conditional one and the prior one-point probability of δ_{DM} as

$$\mathscr{P}(\delta_{\mathrm{G}}, \delta_{\mathrm{DM}}) = \mathscr{P}(\delta_{\mathrm{G}}|\delta_{\mathrm{DM}}) \mathscr{P}(\delta_{\mathrm{DM}}) \,. \tag{2.109}$$

Further, they introduce the *biasing scatter function* $\sigma_{\mathrm{b}}(\delta_{\mathrm{DM}})$, as

$$\sigma_{\mathrm{b}}^2(\langle \delta_{\mathrm{DM}} \rangle) \equiv \langle \epsilon^2 | \delta_{\mathrm{DM}} \rangle \,, \tag{2.110}$$

where ϵ is the *random biasing field*. The *biasing scatter* is defined as the δ-average over the biasing scatter function

$$\sigma_{\mathrm{b}}^2 = \langle \epsilon^2 \rangle \,. \tag{2.111}$$

Dekel & Lahav (1999) find a deviation from the linear deterministic model up to 20 % in their studies. The stochastic bias in context of higher order clustering was analyzed in Scoccimarro (2000) and also was further tested with N-body simulations in Somerville et al. (2001). Interestingly, the authors found that for very massive halos beyond a redshift dependent masss scale, the correlation function turns over at small distances of about $r \sim 1\ h^{-1}\,\mathrm{Mpc}$ due to so-called *exclusion effects* caused by the finite spatial extent of the halos. This means, that a massive halo is likely to have accreted all surrounding matter and thus, the probability to find another halo at typical galactic scales is lower then for a random distribution. Therefore, an anti-correlation is expected at these scales for very massive halos. Together with the decreasing correlation also the biasing scatter (shown in Equation 2.110) will be reduced and the stochastic bias will be less important or can even lead to a smaller scatter compared to a random Poisson process. This prediction agrees in general with our findings, which I will discuss in Chapter 5. Stochastic bias models also have been analyzed in toy models (Baldauf et al., 2013). In all cases (but for extreme cases where halo exclusion is dominating), stochastic bias leads to higher variance of the density field as compared to a pure deterministic bias model.

In this work the stochastic bias is modelled by taking distribution functions into account, that have a larger variance as compared to the Poisson distribution but the same mean, discussed in detail in Section 4.6.2.

2.8.3 Non-local Bias

Besides deterministic and stochastic bias, also non-local contributions are studied in literature, connecting environmental dependencies, described e.b. by the deformation tensor of the density field to the total biasing relation (Baldauf et al., 2012; Saito et al., 2014). I will discuss this issue in Section 2.10, where I introduce the tidal field tensor, which is necessary to follow up environmental studies of the density field.

2.9 Redshift-Space Distortions

The observed redshifts of galaxies are a composition of the Hubble flow and the projection of the peculiar velocity in line-of-sight direction, which can be written as

$$\boldsymbol{s} = \boldsymbol{r} + \frac{\boldsymbol{v}_{\mathrm{p}} \cdot \hat{\boldsymbol{r}}}{aH(a)}\hat{\boldsymbol{r}}\,, \tag{2.112}$$

where $\boldsymbol{s}$ and $\boldsymbol{r}$ are the comoving positions of the galaxy in *redshift-space* and *real-space* respectively, and $\hat{\boldsymbol{r}}$ the unit vector in radial direction. I will use the terms *redshift-space* and *real-space* to stress the mapping from $\boldsymbol{s}$ to $\boldsymbol{r}$ throughout this thesis. The resulting apparent displacement in line-of-sight direction of objects is called *redshift-space distortion* (RSD). The peculiar velocity $\boldsymbol{v}_{\mathrm{p}}$ is generally divided into two parts (Hamilton, 1998). These are the perturbative part (we will concentrate on the linear terms, identical to the bulk flows), discussed in Section 2.9.1 and quasi-virialized or dispersed velocities due to collapsed objects called fingers-of-god (FOG), described in Section 2.9.2.

RDSs can be used to constrain the nature of gravity and cosmological parameters as the peculiar velocity is depending linearly on the growth rate f_Ω (see e.g. Berlind et al., 2001; McDonald & Seljak, 2009; Nesseris & Perivolaropoulos, 2008; Percival & White, 2009; Song et al., 2011; White et al., 2009, for recent studies). Common techniques to measure RSDs are presented in (Beutler et al., 2014; Blake et al., 2013; Chuang & Wang, 2013; Cole et al., 1995; Jennings et al., 2011; Kwan et al., 2012; Okumura et al., 2008, 2012; Peacock et al., 2001; Percival et al., 2004; Reid et al., 2012; Samushia et al., 2012, 2013; Wang, 2014; da Ângela et al., 2008; de la Torre et al., 2013).

2.9.1 Bulk flows

Firstly we describe the *coherent bulk flow*, which is the velocity due to the potential perturbations that dominates at large scales. We repeat Equation 2.52

$$\boldsymbol{\nabla} \cdot \boldsymbol{v}_{\mathrm{p}} = -aHf_{\Omega}\delta\,. \tag{2.113}$$

To solve this equation, we apply $\boldsymbol{\nabla}$ on both sides and yield

$$\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{v}_{\mathrm{p}}) = -aHf_{\Omega}\boldsymbol{\nabla}\delta\,. \tag{2.114}$$

In the linear limit, the peculiar velocity fields are curl free, as primordial curls would have been dissolved with the expansion of the Universe with $\boldsymbol{\nabla} \times \boldsymbol{v}_{\mathrm{p}} \propto 1/a$. Thus we write

$$\boldsymbol{v}_{\mathrm{p}} = -aHf_{\Omega}\boldsymbol{\nabla}\nabla^{-2}\delta\,, \tag{2.115}$$

and commuted the operators $\boldsymbol{\nabla}, \nabla^{-2}$.

In the context of linear theory, Kaiser (1987) gave a solution to the power spectrum under the distortions of bulk flows. We start with the conservation of the number of galaxies in redshift-space and real-space. Therefore we write

$$n_s(\boldsymbol{s})\mathrm{d}^3s = n(\boldsymbol{r})\mathrm{d}^3r \tag{2.116}$$

with n_s being the number density of objects in redshift-space and n the corresponding number density in real-space in an infinitesimal volume element d^3s and d^3r.

We can relate these number densities via

$$\begin{aligned} n_s(\boldsymbol{s})\mathrm{d}^3s &= n(\boldsymbol{r})\mathrm{d}^3r && (2.117)\\ n_s(\boldsymbol{s}) = n(\boldsymbol{r})\mathbf{J} \qquad\qquad J &= \left|\frac{\mathrm{d}^3r}{\mathrm{d}^3s}\right| = \frac{\mathrm{d}r}{\mathrm{d}s}\frac{r^2}{s^2}\,. && (2.118)\end{aligned}$$

The Jacobian $\mathbf{J}$ describes the coordinate transformation and can be written as a scalar J of the radial components of $\boldsymbol{s}$ and $\boldsymbol{r}$, as the angular components are not affected by RSDs. Now we apply Equation 2.112 to the Jacobian and obtain

$$J \simeq \left(1 - \frac{\partial}{\partial r}\left(\frac{\boldsymbol{v}_{\mathrm{p}} \cdot \hat{\boldsymbol{r}}}{Ha}\right)\right)\,. \tag{2.119}$$

If we further assume the same mean number $\bar{n}$ in real- and redshift-space, we can write

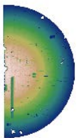
$$\delta_s = \delta\left(1 - \frac{\partial}{\partial r}\left(\frac{\boldsymbol{v}_{\mathrm{p}} \cdot \hat{\boldsymbol{r}}}{Ha}\right)\right)\,, \tag{2.120}$$

where δ_s and δ are the density contrasts in redshift-space and real-space, respectively. We can use now the approximation of a distant observer, setting the product $\hat{\boldsymbol{r}} \cdot \boldsymbol{v} = \hat{\boldsymbol{z}} \cdot \boldsymbol{v}$. This approximation assigns the coordinate $\hat{\boldsymbol{z}}$ to the line-of-sight direction, being a good approximation for very distant galaxies. Finally we can give an expression of the density in Fourier space as

$$\delta_s(\boldsymbol{k}) = \int \mathrm{d}^3r\mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\left(\frac{\partial}{\partial r}\left(\frac{\boldsymbol{v}(\boldsymbol{r}) \cdot \hat{\boldsymbol{z}}}{Ha}\right)\right)\,. \tag{2.121}$$

The integral is solved using the distant observer approximation and thus writes as

$$\delta_s(\boldsymbol{k}) = \left[1 + f_{\Omega}\mu^2\right]\delta(\boldsymbol{k})\,, \tag{2.122}$$



where μ is the cosine of the angle of the wave vector $\boldsymbol{k}$ and the line of sight vector; f_Ω is the growth rate. We can immediately see that as $\mu^2 f_\Omega \geq 0$, the density contrast in redshift-space (on scales where bulk flows dominate) is apparently enhanced. The power spectrum is now written as

$$P_s(\boldsymbol{k}) = P(k)(1 + \beta\mu^2)^2 \,, \tag{2.123}$$

where we introduced $\beta = f_\Omega / b$, with the linear galaxy bias factor b. We can see that due to RSDs the power spectrum is not isotropic anymore. One way to estimate the difference of the power spectra in real and redshift-space is done by calculating the ratio of the isotropic power spectrum, called monopole, and an anisotropic power spectrum, called quadrupole (at lowest anisotropic order). We use the Legendre polynomials of the order l, $L_l(\mu)$, where μ is again the cosine of the angle between the separation $\boldsymbol{k}$ and the line-of-sight direction. $l = 0 \to L_0 = 1$, is used for the monopole whereas $l = 2 \to L_2 = (3\mu^2 - 1)/2$ is the second order term for the quadrupole. The result assumes the form

$$\begin{aligned} P_s^0(k) &= \left[1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right] P(k) \,, & (2.124) \\ P_s^0(k) &= K(f_\Omega, b) P(k) \,, & (2.125) \end{aligned}$$

where K is called Kaiser factor/boost in literature. The equation tells us that by blindly calculating the monopole power spectrum in redshift-space, the power will be boosted by the Kaiser factor. The intuitive explanation is as follows: as the galaxy velocities within clusters are mostly directed to the center of a denser regions, the circumference of this cluster will appear to be squashed in line-of-side direction in redshift-space. Thus, we will gain more power in Fourier space from this region due to RSDs.

2.9.2 Fingers-of-God

The Fingers-of-god effect (FOG) was first described in Jackson (1972), where a radial arrangement of galaxies pointing towards the earth, were observed. FOG are due to the random motion in quasi-virialized objects so that the structures appear to be elongated in redshift-space in line-of-sight direction. This is a non-linear effect that we can solve statistically. FOG become relevant on smaller non-linear scales (see e.g. Reid et al., 2014) in the power spectrum, where we observe a decrease of power.

2.10 Cosmic web classifications

The cosmic web is an expression to characterize the structures on large scales when analyzing the matter distribution in the Universe. As discussed above the coherent part of the peculiar velocity can be determined from the potential as $\boldsymbol{v} \propto \boldsymbol{\nabla}\Phi$. Utilizing this relation, we can estimate the velocity divergence at a given position within the cosmic web. This procedure reveals environmental information of the cosmic structures at this position. In literature, environmental effects on galaxy properties are studied in detail (see e.g. Kauffmann et al., 2004).

Aspherical overdensities, on any scale and in any scenario, will contract such that they become increasingly anisotropic. At first they turn into a flattened pancake, rapidly followed by contraction into an elongated filament and possibly, total collapse into a galaxy or a cluster may follows. It is evidently the major agent in shaping the web-like cosmic geometry.

To classify the structures of the cosmic web, we follow the description in Hahn et al. (2007). We define the *tidal field tensor* at a position $\boldsymbol{x}$ as

$$T_{ij}(\boldsymbol{x}) \equiv \frac{\partial^2 \Phi}{\partial x_i \partial x_j} \,, \tag{2.126}$$

with the peculiar gravitational potential $\Phi = 4\pi G\bar{\varrho}\nabla^{-2}\delta$ and i, j denoting the three coordinates x, y, z. The three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of T_{ij} define the dynamic behaviour of the particular structure. In agreement with Zel'dovich (1970) one can define:

1. *voids:*
   $T_{ij}(\boldsymbol{x})$ has no positive eigenvalues. All components of the peculiar velocity are directed outwards,

2. *sheets:*
   $T_{ij}(\boldsymbol{x})$ has one positive and two negative eigenvalues. This is the famous Zel'dovich pancake, meaning that in one direction the peculiar velocity is pointing inward the structure,

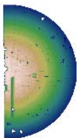
3. *filaments:*
   $T_{ij}(\boldsymbol{x})$ has two positive and one negative eigenvalues. We can picture this structure as a tube like configuration, where the peculiar velocity is streaming in perpendicular to the filament and in one direction outwards, along the filament,

4. *knots:*
   $T_{ij}(\boldsymbol{x})$ has three positive eigenvalues, all velocity components direct inside the structure.

This method is also called the T-web analysis (according to the tidal field tensor). A comparable approach is the *div V-web* study, where the density field is obtained from the divergence of the velocity field as

$$\delta_{\mathrm{v}} = -\frac{\boldsymbol{\nabla} \cdot \boldsymbol{v}_{\mathrm{p}}}{f_{\Omega} H a}\,, \tag{2.127}$$

and the potential is calculated from δ_{v}. These kinds of analyses not only help to visually trace the structure-forming regions within the cosmic web, but also allow us to use the statistical properties of the structures. Further analyses of Forero-Romero et al. (2009) and Nuza et al. (2014) showed that the threshold for a chosen structure must not be necessarily at $\lambda_{\mathrm{th}} = 0$ and already at $\lambda_{\mathrm{th}} = 0.1$ the classification changed dramatically, putting emphasis to carefully choose the threshold scales within each analysis. We are relying on the characterization of cosmic structures within our reconstruction method to e.g. assign galaxies that are distorted by FOG to high density regions, i.e. knots. The description is outlined in Section 5.3. Cosmic web classifications are therefore beneficial to achieve a higher goodness of our reconstruction method.



3

Galaxy Redshift Surveys & N-body Simulations

In this chapter, I will describe the different sources of data that have been used as inputs for our reconstruction method. In Section 3.1 I briefly describe the galaxy survey named Sloan Digital Sky Survey (SDSS), and the particular galaxy redshift catalogue derived from this survey. In Section 3.3 I also describe the N-body simulations on which the mock galaxy data is based on, used in the following analyses. I highlight the Baryon Oscillation Spectroscopic Survey (BOSS) and the CMASS (for Constant (stellar) Mass) galaxy catalogue in detail in Section 3.1.1.

3.1 Sloan Digital Sky Survey

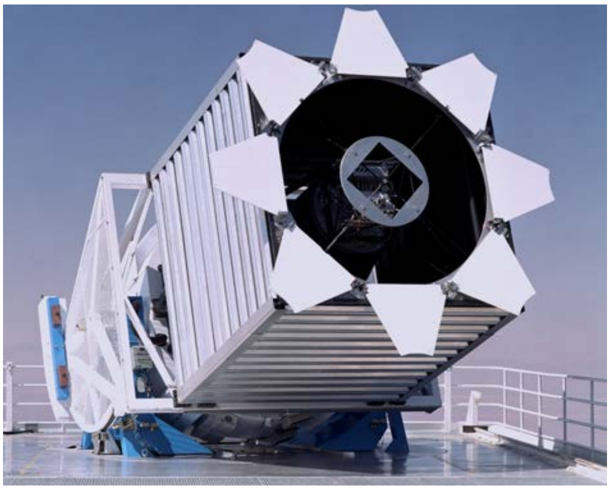
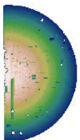


Figure 3.1: Picture of the SDSS 2.5-meter telescope at the Apache Point Observatory, New Mexico (SDSS Collaboration, 2000).

Starting from the year 2000 till 2014 the SDSS collaboration observed in total a quarter of the sky corresponding to about 10000 deg^2 including photometric observations of around 500 million objects of which about 3 million objects have been spectroscopically observed. Up to 2014, the SDSS has been operating divided into three survey periods, namely SDSS-I: 2000-2005, SDSS-II: 2005-2008 and SDSS-III: 2008 till 2014. In this time, 12 Data Releases (DR) have been made public. Since 2014, SDSS-IV is operating.

The photometric system of the SDSS telescope (Gunn et al., 1998) uses a CCD camera which is composed of an array of 30 CCDs arranged in six columns of five CCDs each. The telescope moves along great circles on the sky so that images of objects move along the columns of the CCDs at the same rate. Therefore, the camera produces five images of a given object, all from the same column of CCDs, one from each CCD in that column. Each row corresponds to a different filter, so each object has one image in each filter. SDSS utilizes 5 different filters, called u, g, r, i, z bands (Fukugita et al., 1996).

The operating (mean) wavelengths, magnitude limits, and calibration errors (RMS) of the filters are given in table 3.1



Filter	Wavelength [nm]	Magnitude limit	RMS [%]
u	355.1	22	1.3
g	468.6	22.2	0.8
r	616.5	22.2	0.8
i	748.2	21.3	0.7
z	893.1	20.5	0.8

Table 3.1: Photometric filter wavelengths, operating magnitudes, and accuracy for the SDSS CCD camera. Calibration accuracy (RMS) is described in detail in Padmanabhan et al. (2008).

3.1.1 Baryon Oscillation Spectroscopic Survey

Additionally to the imaging system, SDSS-III also hosts three spectroscopic surveys that are focused on different scientific goals[1]. The APOGEE survey focuses on the structure and evolution of our own Milky Way galaxy using high-resolution infrared spectroscopy while the SEGUE survey was build to study the kinematics and stellar populations at the high-latitude thick disk and halo of the Milky Way. This work used the data taken by the Baryon Oscillation Spectroscopic Survey (BOSS) spectroscopic survey. BOSS observes the spectra of its targets using the double-armed BOSS spectrograph (Smee et al., 2013). The algorithms that are used to perform a spectral classification and the redshift measurements are described in Bolton et al. (2012).

Different from the imaging system, spectroscopy relies on the usage of fibers. In BOSS these fiber positions are stamped in plates so that a fiber of 2 arcsec diameter can be placed on top (up to 1000 fibers per plate). The positions of the fibers are chosen by a primary target selection of the imaging system based on two colors, $g - r$ and $r - i$. The fibers then are read out with CCDs with extended wavelength coverage to span 360 to 1000 nm.

3.1.2 CMASS galaxy catalogue

While the SDSS-I and SDSS-II is purely magnitude limited, CMASS and LOWZ employ color cuts in order to target more distant galaxies. These cuts are designed to obtain a sample with approximately *constant stellar mass* (Reid et al., 2016). The galaxy sample is split in a high-redshift sample called CMASS (between $0.43 \leq z \leq 0.7$) and a low-redshift sample called LOWZ ($z \leq 0.43$) which have different magnitude limits (see Anderson et al. (2014) for details of the targeting strategy). The combined sample of LOWZ and CMASS is discussed and has been analyzed in Alam et al. (2016).

In this thesis I use the North Galactic Cap (NGC) of the CMASS data from the twelfth Data Release (DR12) (Alam et al., 2015) of the SDSS-III BOSS catalogue (Dawson et al., 2013) which is focused on the observation of distant luminous red galaxies (LRGs) and thus creating a three dimensional spatial distribution of the large-scale structure of the Universe.

The CMASS sample consists mainly of LRGs, constant in mass and volume limited. The target selection of the CMASS sample together with the algorithms used to create large scale structure catalogues (the MKSAMPLE code), are presented in Reid et al. (2016) and also previously summarized in Eisenstein et al. (2011).

---

[1] `http://www.sdss.org/surveys/`

3.2 Survey geometry

The geometry of a galaxy redshift survey is a tedious task to deal with as the observing strategies of modern galaxy surveys typically produce angular masks with complex boundaries and variable completeness (Hamilton & Tegmark, 2004).

Completeness in terms of survey geometry denotes the ratio of galaxies that have been observed photometrically and spectroscopically. We use the software package MANGLE (Hamilton & Tegmark, 2004; Swanson et al., 2008) to geometrically deal with the completeness. The fundamental idea is to resolve a mask into a union of non-overlapping polygons each of whose edges is part of a circle (not necessarily a great circle) on the celestial sphere. Each polygon then defines an area on the sky with a completeness value between 0 and 1. MANGLE reads in completeness files provided by the SDSS collaboration. These files contain patches of the sky that have been observed or vetoed, i.e. positions that are excluded from the mask, e.g. due to bright stars or fiber collisions.

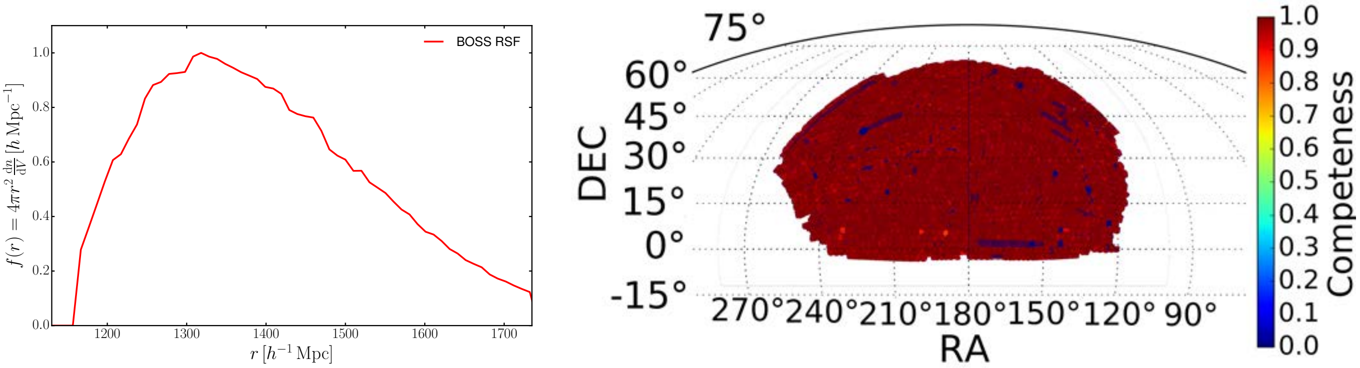


Figure 3.2: Left: Radial selection function (RSF) $f(r) = 4\pi r^2 \,\mathrm{d}n/\mathrm{d}V$. Right: Angular completeness mask of the North Atlantic cap of the BOSS survey projected on a sphere. The color code corresponds to the completeness at each point on the sky. Right ascension RA and declination DEC are shown in the range of the analysed BOSS data.

As the survey operates in a certain color/magnitude range (see Section 3.1.1), we also have to take into account the so-called radial selection function (RSF) $f(r)$. The RSF represents the galaxy counts per spherical shell as function of comoving distance r (or redshift z). In Figure 3.2 on the left the RSF is shown for the DR12 CMASS galaxy sample. The distances correspond to the redshifts between $0.43 \leq z \leq 0.7$, exactly the range of the CMASS sample. The particular shape of the RSF is due to the targeting strategy of the CMASS catalogue, applying magnitude and color cuts to trace predominantly LRGs as described in Sections 3.1.1 and 3.1.2.

On the right of Figure 3.2 we can see the angular completeness mask projected on the sky for the North Galactic Cap (NGC). This figure was created by applying a high-resolution mesh in right ascension - declination plane (RA (α)-DEC(δ)) of 350 million (19000^2) points in the range of the NGC.

Each polygon is allocated with a completeness value. We apply our 350 million points of our mesh on top of the polygons and screen for each point the completeness value with MANGLE. The different completeness values can be seen in Figure 3.2, where also the polygon-like structure of the mask is visible. Areas that have zero completeness within the observed region can either be caused by not having been observed, fiber collisions on the plate or bright stars in front of the targeted galaxy, that spoiled the spectroscopy in these regions. These effects are all incorporated into the angular mask.

In Sections 4.5.4 and 4.5.5, I describe the handling of the survey geometry and the observational galaxy weights in our reconstruction algorithm.

3.3 Dark Matter Simulations

In this section the main cosmological simulations that have been used for the reconstruction analyses are briefly described. I will focus on dark matter only N-body simulations in the following.

3.3.1 Fundamentals of N-body Simulations

N-body solvers in simulations are used to sample the dynamics of the Universe. Dark matter particles are collisionless, so the evolution of the Universe is driven by the mean potential rather than two-body interactions of N dark matter particles. Therefore it is convenient to describe the phase-space distribution of the particles within the simulation by a distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$ so that $f(\boldsymbol{x}, \boldsymbol{p}, t)\mathrm{d}^3x\mathrm{d}^3p$ represents the probability of a particle to be found in the volume between V and $V + \mathrm{d}V$, with a momentum range of $\boldsymbol{p}$ and $\boldsymbol{p} + \mathrm{d}\boldsymbol{p}$. This is a $6N + 1$ dimensional equation for which we can define the continuity equation in the collisionless limit. Therefore, we define the phase-space derivative $\boldsymbol{\nabla}_{\mathrm{ph}} = (\boldsymbol{\nabla}_{\boldsymbol{x}}, \boldsymbol{\nabla}_{\boldsymbol{p}})$ and phase-space velocity $\boldsymbol{v}_{\mathrm{ph}} = (\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{p}})$, which are the derivatives of the phase-space positions w.r.t. time. The continuity equation then writes as

$$\begin{aligned}\frac{\partial f}{\partial t} + \boldsymbol{\nabla}_{\mathrm{ph}}(f\boldsymbol{v}_{\mathrm{ph}}) &= 0\,, &(3.1)\\ \frac{\partial f}{\partial t} + \boldsymbol{\nabla}_{\boldsymbol{x}} \cdot (f\boldsymbol{v}_{\boldsymbol{x}}) + \boldsymbol{\nabla}_{\boldsymbol{p}} \cdot (f\boldsymbol{v}_{\boldsymbol{p}}) &= 0\,. &(3.2)\end{aligned}$$

As most cosmological simulations rely on a Newtonian description of gravity, we can thus specify the time derivatives of our phase-space vector $(\boldsymbol{x}, \boldsymbol{p})$ as

$$\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\right) = (\boldsymbol{v}, -\boldsymbol{\nabla}\Phi(\boldsymbol{x}, t))\;, \tag{3.3}$$

where $\Phi(\boldsymbol{x}, t)$ is the potential function satisfying the Poisson equation of a given matter density $\varrho(\boldsymbol{x}, t)$

$$\nabla^2\Phi(\boldsymbol{x}, t) = 4\pi G\varrho(\boldsymbol{x}, t) = 4\pi G\int f(\boldsymbol{x}, \boldsymbol{v}, t)\mathrm{d}^3v\,. \tag{3.4}$$

Combining now Equation 3.1 with 3.3 we obtain the collisionless Boltzmann equation (CBE)

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} f - \boldsymbol{\nabla}\Phi \cdot \boldsymbol{\nabla}_{\boldsymbol{v}} f = 0\,. \tag{3.5}$$

Equation 3.5 fulfills $\mathrm{d}f/\mathrm{d}t = 0$, which means that the local phase-space density around a particle is constant and does not change along the particles' trajectories. This is also known as *Lioville's theorem.*

Equation 3.5 has to be fulfilled by all N particles simultaneously. In other words, for any time t the CBE must be solved by a set of phase-space trajectories $(\boldsymbol{x}(t), \boldsymbol{v}(t))$. These solutions can be found by setting up an Initial Condition (IC) of the phase-space density and evolving the system of particles according to Equation 3.3. So a N-body simulation can be dealt with as Monte Carlo mapping of a probability function $f(\boldsymbol{x}, \boldsymbol{v}, t)$ with N tracers for different time steps.

The matter density $\varrho(\boldsymbol{x}, t)$ of a particle with mass m is described by a Dirac delta function

$$\varrho(\boldsymbol{x}) = \sum_j m_j\delta(\boldsymbol{x} - \boldsymbol{x}_j)\,. \tag{3.6}$$

In order to avoid divergences, namely for two particles to be at zero separation, a smoothing kernel $W(\boldsymbol{x} - \boldsymbol{x'}, \epsilon)$ (typically a Plummer sphere) needs to be applied the particle density, so that the force is softened at a certain length scale ϵ. The smoothing guarantees, that the force

$$\boldsymbol{F}(\boldsymbol{x}_i) \propto \sum_{j \neq i} \frac{1}{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2 + \epsilon^2} \tag{3.7}$$

on particle i has a minimum separation of ϵ instead of the pure inverse square law that would diverge at close encounter. The time evolution is done numerically with the leapfrog scheme described in Section 4.3.3.

During the years different methods have been developed, that are describing how to sum the particles' potential or the force acting on it respectively. The very straight forward method is the direct summation of all particles' contributions at each point $\boldsymbol{x}$ in space. However this method is not feasible for large N-body simulations as the computing time scales like $\mathscr{O}(N^2)$. Therefore, I will discuss briefly methods that speed up the computation time.

- **Tree codes**
  In order to reduce the number of particle-particle summations, Appel (1985) and later in Astrophysics Barnes & Hut (1986) suggested not to calculate all direct connections between two particles but sum up many clustered particles if the *Multipole-Acceptance-Criterion* (MAC) is fulfilled in a clustering and connect them in a tree like structure. The MAC states that particles are encapsulated to a single point if the distance is much larger compared to the diameter of the cluster. The MAC is a free parameter.

  This means that the structure is approximated by a point source for another distant structure and so, all particles within this structure do not need to be summed up separately. This reduces the number of computing steps to $\mathscr{O}(N \log N)$.

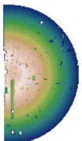
- **Particle-Mesh**
  For this method a grid is defined on top of the volume so that the particles are assigned to a particular cell. Now the Poisson equation is solved on a grid using Fast Fourier Transformations (FFTs) as

  $$\Delta\Phi = 4\pi G \varrho \tag{3.8}$$
  $$\hat{\Phi} = -4\pi G \frac{\hat{\varrho}}{k^2}\,, \tag{3.9}$$

  where $\boldsymbol{k}$ is the wave vector and the variables in Fourier space are denoted with the $\hat{}$-symbol.

- **TreePM**
  This method replaces the pure tree ansatz by a hybrid method consisting of a synthesis of the particle-mesh method and the tree algorithm. The potential of Equation 3.8 is split in Fourier space into a long-range and a short-range part. The long-range part is solved with the Fourier method (Particle Mesh) whereas the short range part is solved with the tree method. This is the method the GADGET-2 (Springel, 2005) code is using.

- **Adaptive Mesh**
  The dynamic range of particle-mesh codes can be increased by using an adaptive rather than a static grid to solve the Poisson Equation. In the Adaptive Mesh Refinement (AMR) (Kravtsov et al., 1997) (parallel version in a Adaptive-Refinement-Tree Gottloeber & Klypin (2008)) method, the grid elements are refined (a higher resolution is applied), e.g. around the highest density regions where the particle number (potential gradient) is high.



3.3.2 Bolshoi simulation

The Bolshoi simulation (Klypin et al., 2011) is a ΛCDM dark matter simulation run with the ART code, which is an AMR code described in Section 3.3.1. Bolshoi was run with a volume of 250 h^{-1} Mpc per side with 2048^3 (~ 8 billion) particles. Bolshoi used a set of cosmological parameters that are compatible with WMAP measurements (Komatsu et al., 2009), namely $H_0 = 100$ km s^{-1} Mpc^{-1}, $h = 0.7, \Omega_{\mathrm{M}} = 0.27, \Omega_{\mathrm{B}} = 0.047, \Omega_{\Lambda} = 0.73, n_{\mathrm{s}} = 0.95, \sigma_8 = 0.82$. The mesh is gridded into 256^3 cells and the comoving force resolution is $\epsilon = 1\,h^{-1}$ kpc.

3.3.3 MultiDark simulations

The BigMultiDark Planck (BigMDPL) simulation is one of a series of simulations, called the Multidark simulations (Klypin et al., 2016). It was run with the GADGET-2 code, consisting of 3840^3 particles, a force smoothing of $\epsilon = 10\,h^{-1}$ kpc and Planck cosmological parameters (Planck Collaboration, 2015): $H_0 = 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}, h = 0.678, \Omega_{\mathrm{M}} = 0.307, \Omega_{\mathrm{B}} = 0.048, \Omega_{\Lambda} = 0.693, n_{\mathrm{s}} = 0.96, \sigma_8 = 0.83$. The volume of BigMDPL is $2.5\,h^{-1}$ Gpc per side.

3.3.4 Halo finding

As discussed in Section 2.6.7, dark matter halos play an vital role in ΛCDM cosmology. Halos are considered to function as a host where baryonic matter condenses and thus star formation can take place. Identifying dark matter halos within a (quasi-)continuous matter density field of a simulation needs exact definitions (White, 2001). Although there were many different halo finding algorithms developed over the past years, I will focus on the algorithms Bound-Density-Maxima (BDM) (Klypin & Holtzman, 1997) and RockStar (Behroozi et al., 2013) .

The BDM algorithm finds the maxima of the density of dark matter particles. The radius of the spherical halo is defined with the virial radius R_{vir} as the radius within the mean density is the virial overdensity δ_{vir} times the mean universal matter density $\varrho_{\mathrm{M}} = \Omega_{\mathrm{M}}\varrho_{\mathrm{crit}}$. Thus, the virial mass is given by

$$M_{\mathrm{vir}} = \frac{4}{3}\pi\delta_{\mathrm{vir}}\varrho_{\mathrm{M}}R_{\mathrm{vir}}^3\,. \tag{3.10}$$

Also widely used conventions are, δ_{200} (R_{200}, M_{200}), denoting an overdensity 200 times the critical density ϱ_{crit} at a given redshift z. Unbound particles that can not be assigned to a halo are removed. In this way, the BDM algorithm provides several statistics for halos including virial mass and radius, as well as velocity.

RockStar uses a friends-of-friends algorithm (Davis et al., 1985) for a first classification of clustered particles. Then spherical dark matter halos and subhalos are identified using an approach based on adaptive hierarchical refinement of friends-of-friends groups in six phase-space dimensions and one time dimension. RockStar computes the halo mass using spherical overdensities of a virial structure. Before calculating halo masses and circular velocities, the halo finder removes unbound particles from the final mass of the halo.

3.3.5 Mock Catalogues

Assigning galaxies to halos is a key point for precise cosmological simulations. Numerous methods have been developed to address this problem. The most enclosed method relies on the joint numerical simulation of dark and baryonic matter with a hydrodynamic framework. Most prominent simulations using this approach are the Illustris (Vogelsberger et al., 2014) and Eagle (Schaye et al., 2015) simulations. However, this approach is not yet computationally suited for volumes observed by BOSS.

If the underlying simulations contain dark matter only, the questions arises how the observable galaxies can be predicted with all their properties, such as mass, position, velocity, and luminosity. Furthermore, in order to fully imitate an observation from a galaxy redshift survey, these catalogues must include so-called light-cone effects in geometry, structure growth with redshift, and selection effects. In Rodríguez-Torres et al. (2016) the authors present the SUGAR code which they used to create mock galaxy redshift survey catalogues based on the BigMDPL simulation. Within the SUGAR framework, the authors used Halo Abundance Matching (HAM) (Conroy et al., 2006; Guo et al., 2010; Kim et al., 2008; Nuza et al., 2013; Tasitsiomi et al., 2004; Trujillo-Gomez et al., 2011; Vale & Ostriker, 2004) to assign galaxies to a dark matter halo that was previously defined with BDM.

HAM supposes a correspondence between the luminosity or stellar and dynamical masses: galaxies with more stars are assigned to more massive halos or subhalos. The luminosity in a red-band is sometimes used instead of stellar mass. Also methods that include some degree of stochasticity in the relation between stellar and dynamical masses due to deviations in the merger history, angular momentum, halo concentration, and even observational errors have been elaborated (Behroozi et al., 2010; Leauthaud et al., 2011).

Beside HAM, also another method is commonly used, called semi-analytic models (SAMs). These methods serve to populate model galaxies with the evolving of dark matter halos. The SAMs use simplified, partly phenomenological recipes to trace the evolution of model galaxies, so attempt to predict the detailed properties of galaxies (White & Frenk (1991); Kauffmann et al. (1993); Somerville & Primack (1999); Kauffmann et al. (2003); Kang et al. (2005)). In these models, adjustable parameters are needed be calibrated against the observations for the efficiencies of poorly understood physical processes such as star formation or feedback of Supernovae and active galactic nuclei. In recent years, more details have been added into these models, increasing the accuracy of the representation of the observed galaxy populations. The faithful representation of the observations and the relative low computational costs make SAMs also suitable for building mock catalogues to compare them with the large galaxy surveys.

4

Implementation of ARGO

In this chapter I describe the reconstruction algorithm code ARGO (**A**lgorithm for the **R**econstruction of the **G**alaxy traced **O**verdensities) (Kitaura & Enßlin, 2008). I begin with the basic prerequisite of the statistical framework ARGO is build upon in Section 4.1. In modern cosmology statistical methods are used in a broad field and I will mostly adopt the description given by Heavens (2009); Verde (2007, 2010) who summarized the statistical concepts in his context. Afterwards I discuss the physical model of the reconstructed density field in Section 4.4. Finally I derive all quantities that the inference analysis requires in detail in Section 4.5. I further describe the Bayesian inference method and advantages that arise of a Bayesian treatment of parameter estimation.

4.1 Statistical background

Before I describe the method used in ARGO, I will firstly discuss the fundamental statistical background necessary for a thorough understanding of the reconstruction procedure.

4.1.1 Bayesian Approach

The two standard approaches to statistics, *frequentist* and *Bayesian* interpretation, differ in the way probability is perceived. The frequentist (originating from the term frequency) approach assumes that a certain probability $\mathscr{P}(A)$ for an event A can be calculated by the ratio of number of successful (w.r.t event A) trials n_A and total number of trials N, thus $\mathscr{P} = n_A/N$.

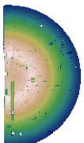
This is a robust definition, independent of assumptions of the underlying random process. Nevertheless, this interpretation has shortcomings if the probability can not be expressed in a frequency of trials. Therefore the Bayesian interpretation of probabilities has a broader validity and is colloquially stated as *degree of belief* (Nußbaum, 1975) in a hypothesis. It is built upon the Bayes' theorem which writes for the parameter vector $\boldsymbol{\theta}$ and the data vector $\boldsymbol{x}$ as

$$\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathscr{P}(\boldsymbol{x}|\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta})}{\mathscr{P}(\boldsymbol{x})} \,. \tag{4.1}$$

In this context:

- $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$ is the probability (density) of the ***posterior*** function. The conditional probability of the parameter vector $\boldsymbol{\theta}$ given the data vector $\boldsymbol{x}$;

- $\mathscr{P}(\boldsymbol{x}|\boldsymbol{\theta})$ is the probability of the data $\boldsymbol{x}$ given the parameter vector $\boldsymbol{\theta}$. This function is also called ***likelihood*** and will be denoted with $\mathscr{L}(\boldsymbol{x}|\boldsymbol{\theta})$;



- $\mathscr{P}(\boldsymbol{\theta})$ is the probability of the parameter vector $\boldsymbol{\theta}$, i.e.. the underlying distribution from which $\boldsymbol{\theta}$ is realized. As this function is already known without prior[1] knowledge of the data, this function is called ***prior*** and will be denoted with $\pi(\boldsymbol{\theta})$;
- $\mathscr{P}(\boldsymbol{x})$ is called ***evidence*** and represents the often inaccessible probability of the data vector $\boldsymbol{x}$.

We note that in many statistical problems a distinction is useful, whether parameters are to be inferred by the analysis or are previously known. This knowledge is called ***information*** $\mathscr{I}$. It is written in a conditional probability, e.g. $\pi(\boldsymbol{\theta}, \mathscr{I})$. This notation is useful to update the prior given the posterior of a Bayesian analysis and the previous information, if the prior and likelihood are conjugate (of the same form). It is also important to note that the evidence acts as the normalization of the posterior for parameter estimation problems. Thus we can express the evidence by marginalizing over $\boldsymbol{\theta}$ as :

$$\mathscr{P}(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \mathscr{L}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \,. \tag{4.2}$$

We also immediately see that $\mathscr{P}(\boldsymbol{x}|\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta})$ is the joint probability of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, that is written as $\mathscr{P}(\boldsymbol{x}, \boldsymbol{\theta})$.

4.1.2 Parameter estimation

In Equation 4.1 the pivotal concept of Bayesian statistics is shown. It gives the probability density for a set of parameters $\boldsymbol{\theta}$ conditioned on the data $\boldsymbol{x}$. The upcoming question is now how the set of parameters $\boldsymbol{\theta}$ is estimated. Therefore, it is adequate to discuss the basic difference of Bayesian methods to *frequentist* approaches like the *Maximum Likelihood Estimate*[2] (MLE).

MLE is used to find the best estimate for $\boldsymbol{\theta}$, called $\hat{\boldsymbol{\theta}}$, given the data $\boldsymbol{x}$, that maximizes the likelihood function $\mathscr{L}(\boldsymbol{x}|\boldsymbol{\theta})$. As no distribution of $\boldsymbol{\theta}$ is assumed, $\hat{\boldsymbol{\theta}}$ will be a point estimate. So in a Bayesian formalism the MLE relies on the assumption that the ratio of prior and evidence is constant

$$\frac{\pi(\boldsymbol{\theta})}{\mathscr{P}(\boldsymbol{x})} = const \,, \tag{4.3}$$

and thus does not contribute in the maximization. So the MLE is able to estimate the most probable value of $\boldsymbol{\theta}$ given the data, but does not predict the outcome of $\boldsymbol{\theta}$ as such in a random process, meaning that frequentist approached cannot estimate the probability density of $\boldsymbol{\theta}$, but only define a *best guess* value. The likelihood is not a probability density in a strict sense as it is not necessarily normalized to 1 and does not assign a probability to a set of parameters. Nevertheless a likelihood is a useful tool if it is used as a ratio.

A key difference between a Bayesian approach and frequentist parameter estimation methods, such as χ^2 or maximum likelihood, is that Bayesian inference is utilizing the whole support of $\pi(\boldsymbol{\theta})$ and not only restricted to a single point estimate of $\boldsymbol{\theta}$, as done by MLE. This will however make a difference in the conclusion drawn from a parameter estimation regarding the choice of the prior distribution (Gilks et al., 1995).

The pendant to the MLE method in Bayesian statistics is the so-called Maximum a posteriori (MAP) estimate. MAP estimates a point $\boldsymbol{\theta}_{\mathrm{MAP}}$ for which $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$ assumes its maximum. In practice it is only necessary to maximize a function that is proportional to the normalized posterior $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$, thus

$$\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathscr{L}(\boldsymbol{x}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \tag{4.4}$$

---

[1] This is not fully correct as the prior can be updated with constraints from data.

[2] In literature we usually encounter problems that are solved by minimizing the negative logarithm. This is numerically more accessible than a (presumably) diverging maximization. Also this function then coincides with $\chi^2 = -2\log\mathscr{L}$ parameter estimation for Gaussian PDFs.

can be a better choice to be maximized.

4.1.3 Prior choice

Bayesian statistics relies on the assumption that a prior knowledge is given to a problem which is received either due to a subjective estimate or a previous analysis. If there is good knowledge, e.g. about an estimated parameter, this knowledge can be incorporated into the prior and thus will yield more precise estimates (Gelman, 2010).

However, if there is just little known about the statistical occurrence and thus the choice of the prior is not specific, empirical Bayesian methods may have benefits. They are constructed in a way that the prior itself is derived from data. Furthermore, there is also the possibility to set up the analysis with so-called uninformative priors that do not contribute any knowledge about the parameters of interest. Two important types of priors that belong to this class are uniform distributions and Jeffrey's priors (Lee, 2012).

It is obvious that the choice of the prior will lead to different interpretations and results as compared to *frequentist* methods. To some extend the arbitrary choice of a prior is a main point of criticism of Bayesian statistics.

4.1.4 Hierarchical models

Many statistical applications contain multiple parameters that can be related or connected in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect their dependence.

This problem can be addressed with a hierarchical probability model. If we recall Equation 4.1, we face a single parameter vector $\boldsymbol{\theta}$. This parameter now may be depending on another parameter $\boldsymbol{\phi}$, called *hyperparameter*. Straight on modelling this problem would yield up to the normalization

$$\mathscr{P}(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathscr{L}(\boldsymbol{x}|\boldsymbol{\phi},\boldsymbol{\theta}) \times \pi(\boldsymbol{\phi},\boldsymbol{\theta})\,. \tag{4.5}$$

This is not only technically challenging due to dimensionality, moreover, this formalism also does not mirror the knowledge of how $\boldsymbol{\theta}$ depends on $\boldsymbol{\phi}$. Using Bayes rule we obtain

$$\begin{aligned}
\mathscr{P}(\boldsymbol{\phi},\boldsymbol{\theta}) &= \mathscr{P}(\boldsymbol{\phi}) \times \mathscr{P}(\boldsymbol{\theta}|\boldsymbol{\phi})\,, && (4.6)\\
\mathscr{P}(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{x}) &\propto \mathscr{P}(\boldsymbol{\phi},\boldsymbol{\theta}) \times \mathscr{P}(\boldsymbol{x}|\boldsymbol{\phi},\boldsymbol{\theta})\,, && (4.7)\\
&= \mathscr{P}(\boldsymbol{\phi},\boldsymbol{\theta}) \times \mathscr{P}(\boldsymbol{x}|\boldsymbol{\theta})\,, && (4.8)
\end{aligned}$$

where the last simplification holds, if parameter $\boldsymbol{\phi}$ has no direct influence on $\boldsymbol{x}$ but through $\boldsymbol{\theta}$. In Figure 4.1 a more complex example is shown. Here parameters $\boldsymbol{A}$ and $\boldsymbol{\theta}$ have direct influence on the outcome of the data $\boldsymbol{x}$. In the hyperplane there are still $\boldsymbol{B}$ and $\boldsymbol{C}$ modulating the parameter $\boldsymbol{A}$, and $\boldsymbol{\phi}$ modulating parameter $\boldsymbol{\theta}$. Applying Equation 4.8, this posterior can be expressed up to a normalization as

$$\mathscr{P}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{x}) \propto \underbrace{\mathscr{L}(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\boldsymbol{\phi}) \times \pi(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{C})}_{\text{direct plane}} \times \underbrace{\pi(\boldsymbol{B}) \times \pi(\boldsymbol{C}) \times \pi(\boldsymbol{\phi})}_{\text{hyperplane}}\,. \tag{4.9}$$

We will show how to use this representation of the products of conditional probabilities in Section 4.3.2 to sample each parameter separately. This is an important scheme within our reconstruction method.

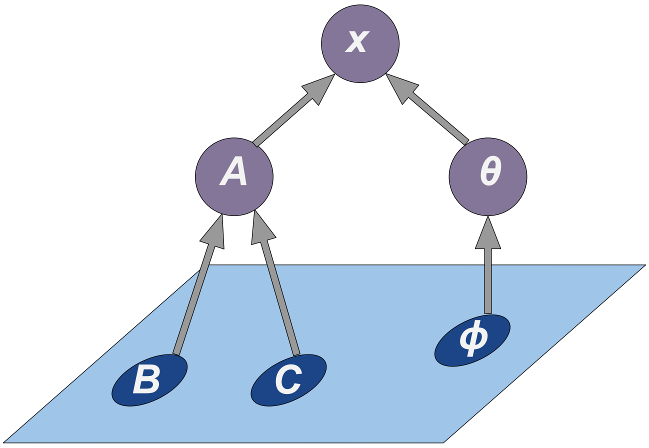


Figure 4.1: Concept of hierarchical modelling given a data set $\boldsymbol{x}$ and parameters $\boldsymbol{A}$, $\boldsymbol{\theta}$. These parameters themselves depend on $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{\phi}$ respectively. $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{\phi}$ only influence the data $\boldsymbol{x}$ via $\boldsymbol{A}$ and $\boldsymbol{\theta}$, illustrated in the figure as no direct arrow is shown. The *hyperparameter space* is shown in the blue region. This represents a typical setting where Bayesian hierarchical modelling is convenient to apply.

4.2 Sampling techniques

Having discussed the theoretical background of statistical parameter estimation (Section 4.1.2) with Bayesian statistics (Section 4.1), I will now focus on methods to generate samples from a probability density distribution (Murray et al., 2012), which is a key ingredient of the shown works. Attaching importance to Markov Chain Monte-Carlo (MCMC) (Neal, 1993; Newman & Barkema, 1999) I introduce different sampling schemes of which Section 4.3.2 and Section 4.3.3 are most relevant for ARGO.

4.2.1 Concept of Monte-Carlo samplers

Monte-Carlo (MC) algorithms are stochastic methods that rely on repeated random sampling of independent and identically distributed (*i.i.d.*) random numbers to gain numerical results. As MC algorithms have a very broad range of applications such as optimization, I will limit this discussion to sampling and integration methods. Each of these methods are illustrated with an example.

- **Rejection sampling:**
  Suppose we want to draw samples from the posterior $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$ which must not be necessarily normalized (no constrains on $\int \mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$). If there is no closed analytic form of $\mathscr{P}$ or the given function is very complicated to sample from, rejection sampling is a method to overcome this problem. Therefore, we choose a distribution $g(\boldsymbol{\theta})$ which we can sample from (proposal function). Moreover, $g(\boldsymbol{\theta})$ multiplied with a constant c must be greater than $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$ at all $\boldsymbol{\theta}$. $c \cdot g(\boldsymbol{\theta})$ then is called envelope function. Now we draw θ_g from a distribution proportional to $g(\boldsymbol{\theta})$. Afterwards the importance ratio $R = \frac{\mathscr{P}(\boldsymbol{\theta}_g|\boldsymbol{x})}{c \cdot g(\boldsymbol{\theta}_g)}$ is calculated and accepted, if the ratio R is less then a uniform random number between 0 and 1, $R \leq U[0,1]$, otherwise rejected and a new random point θ_g is drawn. The schematic procedure is shown in Figure 4.2.

- **Importance sampling:**
  Different from rejection sampling, importance sampling cannot be used to draw samples of $\mathscr{P}$.

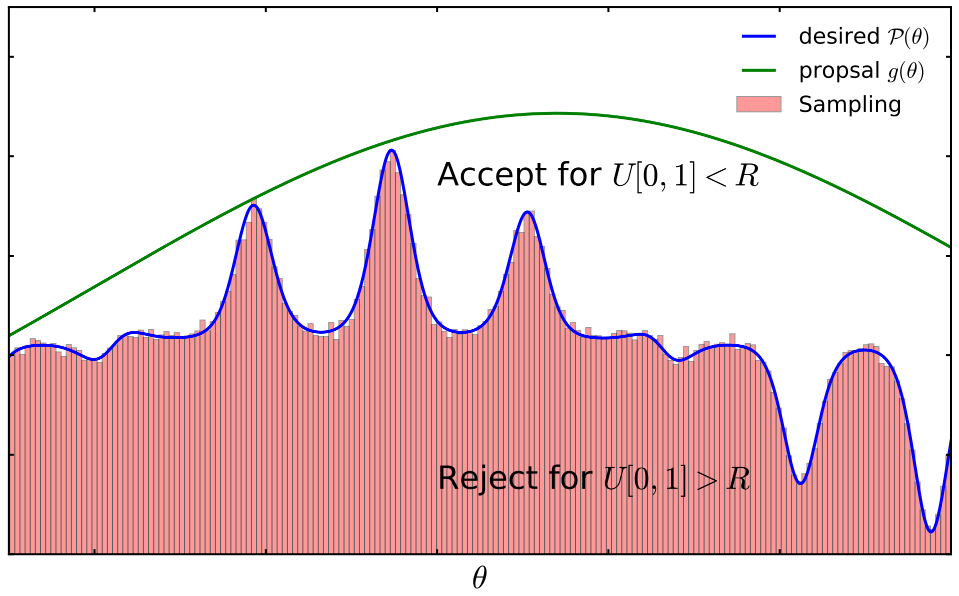


Figure 4.2: Concept of rejection sampling shown in one dimension. In blue color the desired PDF is plotted that we want to draw samples from, $\mathscr{P} \propto \exp\left(\left(\cos\theta^{1.5}\right)^{1.2}\sin\theta + 2\exp\left(-\theta\right)^{1.9}\right)$. The proposal function $g(\boldsymbol{\theta})$ is shown in green, basically being a normal distribution with selected mean and variance, $g(\boldsymbol{\theta}) \propto \mathscr{N}(\mu,\sigma)$. The Importance ratio $R = \frac{\mathscr{P}(\boldsymbol{\theta_g}|\boldsymbol{x})}{c\cdot g(\boldsymbol{\theta_g})}$ is compared to a uniform random number $U[0,1]$ between 0 and 1. The accepted draws of $\boldsymbol{\theta}$ from g are then effectively samples of $\mathscr{P}$.

Instead, it is an efficient method to compute integrals of the form

$$\mathbb{E}_{\mathscr{P}}[h(\boldsymbol{\theta}|\boldsymbol{x})] = \int h(\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}\,, \tag{4.10}$$

where $\mathbb{E}_{\mathscr{P}}[h(\boldsymbol{\theta}|\boldsymbol{x})]$ is the expectation of the function $h(\boldsymbol{\theta})$ given the normalized probability distribution $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$. The dependency of $h(\boldsymbol{\theta})$ on $\boldsymbol{x}$ is therefore induced by $\mathscr{P}$. If we are able to draw samples $\boldsymbol{\theta}_{\mathscr{P}}$ of $\mathscr{P}$, we can immediately solve the integral by

$$\mathbb{E}_{\mathscr{P}}[h(\boldsymbol{\theta}|\boldsymbol{x})] = \int h(\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \approx \frac{1}{N}\sum_{n=1}^{N} h(\boldsymbol{\theta}^n_{\mathscr{P}})\,, \tag{4.11}$$

summing up $n = 1, ..., N$ draws. This is called Monte-Carlo integration. But what if the direct sampling from $\mathscr{P}$ is not possible, e.g. due to dimensionality or not having a closed analytical form? One possible solution to this problem is to rely on another normalized probability density function (PDF) $g(\boldsymbol{\theta})$ of which we can directly draw samples from. Thus we write

$$\begin{aligned}
\mathbb{E}_{\mathscr{P}}[h(\boldsymbol{\theta}|\boldsymbol{x})] &= \int h(\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \\
&= \int h(\boldsymbol{\theta})\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})\frac{g(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}d\boldsymbol{\theta} \\
&= \int \left[h(\boldsymbol{\theta})\frac{\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})}{g(\boldsymbol{\theta})}\right] g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \sum_{n=1}^{N} h(\boldsymbol{\theta}^n_g)w(\boldsymbol{\theta}^n_g)\,,
\end{aligned} \tag{4.12}$$

where $w(\boldsymbol{\theta}^n_g) = \frac{\mathscr{P}(\boldsymbol{\theta}^n_g|\boldsymbol{x})}{g(\boldsymbol{\theta}^n_g)}$ and $\boldsymbol{\theta}^n_g$ is the n-th sample of $\boldsymbol{\theta}$ drawn from PDF $g(\boldsymbol{\theta})$.

4.3 Markov Chain Monte-Carlo

Rejection and importance sampling are grossly inefficient with soaring dimensionality of the inference problem. It will be computationally unfeasible to explore a multidimensional PDF with pure rejection sampling. However, instead of drawing *i.i.d.* samples from a single PDF, I will now introduce a method sequentially sampling from a so-called Markov chain. A Markov chain is defined as a sequence of random variables $\{\boldsymbol{\theta}^t\}$ (i.e. parameters in Bayesian statistics, Section 4.1) whose current draw (or state) $\boldsymbol{\theta}^t$ only depends on the most recent draw $\boldsymbol{\theta}^{t-1}$. We can express this *Markov property* in terms of probability as:

$$\mathscr{P}(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ..., \boldsymbol{\theta}^t) = \mathscr{P}(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t)\,. \tag{4.13}$$

The probability to move from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^{t+1}$ is induced by the *transition kernel*, a transition distribution for a continuous case or transition matrix for a discrete case, frequently denoted with $T_{t+1}(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t)$ ($k \times k$ matrix $\mathbf{T}$ for a k dimensional problem). T_{t+1} has to stay the same for any t so that Equation 4.13 holds. The key point of Markov chains is the *stationary distribution* $\mathscr{P}_{st}$. This distribution is constructed to satisfy the equation

$$\int T_{t+1}(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t)\mathscr{P}_{st}(\boldsymbol{\theta}^t)d\boldsymbol{\theta} = \mathscr{P}_{st}(\boldsymbol{\theta}^{t+1})\,, \tag{4.14}$$

expressing that once the chain has moved to the stationary distribution, the transition kernel has no effect on $\mathscr{P}_{st}$ anymore, so that $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t+1}$ are drawn from the same PDF.

Markov chain Monte-Carlo (MCMC) now uses Equation 4.14 in a way that the distribution we want to sample from, called *target distribution*, is the stationary distribution of a Markov chain. Once the desired distribution $\mathscr{P}_{st}$ is obtained, MC methods are used to draw samples from. MCMC came to newly popularity in the 90's with Gelfand & Smith (1990) and so encouraged Bayesian statistical analyses, which have been inaccessible due to computational reasons before. Yet there is still one caveat using MCMC. As we mentioned before, samples from a MCMC are not *i.i.d.* random variables. They slightly depend on each other given by condition 4.13, even if the stationary distribution of the MCMC is reached[3].

However, to have a well defined stationary distribution, a Markov chain must come up with the following properties:

- **Irreducibility:**
  This means, that the probability for any state moving from θ_i to θ_j is greater than 0,
  $$T_{ij} = \mathscr{P}(\theta^{t+1} = \theta_i|\theta^t = \theta_j) > 0\,,$$
  which means that at any time step (iteration) the chain can assume any allowed state and $\mathbf{T}_{ij}$ can not be reduced into smaller matrices.
- **Aperiodicity:**
  The chain must not get trapped into cycles. So the probability for θ_i moving to θ_i at iteration n is $\mathscr{P}^n_{i,i} > 0$.
- **Positive recurrency:**
  A state i is recurrent if the expected time until the process returns to state θ_i is finite.

If these conditions are satisfied all states of the chain are ergodic, which means that the chain is ergodic itself. Ergodicity for a chain means that the average over sufficient large time steps t is the same as averaging over the ensemble of samples, which is effectively the same as *i.i.d.* sampling from a PDF.

---

[3] Also ergodicity is a condition for a MCMC to have a unique stationary distribution.

4.3.1 Metropolis-Hastings sampling

Metropolis-Hastings algorithm (M-H) (Hastings, 1970) is the generalization of Metropolis et al. (1953) method, useful for sampling from Bayesian posterior distributions. The basic idea of M-H is to sample from the PDF of interest that corresponds to the stationary distribution of a Markov chain. We set up an arbitrary starting state in parameter space $\boldsymbol{\theta}^0$, so that the starting distribution $\mathscr{P}(\boldsymbol{\theta}^0|\boldsymbol{x}) > 0$. To evolve the chain from state t to $t+1$, we draw a *new proposal* sample $\boldsymbol{\theta}'$ from a *proposal distribution (jumping distribution)* at time $t+1$, $\mathscr{J}_{t+1}(\boldsymbol{\theta}'|\boldsymbol{\theta}^t)$. This new proposal $\boldsymbol{\theta}'$ is used to calculate the ratio of the PDFs

$$r = \frac{\mathscr{P}(\boldsymbol{\theta}'|\boldsymbol{x}) / \mathscr{J}_{t+1}(\boldsymbol{\theta}'|\boldsymbol{\theta}^t)}{\mathscr{P}(\boldsymbol{\theta}^t|\boldsymbol{x}) / \mathscr{J}_{t+1}(\boldsymbol{\theta}^t|\boldsymbol{\theta}')} \,. \tag{4.15}$$

Thus, the transition probability $T_{t+1}(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t)$ for the state $\boldsymbol{\theta}^t$ is a mixture of its probability and the proposal distribution. In Equation 4.15 it can be seen that for a symmetric proposal distribution the ratio is simplified to $r = \dfrac{\mathscr{P}(\boldsymbol{\theta}'|\boldsymbol{x})}{\mathscr{P}(\boldsymbol{\theta}^t|\boldsymbol{x})}$.

We finally decide whether to keep the new proposal state $\boldsymbol{\theta}'$ or not by

$$\boldsymbol{\theta}^{t+1} = \begin{cases} \boldsymbol{\theta}' & \text{if } U[0,1] < min(r,1) \\ \boldsymbol{\theta}^t & \text{else} \end{cases} \,. \tag{4.16}$$

As Equation 4.16 already states, every iteration of M-H produces a sample, whether the chain assumes a new state $\boldsymbol{\theta}'$ or not. Another very interesting (and totally necessary in many applications) point of the presented MCMC methods is that the *evidence*, the normalization of the posterior, is cancelling out as we are only interested in the ratio.

Equation 4.16 is derived from the joint probability of two states $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ which are drawn from the stationary distribution as

$$\begin{aligned} \mathscr{P}(\boldsymbol{\theta}^t = \boldsymbol{\theta}^a, \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^b) &= \mathscr{P}(\boldsymbol{\theta}^a|\boldsymbol{x}) \mathscr{J}_{t+1}(\boldsymbol{\theta}^b|\boldsymbol{\theta}^a) \,, & (4.17) \\ \mathscr{P}(\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^a, \boldsymbol{\theta}^t = \boldsymbol{\theta}^a) &= \mathscr{P}(\boldsymbol{\theta}^a|\boldsymbol{x}) \mathscr{J}_{t+1}(\boldsymbol{\theta}^a|\boldsymbol{\theta}^b) \,, & (4.18) \end{aligned}$$

where the joint probability is symmetric.

The choice of the proposal distribution is rather arbitrary. For many inference problems a proposal function that relies on *random walks* is satisfactory, that can be proposals from a normal distribution that are centered on the previous iteration with arbitrary width or even uniform distributions. However, as the target distribution we want to sample from can be fairly complex, methods were developed that do not rely on random walks but utilize the knowledge of the parameter space, such as steepness (see Robert & Casella (2005) for more examples). One of these modifications are presented in the Langevin algorithm, where the proposal state $\boldsymbol{\theta}'$ is given by

$$\boldsymbol{\theta}' \sim \boldsymbol{\theta}^{t-1} + \mathscr{N}(\epsilon) + \boldsymbol{\nabla} \log \mathscr{P} \,, \tag{4.19}$$

where $\mathscr{N}(\epsilon)$ is the normal distribution centered at $\boldsymbol{\theta}^{t-1}$ and width ϵ. It is obvious that this proposal distribution will prefer phase-space regions with high gradient and thus help the algorithm to reach the stationary distribution. An example for a two dimensional problem is shown in Figure 4.3.

4.3.2 Gibbs sampling

Proposed in Geman & Geman (1984) the Gibbs sampler is a particular interesting MCMC. It relies on the ability to express a conditional probabilistic model for each parameter, nevertheless the Gibbs

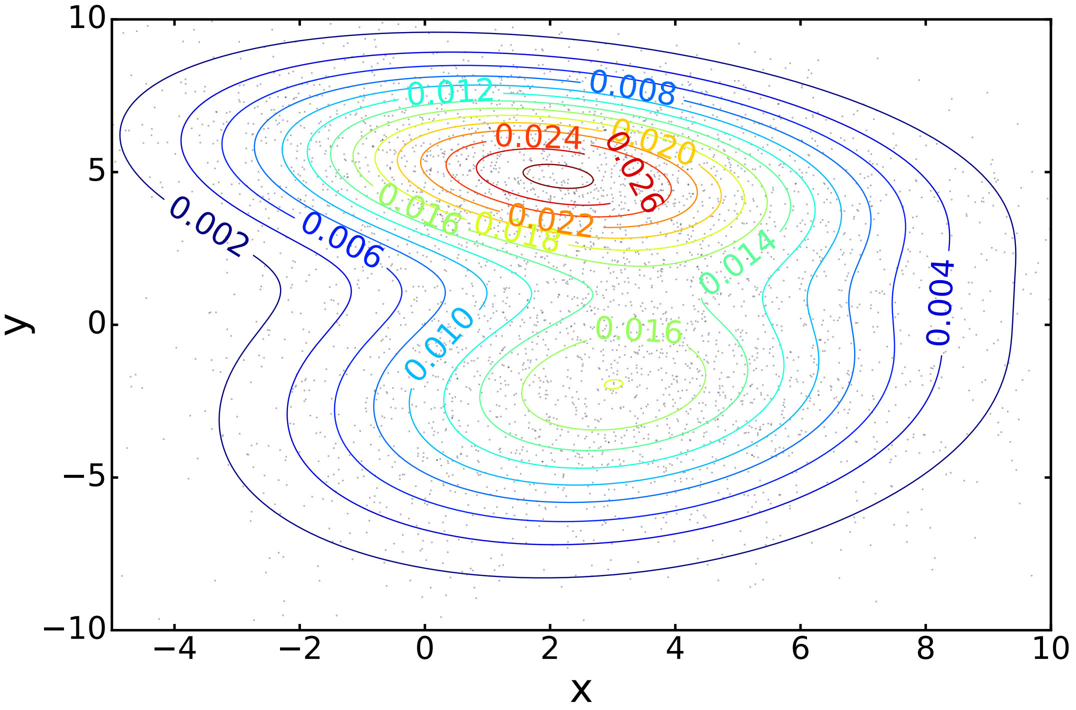


Figure 4.3: Example of a Metropolis-Hastings scheme shown for a two dimensional sampling from a bi-variate normal distribution $\mathscr{N} = \mathscr{N}_1(\mu_x = 2, \mu_y = 5, \sigma_x = 3, \sigma_y = 2, \sigma_{xy} = -1.6) + \mathscr{N}_2(\mu_x = 3, \mu_y = -2, \sigma_x = 3, \sigma_y = 3, \sigma_{xy} = 1.6)$. The proposal distribution was chosen to be a normal distribution centered on the current state with unity width. The color code with the corresponding numbers represent the density contours.

algorithm will not be applicable if there is no possibility to impose conditional probabilities. Denoting the parameter vector $\boldsymbol{\theta}^t = (\theta_1^t, ..., \theta_D^t)$ of the t-th iteration, the Gibbs sampler will evaluate D times each corresponding conditional probability for every iteration t. Starting with $\boldsymbol{\theta}^0$, we can now write the probability for $\boldsymbol{\theta}^1$ element-wise as

$$\theta_1^1 \curvearrowleft \mathscr{P}(\theta_1^1|\theta_2^0, ..., \theta_D^0, \boldsymbol{x})\,. \tag{4.20}$$

As the chain goes on through each vector component

$$\begin{array}{rcl}
\theta_2^1 & \curvearrowleft & \mathscr{P}(\theta_1^1|\theta_3^0, ..., \theta_D^0, \boldsymbol{x})\,, \\
 & \vdots & \\
\theta_D^1 & \curvearrowleft & \mathscr{P}(\theta_1^1|\theta_2^1, ..., \theta_{D-1}^1, \boldsymbol{x})\,, \\
\cline{1-2}
\theta_1^2 & \curvearrowleft & \mathscr{P}(\theta_2^1|\theta_3^1, ..., \theta_D^1, \boldsymbol{x})\,, \\
 & \vdots & \\
\cline{1-2}
\theta_D^N & \curvearrowleft & \mathscr{P}(\theta_1^N|\theta_2^N, ..., \theta_{D-1}^N, \boldsymbol{x})\,,
\end{array}$$

it will finally reach the stationary distribution after N sufficient large iteration steps so that effectively $\boldsymbol{\theta}$ is sampled from the target distribution $\mathscr{P}(\boldsymbol{\theta}|\boldsymbol{x})$.

Therefore, in the Gibbs sampling scheme an arbitrary choice of the proposal distribution is obsolete, as the model itself imposes the conditional distributions. This scheme is particularly helpful for hierarchical models as discussed in Section 4.1.4.

4.3.3 Hamiltonian Monte-Carlo

Duane et al. (1987) presented a method they called *Hybrid Monte Carlo* (HMC) which was used in lattice quantum chromodynamics. The name was chosen because they auxiliary included variables that *guided* the Monte-Carlo simulation. Later, HMC arouse interest by the scientific community (Neal, 1993, 2012).

The model itself relies on Hamiltonian mechanics. The *phase-space* dynamics is described by the Hamilton function (Hamiltonian) $\mathscr{H}(\boldsymbol{q}, \boldsymbol{p}, t)$, where $\boldsymbol{q}$ is the *canonical* coordinate, $\boldsymbol{p}$ the conjugate momentum and t the time. The Hamiltonian also equals to the total energy in general cases [4]. Any quantity (e.g. $f(\boldsymbol{q}, \boldsymbol{p}, t)$) within this system is evolving in time according to the Hamiltonian (Hand & Finch, 1998) as

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}t} &= \frac{\partial f}{\partial q}\frac{\partial \mathscr{H}}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial \mathscr{H}}{\partial q} + \frac{\partial f}{\partial t}\,, \\ &= \{f, \mathscr{H}\} + \frac{\partial f}{\partial t}\,, \end{aligned} \tag{4.21}$$

where $\{\}$ is called *Poisson bracket*. From Equation 4.21 we immediately derive the *Hamiltonian equations of motion*:

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\frac{\partial \mathscr{H}}{\partial \boldsymbol{q}}\,, \tag{4.22}$$

$$\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}t} = \frac{\partial \mathscr{H}}{\partial \boldsymbol{p}}\,. \tag{4.23}$$

Equations 4.22 only hold if the Hamiltonian $\mathscr{H} \neq \mathscr{H}(t)$ is not explicitly depending on time t as

$$\begin{aligned} \frac{\mathrm{d}\mathscr{H}}{\mathrm{d}t} &= \underbrace{\{\mathscr{H}, \mathscr{H}\}}_{=0} + \frac{\partial \mathscr{H}}{\partial t}\,, \\ &= 0\,, \end{aligned} \tag{4.24}$$

which also means that the Hamiltonian equations of motion are energy conserving. We further derive the derivative for any quantity that is independent of time, e.g. of the form $\varrho(\mathscr{H}(\boldsymbol{q}, \boldsymbol{p}))$, thus $\partial\varrho/\partial t = 0$,

$$\frac{\mathrm{d}\varrho}{\mathrm{d}t} = \sum_i \left(\frac{\partial \varrho}{\partial p_i}\frac{\partial \mathscr{H}}{\partial q_i} - \frac{\partial \varrho}{\partial q_i}\frac{\partial \mathscr{H}}{\partial p_i} \right)\,, \tag{4.25}$$

$$\begin{aligned} &= \sum_i \left(\frac{\partial \varrho}{\partial \mathscr{H}}\frac{\partial \mathscr{H}}{\partial p_i}\frac{\partial \mathscr{H}}{\partial q_i} - \frac{\partial \varrho}{\partial \mathscr{H}}\frac{\partial \mathscr{H}}{\partial q_i}\frac{\partial \mathscr{H}}{\partial p_i} \right)\,, \\ &= 0\,, \end{aligned} \tag{4.26}$$

for which $\varrho = \exp\left(-\frac{\mathscr{H}(\boldsymbol{q}, \boldsymbol{p})}{k_\mathrm{b}T}\right)$ is a solution. This is the Boltzmann distribution, assigning a probability to a state with Energy $E = \mathscr{H}$, and temperature T, where k_b is the Boltzmann constant.

---

[4] See Goldstein et al. (2001) for detailed discussions. We will only consider cases where $\mathscr{H}$ equals the energy of the system.

Now that we elaborated the Hamiltonian dynamics, we need to build the link to a probabilistic measure. This is done by defining the *canonical distribution* $\mathscr{P}(\boldsymbol{q},\boldsymbol{p})$ to sample from, and the Hamiltonian $\mathscr{H}(\boldsymbol{q},\boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p})$, where U is the potential of Hamiltonian, depending on the spatial coordinate $\boldsymbol{q}$ only, and the kinetic term K solely depending on the momentum $\boldsymbol{p}$. Putting this together the canonical distribution can be written as

$$\begin{aligned} \mathscr{P}(\boldsymbol{q},\boldsymbol{p}) &= \frac{1}{Z}\exp\left(-\frac{\mathscr{H}(\boldsymbol{q},\boldsymbol{p})}{k_{\mathrm{b}}T}\right), & (4.27)\\ \Rightarrow \mathscr{P}(\boldsymbol{q})\mathscr{P}(\boldsymbol{p}) &= \frac{1}{Z}\exp\left(-\frac{U(\boldsymbol{q})}{k_{\mathrm{b}}T}\right)\exp\left(-\frac{K(\boldsymbol{p})}{k_{\mathrm{b}}T}\right), & (4.28) \end{aligned}$$

where Z is the normalization of the distribution function. Thus we see that $U(\boldsymbol{q})$ and $K(\boldsymbol{p})$ are factorizing into separate probabilities $\mathscr{P}(\boldsymbol{q})$ and $\mathscr{P}(\boldsymbol{p})$. Hamiltonian Monte-Carlo now defines the (pseudo) spatial vector $\boldsymbol{q}$ to be the variable which we will sample, and $\boldsymbol{p}$ the auxiliary variable that will allow the Hamiltonian to explore the phase-space. Therefore, we can write the above Equation 4.27 as

$$U(\boldsymbol{q}) = -\ln\mathscr{P}(\boldsymbol{q}), \tag{4.29}$$

linking the desired posterior probability distribution $\mathscr{P}(\boldsymbol{q})$ to the potential $U(\boldsymbol{q})$, actually sampled by HMC.

Equation 4.27 imposes that $\boldsymbol{q}$ and $\boldsymbol{p}$ are sampled independently. As we are only interested in drawing samples from the posterior and not overmuch care about the exact probability of each sample, we drop the normalization term Z and also set the temperate $k_{\mathrm{b}}T = 1$. We will later see anyway that all constant terms will vanish due to the HMC method. The kinetic term corresponds to the kinetic energy of a free particle

$$K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{p}, \tag{4.30}$$

where $\mathbf{M}$ is the symmetric positive semidefinite mass variance matrix (covariance of the momenta $\left\langle\boldsymbol{p}^{\mathsf{T}}\boldsymbol{p}\right\rangle$). In Duane et al. (1987) this matrix has been chosen to be the identity matrix $\mathbf{I}$. However Jasche & Kitaura (2010) and Neal (2012) elaborated methods to keep the diagonal shape of the mass matrix but include preconditions to enhance the performance of the HMC.

Moreover, the form for the kinetic energy corresponds to minus logarithm of a Gaussian probability density with mean zero and covariance matrix $\mathbf{M}$, so that $\boldsymbol{p}$ can be drawn from a multivariate Gaussian distribution $\mathscr{P}(\boldsymbol{p})$ for each step.

Now the Hamiltonian equations of motion in Equation 4.22 can be formulated for the phase-space variables $\boldsymbol{q}$ and $\boldsymbol{p}$ as

$$\begin{aligned} \frac{\mathrm{d}q_i}{\mathrm{dt}} &= \frac{\partial\mathscr{H}}{\partial p_i} = \sum_j M_{ij}^{-1}p_j, & (4.31)\\ \frac{\mathrm{d}p_i}{\mathrm{dt}} &= -\frac{\partial\mathscr{H}}{\partial q_i} = -\frac{\partial U(q)}{\partial q_i}. & (4.32) \end{aligned}$$

It is not possible to follow the Hamiltonian dynamics exactly due to the discretization of the equations of motion. To evolve the Hamiltonian in (pseudo-) time, it is convenient to use the *leapfrog* scheme, which has the properties of being *time*-reversible and to conserve the phase-space volume being necessary conditions to ensure *ergodicity*:

$$
\begin{aligned}
p_i\left(t+\frac{\epsilon}{2}\right) &= p_i(t) - \frac{\epsilon}{2}\,\frac{\partial U(\boldsymbol{q})}{\partial q_i}\bigg|_{q_i(t)} , && (4.33)\\
q_i\left(t+\epsilon\right) &= q_i(t) + \epsilon \sum_j M^{-1}_{ij}\, p_j\left(t+\frac{\epsilon}{2}\right) , && (4.34)\\
p_i\left(t+\epsilon\right) &= p_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\,\frac{\partial U(\boldsymbol{q})}{\partial q_i}\bigg|_{q_i(t+\epsilon)} . && (4.35)
\end{aligned}
$$

The dynamics of this system is now calculated for a period of *time* $t = [0...\tau]$, with a discrete value of ϵ small enough to give acceptable errors and for $N_\tau = \Delta\tau/\epsilon$ leapfrog steps. In practice ϵ and N_τ are randomly drawn from a uniform distribution to avoid resonant trajectories (see Neal, 1993). The leapfrog method is a good compromise of accuracy, stability and efficiency (Athanassoula, 1993). The solution of the equations of motion will move the system from an initial state $(\boldsymbol{q}, \boldsymbol{p})$ to a final state $(\boldsymbol{q}', \boldsymbol{p}')$ after each sampling step. Although the Hamiltonian equations of motion are energy conserving, our approximate solution is not. For this reason a Metropolis-Hastings acceptance step has to be introduced in which the new phase-space state $(\boldsymbol{q}', \boldsymbol{p}')$ is accepted with a certain probability:

$$
P_A = min\left[1, \exp(-\Delta\mathscr{H})\right] , \tag{4.36}
$$

with $\Delta\mathscr{H} \equiv \mathscr{H}(\boldsymbol{q}', \boldsymbol{p}') - \mathscr{H}(\boldsymbol{q}, \boldsymbol{p})$. According to Gelman et al. (2013) the number of steps N_τ and the step size ϵ should be chosen to have approximately 65% acceptance rate so that the HMC is maximally efficient. Moreover, the initial guess distribution that the HMC is starting from is in general not the target distribution. Thus, the equations of motion will move the system to the correct target distribution of which we want samples from. The initial samples, before reaching the target belong to the *burn-in* phase. The length of the burn-in phase depends on the target distribution. We will show statistical criteria that help us to estimate the convergence behavior of the HMC.

4.4 Density field transformation

Transformation of density field of the dark matter is a well studied method to extract specific information from the cosmic web. As we discussed in Section 2.3.3 the current picture of ΛCDM predicts early Universe's initial conditions to follow Gaussian statistics. This is very much encouraged by the highly Gaussian CMB. Linear growth preserves the Gaussianity as each mode of the density field (i.e. the wave vector $\boldsymbol{k}$) evolves independently and thus the covariance matrix of the density field stays diagonal. Non-linear growth of structures due to gravitational attraction however, yields mode coupling and therefore introduces off-diagonal entries into the covariance. Density field transformations are beneficial to reduce these unpleasant effects to recover the Gaussian statistics. Local field transformations have been studied recently in literature (see Neyrinck, 2011; Neyrinck et al., 2011). These works emphasize that a logarithmic local transformation of the density field approximates observables extremely well and extracts vital information from the matter field, called "sufficient statistics" in literature (Carron & Szapudi, 2014b). It can also been shown that the logarithmic transformation to some extent restores the Gaussian statistics of the density field (Neyrinck et al., 2009), unleashes cosmological information from the two-point power spectrum and is a reasonably good estimate for the linear displacement field (Falck et al., 2012; Kitaura & Angulo, 2012).

Methods like in Zhang et al. (2011) decompose the density field $\boldsymbol{\delta}$ into Gaussian and non-Gaussian components using a discrete wavelet transformation (DWT). In this way non-linearities in the power spectrum can be reduced relying on the distribution of the wavelet components. The most promising Gaussianization is achieved when a transformation is applied to the density field so that the one-point distribution of the density is being removed by all higher-order moments. In Neyrinck et al. (2009) the authors present two promising Gaussianization schemes:

- **Gaussian mapping** $G(\delta)$:
  This transformation is written as

  $$G(\delta) \equiv \sqrt{2\sigma^2}\mathrm{erf}^{-1}\left(2f_{<\delta} - 1 + \frac{1}{N}\right) , \tag{4.37}$$

  where $f_{<\delta}$ is the fraction of volume elements less dense than δ, σ^2 the variance of the Gaussian, N the number of volume elements and erf the error-function.

- **Logarithmic transform** $\log(\delta)$:
  Logarithmic transformation is a local mapping of the density field to its logarithm written as:

  $$\log(\delta) \equiv \ln(1+\delta) + C\,, \tag{4.38}$$

  where C is a constant of integration.

The physical meaning of the logarithmic transform was derived in 2.7.1. It is a direct consequence of linear evolution in conformal time given by the continuity equation (Coles & Jones, 1991). We review the lognormal distribution as the solution of the continuity equation for density $\frac{\varrho}{\langle\varrho\rangle} = 1+\delta$ written in conformal time $\mathrm{d}\tau = \frac{\mathrm{d}t}{a}$ as

$$\int \frac{1}{\varrho}\frac{\mathrm{d}\varrho}{\mathrm{d}\tau} = -\int \mathrm{d}\tau \boldsymbol{\nabla}\cdot\boldsymbol{u}\,, \tag{4.39}$$

$$\ln(1+\delta) = -\int \mathrm{d}\tau \boldsymbol{\nabla}\cdot\boldsymbol{u}\,. \tag{4.40}$$

In the Zel'dovich approximation 2.7.2, the evolution of structure is described by Lagrangian initial coordinates $\boldsymbol{q}$ and Eulerian comoving coordinates $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{q} + \boldsymbol{\Psi}\,, \tag{4.41}$$

with the linear displacement field is given by $\boldsymbol{\Psi} = -D\boldsymbol{\nabla}\Phi$. D denotes the linear growth factor and Φ the peculiar gravitational potential. The linear Poisson equation still relates the linear density contrast δ_{L} to the gravitational potential with

$$\delta_{\mathrm{L}} = D\boldsymbol{\nabla}^2\Phi\,, \tag{4.42}$$

so that we find

$$\delta_{\mathrm{L}} = -\boldsymbol{\nabla}\cdot\boldsymbol{\Psi}\,. \tag{4.43}$$

The velocity in comoving frame $\boldsymbol{x}$ can be expressed via displacement as

$$\boldsymbol{u} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} = \frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}\tau}\,, \tag{4.44}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = \boldsymbol{\nabla}\cdot\frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}\tau}\,. \tag{4.45}$$

Combining Equations 4.40 and 4.45 we find

$$\boldsymbol{\nabla} \cdot \boldsymbol{\Psi} = -\log\left(1+\delta\right) + C\,, \tag{4.46}$$

where C is a constant of integration and well approximated by $\langle \log\left(1+\delta\right) \rangle$ (Kitaura & Angulo, 2012). Equation 4.46 shows that the logarithmic transform of the density contrast yields a reasonable estimate of the linear density, which is Gaussian distributed. All higher orders are comprised in the constant mean, $C = \delta_{+}$ as

$$\log\left(1+\delta\right) = \delta_{\mathrm{L}} + \delta_{+}\,. \tag{4.47}$$

Also it has been shown in Carron & Szapudi (2014a,b) that the logarithmic density transformation are statistically very close to the sufficient statistics of the one-point probability density function of the dark matter density field.

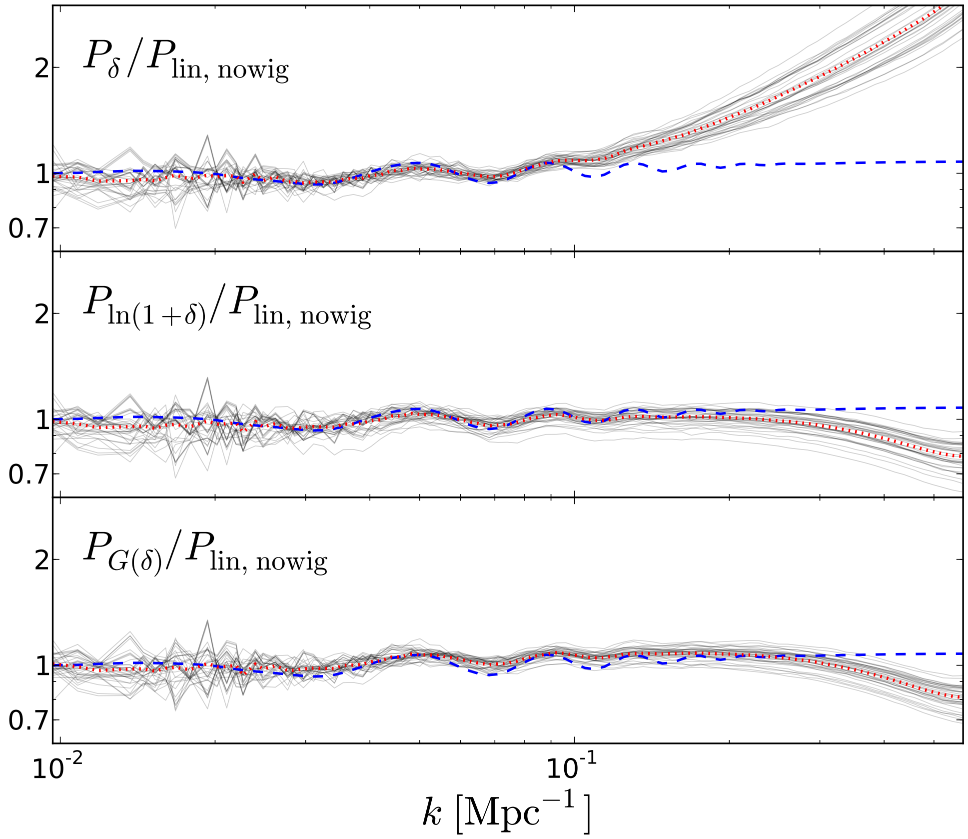


Figure 4.4: Neyrinck et al. (2011): Ratio of the linear power spectrum without wiggles $P_{\mathrm{lin,nowig}}$ with: (top) full non-linear power spectrum P_δ, (middle) the power spectrum of the logarithmic transformed density $P_{\ln(1+\delta)}$, (bottom) the power spectrum of the gaussianized density $P_{G(\delta)}$. The non-linear increase of power at scales $k = 10^{-1}$ Mpc^{-1} is efficiently resolved up to $k = 3 \cdot 10^{-1}$ Mpc^{-1}.

The results of Gaussianization are summarized in Figure 4.4. The power spectra of 37 simulations have been measured and compared to the linear power spectrum (covariance of the linear density constrast) of which the BAO wiggles have been removed ($P_{\mathrm{lin,nowig}}$). It can imposingly be seen that non-linear deviations of the power pectrum P_δ are recovered mostly when Gaussianizations $P_{\ln(1+\delta)}$ or $P_{G(\delta)}$ are computed.

Gaussianiazation is not uniquely used for LSS but also is applied in weak lensing studies (Joachimi et al., 2011; Yu et al., 2011) to obtain improved cosmological constraints. Also, linearised density fields can yield improved displacement and peculiar velocity fields (Kitaura et al., 2012b).

4.5 Inference analysis with ARGO

The work presented here has been performed with a fundamentally refined and extended implementation of the ARGO code as presented in Kitaura & Enßlin (2008); Kitaura et al. (2009, 2010). The philosophy of ARGO follows Bayesian statistical inference. We interpret the underlying dark matter density field as a probability distribution for which the resulting halo or galaxy distributions are statistical discrete representations. The purpose of an inverse analysis, as the one presented, is to infer the underlying distribution if only a statistical realization is known.

Obviously, galaxies/halos only form where the dark matter density contrast is high enough so that the structures can collapse. Still, we can observe how the tracers follow the smooth dark matter distribution on large enough scales. This means, that the number density of tracers is prone to a statistical bias.

4.5.1 Bayesian analyses in cosmology

Bayesian inference analyses and reconstruction techniques are widely used in cosmology, especially LSS and CMB analyses (Bunn et al., 1996; Fisher et al., 1995; Zaroubi et al., 1995). Cosmological inference is about to undergo further development when joint inference studies are to be performed combining CMB (Doré et al., 2001), LSS and also upcoming weak lensing (Alsing et al., 2016; Bull et al., 2016; Heavens et al., 2016; Kitching et al., 2015).

Also the reconstruction of the large scale gravitational potential, from which the coherent peculiar velocities can be derived, are studied. This is the case of the integrated Sachs-Wolfe effect (see e.g., Granett et al., 2008), the kinematic Sunyaev-Zeldovich effect (see e.g., Hernández-Monteagudo et al., 2015; Planck Collaboration et al., 2016; Schaan et al., 2015), the cosmic flows (e.g. Branchini et al., 2012; Courtois et al., 2012; Heß & Kitaura, 2016; Kitaura et al., 2012c; Lavaux et al., 2010; Watkins et al., 2009), or the baryon acoustic oscillations (BAO) reconstructions (see e.g. Eisenstein et al., 2007; Padmanabhan et al., 2012; Ross et al., 2015). Environmental studies of galaxies also demonstrated to benefit from accurate density and velocity reconstructions (Erdoğdu et al., 2004; Nuza et al., 2014).

Nevertheless, all these studies are affected by redshift-space distortions and the sparsity of the signal, which must be handled carefully (McCullagh et al., 2016). Indeed, Seljak (2012) has pointed out that if not properly modeled, non-linear transformations on density fields including redshift-space distortions can lead to biased results. The majority of previous Bayesian density field reconstructions applied to galaxy redshift surveys did not correct for the anisotropic redshift-space distortions (see e.g. Jasche et al., 2010, 2015).

The RSD modelling is especially emphasized by the analysis presented in this thesis. A recent study suggested to measure the growth rate from density reconstructions (Granett et al., 2015). However, instead of correcting redshift-space distortions, these were included in the power spectrum used to recover the density field in redshift-space.

4.5.2 Defining the posterior for ARGO

This work is entirely focused on LSS inference and the challenges within this charge. In order to infer the underlying dark mater given a set of galaxy or halo (tracers hereafter) coordinates, we build up a statistical model that links the observed number density of the tracers to the sought dark matter density. In the Bayesian framework, we define posterior distributions whose samples are statistical representations of the quantity that we would like to infer. ARGO is designed to jointly infer several conditional probabilities densities in a Gibbs sampling scheme. The presented work will mainly focus on the inference of the the dark matter density and the real-space positions of the tracers as galaxies in redshift surveys are observed in redshift-space (see Section 2.9). We will also discuss ARGO's extensions, e.g. in Section 4.8, the selection function sampling to overcome the Kaiser-rocket effect. This effect is a systematic maladjustment of the radial selection when calculated from galaxies in redshift-space.

The relation of real-space positions $\boldsymbol{r}$ and observed positions in redshift-space $\boldsymbol{s}^{\mathrm{obs}}$ are obtained by the following mapping (see Section 2.9):

$$\boldsymbol{r} = \boldsymbol{s}^{\mathrm{obs}} - \left(\frac{\boldsymbol{v}(\boldsymbol{r}, a) \cdot \hat{\boldsymbol{r}}}{H(a)a} \right) \hat{\boldsymbol{r}} \,, \tag{4.48}$$

where $\hat{\boldsymbol{r}}$ is the unit velocity vector in line-of-sight direction and $\boldsymbol{v}(\boldsymbol{r}, a)$ the velocity w.r.t. position $\boldsymbol{r}$ and scale factor a.

Having defined this mapping we can now use velocity and real-space positions as synonyms in our reconstruction model, as we will need the velocity to infer the real-space positions and vice versa. The joint posterior for the N_{c} dimensional linear density field δ_{L} and real-space positions then is defined as:

$$\mathscr{P}_{\mathrm{joint}} \left(\boldsymbol{\delta}_{\mathrm{L}}, \{\boldsymbol{r}_1 ... \boldsymbol{r}_N\} | \{\boldsymbol{s}_1^{\mathrm{obs}} ... s_N^{\mathrm{obs}}\}, w(\boldsymbol{r}), \mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}}), \boldsymbol{p}_{\mathrm{B}}, f_{\Omega} \right) \,, \tag{4.49}$$

where each $\boldsymbol{r}_i \in \mathbb{R}^3$ is the three dimensional vector of the volume element i

This posterior gives a joint probability for the linear density field $\boldsymbol{\delta}_{\mathrm{L}}$ and N three dimensional velocity vectors $\{\boldsymbol{v}_1 ... \boldsymbol{v}_N\}$ thus obtaining the real-space positions $\{\boldsymbol{r}_1 ... \boldsymbol{r}_N\}$, given

- **observed redshift-space positions** $\{\boldsymbol{s}_1^{\mathrm{obs}} ... s_N^{\mathrm{obs}}\}$;

- **completeness** $w(\boldsymbol{r})$,
  a scalar between 0 and 1, denoting observational effects in a given region of the sky. Detailed discussion follows in Section 4.5.4;

- **cosmic covariance** $\mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}})$ of the linear density field,
  defined as $\mathbf{C}_{\mathrm{L}} = \langle \boldsymbol{\delta}_{\mathrm{L}}^{\dagger} \boldsymbol{\delta}_{\mathrm{L}} \rangle$, depending on a set of cosmological parameters $\boldsymbol{p}_{\mathrm{C}}$ (expressed as a vector) within a ΛCDM framework;

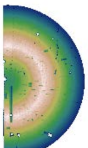
- **bias parameters** $\boldsymbol{p}_{\mathrm{B}}$,
  the vector of the bias parameters described in Section 2.8;

- **growth rate** $f_{\Omega} \equiv \dfrac{\mathrm{d} \log D(a)}{\mathrm{d} \log a}$,
  the logarithmic derivative of the logarithm of the growth factor $D(a)$ w.r.t. the scale factor a.

The direct inference from Equation 4.49 is not viable. That is why a hierarchical model of conditional probabilities as described in Section 4.1.4 is build up, so that the resulting inference analysis can be done in a Gibbs sampling scheme shown in Section 4.3.2.

$$\boldsymbol{\delta}_{\mathrm{L}} \quad \curvearrowleft \quad \mathscr{P}_{\delta} \left(\boldsymbol{\delta}_{\mathrm{L}} | N_{\mathrm{G}} \left(\boldsymbol{r} \right), w(\boldsymbol{r}), \mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}}), \boldsymbol{p}_{\mathrm{B}} \right) \,, \tag{4.50}$$

$$\{\boldsymbol{r}_1 ... \boldsymbol{r}_N\} \quad \curvearrowleft \quad \mathscr{P}_{r} \left(\{\boldsymbol{r}_1 ... \boldsymbol{r}_N\} | \{\boldsymbol{s}_1^{\mathrm{obs}} ... s_N^{\mathrm{obs}}\}, \{\boldsymbol{v}_1 ... \boldsymbol{v}_N\} \left(\boldsymbol{\delta}_{\mathrm{L}}, f_{\Omega} \right) \right) \,, \tag{4.51}$$



where each velocity $\boldsymbol{v}_i$ itself is a function of of the linear density contrast and the growth rate.

Due to statistical advantages we will define all physical quantities in ARGO w.r.t. to the linear density contrast (more details in Section 4.6.3) as this is the signal we are actually sampling. The only exception is the bias relation described in Section 4.6.1, which has to be defined for the non-linear density contrast.

4.5.3 Running ARGO on a mesh grid

We run ARGO on a cubic equilateral mesh grid of N_c ($\sqrt[3]{N_\mathrm{c}}$ per axis) cells. We use the nearest grid point (NGP) algorithm to assign each of the N_G tracers to a particular cell. ARGO reconstructs the underlying dark matter density and also the velocity on this mesh grid, meaning that we will have to solve equations 4.50 and 4.51 for N_c density and $3 \times N_\mathrm{c}$ velocity parameters. We will elaborate the Hamilton equations of motion for each cell identified with index $i \in 0, ..., N_\mathrm{c} - 1$.

4.5.4 Window handling in ARGO

The radial selection function and the angular completeness, both discussed in Section 3.2, are the two constituents for ARGO to be combined to the so-called *window function* w. We need to project the angular completeness (see Figure 3.2) into each cell of a three dimensional cubic volume used for the reconstructions. This is done by transforming equatorial coordinates (right ascension α and declination δ) into comoving Cartesian coordinates x, y, z with

$$\begin{aligned} x &= \cos\alpha\cos\delta\,, \\ y &= \sin\alpha\cos\delta\,, \\ z &= \sin\delta\,, \end{aligned} \tag{4.52}$$

at unit distance. Now having done this transformation we still need to evaluate the completeness in line of sight, as Equation 4.52 only holds for a sphere with unity radius. If no further information is given, we would simply assign all cells in the cubic volume with the completeness value of the angular mask's line-of-sight projection, indiscriminately of their radial distance r. We also need to average the completeness of the projections within one cell, as several points of the angular completeness mask mesh can correspond to the same cell in line-of-sight projection of the cubic box. This is due to different resolutions of the angular mask and the cubic box. The resulting projection in a $z - y$ (top left) and $z - x$ plane (top right) can be seen in Figure 4.5 for a cubic box with 512 cells per side. These projections now assumes the completeness entirely depending on the angular coordinates each cell is projected to, without taking its radial distance into account. The next step is to also consider the variation of the completeness due to the radial selection function (RSF). This is done by multiplying the value of the RSF given the cell's radial distance r to each cell. So the final *window function*, including angular mask and the RSF is shown for $z - y$ (bottom left) and $z - x$ (bottom right) plane in Figure 4.5.

We will later see the advantage of this bipartite procedure as once the angular projection is done on a mesh grid, any RSF can be multiplied trivially afterwards. This is a key advantage for the selection function sampling, shown in Section 4.8.

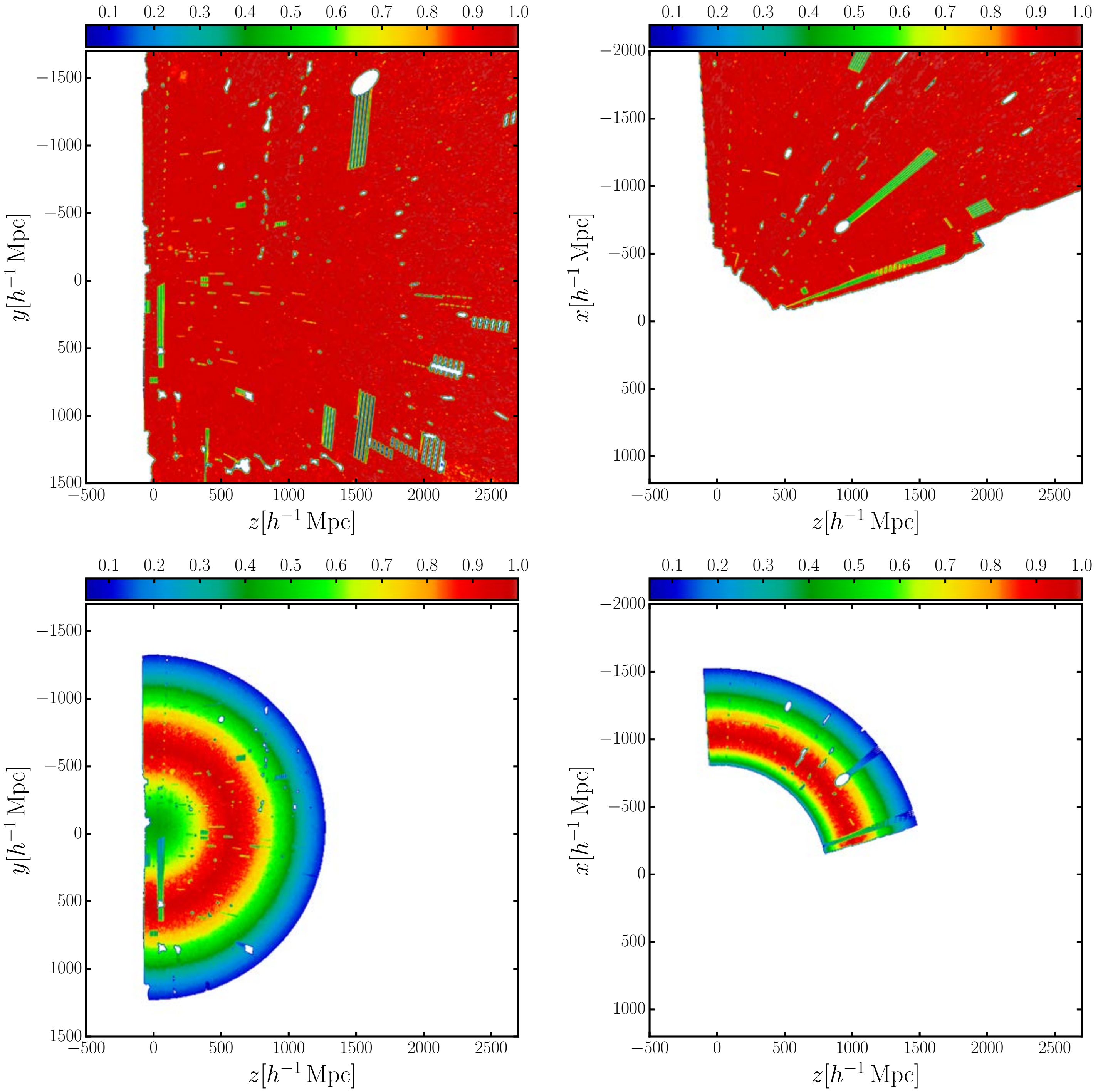
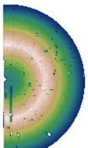


Figure 4.5: Top panel showing two slices of the pure angular completeness mask projected into a cubic box with 512 cells per side. On the bottom panel the same slices are shown but with the RSF multiplied into. Regions that have not been observed or vetoed are drawn in white.



4.5.5 Galaxy weights handling in ARGO

Besides the mask and window handling we also need to apply weights to BOSS galaxies. These weights are described in Reid et al. (2016) and correct for dependencies between target density and both, stellar density and seeing. They are labeled and described in the following:

- **Close Pairs:**
  For separations below 62″ the pairs of targets are too close to obtain both spectra. In this case the nearest neighbor is weighted with w_{cp}.
- **No Redshift:**
  For 1.8% of the CMASS targets, the spectroscopic pipeline fails to obtain a robust redshift. We do not necessarily expect these to be distributed randomly with respect to e.g., plate center or redshift, and so we again adopt a nearest neighbour upweighting scheme to account for these objects. Redshift failure galaxies were permitted to be upweighted because of a nearest neighbour fiber collision. We therefore transfer the total weight to the nearest neighbour of the redshift failure, incrementing a weight w_{noz}.
- **Angular Weights:**
  We remove non-cosmological fluctuations in CMASS target density with stellar density and seeing and condense these weights to the total systematic effect $w_{\mathrm{systot}} = w_{\mathrm{see}} \times w_{\mathrm{star}}$.

The total weight for object i can now be written as

$$w_{\mathrm{tot},i} = w_{\mathrm{systot},i}(w_{\mathrm{cp},i} + w_{\mathrm{noz},i} - 1)\,. \tag{4.53}$$

4.6 Density sampling

To set up a Bayesian framework we define a prior and a likelihood whose product then is proportional to the posterior we want to sample from. As mentioned in Section 4.3, it is not required to explicitly define the normalization of the posterior. This is because we only want to draw samples from this posterior without having to know the exact probability of each drawn sample (implicitly this comes with the density of the drawn samples). We define the total posterior for density sampling (Equation 4.50) up to a normalization as:

$$\begin{aligned}\mathscr{P}_\delta\left(\boldsymbol{\delta}_{\mathrm{L}}|N_{\mathrm{G}}\left(\boldsymbol{r}\right),\boldsymbol{w}(\boldsymbol{r}),\mathbf{C}\left(\boldsymbol{p}_{\mathrm{C}}\right),\boldsymbol{p}_{\mathrm{B}}\right) \propto \\ \pi(\boldsymbol{\delta}_{\mathrm{L}}|\mathbf{C}\left(\boldsymbol{p}_{\mathrm{C}}\right)) \times \mathscr{L}(\boldsymbol{N}_{\mathrm{G}}|\mathscr{B}(\boldsymbol{\lambda}|\boldsymbol{\delta},\boldsymbol{p}_{\mathrm{B}}))\,,\end{aligned} \tag{4.54}$$

with $\boldsymbol{\lambda}$ being the expected number of galaxies per volume element averaged over all realizations of the Markov chain $\boldsymbol{\lambda} \equiv \langle \boldsymbol{\rho}_{\mathrm{G}} \rangle_{\mathrm{Chain}}$. The bias function $\mathscr{B}(\boldsymbol{\lambda}|\boldsymbol{\delta},\boldsymbol{p}_{\mathrm{B}})$ relates the the number count of galaxies to the underlying dark matter density given the bias parameters $\boldsymbol{p}_{\mathrm{B}}$. the bias function is in general non-linear, non-local and stochastic as discussed in Section 2.8.

4.6.1 Likelihood

The likelihood defines the model of the data. It is the probability to draw a certain count of galaxies per cell, $N_{\mathrm{G}i}$, given the expectation value of galaxy counts λ_i for this particular cell. This relation is modelled by a distribution function $f(\lambda, N)$. In general, the choice of the distribution function f is depending on the statistical problem and the random process that has to be modelled. With regard to cosmological reconstructions, the distribution function depends on the galaxy population and the bias function. A natural assumption for any point process is provided by the Poisson distribution, which we will consider in the following. So we define the likelihood for a set of N_{G} observations

$$L(\boldsymbol{N}_{\mathrm{G}}|\boldsymbol{\lambda}, \boldsymbol{p}_{\mathrm{B}}) = \prod_{1}^{N_{\mathrm{c}}} f(\lambda_i, N_{\mathrm{G},i}, \boldsymbol{p}_{\mathrm{B}}) \,. \tag{4.55}$$

In Section 4.6.2 however, we elaborate deviations from Poissonity and introduce over-dispersion via the negative binomial (NB) distribution, meaning that we seek to model a larger variance compared to the Poisson distribution.

We define the expected number of tracers per cell as:

$$\lambda_i \equiv \langle \rho_{\mathrm{G}} \rangle_i = f_{\bar{N}} w(r_i) \mathscr{B}(\delta_i) \,, \tag{4.56}$$

where we defined the global quantity $f_{\bar{N}}$. This is the normalization of our expectation value, ensuring a mean number density given a bias model as

$$f_{\bar{N}} = \frac{\bar{N}}{\int \mathscr{B}(\boldsymbol{\delta}) \mathrm{d}V/V} = \frac{\bar{N}}{\langle \mathscr{B}(\boldsymbol{\delta}) \rangle} \,. \tag{4.57}$$

$\bar{N}$ is the completeness corrected mean number density $\bar{N} = \frac{\sum_i N_{\mathrm{G},i}}{\sum_i w(r_i)}$. This normalization is important as it relates the observed number density of tracers to the dark matter density $\boldsymbol{\delta}$ via our bias model $\mathscr{B}$. As we can see in Equation 4.57, $f_{\bar{N}} = \bar{N}$, if $\langle \mathscr{B}(\boldsymbol{\delta}) \rangle = 1$, meaning that the deterministic bias, described in Section 2.8.1, is unity.

In ARGO, we use a scale dependent power law bias, written as

$$\rho_{\mathrm{G}} \approx \exp\left(\sum_{k=0}^{\infty} b_k \cdot \log(1 + \delta_k) \right) \,, \tag{4.58}$$

$$\rho_{\mathrm{G}} \approx (1 + \delta_{\mathrm{DM}})^b \,, \tag{4.59}$$

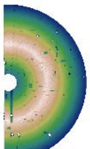
based on the work of Cen & Ostriker (1993) (Equation 2.108).

This model has been further elaborated in de la Torre & Peacock (2013) to increase the resolution in N-body simulations. Also Neyrinck et al. (2014) analyzed many halo mock catalogues from the *Multum In Parvo* (MIP) simulations (Aragon-Calvo, 2012) and found the same relation up to an exponential cut-off term $\exp\left(-\frac{\rho_{\mathrm{DM}}}{\rho_\epsilon}\right)$, where ρ_ϵ is a free parameter to describe the scale of the cut-off. The relation they found is shown in Figure 4.6.

We also account for the peak-background bias models, that we discussed in Section 2.6.6. According to this model, the high density regions collapse first and form galaxies and clusters. The gravitational collapse can only take place if the dark matter density fluctuation exceeds a certain threshold, called critical density δ_{C}. Thus, we expect no collapsed objects below a certain threshold, supported by the findings in Neyrinck et al. (2014). Accounting for this point, we also incorporate a *threshold bias* δ_{th}, described by a Heaviside step function Θ . Summarizing the above Equations 4.56 and 4.59 we write down the expected number of tracers per cell as

$$\lambda_i = f_{\bar{N}} w(r_i) (1 + \delta_i)^b \, \Theta(\delta_i - \delta_{\mathrm{th}}) \,, \tag{4.60}$$

where $\boldsymbol{\delta}$ is the full matter density field, unless it is subscripted with capital L, in that case $\boldsymbol{\delta}_{\mathrm{L}}$ denotes the linear density field. The variable b is the power law bias parameter.



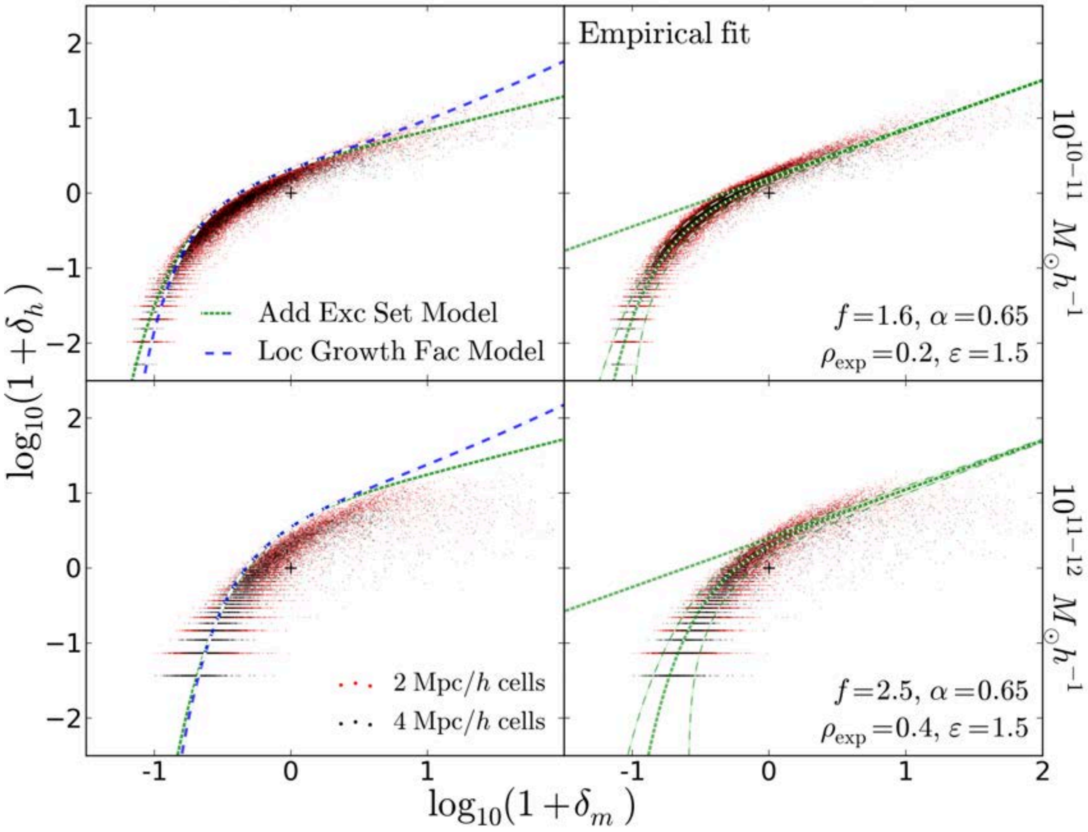


Figure 4.6: Neyrinck et al. (2014): Scatter plots of the MIP-ensemble-mean halo log-density versus the matter log-density, for two mass ranges of halos. Each dot represents a $2\,h^{-1}$ Mpc (red) or $4\,h^{-1}$ Mpc (black) cubic grid cell, with an NGP-estimated density. In the right-hand column, empirical power law-plus-exponential fits to the curve are shown in dotted curves, with and without the exponential. The lighter, dashed curves show vertical one-σ standard deviations in δ_h away from the mean, assuming Poisson statistics with mean given by the fit. The left-hand column shows theory curves from the additive-excursion-set (AES) and local-growth-factor models. The black crosses indicate the origin, $(0, 0)$.

4.6.2 Deviation from Poissonity

The Poisson probability density function

$$f_{\rm P}(\lambda|N) = \frac{e^{-\lambda}\lambda^N}{N!}\,.$$

has a well defined variance $\sigma_{\rm P}^2$ equal to its expectation value $\lambda = \sigma_{\rm P}^2$. Thus, the Poisson distribution is not capable to describe data that are over-dispersed, also called super-Poissonian distributions. Distributions that can describe over-dispersed random processes are e.g. the negative binomial (NB) and the gravitational thermodynamics (GT) (Saslaw & Hamilton, 1984) distribution, written as

$$f_{\rm NB}(\lambda, N, \beta) \quad = \quad \frac{\lambda^N}{N!}\frac{\Gamma(\beta+N)}{\Gamma(\beta)(\beta+\lambda)^N}\frac{1}{\left(1+\frac{\lambda}{\beta}\right)^{\beta}}\,, \tag{4.61}$$

$$f_{\rm GT}(\lambda, N, b) \quad = \quad \frac{\lambda}{N!}{\rm e}^{-\lambda(1-b)-bN}(1-b)\left[\lambda(1-b)+bN\right]^{N-1}\,, \tag{4.62}$$

with the Gamma-function $\Gamma(n+1) = n!$.

The parameters β and b are connected to the expectation value λ and the variance σ^2 by $\beta = \lambda^2/(\sigma^2 - \lambda)$ and $b = 1 - \sqrt{\lambda/\sigma^2}$, respectively. This implies that the over-dispersion term shows a quadratic and a linear dependence of the expected halo number count λ for the NB $\sigma^2_{\rm NB} = \lambda + \lambda^2/\beta$ and the GT $\sigma^2_{\rm GT} = \lambda/(1-b)^2 = \lambda + \lambda\, b(2-b)/(1-b)^2$, respectively (Ata, Kitaura & Müller, 2014). To obtain a different dependence, one could take the NB expression and include a dependence of β on λ. For $\beta \propto \lambda$ we find that the NB and the GT PDFs can equally describe over-dispersion.

In the following we will focus on the NB PDF.

We compare the NB and GT distributions for different β and b factors with expectation value $\lambda = 4$ in Figure 4.7 with the Poisson distribution of the same λ .

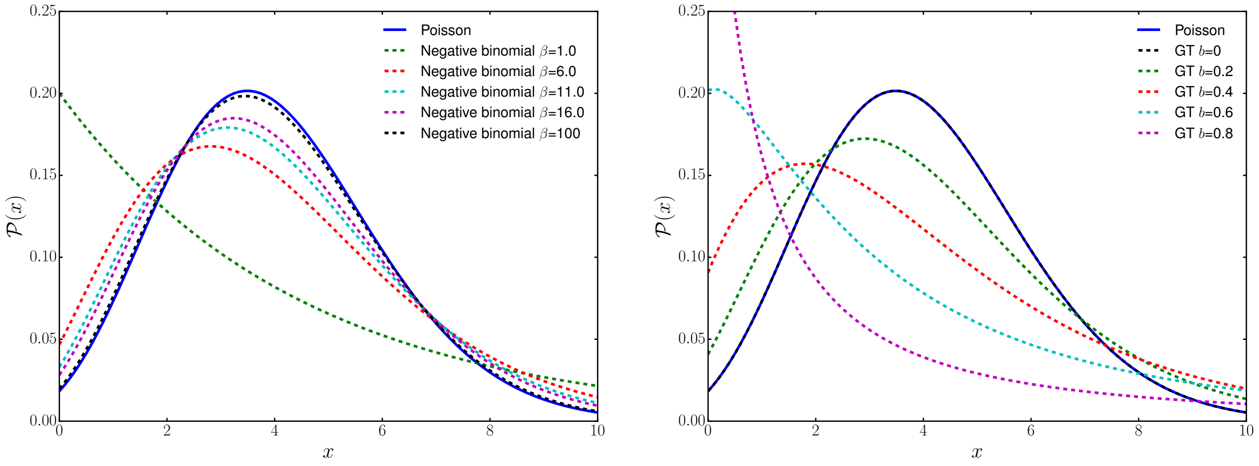


Figure 4.7: Comparison of the NB and GT with the Poisson distribution with expectations value $\lambda = 4$. Left: NB distribution for various β factors. For $\beta \to \infty$, the NB turns into a Poisson distribution. Right: GT distribution for different b parameters. $b = 0$ corresponds to the Poisson distribution.

The variance is always larger than the mean (equal for $\beta \to \infty$) for the negative binomial. Since the Poisson distribution determines the mean and the variance to be equal, it is unsuitable for data with greater variance than mean. Thus the negative binomial PDF is appropriate, as shown in Ata, Kitaura & Müller (2015). The necessity of an over-dispersed modelling of tracers is a direct result of the stochastic bias relation, as shown in Section 2.8.2. Also, it has been shown recently in Neyrinck et al. (2014), that if analyzing an ensemble of simulations with the same initial conditions but different halo finding criteria, the probability distribution of the density for a single cell is likely to deviate from Poissonity, shown in Figure 4.8. A more appropriate modelling needs to take over-dispersion into account. We can see from Figure 4.8 that the deviation from Poissonity is getting enhanced with a rising number of halos per cell. This emphasizes that over-dispersion is therefore resolution and type-of-tracer dependent.

We will discuss this point in more detail in Section 6.3, applying a NB reconstruction to the data.

4.6.3 Prior

The prior is the underlying structure formation model. As we work with the linearized density $\boldsymbol{\delta}_{\rm L}$ (Kitaura & Angulo, 2012), we assume Gaussian statistics. However many different structure formation models can be applied (Heß et al., 2013; Jasche & Wandelt, 2013; Kitaura, 2013; Kitaura et al., 2012c; Wang, 2014; Wang et al., 2013). For the linearized density $\boldsymbol{\delta}_{\rm L} = \log{(1+\boldsymbol{\delta})} - \boldsymbol{\mu}$, with $\boldsymbol{\mu} = \langle \log(1+\boldsymbol{\delta})\rangle$,

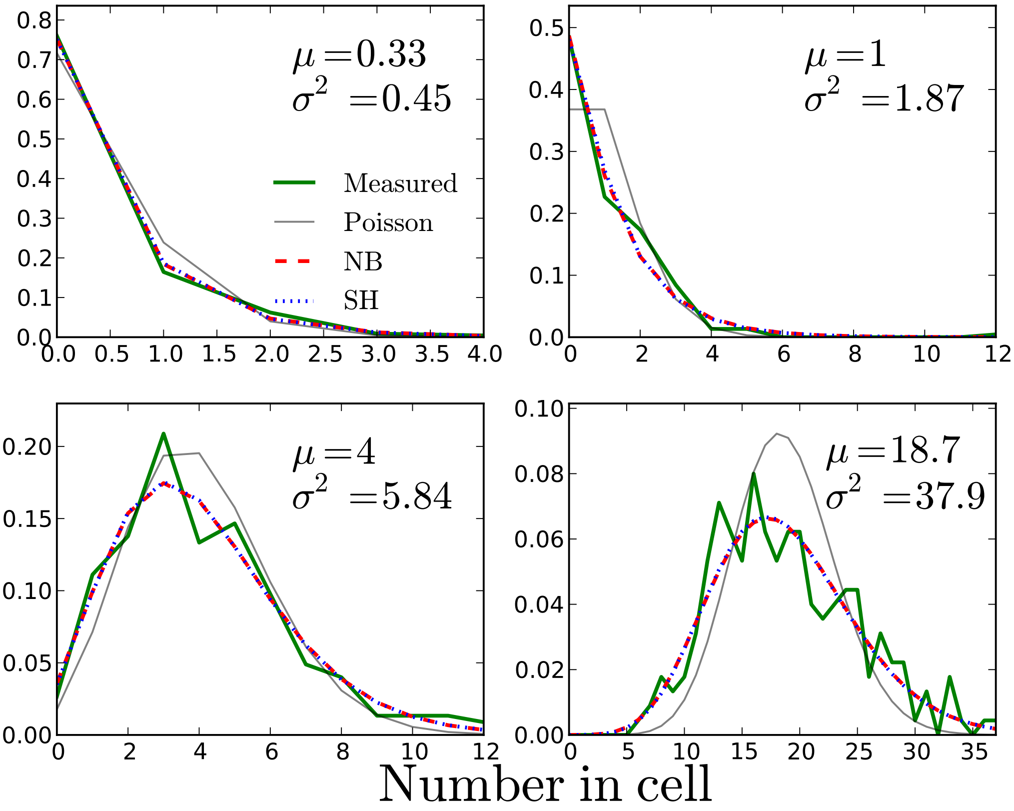


Figure 4.8: Neyrinck et al. (2014): Probability distribution functions across the MIP ensemble of the number of same mass halos in single cells of different mean halo densities. The PDFs (thick solid green) are wider than a Poisson distribution of the same mean (thin solid grey). Also shown are the Saslaw-Hamilton (SH, called GT in our case, dotted blue) and negative-binomial (NB, dashed red) distributions that model the super-Poissonity; both look like good fits. The PDF means and variances are written in each panel.

we write

$$\pi(\boldsymbol{\delta}_{\mathrm{L}}|\mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}})) = \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{c}}}\det(\mathbf{C}_{\mathrm{L}})}} \exp\left(-\frac{1}{2}\boldsymbol{\delta}_{\mathrm{L}}^{\dagger}\,\mathbf{C}_{\mathrm{L}}^{-1}\boldsymbol{\delta}_{\mathrm{L}}\right), \tag{4.63}$$

a multivariate Gaussian with zero mean and given covariance. The $N_{\mathrm{c}} \times N_{\mathrm{c}}$ linear covariance matrix $\mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}}) = \langle \boldsymbol{\delta}_{\mathrm{L}}^{\dagger}\boldsymbol{\delta}_{\mathrm{L}}\rangle$ is diagonal due to the absence of mode coupling processes and can also be expressed in form of the power spectrum.

4.6.4 Sampling with Hamiltonian Monte-Carlo

Reviewing the Hamiltonian Monte-Carlo (HMC) method in Section 4.3.3, we find the necessity to compute the negative logarithm of the posterior and its derivative w.r.t. to the sampled quantity, the linear density contrast $\boldsymbol{\delta}_{\mathrm{L}}$ in our case. Thus, we write

$$-\ln(\mathscr{P}) \quad = \quad -\ln(\pi) - \ln(\mathscr{L})\,. \tag{4.64}$$

The negative logarithm of the prior writes as

$$-\ln\pi(\delta_{\mathrm{L}}|\mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}})) = \frac{1}{2}\boldsymbol{\delta}_{\mathrm{L}}^{\dagger}\,\mathbf{C}_{\mathrm{L}}^{-1}\,\boldsymbol{\delta}_{\mathrm{L}} + c\,, \tag{4.65}$$

with all constants comprised into the term c.

Finally the negative logarithm of the likelihood for the Poisson and NB case can be computed. The Poisson case writes as

$$\begin{aligned}
\mathscr{L}_{\mathrm{P}}(\boldsymbol{N}_{\mathrm{G}}|\boldsymbol{\lambda}) &= \prod_{i}^{N_{\mathrm{C}}} \frac{\mathrm{e}^{-\lambda_i} {\lambda_i}^{N_i}}{N_i!}\,, && (4.66)\\
-\ln \mathscr{L}_{\mathrm{P}} &= \sum_i \lambda_i - N_i \ln \lambda_i - c\,. && (4.67)
\end{aligned}$$

We calculate the corresponding term for the NB distribution:

$$\begin{aligned}
\mathscr{L}_{\mathrm{NB}}(\boldsymbol{N}_{\mathrm{G}}|\boldsymbol{\lambda},\beta) &= \prod_{i}^{N_{\mathrm{C}}} \left(\frac{{\lambda_i}^{N_i}}{N_i!} \frac{\Gamma(\beta+N_i)}{\Gamma(\beta)(\beta+\lambda_i)^{N_i}} \frac{1}{\left(1+\frac{\lambda_i}{\beta}\right)^{\beta}} \right)\,, && (4.68)\\
-\ln \mathscr{L}_{\mathrm{NB}} &= \sum_i -N_i \ln \lambda_i + N_i \ln(\beta+\lambda_i) + \beta \ln(1+\lambda_i/\beta) - c. && (4.69)
\end{aligned}$$

We comprise terms that do not depend on the sampled variables again in the constant term c.

In total, the posterior function with the NB likelihood writes as

$$\begin{aligned}
&\mathscr{P}_\delta\left(\boldsymbol{\delta}_{\mathrm{L}}|N_{\mathrm{G}}\left(\boldsymbol{r}\right),\boldsymbol{w}(\boldsymbol{r}),\mathbf{C}\left(\boldsymbol{p}_{\mathrm{C}}\right),\boldsymbol{p}_{\mathrm{B}}\right) = \\
&\quad \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{c}}}\,,\det(\mathbf{C})}} \exp\left(-\frac{1}{2}\sum_{\alpha\beta}[\log(1+\delta_\alpha)-\mu]\,, C^{-1}_{\alpha\beta}[\log(1+\delta_\beta)-\mu]\right) \times \\
&\quad \prod_{i=1}^{N_{\mathrm{c}}} \left(\frac{[f_{\bar{N}} w(r_i)(1+\delta_i)^b]^{N_i}}{N_i!} \frac{\Gamma(\beta+N_i)}{\Gamma(\beta)(\beta+[f_{\bar{N}} w(r_i)(1+\delta_i)^b])^{N_i}} \frac{1}{\left(1+\frac{[f_{\bar{N}} w(r_i)(1+\delta_i)^b]}{\beta}\right)^{\beta}} \right)\,. \qquad (4.70)
\end{aligned}$$

The next step is now to compute the equations of motion to evolve the Hamiltonian in phase-space. Recalling Equation 4.32 as

$$\frac{\mathrm{d}p_i}{\mathrm{dt}} = -\frac{\partial\mathscr{H}}{\partial q_i} = -\frac{\partial U(q)}{\partial q_i}\,, \qquad (4.71)$$

shows that we need to build the gradient w.r.t. $\boldsymbol{\delta}_{\mathrm{L}}$. Equation 4.31 does not depend on $U(q)$ and therefore the term $\frac{\mathrm{d}q_i}{\mathrm{dt}} = \frac{\partial\mathscr{H}}{\partial p_i}$ can be estimated directly as $\sum_j M^{-1}_{ij} p_j$.

The gradient term for the prior writes as

$$\begin{aligned}
-\frac{\partial}{\partial\delta_{\mathrm{L},i}} \ln\pi &= -\frac{\partial}{\partial\delta_{\mathrm{L},i}} \ln\left(\frac{1}{\sqrt{(2\pi)^{N_{\mathrm{c}}}\det(\mathbf{C}_{\mathrm{L}}(\boldsymbol{p}_{\mathrm{C}}))}} \exp\left(-\frac{1}{2}\boldsymbol{\delta}_{\mathrm{L}}^{\dagger}\mathbf{C}_{\mathrm{L}}^{-1}(\boldsymbol{p}_{\mathrm{C}})\boldsymbol{\delta}_{\mathrm{L}}\right)\right)\,,\\
&= \frac{1}{2}\sum_{ij}(\delta^{\mathrm{K}}_{ik}(C^{-1}_{ij}\delta_{\mathrm{L},j}) + \delta^{\mathrm{K}}_{jk}(\delta_{\mathrm{L},i}C^{-1}_{ij}))\,,\\
&= \frac{1}{2}\left[\sum_j(C^{-1}_{kj}\delta_{\mathrm{L},j}) + \sum_i(\delta_{\mathrm{L},i}\,C^{-1}_{ik})\right]\,,\\
-\frac{\partial}{\partial\boldsymbol{\delta}_{\mathrm{L}}}\ln\pi &= \mathbf{C}_{\mathrm{L}}^{-1}\boldsymbol{\delta}_{\mathrm{L}}\,, \qquad (4.72)
\end{aligned}$$

where we used the Kronecker Delta δ^{K}_{ik}. Finally the derivative of the likelihood w.r.t. the sampled variable $\boldsymbol{\delta}_{\mathrm{L}}$ is computed with chain rule

$$\frac{\partial}{\partial \delta_{\mathrm{L},i}} = \left(\frac{\partial \delta_{\mathrm{L},i}}{\partial \delta_j}\right)^{-1} \frac{\partial \lambda_k}{\partial \delta_j} \frac{\partial}{\partial \lambda_k} \,. \tag{4.73}$$

So we obtain for the Poisson case:

$$-\frac{\partial}{\partial \delta_{\mathrm{L},i}} \ln \mathscr{L}_{\mathrm{P}} = b\lambda_i \cdot \left(1 - \frac{N_i}{\lambda_i}\right) \,. \tag{4.74}$$

Corresponding to Equation 4.74 we can write the derivative of these likelihood functions for the negative binomial case as

$$-\frac{\partial}{\partial \delta_{\mathrm{L},i}} \ln \mathscr{L}_{\mathrm{NB}} = b\lambda_i \cdot \left(\frac{1}{\frac{\lambda_i}{\beta} + 1} + \frac{N_i}{\beta + \lambda_i} - \frac{N_i}{\lambda_i}\right) \,. \tag{4.75}$$

The numerical implementation of the evolution of the Hamiltonian system is done by the leapfrog scheme described in Equations 4.33 and 4.34.

4.7 Velocity sampling

Different approaches have been proposed in the literature to recover the peculiar velocity field from galaxy distributions (Branchini et al., 2002, 2012; Gramann, 1993; Heß & Kitaura, 2016; Kitaura et al., 2012c; Lavaux et al., 2008; Wang et al., 2012; Yahil et al., 1991; Zaroubi et al., 1995), based on various density-velocity relations (see Bernardeau, 1992; Chodorowski et al., 1998; Jennings & Jennings, 2015; Kitaura et al., 2012b; Mohayaee & Tully, 2005; Nadkarni-Ghosh & Singhal, 2016; Nusser et al., 1991).

Our approach relies on an iterative Gibbs-sampling method, as proposed in Kitaura et al. (2012a) and presented in more detail in Kitaura et al. (2016).

We write total velocity field as the sum of the curl-free coherent bulk flow, which can directly be inferred from the large-scale density field within linear theory, and the random dispersion term $\boldsymbol{v}_{\mathrm{disp}}$ for each cell i as

$$\boldsymbol{v}_i \quad = \quad -f_{\Omega}(a)\, H(a)\, a\, \boldsymbol{\nabla}\nabla^{-2}\delta_i + \boldsymbol{v}_{\mathrm{disp}} \,. \tag{4.76}$$

We model the dispersion term as a randomly draw from a Gaussian with a particular standard deviation. The mapping between real-space and redshift-space positions for each individual galaxy is described by Equation 4.48. Within our Gibbs sampler, the velocity in step $j+1$ is derived from the density field that was sampled in the previous step j. Thus the velocity sampling to infer the real-space positions writes as

$$\boldsymbol{r}^{j+1} \quad = \quad \boldsymbol{s}^{\mathrm{obs}} - \left(\frac{\boldsymbol{v}\left(\boldsymbol{r}^j, a\right) \cdot \hat{\boldsymbol{r}}}{H(a)\, a}\right) \hat{\boldsymbol{r}} \,. \tag{4.77}$$

The peculiar velocity needs to be evaluated in real-space, requiring an iterative sampling scheme. The new sets of real-space positions are then used to update the number of tracers per cell $N_{\mathrm{G}}(\boldsymbol{r}^{j+1})$. These new positions are then again used as input for the density sampling scheme.

4.8 Selection function sampling

A novel sampling step we introduced in Ata et al. (2017) is the selection function sampling. As explained in Section 4.5.4, the radial selection function is the normalized number density of galaxies per distance bin. As the velocity sampling yields new positions mapped from redshift- to real-space coordinates, it is straight forward to incorporate these updated positions into the radial selection function. As shown in Figure 3.2, the radial selection function is written in terms of the distance r as

$$f(r) = 4\pi r^2 \frac{\mathrm{d}n}{\mathrm{d}V} \,. \tag{4.78}$$

The newly obtained real-space coordinates from step j are now used to update the radial selection function,

$$f(r)^{j+1} \curvearrowleft \mathscr{P}_f\left(f(r)|\boldsymbol{r}^j\right) , \tag{4.79}$$

where r is the distance to the origin.

4.9 Mean field sampling

The prior's mean field $\boldsymbol{\mu}$ needs to be iteratively adjusted as the linear density $\boldsymbol{\delta}_{\mathrm{L}}$ is sampled according to Equation 4.63 (Kitaura et al., 2012a). This can be done as the mean of the non-linear density contrast should vanish averaging over sufficiently large volumes, $\langle\boldsymbol{\delta}\rangle = 0$. Now within our lognormal assumption we can calculate $\boldsymbol{\mu}$ as

$$\begin{aligned}
\langle\boldsymbol{\delta}\rangle &= 0\,, && (4.80)\\
\langle\exp\left(\boldsymbol{\delta}_{\mathrm{L}} + \boldsymbol{\mu}\right)\rangle - 1 &= 0\,, && (4.81)\\
\langle\exp\left(\boldsymbol{\delta}_{\mathrm{L}}\right)\rangle \cdot \exp\left(\boldsymbol{\mu}\right)\rangle &= 1\,, && (4.82)\\
\boldsymbol{\mu} &= -\ln\langle\exp(\boldsymbol{\delta}_{\mathrm{L}})\rangle \,. && (4.83)
\end{aligned}$$

The mean field will get especially important when large parts of the reconstructed volume are empty and thus dominated by the prior function. We will discuss this issue in Section 6.3.1.

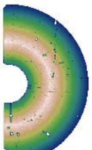
4.10 Normalization sampling

Equation 4.57 needs to be solved in each sampling step as it corrects the normalization of the expectation value and therefore accounts to link the correct number density of tracers with the inferred density contrast given the bias model. This writes

$$f_{\bar{N}} = \frac{\bar{N}}{\int \mathscr{B}(\boldsymbol{\delta})\mathrm{d}V/V} = \frac{\bar{N}}{\left\langle(1+\boldsymbol{\delta})^b\,\Theta(\boldsymbol{\delta} - \boldsymbol{\delta}_{\mathrm{th}})\right\rangle} \,. \tag{4.84}$$

4.11 Covariance sampling

We present two ways to estimate the covariance of the linear density contrast: firstly the direct sampling method, where the individual modes can be obtained from a χ^2-distribution and secondly by analysing the variance of each mode of the power spectrum after the samples have been drawn.



4.11.1 Direct sampling

According to the Wiener-Khinchin theorem the covariance matrix of a homogeneous (and periodic) Gaussian process is diagonal in Fourier space (Loredo, 2012). In Jewell et al. (2004); Wandelt et al. (2004) the authors showed that the signal covariance can be obtained from an inverse gamma distribution where each mode of the power spectrum samples individually. It can also been shown that for a particular variable shift the power spectrum modes can be sampled from a χ^2-distribution of the form

$$\mathscr{P}(x_m|\boldsymbol{\delta}_\mathrm{L}) = \frac{x_m^{\left(\frac{\beta_m}{2}-1\right)}}{\Gamma\left(\frac{\beta_m}{2}\right) 2^{\left(\frac{\beta_m}{2}\right)}} \exp\left(-\frac{x_m}{2}\right) , \tag{4.85}$$

where $x_m = \sigma_m/P_m$ is the power spectrum of the mode m over its spherical part σ_m (due to isotropy this term can be summed over) and $\beta = 2(\alpha + n_m/2 - 1)$, introducing the factor n_m that counts the number of different mode vectors belonging to the mth shell in Fourier space of our volume. α is determined by the prior choice, $\alpha = 0$ corresponds to a flat, $\alpha = 1$ to the Jeffrey's prior.

4.11.2 Post-sampling

Another way to estimate the covariance of the power spectrum is adopted from well tested methods that were applied to mocks catalogues (e.g. Mohammed & Seljak, 2014; Pearson & Samushia, 2016). The covariance is estimated by Equation 4.86, investigating how each individual mode of the power spectrum is varying over the full set of mocks.

$$C_{ij} = \langle P(k_i)P(k_j)\rangle - \langle P(k_i)\rangle\langle P(k_j)\rangle , \tag{4.86}$$

with $P(k_i)$ being the ith mode of the power spectrum regarding the wave vector $\boldsymbol{k}$. The covariance of the power spectrum is a very useful tool to estimate the variance of the cosmological parameters that are inferred from the power spectrum.

4.12 ARGO-Extensions

I discuss two extensions of ARGO. In Section 4.12.1 I describe a reconstruction scheme that will enable us to run ARGO on a *multi-tracer* problem. In Section 4.12.2 I discuss a possible *effective* method, in which I show how to incorporate a growth of structures in an additional bias term.

4.12.1 Multi-tracer analysis

A multi-tracer analysis is required, if the reconstructed volume is populated with heterogeneous tracers, e.g. galaxies with different biases, or if we combine several galaxy catalogues (Bull, 2016) that have been observed with different color-magnitude cuts, meaning that the selection function would be different. For these cases we elaborate a combined likelihood analysis for these tracers within one reconstruction scheme.

Let us assume a set of N galaxy samples $\boldsymbol{N}_{\mathrm{G}1}, \ldots, \boldsymbol{N}_{\mathrm{G}N}$, and each set i to have $\boldsymbol{N}_{\mathrm{G}i}$ galaxies.

We can write the joint problem of inferring the dark matter field δ conditioned on the different halo/galaxy samples by the following posterior PDF

$$\begin{aligned} &\mathscr{P}(\boldsymbol{\delta}|\boldsymbol{N}_{\mathrm{G}1}, \ldots, \boldsymbol{N}_{\mathrm{G}N}, \boldsymbol{p}_\mathrm{C}, \mathscr{B}(\boldsymbol{N}_{\mathrm{G}1}, \ldots, \boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta})) \propto \\ &\quad \pi(\boldsymbol{\delta}|\boldsymbol{p}_\mathrm{C}) \times \mathscr{L}(\boldsymbol{N}_{\mathrm{G}1}, \ldots, \boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}) , \end{aligned} \tag{4.87}$$

with $\mathscr{B}(\boldsymbol{N}_{\mathrm{G}1},\dots,\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta})$ being the joint bias.

If the samples have distinct biasing parameters, we can assume that each of the samples is conditioned on the underlying dark matter field only with:

$$\begin{aligned}&\mathscr{L}(\boldsymbol{N}_{\mathrm{G}1},\dots,\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}1},\dots,\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}))\propto\\&\quad\mathscr{L}(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta}))\times,\dots,\times\mathscr{L}(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}))\,,\end{aligned}\tag{4.88}$$

assuming the form of a product of different likelihood terms. Hence, we can write the posterior PDF as

$$\begin{aligned}&\mathscr{P}(\boldsymbol{\delta}|\boldsymbol{N}_{\mathrm{G}1},\dots,\boldsymbol{N}_{\mathrm{G}N},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta}),\dots,\mathscr{B}(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}))\\&\quad\propto\pi(\boldsymbol{\delta}|\boldsymbol{p}_{\mathrm{C}})\\&\quad\times\mathscr{L}(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta}))\times,\dots,\times\mathscr{L}(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta},\mathscr{B}(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}))\,.\end{aligned}\tag{4.89}$$

For the Hamiltonian sampler we need to compute the potential energy $U=-\ln\mathscr{P}$, which then writes as

$$\begin{aligned}&-\ln\left(\mathscr{P}\left(\boldsymbol{\delta}|\boldsymbol{N}_{\mathrm{G}1},\dots,\boldsymbol{N}_{\mathrm{GN}},\mathscr{B}\left(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta}\right),\dots,\mathscr{B}\left(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}\right)\right)\right)\\&\;=\mathrm{const}-\ln\left(\pi\left(\boldsymbol{\delta}|\boldsymbol{p}_{\mathrm{C}}\right)\right)\\&\;-\ln\left(\mathscr{L}\left(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta},\mathscr{B}\left(\boldsymbol{N}_{\mathrm{G}1}|\boldsymbol{\delta}\right)\right)\right)\\&\;\vdots\\&\;-\ln\left(\mathscr{L}\left(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta},\mathscr{B}\left(\boldsymbol{N}_{\mathrm{G}N}|\boldsymbol{\delta}\right)\right)\right)\,.\end{aligned}\tag{4.90}$$

This expression permits us to incorporate any additional galaxy sample and combine different galaxy catalogues with the presented method. The above calculations demonstrate that the dark matter field serves as a common denominator for different halo/galaxy samples and allows one to perform a multi-tracer analysis.

4.12.2 Effective bias scheme

I will discuss here an approach to describe cosmic growth of structures at different redshifts with a modified radial selection function in an effective description. The linear growth factor has been introduced in Equation 2.49. For a given reference redshift z_{ref} and a redshift z_i we write

$$G_i=\frac{D(z_i)}{D(z_{\mathrm{ref}})}\,.\tag{4.91}$$

Further, we will denote the completeness as function of redshift $w(z)$ in the following. Let $\mathscr{B}(\varrho_{\mathrm{M}}(z))$ be the bias function of the total matter density $\varrho_{\mathrm{M}}(z)$ at redshift z.

Now let us discuss different scenarios for which we write down the galaxy density $\varrho_{\mathrm{G}}(z)$ at redshift z.

Case A: Only considering effects of the selection function.
In this case we do not consider the growth function of each individual tracer, thus no evolution effects are modelled, but we stick to one specific reference redshift z_{ref}. Instead, we incorporate the different growth factors within the reconstruction into the selection function and consequentially into the window function $w(z_i)$. The underlying matter density is therefore at the reference redshift z_{ref}. All variables of this case are super-scripted with **sel** (for selection function).

Our bias model assumes the form

$$\begin{aligned}\varrho_{\mathrm{G}}^{\mathrm{sel}}(z_i) &= w^{\mathrm{sel}}(z_i) f_{\bar{N}}^{\mathrm{sel}} \mathscr{B}^{\mathrm{sel}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\,, && (4.92)\\ &= w^{\mathrm{sel}}(z_i) \frac{\bar{N}^{\mathrm{sel}}}{\langle \mathscr{B}^{\mathrm{sel}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle} \mathscr{B}^{\mathrm{sel}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\,. && (4.93)\end{aligned}$$

Case B: Only considering effects of growth function.
In this case we seek to eliminate the selection function (window) and assume that we can model all selection effects with a modified growth function. We super-script this case with **ev** (for evolution).

$$\begin{aligned}\varrho_{\mathrm{G}}^{\mathrm{ev}}(z_i) &= f_{\bar{N}}^{\mathrm{ev}} \mathscr{B}(\varrho_{\mathrm{M}}(z_i))\,, && (4.94)\\ &= \frac{\bar{N}^{\mathrm{ev}}}{\langle \mathscr{B}^{\mathrm{ev}}(\varrho_{\mathrm{M}}(z_i))\rangle} \mathscr{B}^{\mathrm{ev}}(\varrho_{\mathrm{M}}(z_i))\,. && (4.95)\end{aligned}$$

Case C: Combine evolution and selection function models.
As both upper cases are simplified models, we combine Equation 4.93 & 4.95. We will drop the variables of this combined model and thus use $\varrho_{\mathrm{G}}(z_i) = \varrho_{\mathrm{G}}^{\mathrm{sel+ev}}(z_i)$. The same naming holds for $\bar{N}$ and $\mathscr{B}$.

$$\varrho_{\mathrm{G}}(z_i) = w^{\mathrm{sel}}(z_i) \frac{\bar{N}}{\langle \mathscr{B}(\varrho_{\mathrm{M}}(z_i))\rangle} \mathscr{B}(\varrho_{\mathrm{M}}(z_i))\,. \qquad (4.96)$$

Case D: Effective description.
Now we incorporate the evolution and selection effects in an effective model at reference redshift z_{ref}. We denote the terms with superscript **eff** (for effective).

$$\begin{aligned}\varrho_{\mathrm{G}}^{\mathrm{eff}}(z_i) &= w^{\mathrm{eff}}(z_i) \frac{\bar{N}^{\mathrm{eff}}}{\langle \mathscr{B}^{\mathrm{eff}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle} \mathscr{B}^{\mathrm{eff}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\,, && (4.97)\\ \Rightarrow w^{\mathrm{eff}}(z_i) &= \varrho_{\mathrm{G}}^{\mathrm{eff}}(z_i) \frac{\langle \mathscr{B}^{\mathrm{eff}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle}{\bar{N}^{\mathrm{eff}} \mathscr{B}^{\mathrm{eff}}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))}\,. && (4.98)\end{aligned}$$

So far we have created idealized models. However, in order to link the models to a meaningful formalism, we need to **presume the following relations**:

$$\begin{aligned}\varrho_{\mathrm{G}}^{\mathrm{eff}}(z_i) &\overset{!}{=} \varrho_{\mathrm{G}}(z_i)\,, && (4.99)\\ \mathscr{B}^{\mathrm{eff}}(\varrho_{\mathrm{G}}(z_i)) &\overset{!}{=} \mathscr{B}(\varrho_{\mathrm{G}}(z_i))\,, && (4.100)\\ \bar{N}^{\mathrm{eff}} &\overset{!}{=} \bar{N}\,. && (4.101)\end{aligned}$$

Basically, these relations link the effective model to the combined one. Thus, we use Equation 4.98 inside Equation 4.96 and yield

$$w^{\mathrm{sel}}(z_i) \overset{!}{=} w^{\mathrm{eff}}(z_i) \frac{\mathscr{B}(\varrho_{\mathrm{M}}(z_i))}{\mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))} \frac{\langle \mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle}{\langle \mathscr{B}(\varrho_{\mathrm{M}}(z_i))\rangle}\,. \qquad (4.102)$$

Further, we put Equation 4.102 into Equation 4.96:

$$\varrho(z_i) = w^{\mathrm{eff}}(z_i) \bar{N} \frac{\mathscr{B}(\varrho_{\mathrm{M}}(z_i))^2}{\mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))} \frac{\langle \mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle}{\langle \mathscr{B}(\varrho_{\mathrm{M}}(z_i))\rangle^2}\,. \qquad (4.103)$$

Now we write down Equation 4.60 for the combined and effective model as

$$\mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}})) \;=\; \Theta(\delta_{\mathrm{M}} - \delta_{\mathrm{th}})(1+\delta_{\mathrm{M}})^{b_{\mathrm{eff}}}\,, \tag{4.104}$$

$$\mathscr{B}(\varrho_{\mathrm{M}}(z_i)) \;=\; \Theta(\delta_{\mathrm{M}} - \delta_{\mathrm{th}})(1+\delta_{\mathrm{M}})^{G_i b}\,. \tag{4.105}$$

In the last equation we also made the assumption that we can apply the linear growth factor G_i to the exponential of the matter density. This is valid for the linear expansion term.

The resulting model now is written as

$$\varrho(z_i) \;=\; w^{\mathrm{eff}}(z_i)\bar{N}\frac{\langle\mathscr{B}(\varrho_{\mathrm{M}}(z_{\mathrm{ref}}))\rangle}{\langle\mathscr{B}(\varrho_{\mathrm{M}}(z_i))\rangle^2}\Theta(\delta_{\mathrm{M}} - \delta_{\mathrm{th}})(1+\delta_{\mathrm{M}})^{2G_i b - b_{\mathrm{eff}}}\,. \tag{4.106}$$

The exponential $2G_i b - b_{\mathrm{eff}}$ is a free parameter in total, although G is well defined within linear theory.

4.13 Remarks

We showed in detail the principles of the ARGO code and its strategy to infer the underlying dark matter density and velocity field given a set of tracers. The merits of a Bayesian treatment are numerous. Each sample represents the statistical realization of the *full* posterior, sampled in different sequencing Gibbs sampling steps, shown schematically in Figure 4.9.

The density sampling is the most tedious step. In order to accept an iteration of the Hamilton Monte-Carlo, the densities of all cells within the reconstructed volume must be consistent with the posterior of the density, no matter whether the individual cell of the mesh grid is dominated by the prior or likelihood. The likelihood is constrained by the number counts of tracers and the bias parameters we are applying, deterministic and stochastic. However, the cells where the observed number count of tracers is zero totally rely on our structure formation model. This augmented field is dominated by the prior and will be consistent with the given covariance of the prior, which is an equivalent description for density fluctuations compared to the power spectrum. Only if the weights of prior and likelihood are balancing each other, a decent underlying dark matter density field can be sampled. Otherwise there will be an artificial gradient in the density that will be noticeable in the total power spectrum as a discontinuity.

The physical meaning of the normalization factor $f_{\bar{N}}$ is particularly important. If we would not iteratively sample the normalization, but leave $f_{\bar{N}}$ (Equation 4.57) as a free parameter or, as in Jasche et al. (2010), fix this number to the mean number density of tracers, our inferred dark matter density field would not correspond to the mean number of tracers given the bias model.

$$\langle(1+\boldsymbol{\delta})^b\rangle \neq 1 \qquad \text{if} \quad b \neq 1\,, \tag{4.107}$$

$$f_{\bar{N}} \neq \bar{N} \qquad \text{if} \quad b \neq 1\,. \tag{4.108}$$

The incorrect normalization then will correspond to an additional linear bias term and thus, the applied physical bias b would be meaningless, as it would be degenerate with the incorrect normalization. Only if $f_{\bar{N}}$ is computed from the density field with given bias parameter b, and then applied to the expectation value λ in this consistent way, the inferred dark matter density field will correspond to the given set of galaxy tracers and their abundance.

The effective bias scheme with the additional parameter b_{eff} has no physical meaning as opposed to the scale-dependent bias parameter b. However, the elaborated models shown in Section 4.12.2 may be a fruitful method if the exact growth or bias of the galaxy catalogue is not entirely clear. This still remains an effective description.

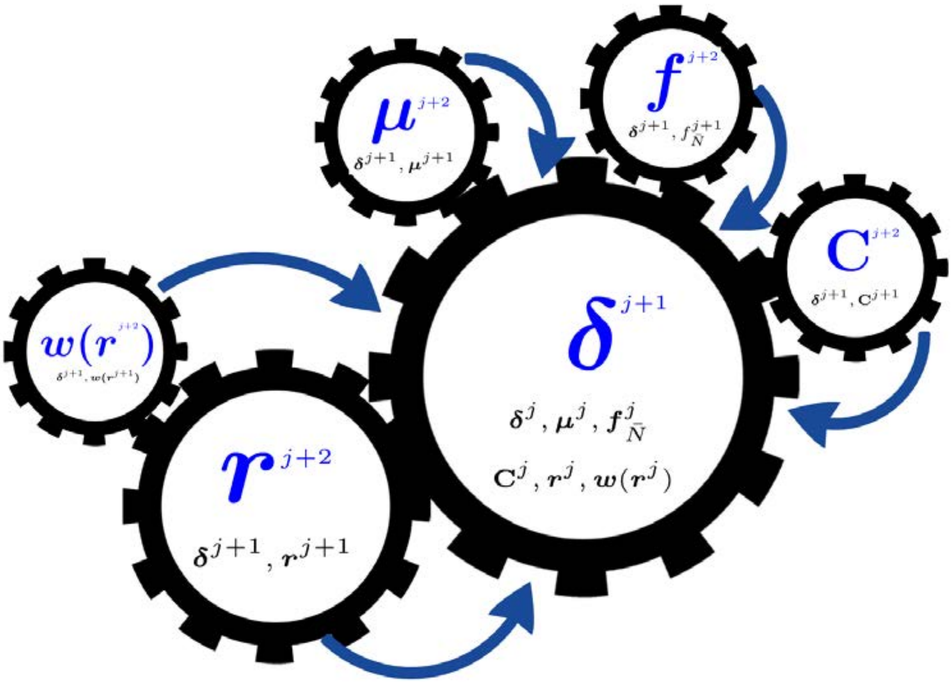


Figure 4.9: Working cycle of ARGO. In blue the inferred quantities are shown, depending on the quantities shown in black below. Gears that are not in direct contact still influence each other as ARGO follows a hierarchical sampling model.

5

Phase-Space Reconstructions with N-body Simulations

In this chapter I present the applications of ARGO that have been published in Ata et al. (2014, 2015) and Kitaura, Ata, Angulo et al. (2016). In each section I emphasize different aspects of our reconstruction strategy and the refinements we applied. The aim of these works is to firstly apply a cosmological density reconstruction using a scale-dependant and stochastic bias, and secondly to test our Bayesian redshift-space distortions corrections. In both cases we validate the outcomes of our reconstruction techniques by comparing them to the N-body simulations, on which the reconstruction was applied.

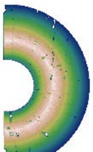
We run ARGO on the Bolshoi simulation, described in Section 3.3.2. It is a dark matter only N-body simulation with a cubic volume of $250\,h^{-1}$ Mpc per side. We choose a snapshot of the simulation at redshift $z = 0$.

5.1 Halo subsamples based on the Bolshoi simulation

In order to validate our method and evaluate the results of our reconstructions, we use the density field of all dark matter particles of the N-body simulation on the one side and two samples of halos created with the *bound-density-maximum* (BDM) halo finder (Section 3.3.4) on the other side. In this way we have full knowledge of the entire dark matter density field but also created catalogues of halos, that trace the dark matter field. The two subsamples distinguish themselves in the mass range of halos and the number density. Subsample S_1 is created by randomly picking 2×10^5 halos in the full mass range of 10^9 to 10^{15} $M_\odot$. As most of the halos of S_1 have low masses, the overall bias is smaller than one.

Subsample S_2 is created by demanding a lower mass limit of 3×10^{12} $M_\odot$, resulting in 1.6×10^4 halos, yielding a sample with bias larger than one. The cuts and the cumulative number of halos are shown in Figure 5.1.

The creation of S_2 is particularly interesting, as this subsample is not only down sampling the catalogue but also selects more massive halos.



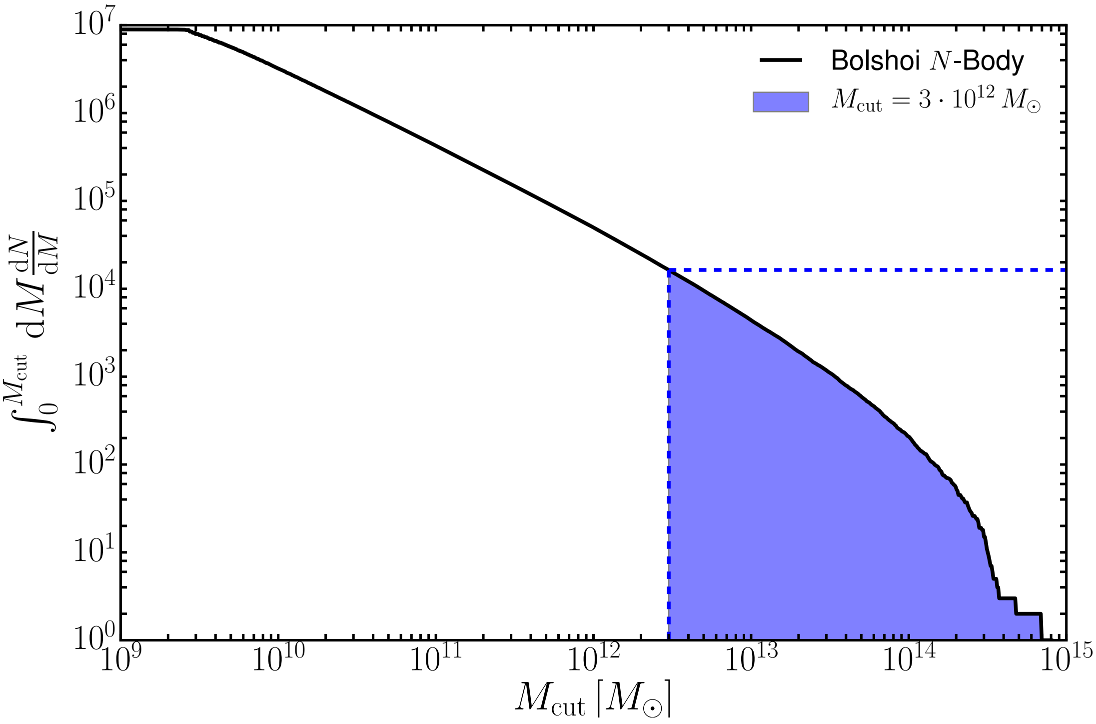


Figure 5.1: Cumulative mass function of the BDM halo catalogue based on the Bolshoi simulation at redshift $z = 0$. The number of halos are shown as a function of the mass cut M_{cut}. A particular choice of $M_{\mathrm{cut}} = 3 \cdot 10^{12}\ M_\odot$ is shown, yielding a number of 1.6×10^4 halos. The blue colored area corresponds to the subsample S_2 whereas S_1 contains the whole mass range of the Bolshoi simulation.

5.2 Density inference with ARGO from the Bolshoi simulation

We defined the expectation value of tracers per cell λ_i in Equation 4.60 as

$$\lambda_i = f_{\bar{N}} w(r_i)(1+\delta_i)^b \Theta(\delta_i - \delta_{\mathrm{th}})\,. \tag{5.1}$$

As discussed in Section 2.6.6, this model is in concordance with Press & Schechter (1974) and the peak background split formalism (Kaiser (1984), Bardeen et al. (1986), Cole & Kaiser (1989), Mo & White (1996), Sheth et al. (2001)), which permit the collapse of halos only above a certain density threshold. Kitaura et al. (2015) also showed that a threshold bias model is crucial to have a decent match of the three-point statistics of the density field.

We expect the threshold bias to be relevant for subsample S_2 as the very massive halos represent the peaks of the initial density field, discussed in Section 2.6.6. This has been illustrated in Figure 4.6. Halos do not form below a certain threshold of underlying dark matter density contrast (overdensity) δ_{DM}.

However, Equation 5.1 does only account for the deterministic bias. The stochastic bias, which is altering the variance of the number counts of tracers per cell, is modelled with the negative binomial (NB) distribution. The NB is able to describe distributions with the same mean but a larger variance compared to the Poisson distribution.

The reconstructions are performed on a regular cubic 128^3 mesh grid. We assign the dark matter particles of the Bolshoi simulation on a 128^3 mesh grid as well using the CiC and the halos with the NGP mass assignment scheme, respectively. We utilize the linear power spectrum created with CAMP (Lewis et al., 2000) as covariance matrix $\mathbf{C}_{\mathrm{L}}$ for the linear density field $\boldsymbol{\delta}_{\mathrm{L}}$ of the prior.

5.2.1 Description of run chains

We apply the ARGO code by running a series of Hamiltonian Monte-Carlo (HMC) chains with different likelihoods and bias settings on both subsamples:

1. Poisson likelihood and unity bias $\mathscr{B} = 1$ with subsample S_1,
2. Poisson likelihood and power law bias $\mathscr{B} = (1+\boldsymbol{\delta})^b$ with subsample S_1,
3. NB likelihood and power law bias $\mathscr{B} = (1+\boldsymbol{\delta})^b$ with subsample S_1,
4. NB likelihood and power law bias including thresholding $\mathscr{B} = (1+\boldsymbol{\delta})^b \Theta(\boldsymbol{\delta} - \boldsymbol{\delta}_{\mathrm{th}})$ with subsample S_2.

We further analyze the different models in two ways:

- **Firstly**, we compute the power spectra of the ARGO reconstructions. To have a realistic comparison, we compute the power spectrum of the BDM halo catalogue together with the power spectrum of the dark matter field from the Bolshoi simulation before the halo finder was applied. We refer to this field the N-body dark matter field in opposite to the halo density field;
- **Secondly**, we validate the ARGO reconstructions by comparing the inferred density fields with the N-body dark matter density field cell-to-cell. This will allow us to value the mean and scatter of the reconstructions.

5.2.2 Power spectrum analysis of subsample S_1

In Figure 5.2 the power spectra of the various density fields are shown. As we are focused on the precise modelling of the power spectra on small scales, we must be sure that the power spectra which we are comparing are not spoiled by observational effects. Therefore, we correct the N-body and the halo power spectrum for the mass assignment kernels (CiC for the N-body and NGP for the halo power spectrum) and also remove the shot-noise (see e.g. Jing, 2005). This makes sure that the power we measure on small scales is due to our physical bias model. Because of these operations, the power spectrum can be trusted up to approximately $k = 0.8\,h\,\mathrm{Mpc}^{-1}$, which corresponds to about 50% of the Nyquist frequency. The power spectrum of the N-body density field is shown as the solid black line. The power spectrum of the halo density (solid green line) has clearly a bias smaller than one. The Poisson reconstruction with unity bias (dashed magenta line) is following closely the halo power spectrum with no visible deviation. The Poisson reconstruction with power law bias and a bias parameter of $b = 0.79$ (dashed red line) shows an agreement with the N-body power spectrum only at the smallest wavenumbers. Already at $k = 0.2\,h\,\mathrm{Mpc}^{-1}$ we note a deviation of more than 10%, increasing steadily with higher wavenumbers. This lack of power can be modelled with the NB reconstruction including a power law bias (dashed blue line). We note an excellent agreement of the NB reconstruction with the N-body power spectrum up to $k = 0.95\,h\,\mathrm{Mpc}^{-1}$ within the scatter, where we used a bias parameter of $b = 0.92$ and deviation from Poissonity of $\beta = 1.47$. The higher bias parameter is necessary as the NB increases the variance also to some extent at large scale, but overproportionally on small scales, which can be seen in the power spectrum.

The scatter has been estimated by running each HMC chain for 12000 iterations, where the first 2000 so-called burn-in iterations have been discarded. The convergence behavior is estimated through the Gelman & Rubin (1992) test and covariance of the overdensity in a scatter plot, shown in Section 5.3.2. The 10000 remaining iterations are then used to estimate the reconstruction's mean and scatter in the power spectrum. These are shown in a one and two σ band for Poisson (beige and brown) and NB (cyan and dark blue). We can also see that the NB method leads to a higher variance in the power spectrum as compared to the Poisson reconstruction. This is expected due to the higher dispersion of the NB distribution. This result was firstly shown in Ata et al. (2015), in which the NB model and also the power law bias have been introduced within a cosmological reconstruction algorithm.

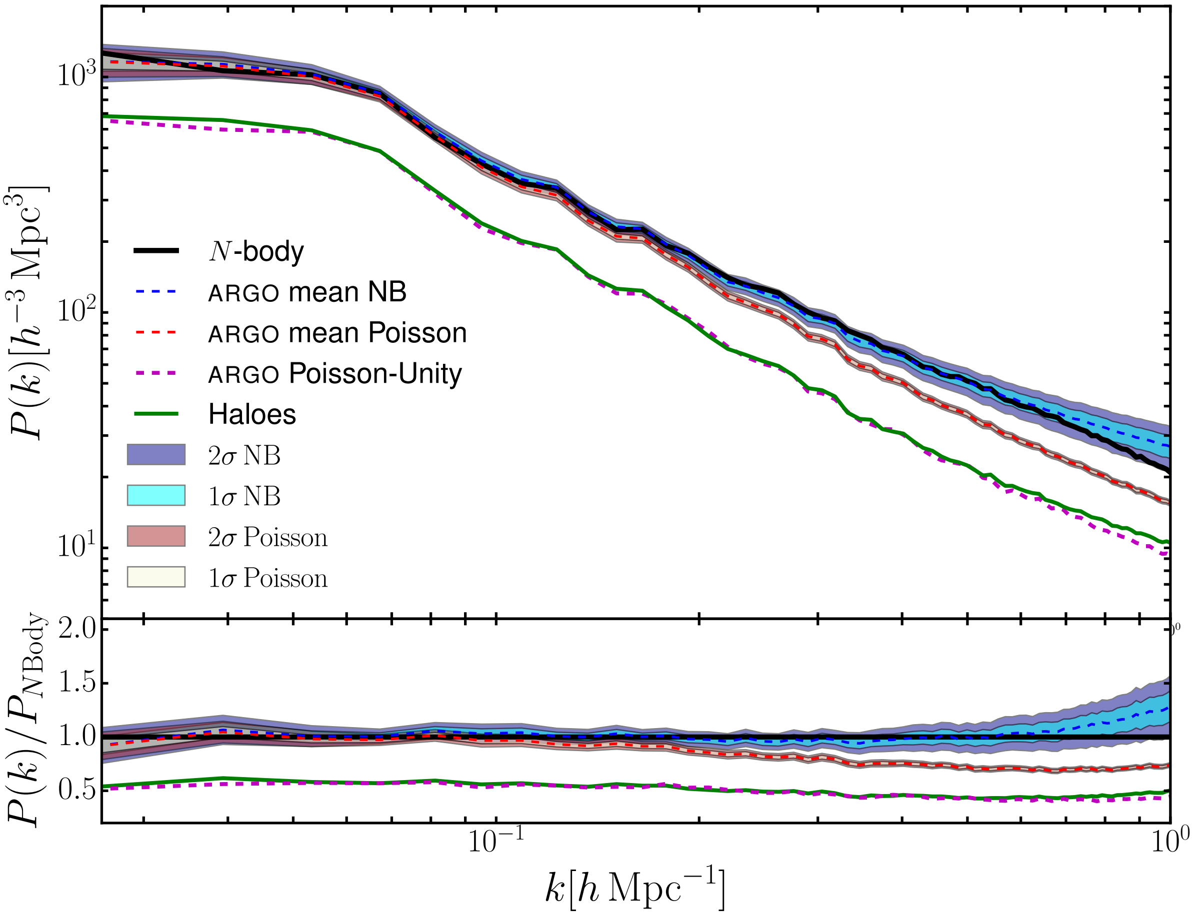


Figure 5.2: **Top:** Power spectra of the different density fields. N-body power spectrum shown as the solid black line. Dashed blue line corresponds to the mean of the ARGO reconstructions with NB likelihood and power law bias with one and two σ contours (cyan and dark blue). In red color the mean of the ARGO reconstructions with Poisson likelihood and power law bias is shown with negligible variance, plotted in brown and beige, respectively. The power spectrum of the BDM halos taken from the Bolshoi simulation is shown in solid green. The dashed magenta line represents the mean of the ARGO reconstructions with Poisson likelihood and unity bias. **Bottom:** Ratios of the above mentioned power spectra and the Bolshoi N-body power spectrum. The power spectrum of the N-body density field has been corrected for shot-noise and deconvolved with the mass assignment kernel.

5.2.3 Cell-to-cell analysis of subsample S_1

To further asses the accuracy of the shown reconstructions, we test these against the N-body density field within a cell-to-cell comparison shown in Figure 5.3. To have a fair comparison we smooth each catalogue with a Gaussian kernel with a smoothing radius of $r_\mathrm{S} = 6\,h^{-1}\,\mathrm{Mpc}$. This smoothing is applied in order to compensate for the different number densities of halos and dark matter particles. Each cell-to-cell correlation of an ARGO reconstruction (excluding the halo density field) is shown for the average of 10000 iterations. The halo-N-body density field correlation plot is reproducing the findings of Neyrinck et al. (2014), shown in Figure 4.6. As we only considered one realization of the halo field, the dispersion in our case is higher. We find that the halo density field and the Poisson reconstruction with unity bias show very similar cell-to-cell correlations when compared to the N-body density field, which are strongly biased towards high densities. The Poisson reconstruction with power

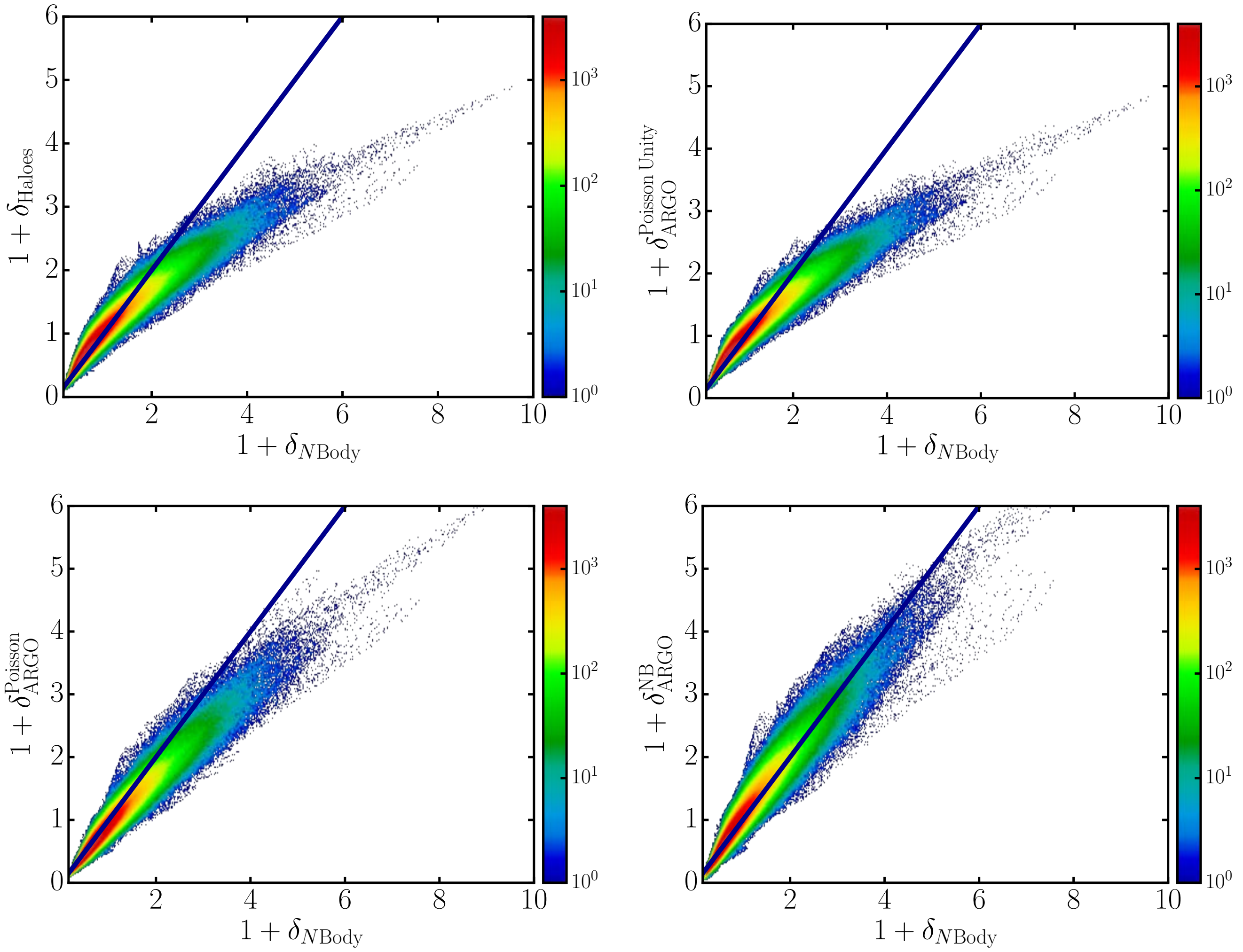


Figure 5.3: Cell-to-cell correlation comparison of the various density fields with the N-body density field after applying a Gaussian smoothing with radius $r_\mathrm{S} = 6\,h^{-1}\,\mathrm{Mpc}$. **Top left:** Comparison with the halo density field of S_1. **Top right:** Comparison to the reconstructed density with Poisson likelihood and unity bias, averaged over 10000 samples. **Bottom left:** Comparison to the reconstructed density with Poisson likelihood and power law bias, averaged over 10000 samples. **Bottom right:** Comparison to the reconstructed density with NB likelihood and power law bias, averaged over 10000 samples. The color bar represents the number density of cells.

law bias shows a more concurrent correlation, however, it is still clearly biased towards high densities. The correlation for the NB reconstruction shows an almost entirely unbiased shape up to the very high density cells. Another very interesting insight from the correlation plots is the scatter. While the Poisson reconstruction show a more narrow correlation behaviour, we note a higher scatter in the NB case. This finding is in agreement with the findings from the power spectrum analysis in Section 5.2.2.

5.2.4 Slice plots of subsample S_1

In Figure 5.4 slices of the different density fields are shown on a 128^3 mesh grid all with a thickness of $10\,h^{-1}$ Mpc In each case we show two different planes, $x-y$ and $y-z$, respectively. On the top, the slices of the N-body simulation are placed, where the smooth filamentary structure is visible. In the middle the halo density of the same slices are shown. Finally on the bottom we eye the reconstructed density slices obtained with a NB likelihood and a power law bias averaged over 10000 iterations. The slices for the Poisson reconstruction are omitted as they are visually not very different from the NB reconstructions. The averaging is an excellent tool to reduce the noise and emphasize structures with high number density. Thus, the filamentary structure of our reconstructions pop up more significantly in the averaged plot as compared to single reconstructions and also the noise is reduced (see Appendix A.3 Figure A.3).

5.2.5 Power spectrum analysis of subsample S_2

The purpose of the subsample S_2 has been explained in Section 5.2 and is an interesting scenario not only to test ARGO's performance for a lower number density of tracers but also validate the additional threshold bias, shown in Equation 5.1. This threshold is motivated e.g. by Neyrinck et al. (2014), as shown in Figure 4.6. Although for higher densities there is a logarithmic proportionality of the halo density and the N-body matter density, at densities lower that $\delta_m \approx 0.5$, the exponential cut-off is visible, meaning that halos do not form if the background matter density is too low. This exponential cut-off, shown in Figure 4.6, is approximated by a sharp Heaviside step function in our bias model.

In Figure 5.5 we find the corresponding power spectra plotted for the S_2 subsample. We can clearly see that subsample S_2 has a bias greater than one as compared the the power spectrum of the full Bolshoi N-body. The scatter of the reconstruction, which we run with $\delta_{\mathrm{th}} = 1.5$, is significantly larger than in Figure 5.2 for S_1. Further we found that overdisperion is negligible in this case, thus we set β to a very high value in the NB. The large scatter is because the number density of tracers is more than one order of magnitude smaller compared to S_1, so that the reconstruction is less stable in low density areas, and consequently the variance of the reconstruction is higher. Still we are mainly well within the one σ error band for the mean of our reconstructions. We find the threshold model to be a useful extension to our previously introduced bias model.

5.2.6 Cell-to-cell analysis of subsample S_2

Now we will also have a look at the cell-to-cell comparison regarding subsample S_2. This is shown in Figure 5.6 after a Gaussian smoothing with radius $r_{\mathrm{S}} = 6\,h^{-1}$ Mpc was applied. We show on the left the cell-to-cell for the halos of S_2 and on the right the averaged reconstruction obtained from 10000 samples. It is clearly visible that the halo density is tilted w.r.t. the N-body density, but opposed to the trend in S_1. This is in accordance to the bias found for S_1 to be less than one and for S_2 to be larger than one. Also, the spreading towards high densities is much larger compared to S_1. Another interesting realizing from this plot is the shift in the low density region. It can be seen that the density

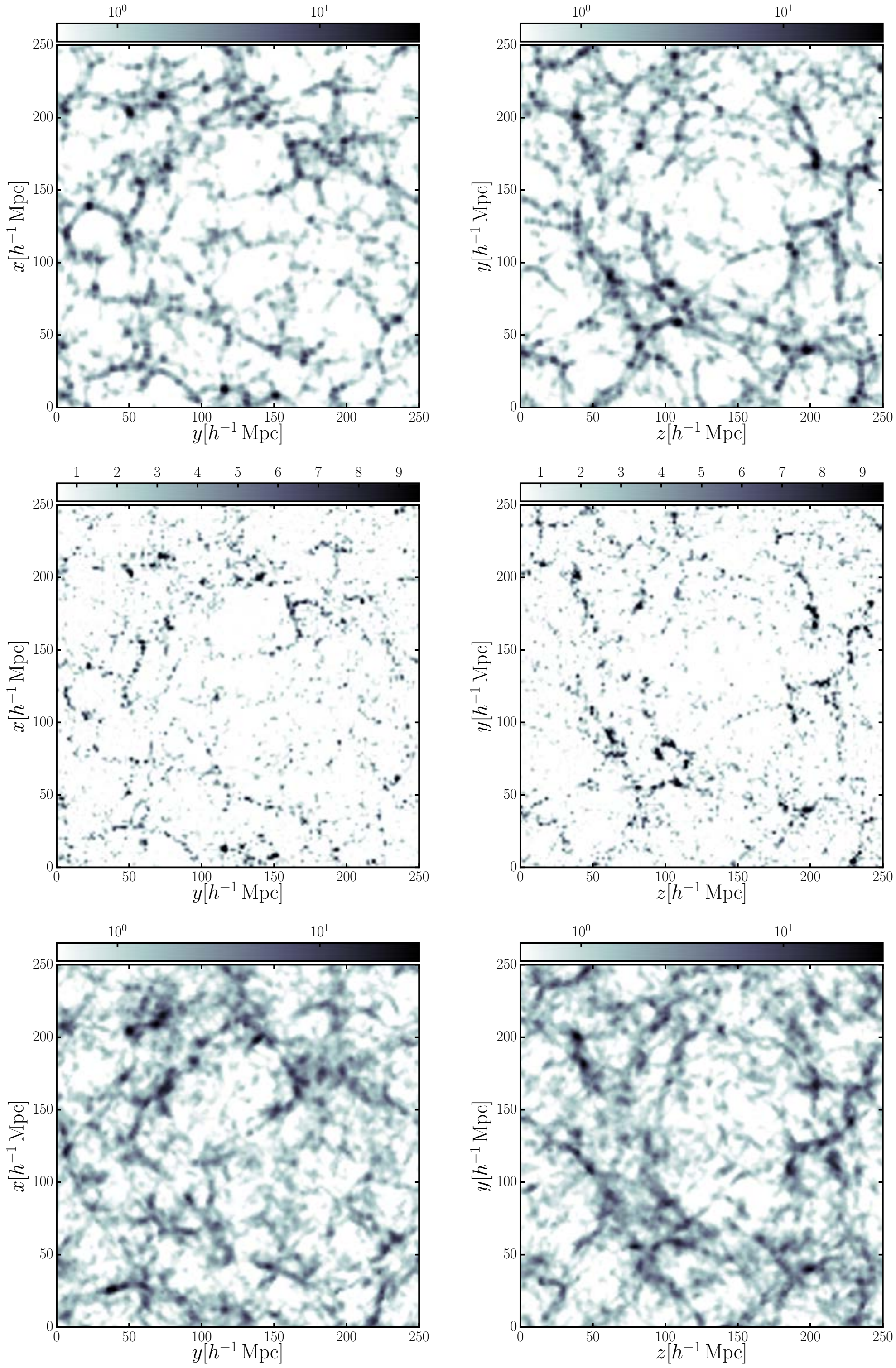


Figure 5.4: Slices of the $1+\delta$ density with a thickness of $10\,h^{-1}$ Mpc, shown for three different planes. **Top:** N-body density from the Bolshoi simulation, **Middle:** halo density of subsample S_1, **Bottom:** ARGO NB reconstruction with power law bias of subsample S_1, averaged over 10000 iterations.

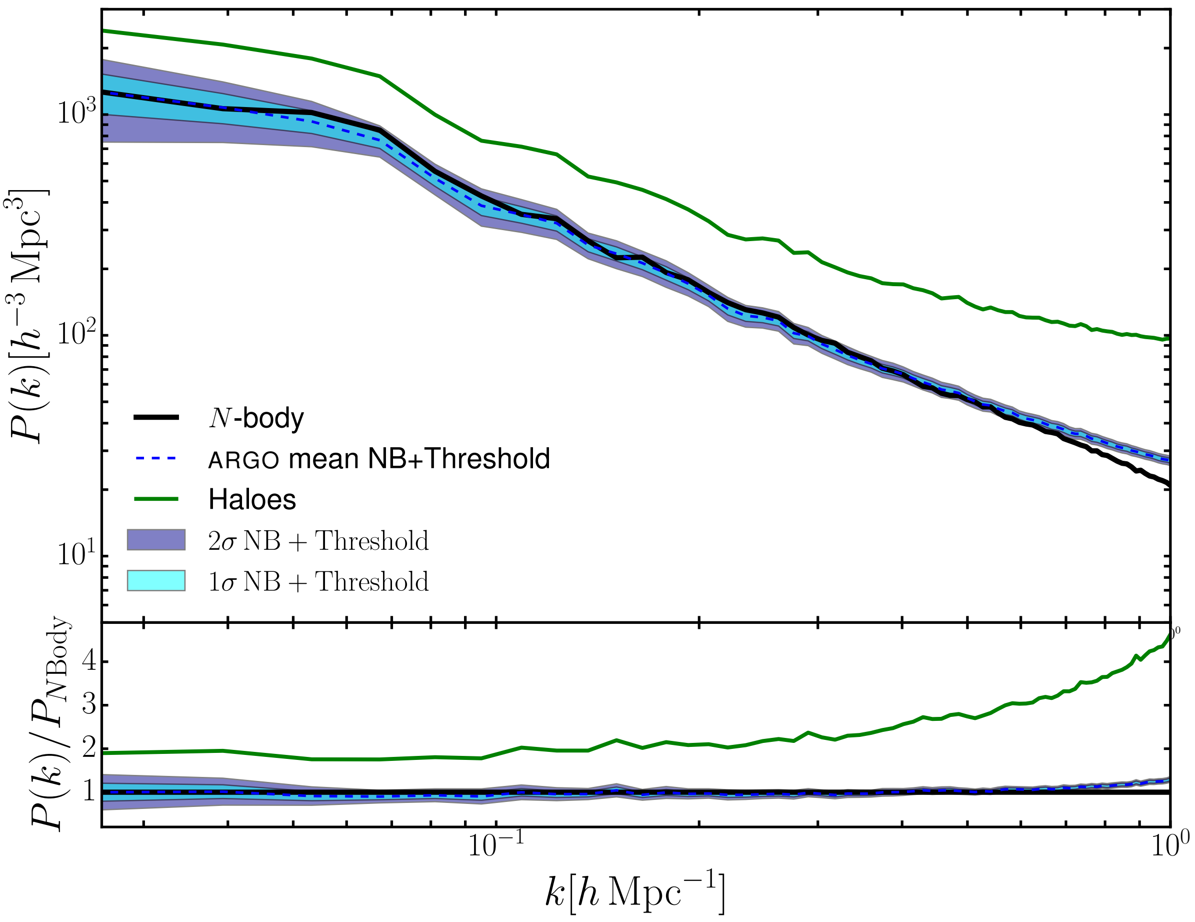


Figure 5.5: **Top:** Power spectra of the different density fields. N-body power spectrum shown in solid black line. In dashed blue the mean of the ARGO-reconstruction of the NB+Threshold model is plotted with one and two σ contours. The power spectrum of the BDM halos is shown in solid green. As we restricted the mass range of the halos in S_2, we now created a sample with bias greater than one. **Bottom:** Ratio of the reconstructed power spectra and the Bolshoi N-body power spectrum, that has been corrected for shot-noise and deconvolved with the mass assignment kernel.

of halos drops much faster than the N-body density when compared to the line through the origin. This is in agreement with the findings of Figure 4.6. The low and high density shifts are corrected within our reconstruction and thus yielding unbiased realizations of the underlying dark matter field. However, we also notice that the overall variance of the reconstruction in S_2 is higher than in both reconstruction cases of S_1, the Poisson and NB likelihood with power law bias. This is because of the lower number density of tracers.

5.2.7 Slice plots of subsample S_2

We visualize the results of the S_2 analysis in the same slices as done for subsample S_1. As we can see in Figure 5.7 the number density of halos is reduced significantly. Only the massive halos, that are located in the highest density regions, are contained in subsample S_2, due to the mass cut. The information for the low density filaments get lost, for this reason the reconstruction of S_2 is much

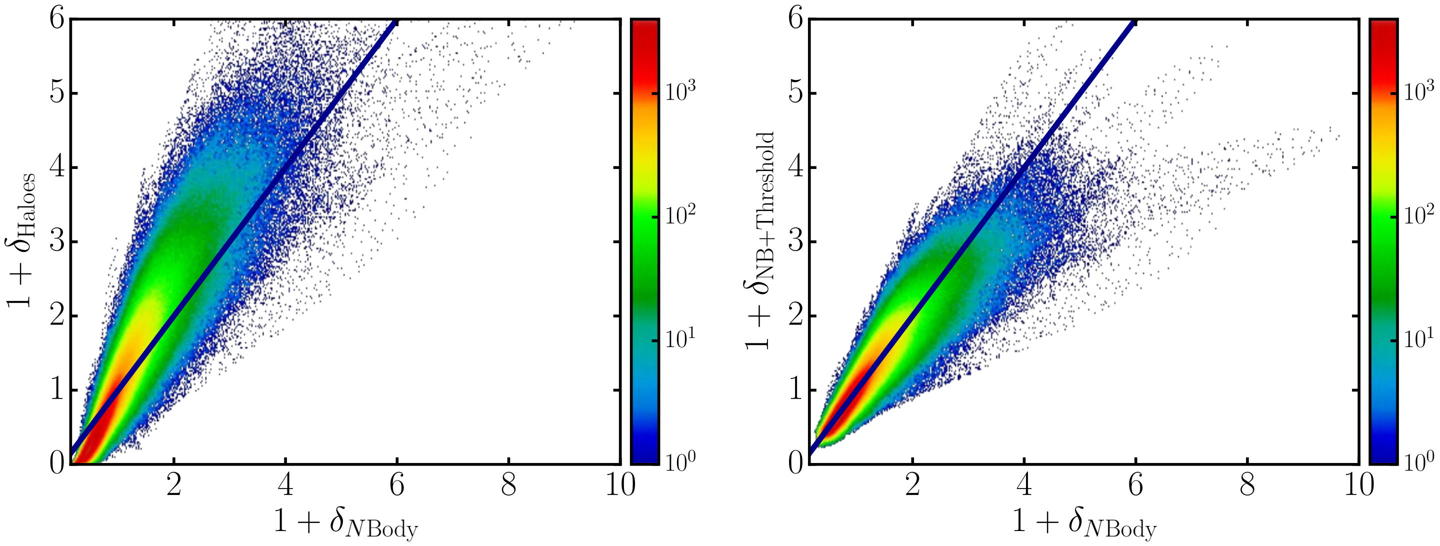


Figure 5.6: Cell-to-cell correlation of the density fields with the N-body density field after applying a Gaussian smoothing with radius $r_{\mathrm{S}} = 6\,h^{-1}\,\mathrm{Mpc}$. **Left:** Comparison with the halo density field of S_2. **Right:** Comparison with the reconstructed density obtained by a NB likelihood and power law bias including a thresholding term, averaged over 10000 samples. The color bar represents the number density of cells.

more challenging than for subsample S_1. What is also clearly visible through all slices is the higher variance in the reconstructions. The reconstructed structures in S_1 have been more evidently tracing the N-body structures, whereas in this case we see more variance in the mean of 10000 realizations. This is a natural statistical behavior as less tracers reduce the signal-to-noise ratio.

5.2.8 Remarks & Conclusion

We have shown the results from the application of ARGO to the associated halo catalogues derived from the Bolshoi N-body simulation (Ata et al. (2014, 2015)). The main emphasis of this analysis was to establish and test a sophisticated bias model for two different scenarios of halo tracers with different bias and also different number density. The novel bias model consist of two parts. The first part is the deterministic definition of the expectation value of a halo density, given the dark matter density described in Equation 4.60. Both, the power law ansatz and the threshold term in our bias model implemented in ARGO are novelties in a reconstruction algorithm. To implement the second part, the stochastic bias, we refined the likelihood model. Although a Poisson likelihood is a reasonable choice to deal with discrete number counts from an underlying smooth distribution, it is inadequate for populations whose dispersion is larger than the mean value. Our stochastic bias model therefore relies on the sampling from the NB distribution and adjusting the additional parameter β to control the overdispersion. We showed that within our bias model, the ARGO code is able to infer unbiased sampled of the underlying dark matter given both, a halo catalogue with bias smaller than one, using a power law bias and deviation from Poissonity, and also given a halo catalogue with bias greater than one, using an additional threshold bias with negligible deviations from Poissonity. We note however that the threshold bias and the power law bias may yield indistinguishable results regarding the two point statistics. This can be shown using subsample S_2. Nevertheless, Kitaura et al. (2015) showed that threshold bias represents an indispensable ingredient also to have a reasonable matching with the dark matter density in the three-point statistics. This is an ongoing development in ARGO and

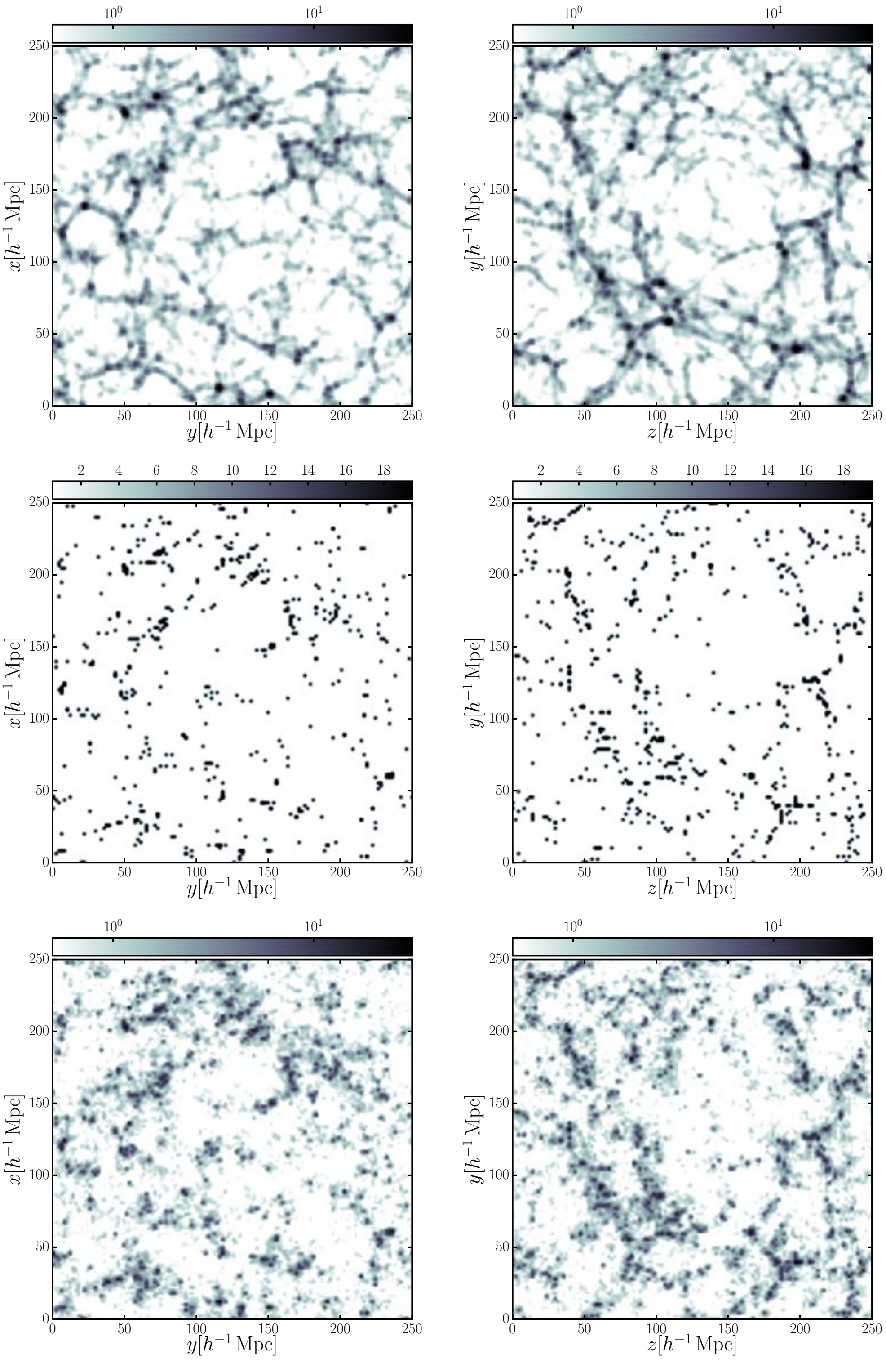


Figure 5.7: Slices of $1+\delta$ density with a thickness of $10\,h^{-1}$ Mpc shown for different planes. **Top:** N-body density of the Bolshoi simulation, **Middle:** halo density of S_2, **Bottom:** NB reconstructions with threshold bias, averaged over 10000 iterations.

may include a skewed prior and also non-local bias terms. Another remark concerns the numerical deficiency within the HMC method including the threshold term. The density threshold component of our bias model introduces a numerical instability, since the additional gradient terms diverge around the density threshold and also represent a rigorous constrain on the density field. Reviewing Equation 4.73 we need to consider the additional derivative $\frac{\partial \lambda}{\partial \delta}$, induced by the step function $\Theta(\delta - \delta_{\mathrm{th}})$. Up to a constant we write

$$\frac{\partial \lambda}{\partial \delta} \propto b(1+\delta)^{b-1}\Theta(\delta - \delta_{\mathrm{th}}) + (1+\delta)^{b}\delta^{\mathrm{D}}(\delta - \delta_{\mathrm{th}})\,, \tag{5.2}$$

where δ^{D} is the Dirac delta-function. Also, Equations 4.74 and 4.75 include logarithmic terms of the expectation value λ. Therefore the Θ-function will cause severe numerical problems. We overcome this numerical instability by introducing another Gibbs-sampling strategy and consider the step function to be constant within one density sampling step. If this is the case, we can neglect all terms including the derivative and logarithm of the threshold bias term, e.g. shown in Equation 4.74. The Gibbs strategy to sample the density field $\boldsymbol{\delta}$ then writes as

$$\boldsymbol{\delta} \curvearrowleft \mathscr{P}_{\delta}\left(\boldsymbol{\delta} | N_{\mathrm{G}}\left(\boldsymbol{r}\right), w(\boldsymbol{r}), \mathbf{C}\left(\boldsymbol{p}_{\mathrm{C}}\right), \boldsymbol{p}_{\mathrm{B}}\right)\,, \tag{5.3}$$

$$\Theta \curvearrowleft \mathscr{P}_{\Theta}(\Theta | \boldsymbol{\delta}, \delta_{\mathrm{th}})\,. \tag{5.4}$$

For simplicity we assume no uncertainty on the threshold term which means that $\mathscr{P}_{\Theta}$ is a Dirac delta-function. However this assumption may be further elaborated. Numerical tests show this method to be stable but the sampling itself is highly inefficient. The analysis of subsample S_2 with threshold bias appears to have a significantly higher rejection rate in the HMC sampling step by a factor of ten (see Section 5.3.2) and also equivalently increases the computational time by this factor.

Another aspect we did not mention in detail yet is addressing the aliasing effects. We write the exact equations for the inferred density per grid cell in Equation 4.70, which means that the reconstructed density field is not affected by noise as it is the case for the galaxy (halo) density field we assign to the grid. In this way we are not limited by modes that are beyond the Nyquist-frequency, called corner modes (e.g. Falck et al., 2016). Thus, the inferred density field does not need be deconvolved with a mass assignment kernel and already yields the correct number density per cell. Finally an interesting point worth mentioning in the cell-to-cell comparison is the low density regime of our shown reconstruction using subsample S_2. We note that ARGO reconstructs less low density cells as compared to the full N-body density field. This is a consequence of the threshold bias term, which is not the case for the run reconstructions with subsample S_1. However, an imperfect adjustment of the combination of power law bias and threshold bias parameter may have exaggerated this effect, as we already noted a degeneracy in the two-point statistics. However, a detailed study including the higher-order statistics will clearly help to understand this fact.

5.3 Velocity sampling with the BigMultiDark Simulation

After we discussed the density inference with ARGO in detail, we now focus on the self-consistent description of the velocity sampling techniques within our algorithm. The analysis presented in Kitaura, Ata, Angulo et al. (2016) aims to correct for redshift-space distortions (RSDs), using a joint sampling of the density and the velocity field with ARGO, applied on a mock galaxy catalogue at redshift $z = 0.57$ (Rodríguez-Torres et al., 2016) based on a snapshot of the BigMultiDark (BigMD) simulation, that we introduced in Section 3.3.3. The mock catalogue provides galaxy positions and also the three dimensional velocities of the galaxies, ideal to test our method. We follow Kitaura et al. (2012a) with several refinements. According to Hamilton (1998), we distinguish between two different components of the RSDs, namely the coherent bulk velocity $\boldsymbol{v}_r^{\rm bulk}(\boldsymbol{r})$, being the velocity due to the peculiar potential $\Phi(\boldsymbol{r})$, and the dispersed components due to quasi-virialization $\boldsymbol{v}_r^{\sigma}(\boldsymbol{r})$. Thus we write

$$\boldsymbol{s}^{\rm obs} = \boldsymbol{r} + \boldsymbol{v}_r^{\rm bulk}(\boldsymbol{r}) + \boldsymbol{v}_r^{\sigma}(\boldsymbol{r})\,. \tag{5.5}$$

The subscript r illustrates that we are working with the projections in line-of-sight direction. Now according to Equation 4.48, we can write this expression inside our sampling method for iteration step i as

$$\boldsymbol{r}_i^{\rm bulk} = \boldsymbol{s}^{\rm obs} - (\boldsymbol{v}_i^{\rm bulk}(\boldsymbol{r}_{i-1}) \cdot \hat{\boldsymbol{r}})\hat{\boldsymbol{r}}/(Ha)\,, \tag{5.6}$$

with the projection $\boldsymbol{v}_r^{\rm bulk}(\boldsymbol{r}) = (\boldsymbol{v}^{\rm bulk}(\boldsymbol{r}) \cdot \hat{\boldsymbol{r}})\hat{\boldsymbol{r}}/(Ha)$.

The important fact here is the necessity that the coherent bulk velocities have to be evaluated at the real-space position of each individual tracer, obtained in the previous Gibbs step, written as $\boldsymbol{r}_{i-1}$. Only with this procedure the sampler will be able to converge as described in Equation 4.50.

The dispersed component demands a statistical treatment as at those scales linear theory can not be applied and on the local matter density contrast (Matsubara, 2008; Peacock & Dodds, 1994; Scoccimarro, 2004). Therefore, it can be successfully modelled with a Gaussian distribution. In Kitaura et al. (2014) the authors propose a velocity dispersion term given by a Gaussian $\mathscr{G}$ with zero mean and a variance given by

$$\sigma(\boldsymbol{r}_i^{\rm bulk}) \propto (1+\delta_{\rm DM})^{\epsilon}\,, \tag{5.7}$$

where ϵ is a parameter depending on the amount of (quasi-)virialized clusters and the candidate position $\boldsymbol{r}_i^{\rm bulk}$ is inferred from the sampling of the coherent bulk flow, shown in Equation 5.6. As discussed in Section 2.9.2, the fingers-of-god (FOG) cause a decrease of apparent clustering therefore, also a decrease in the power spectrum at these scales is expected.

In order to compensate for this effect, we run a web classification (Hahn et al., 2007) explained in Section 2.10, to find knots that are characterized by three positive eigenvalues $\lambda_{1,2,3} > 0$ of the Hessian $\mathscr{H}$. So in line-of-sight direction we write

$$\boldsymbol{v}_r^{\sigma}(\boldsymbol{r}_i^{\rm bulk}) = (\boldsymbol{v}^{\sigma} \cdot \hat{\boldsymbol{r}})\hat{\boldsymbol{r}}/(Ha) = \mathscr{G}(\boldsymbol{v}_r^{\sigma}(\boldsymbol{r}_i^{\rm bulk})|\sigma(\boldsymbol{r}_i^{\rm bulk}), \forall\lambda(\mathscr{H}(\delta_{\rm DM}) > 0))\,. \tag{5.8}$$

The new candidate position then is $\boldsymbol{r}_{{\rm new},i}^{\rm bulk} = \boldsymbol{r}_i^{\rm bulk} + \boldsymbol{v}_r^{\sigma}(\boldsymbol{r}_i^{\rm bulk})$, where $\boldsymbol{v}_r^{\sigma}(\boldsymbol{r}_i^{\rm bulk})$ was chosen in line-of-sight direction to coincide with a knot type structure. Consecutively, we can select the minimum of the position obtained from the steps of Equations 5.6 and 5.8: $min(|\boldsymbol{r}_i^{\rm bulk} - \boldsymbol{r}_{{\rm new},i}^{\rm bulk}|)$. Now we have the position of the particular tracer in real-space and also modelled the dispersion according to the local density. Finally, we make sure that the density at the new iteration is larger or equal as compared to the previous step by $\delta(\boldsymbol{r}_{{\rm new},i}^{\rm bulk}) \geq \delta(\boldsymbol{r}_{{\rm new},i-1}^{\rm bulk})$. In this way we attempt to assign a galaxy to a overdense region statistically and sequentially increase the power at small scales.

5.3.1 Results

We run ARGO on a 128^3 mesh grid on the BDM mock galaxy catalogue obtained from the BigMD simulation as described above. We use a subsample with a side length of $1.25\,h^{-1}$ Gpc with a galaxy number density of $n = 3.29 \times 10^4\,h^3\,{\rm Mpc}^{-3}$. We use the power law bias description, shown in Ata et al. (2015) and a NB likelihood with negligible overdispersion. We run two chains:

1. **First chain.**
   The first RSDs corrections are based on coherent peculiar bulk motions directly derived from the density field on a mesh using linear theory as described in Section 4.7, Equation 4.76. We apply an additional Gaussian smoothing with radius $r_{\rm S} = 7\,h^{-1}\,{\rm Mpc}$.

2. **Second chain.**
   We add the velocity dispersion described in Equation 5.7 and thus correct for the FOG.

We analyze the chains with the following methods:

- **Fourier space:**
  We calculate the monopole power spectra and show the deviations of the real-space and redshift-space power due to RSDs and also the corrections we achieve. We calculate the 2D power spectra $P(k_\perp, k_\parallel)$, with which we test our method's capability to overcome the anisotropy of the distorted line-of-sight direction $k_\parallel$.

- **Configuration space:**
  We calculate the corresponding monopole two-point correlation function $\xi(r)$ for the input redshift-space catalogue as well as the corrected real-space catalogue obtained from the two chains described above. Finally, we calculate the quadrupole correlation function $Q(r)$ to estimate the goodness of our method considering the isotropy in configuration space.

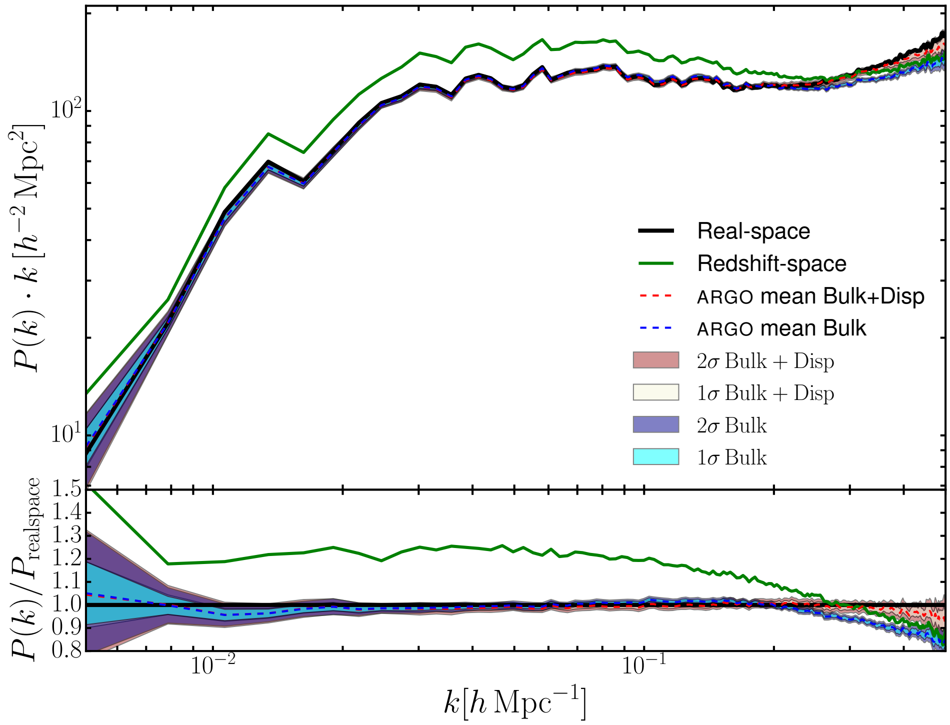


Figure 5.8: Monopole power spectra shown for real-space positions of the mock galaxies in black, redshift-space in green color. The dashed blue line corresponds to corrections only concerning the coherent bulk flows (*Bulk*) obtained with linear theory. Finally, the dashed red line shows the RSDs corrections, if the treatment of the velocity dispersion is added (*Bulk+Disp*). The reconstructions are plotted with their mean and one and two σ variances, obtained by averaging over 1000 iterations.

In Figure 5.8 we show the monopole power spectra of the mock galaxies in real-space, redshift-space, and also the power spectra of the two different reconstruction schemes. The real-space power spectrum is shown as the black solid line, redshift-space is shown in solid green. We see an overall increase of power in redshift-space up to $k \sim 0.25\,h\,\mathrm{Mpc}^{-1}$, as predicted by the Kaiser's linear theory of RSDs. For scales greater than $k \geq 0.25\,h\,\mathrm{Mpc}^{-1}$, we witness the FOG effect, causing a decrease of power. The corrected positions obtained from ARGO are shown as dashed lines, dashed blue for corrections only concerning the coherent bulk flows (*Bulk*) obtained with linear theory, and dashed red where we also applied the correction of the velocity dispersion (*Bulk+Disp*). Both chains are shown with the one and two σ variances averaged over 1000 iterations after the burn-in phase. The two chains overlap completely for the largest scales for our reconstructions, showing that ARGO is capable to correct large-scale bulk motions if linear theory with a certain smoothing is applied. At the small scales, when FOG are dominating however, we see a significant difference. The *Bulk* method is not able to recover the lack of power due to FOG. This is improved clearly when the *Bulk+Disp* method is applied. In that case we achieve a matching up to the smallest scales within our reconstructed volume in the one σ band. This is evidently an improvement.

In Figure 5.9 the 2D power spectra are shown. On the left we only show the input catalogues; color-coded contour plots correspond to the real-space mock catalogue. It can be seen, that on all scales the power spectrum is isotropic. However, the anisotropic pattern in redshift-space (dotted lines) can be clearly seen. In particular, the enhancement of power due to the Kaiser effect is very prominent for small wave numbers (large scales) in $k_{\|}$, whereas FOG tend to decrease the power on small scales. We can see that the coherent bulk flow corrections alone (solid line), shown in the central plot perform very well to correct for the Kaiser boost, wherras the the additional dispersion correction is fruitful especially for smaller scales, seen in the right plot (solid line).

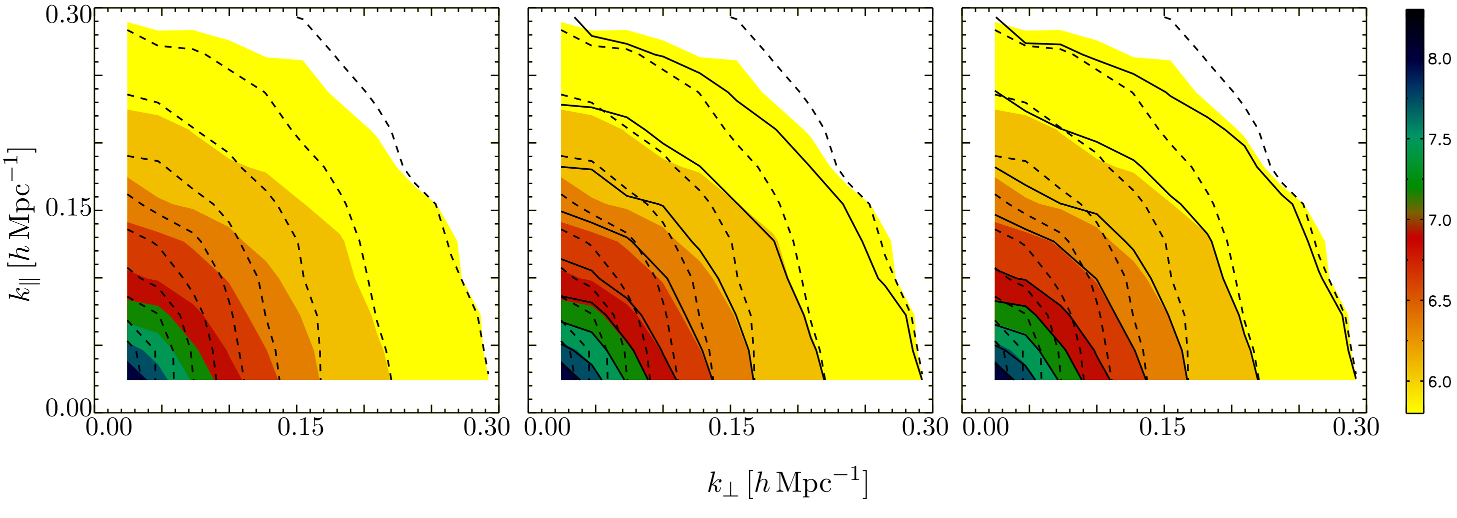


Figure 5.9: 2D power spectra $P(k_{\perp}, k_{\|})$ of the mock galaxy catalogue in real-space (color-coded contour regions), in redshift-space (dashed lines), including the reconstructed galaxy density field in real-space (solid lines): with an smoothing of $r_S = 7\,h^{-1}\,\mathrm{Mpc}$ (2nd panel), including virialized RSD correction (3rd panel).

In Figure 5.10 the monopole and quadrupole correlation functions are shown. From these plots we can see that the two different reconstructions, *Bulk* and *Bulk+Disp*, both with Gaussian smoothing, yield similar results. The left panel shows how accurately the real-space BAO can be obtained from redshift-space. Interestingly, the full *Bulk+Disp* RSD correction algorithm also shows slightly better agreement with the true real-space quadrupole not only on small scales, but also displaying less artificial spikes at large scales. Still, at the relevant scales of the BAO already the *Bulk* is indistinguishable from *Bulk+Disp*. The ratios with the real-space catalogue show an excellent agreement within the error bands.

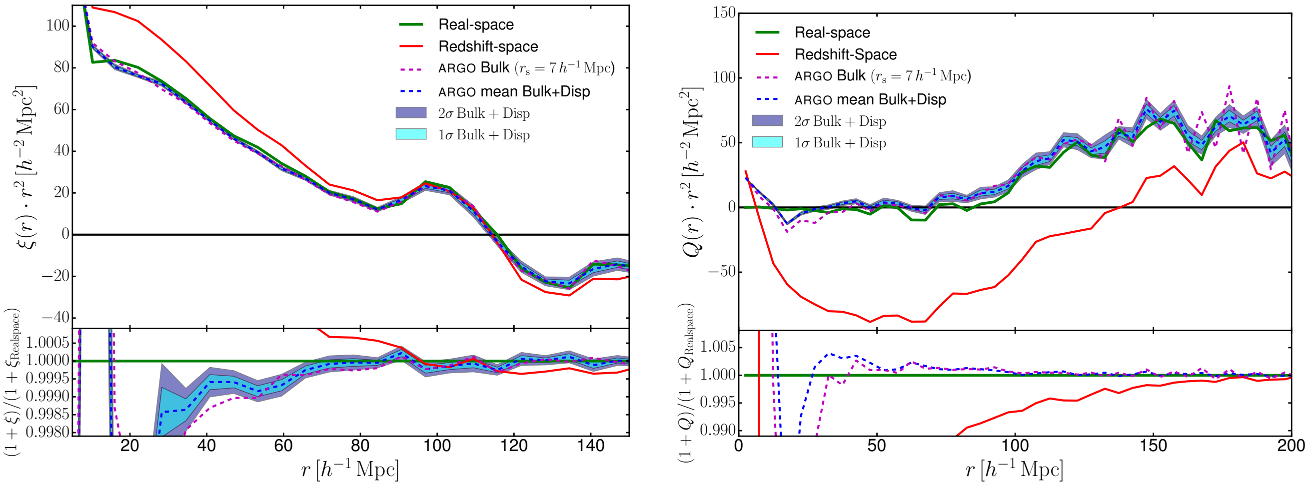


Figure 5.10: Left panel: two-point correlation function, **Right panel**: corresponding quadrupoles. The green solid line corresponds to the real-space positions of the mock catalogue, whereas the redshift-space positions are shown as red solid line. The magenta dashed line displays the *Bulk* reconstruction, where no dispersion term was applied, and finally the blue dashed line indicates to the *Bulk+Disp* reconstruction with one and two σ contour. The sampling of the coherent bulk velocities have been performed after a Gaussian smoothing with radius $r_S = 7\,h^{-1}$ Mpc. The lower panel shows the ratio of each plotted correlation function with the real-space correlation function.

5.3.2 Convergence

In this Section I show the behavior of ARGO exemplary for one reconstruction, how our code convergences.

As we already mentioned in Section 5.2.2, we need to estimate the behavior of each Markov chain's evolution to ensure that the realizations that we obtain from the sampling procedure actually belong to the desired target distribution. Different estimators have been proposed in literature of which we choose the Potential Scale Reduction Factor (PSRF) $\hat{R}$, introduced in Gelman & Rubin (1992). The exact definition is shown in Appendix A.2. The PSRF estimator is applied by comparing m chains that are supposed to have the same target distribution but have started at different initial points. Now we can compare the variance of the m means of the different chains to the mean of the variance of each individual chain.

A value of $\hat{R} = 1.1$ is assumed to represent a converged chain. Also, we estimate the covariance of the density field given the first 500 iterations. We therefore use a similar definition of the covariance given in Equation 4.86. However, we want the covariance to be expressed differently. Equation 4.86 averages over many realizations to estimate the covariance of the power spectrum's mode k_a with another mode k_b, assuming in total s modes $(a, b \in \{1, ..., s\})$. The resulting covariance will therefore be a symmetric matrix with N_s^2 entries, where N_s is the total number of modes measured from each power spectrum. For our convergence tests we however, seek to estimate the evolution of the density field as a whole considering the first 500 iterations. Thus we write for two different iterations i, j

$$C_{ij} = \langle \delta^i \delta^j \rangle - \langle \delta^i \rangle \langle \delta^j \rangle \,. \tag{5.9}$$

In the exact same way we can also run this diagnostics for the evolution of the power spectra of each individual iteration, averaging over the modes

$$C_{ij} = \langle P^i P^j \rangle - \langle P^i \rangle \langle P^j \rangle \,, \tag{5.10}$$

where now P^i corresponds to the power spectrum of iteration i as a whole, including all modes, and must not be confused with the power of mode i. Nevertheless, in this chapter we will focus on PSRF and density covariance entirely. These are shown in Figure 5.11. As the first 500 iterations shown in the right panel show increasing covariance (correlation, when divided by the individual variances σ_i, σ_j), we discard the first 1000 to be safe. The left panel shows the PSRF after discarding the first 1000 burn-in iterations for each density cell. $\hat{P} - 1$ is everywhere less than 0.1, which gives us the sureness that our chain converged.

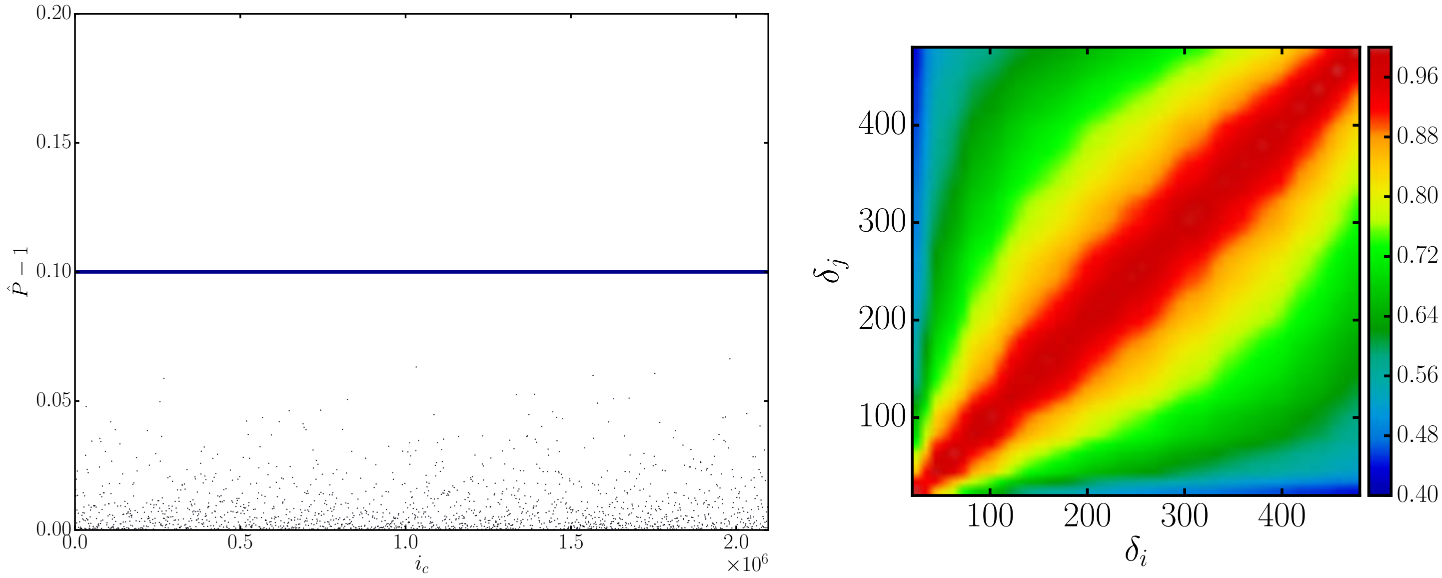


Figure 5.11: **Left panel**: the PSRF of each density cell is plotted. For better visualization we thinned out the points by a factor of 1000. $\hat{P} - 1$ is shown on the y-axis, so that values below 0.1 satisfy a convergence. **Right panel**: the covariance for the first 500 iterations of all density cells. The correlation increases towards higher iterations, indicated by the color bar.

It is beneficial to have a good insight about the convergence behavior of a run chain in order to adapt the parameters of the Hamilton equations of motion, such as the step size shown in the leap-frog scheme, shown in Section 4.3.3, Equations 4.31 and 4.33. This will help the chain to converge faster and also to draw more independent samples when converged. Therefore, an adaptive choice of the step size ϵ is preferable.

5.3.3 Remarks & Conclusion

In this application of ARGO, we aimed to jointly sample the density and velocity fields given a mock galaxy catalogue. We have shown that ARGO is capable of reconstructing the real-space positions on large scales with linear theory, modelling the coherent bulk flows, called *Bulk* in our method, and a stochastic description called *Bulk+Disp* to correct for the FOG effect, as the latter one is a highly non-linear phenomenon. As *Bulk+Disp* requires a full web classification in order to assign a mock galaxy position to a high density region, it is computationally still affordable on small mesh sizes, but will be much more expansive for higher resolutions though. Also, we showed that the gains of *Bulk+Disp* at the BAO scales are not significant, and the *Bulk* reconstructions already correct the RSD sufficiently. On small scales however, at least up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$, shown in the isotropization of the 2D power spectrum, the FOG corrections are indispensable to yield the correct power spectra.

6

Phase-space Reconstructions with Galaxy Redshift Surveys

In this chapter I present the applications of ARGO on the BOSS galaxy redshift survey. The results of the reconstructions are published in Ata et al. (2017), whereas in Alam et al. (2016) the general analysis of the galaxy survey is presented. Several improvements were implemented to ARGO during this work. Tests done without these refinements can be partly seen in Appendix A.1.

6.1 Applying ARGO on galaxy surveys

In order to apply ARGO on a galaxy catalogue obtained from a redshift survey like SDSS, as described in Section 3.1, we need several ingredients that are not required by a mock catalogue. To use the survey data, like the CMASS catalogue (Section 3.1.2) of BOSS (Section 3.1.1) within a cubic volume as done by ARGO, we translate the positions of the galaxies from a equatorial coordinate system into Cartesian coordinates. This is done by (in concordance to Equation 4.52)

$$
\begin{aligned}
x &= r \cos\alpha \cos\delta \,, \\
y &= r \sin\alpha \cos\delta \,, \\
z &= r \sin\delta \,, \\
r &= \frac{H_0}{c} \int\limits_0^z \frac{dz'}{\sqrt{\Omega_\mathrm{M}(1+z')^3 + \Omega_\Lambda}} \,,
\end{aligned}
\tag{6.1}
$$

where the parameters Ω_M and Ω_Λ are depending on the chosen cosmological model. We use the parameters obtained from Planck Collaboration (2015) as fiducial cosmological model. Important to note still is the impact of an incorrect assumption of a cosmological model to map redshift-space positions into comoving real-space coordinates, called Alcock-Paczynski (AP) effect (Alcock & Paczynski, 1979). These distortions are due to the fact, that measured distances along line-of-sight and perpendicular to line-of-side are significantly different. Induced distortions by the AP effect are to a large extent degenerate with the RSD, causing similar anisotropy in the clustering. Studies of anisotropies with galaxy surveys (Peacock et al., 2001) including the AP effect are challenging and have been studied in literature (see e.g. Ballinger et al., 1996; Matsubara & Suto, 1996). However, this degeneracy can be overcome when analyzing the clustering at various distinct redshift ranges utilizing the fact that the evolution of RSDs and the AP effect is different with redshift (Li et al., 2015). Other methods that seek

to constrain cosmological parameters, such as the growth rate f_Ω or bias in a reconstruction method yet do not break this degeneracy caused by RSDs and the AP effect (see e.g. Granett et al., 2015).

As discussed in Section 4.5.4 the survey geometry is a highly challenging task to deal with, and needs careful treatment. Most demanding is the steep decrease of completeness at the edges of the observed region. A very small mistreatment may cause a gradient in the reconstructed density field, which in turn may lead to a large velocity field in the reconstruction that is unphysical.

6.1.1 Renormalized lognormal priors

In the previous reconstructions the entire volume was populated with (mock) galaxies. This is not longer the case when ARGO is applied to a galaxy redshift survey catalogue. As the reconstructed volume needs to be cubic to preserve isotropy of the prior, large parts of the reconstructed volume may be empty. The knowledge of the completeness of the cells within our reconstructed volume is comprised in the window function, assuming values between $w = 0$ (not observed) and $w = 1$ (completely observed). Problems still arise due to the random lognormal field that is used to fill these empty regions. The lognormal model for structure formation yields a poor description of the three-point statistics (see Chuang et al., 2015; White et al., 2014), and will have a different mean field $\boldsymbol{\mu}$ which depends on the higher order statistics of the dark matter field, shown in Equation 4.47.

This can be seen if we consider the one-point distribution of the lognormal distribution

$$P^{(1)}_{\rm LN}(1+\delta) = \frac{1}{\sqrt{2\pi\sigma^2}}\left[\frac{\ln(1+\delta)-\mu}{2\sigma^2}\right]\frac{1}{1+\delta}\,, \tag{6.2}$$

where σ^2 is the variance and μ the expectation value of $\ln(1+\delta)$. Further we find that the expectation value $E[1+\delta]$ assumes the form

$$E[1+\delta] = \mathrm{e}^{\mu+\frac{\sigma^2}{2}}\,. \tag{6.3}$$

Using the mean of $\langle(1+\delta)\rangle = 1$, we write

$$\mathrm{e}^{\mu+\frac{\sigma^2}{2}} = 1 \Rightarrow \mu = -\sigma^2/2\,. \tag{6.4}$$

Equation 6.4 is the theoretical prediction of the mean for a perfect lognormal field. In Section 4.4, we found that the mean μ is comprising all higher orders beyond the Gaussianization done with the lognormal transformation. However, due to the different higher order contributions of the fully non-linear evolved density contrast $\boldsymbol{\delta}$ in the Universe, we expect $\boldsymbol{\mu}$ to be different from the theoretical predictions of (one-dimensional) $\mu = -\sigma^2/2$. In Ata et al. (2017) we therefore introduced a significant improvement dedicated to this problem. According to Section 4.9 Equation 4.83 we can estimate the mean field $\boldsymbol{\mu}$ from the linear density field $\boldsymbol{\delta}_{\rm L}$. We can extend this ansatz now and calculate $\boldsymbol{\mu}$ in dependency of the window function $w(\boldsymbol{r})$. In this way $\boldsymbol{\mu}$ is supposed to be more precisely renormalizing the lognormal prior of the density field. We write consequently

$$\boldsymbol{\mu}_{w(r)} \quad = \quad -\ln\left\langle \exp(\boldsymbol{\delta}_{\rm L})\right\rangle_{w(r)}\,. \tag{6.5}$$

We expect the mean field to be very close to the theoretical expectation $\boldsymbol{\mu} = -\boldsymbol{\sigma}^2/2$ in the empty regions of the reconstructed volume, whereas $\boldsymbol{\mu}$ in the data dominated regions may strongly deviate from this estimation. We will use up to 10 completeness bins in the latter work.

6.1.2 Window function weighted normalization

Another important issue is concerning the redefinition of the normalization factor $f_{\bar{N}}$ introduced in Equation 4.57. As we will yield a reconstruction, which is a composition of prior and likelihood dominated regions, we must take care that both regions are equally weighted in terms of Hamilton Monte-Carlo. Therefore we modify the normalization and apply the normalization to those cells of our mesh grid only where the window function is larger than 0, $w(r) > 0$. Thus we write

$$f_{\bar{N}|w(r)>0} = \bar{N} \Bigg/ \left[\frac{\sum\limits_{i=1}^{N_{w(r)>0}} (1+\delta_i)^b}{N_{w(r)>0}} \right] , \tag{6.6}$$

and the summation is applied to all density cells in which $w(r) > 0$ so that $N_{w(r)>0} \leq N_c$, with N_c being the total number of cells in the volume. Only if this normalization is applied, the reconstruction will successfully normalize the likelihood function in concordance to the prior.

6.1.3 Light-cone effects

When working with the data of a galaxy redshift survey, physical effects are arising that I summarized as light-cone effects. I will explain these points in this Section separately. The main reason for light-cone effects are the different redshifts each individual galaxy is observed at. Different redshifts mean different stages of cosmic evolution. In linear theory and within ΛCDM this is described by the growth factor:

$$D(z) = \frac{H(z)}{H_0} \frac{\int\limits_z^\infty dz' \frac{(1+z')}{H^3(z')}}{\int\limits_0^\infty dz' \frac{(1+z')}{H^3(z')}} , \tag{6.7}$$

relating the density field at a given redshift to a reference redshift z_{ref}: $\delta_i(z_{\mathrm{ref}}) = G(z_{\mathrm{ref}}, z_i)\, \delta_i(z_i)$ with

$$G(z_{\mathrm{ref}}, z_i) \equiv D(z_i)/D(z_{\mathrm{ref}}) \,. \tag{6.8}$$

Given a redshift z one can define the ratio between the galaxy correlation function in redshift-space at z, called $\xi^s_{\mathrm{G}}(z)$ and the matter correlation function in real-space at z_{ref}, called $\xi_{\mathrm{M}}(z_{\mathrm{ref}})$ as

$$c^s_{\mathrm{L}}(z) \equiv \sqrt{\xi^s_{\mathrm{G}}(z)/\xi_{\mathrm{M}}(z_{\mathrm{ref}})} \,. \tag{6.9}$$

The quantity $\xi^s_{\mathrm{G}}(z)$ can be obtained from the data without having to assume any bias, nor growth rate. Furthermore, one can use the Kaiser factor

$$K = 1 + \frac{2}{3}\frac{f_\Omega}{b_{\mathrm{L}}} + \frac{1}{5}\left(\frac{f_\Omega}{b_{\mathrm{L}}}\right)^2 , \tag{6.10}$$

with f_Ω being the growth rate and b_{L} the linear bias factor (Kaiser, 1987)) to relate the galaxy correlation function in redshift-space to the matter real-space correlation function

$$\begin{aligned} \xi^s_{\mathrm{G}}(z) &= K(z)\, \xi_{\mathrm{G}}(z) \\ &= K(z)\, b^2_{\mathrm{L}}(z)\, G^2(z, z_{\mathrm{ref}})\, \xi_{\mathrm{M}}(z_{\mathrm{ref}}) \,. \end{aligned} \tag{6.11}$$

From the last two equations we find a quadratic expression for $b_{\rm L}(z)$ for each redshift z

$$b^2_{\rm L}(z)+\frac{2}{3}f_\Omega(z)b_{\rm L}(z)+\frac{1}{5}f^2_\Omega(z)-\frac{(c^s_{\rm L}(z))^2}{G^2(z,z_{\rm ref})}=0\,, \tag{6.12}$$

with only one positive solution.

$$b_{\rm L}(z)=-\frac{1}{3}f_\Omega(z)+\sqrt{-\frac{4}{45}f_\Omega(z)^2+(c^s_{\rm L}(z))^2\left(\frac{D(z_{\rm ref})}{D(z)}\right)^2}\,. \tag{6.13}$$

We assume the bias measured in redshift-space on large scales $c^s_{\rm L}(z)$ (with respect to the dark matter power spectrum) to be constant for CMASS galaxies across the considered redshift range (see, e.g., Rodríguez-Torres et al., 2016). Nevertheless, the (real-space) linear bias $b_{\rm L}(z)$ is not, as it needs to precisely compensate for the growth of structures (growth factor) and the evolving growth rates. Now we have elaborated an expression for the linear bias $b_{\rm L}$. Our bias model however is relying on a power law bias description. This is why we also correct this by multiplying a non-linear bias correction factor $f_{\rm b}$. The factor $f_{\rm b}$ is expected to be less than unity, since we are using the linear bias in the power law and thus we have to cancel the higher order terms.

As we seek to reconstruct the underlying dark matter field for a particular redshift $z_{\rm ref}$, we have to redefine Equation 4.60 with the refinements we needed to perform for a light-cone. In cell i we therefore write the expectation value λ as

$$\lambda_i=f_{\bar N}w(r_i)(1+G(z_i)\delta_i)^{b_{\rm L}(z_i)f_{\rm b}}\,, \tag{6.14}$$

where we neglect the threshold term in the following analysis.

Revising Equation 4.73

$$\frac{\partial}{\partial\delta_{{\rm L},i}}=\left(\frac{\partial\delta_{{\rm L},i}}{\partial\delta_j}\right)^{-1}\frac{\partial\lambda_k}{\partial\delta_j}\frac{\partial}{\partial\lambda_k}\,, \tag{6.15}$$

we have to recalculate the term $\frac{\partial\lambda}{\partial\delta}$, including the modifications in Equation 6.14.

$$\frac{\partial\lambda_k}{\partial\delta_j}=G(z_k)b_{\rm L}(z_k)f_{\rm b}\frac{\lambda_k}{1+\delta_k}\delta^{\rm D}_{jk}\,. \tag{6.16}$$

Now the derivative of the Poisson likelihood function writes in total as

$$-\frac{\partial}{\partial\delta_{{\rm L},i}}\ln\mathscr{L}_{\rm P}=G(z_i)b_{\rm L}(z_i)f_{\rm b}\lambda_i\left(\frac{1+\delta_i}{1+G(z_i)\delta_i}\right)\left(1-\frac{N_i}{\lambda_i}\right)\,, \tag{6.17}$$

and consequently for the negative binomial likelihood function as

$$-\frac{\partial}{\partial\delta_{{\rm L},i}}\ln\mathscr{L}_{\rm NB}=G(z_i)b_{\rm L}(z_i)f_{\rm b}\lambda_i\left(\frac{1+\delta_i}{1+G(z_i)\delta_i}\right)\left(\frac{1}{\frac{\lambda_i}{\beta}+1}+\frac{N_i}{\beta+\lambda_i}-\frac{N_i}{\lambda_i}\right)\,. \tag{6.18}$$

6.1.4 Bias prediction from renormalized perturbation theory

One can predict $f_{\rm b}$ from renormalized perturbation theory, which in general will be a function of redshift. Let us Taylor expand our bias expression (Eq. 6.14) to third order

$$\begin{aligned}\delta_{\rm g}(z_i)\equiv\frac{\varrho_{\rm g}}{\bar\varrho_{\rm g}}(z_i)-1\simeq b_{\rm L}(z_i)f_b(z_i)\delta(z_i)\\ +\frac{1}{2}b_{\rm L}(z_i)f_b(z_i)(b_{\rm L}(z_i)f_b(z_i)-1)\left((\delta(z_i))^2-\sigma^2(z_i)\right)+\\ \frac{1}{3!}b_{\rm L}(z_i)f_b(z_i)(b_{\rm L}(z_i)f_b(z_i)-1)(b_{\rm L}(z_i)f_b(z_i)-2)\left(\delta(z_i)\right)^3\,,\end{aligned} \tag{6.19}$$

with $\delta(z_i) = G(z_i, z_{\rm ref})\delta(z_{\rm ref})$. The usual expression for the perturbatively expanded overdensity field to third order ignoring non-local terms is given by

$$\delta_{\rm g}(z_i) = c_\delta(z_i)\delta(z_i) + \frac{1}{2}c_{\delta^2}(z_i)(\delta^2(z_i) - \sigma^2(z_i)) + \frac{1}{3!}c_{\delta^3}(z_i)\delta^3(z_i)\,. \tag{6.20}$$

Correspondingly, one can show that the observed, renormalized, linear bias is given by (see McDonald & Roy, 2009)

$$b_\delta(z_i) = c_\delta(z_i) + \frac{34}{21}c_{\delta^2}(z_i)\sigma^2(z_i) + \frac{1}{2}c_{\delta^3}(z_i)\,. \tag{6.21}$$

By considering that in our case the observable linear bias is expected to be given by $b_{\rm L}(z_i)$ and identifying the coefficients $\{c_\delta = f_{\rm b}b_{\rm L},\ c_{\delta^2} = f_{\rm b}b_{\rm L}(f_{\rm b}b_{\rm L}-1),\ c_{\delta^3} = f_{\rm b}b_{\rm L}(f_{\rm b}b_{\rm L}-1)(f_{\rm b}b_{\rm L}-2)\}$ from Eqs. 6.19 and 6.20 one can derive the following cubic equation for $f_{\rm b}$

$$\begin{aligned} & f_{\rm b}^3\left(\frac{1}{2}b_{\rm L}^3(z_i)\sigma^2(z_i)\right) \\ & +f_{\rm b}^2\left(-\frac{3}{2}b_{\rm L}^2(z_i)\sigma^2(z_i) + \frac{34}{21}b_{\rm L}^2(z_i)\sigma^2(z_i)\right) \\ & +f_{\rm b}\,b_{\rm L}(z_i)\left(1 + \left(-\frac{34}{21}+1\right)\sigma^2(z_i)\right) - b_{\rm L}(z_i) = 0\,. \end{aligned} \tag{6.22}$$

Let us consider the case of a cell resolution of $6.25\,h^{-1}\,$Mpc. The only real solutions for redshift $z = 0.57$ ($G = 0.78$) and $b_{\rm L} = 2.1 \pm 0.1$, are $f_{\rm b} = 0.62{\pm}0.01$ including the variance from the non-linear transformed field ($\sigma^2(\delta) = 1.75$), and $f_{\rm b} = 0.71 \pm 0.02$ including the variance from the linear field ($\sigma^2(\delta_{\rm L}) = 0.91$). Let us, hence, quote as the theoretical prediction for the bias correction factor the average between both mean values with the uncertainty given by the difference between them $f_{\rm b} = 0.66 \pm 0.1$. These results show little variation (± 0.01) across the redshift range. Leaving $f_{\rm b}$ as a free parameter and sampling it to match the power spectrum on large scales yields $f_{\rm b} = 0.7 \pm 0.05$. However, there is an additional uncertainty associated to this measure, since the result depends on the particular k mode range used in the goodness of fit. Therefore, one can conclude that the theoretical predictions account for the non-linear correction within the associated uncertainties.

6.2 Cosmic Web and Cosmic flows from SDSS

In this Section I present the results of ARGO run with the CMASS catalogue (Section 3.1.2) obtained from the BOSS galaxy redshift survey (Section 3.1.1) that were published in Ata et al. (2017). For validation we also run the exactly same configuration on a galaxy light-cone mock catalogue (Rodríguez-Torres et al., 2016), which has the same number density and geometry compared to the galaxy catalogue.

The runs are performed on a mesh grid of 512^3 cells for a cubical volume of $3200\,h^{-1}\,$Mpc side length. In total we have $N_{\rm G} = 571372$ galaxies in the CMASS sample and $N_{\rm G} = 612937$ in the mock catalogue that we assign to the mesh grid with the NGP kernel. This choice leads to a cell resolution of $d_{\rm L} = 6.25\,h^{-1}\,$Mpc .

We choose the minima of the Cartesian coordinates to fit all galaxies inside the reconstructed volume, leaving 90% of the cells empty

$$\begin{aligned} x_{\rm min} &= -2000\,h^{-1}\,{\rm Mpc}\,, \\ y_{\rm min} &= -1650\,h^{-1}\,{\rm Mpc}\,, \\ z_{\rm min} &= -500\,h^{-1}\,{\rm Mpc}\,. \end{aligned} \tag{6.23}$$

We choose the reference redshift, the redshift at which we reconstruct the density field, to be $z = 0.43$. This choice guarantees that the growth ratio $G(z_{\rm ref}, z_i) \equiv D(z_i)/D(z_{\rm ref})$ remains below one for all galaxies within our volume. Otherwise, negative densities (density contrasts below -1) will arise in low density cells, causing singularities in the model.

Further, we discard the first 1000 samples to be safe against the burn-in phase of the Monte-Carlo sampling. On average ARGO requires 10 Gb memory and approximately six minutes per Gibbs-iteration after convergence, running on eight cores parallel. Only in each 50th iteration the velocity sampling is applied, as in that sampling step the time consumption goes up beyond ten minutes due to the computational costs of the Fast-Fourier-Transformations (FFTs).

6.3 Cosmic Web reconstructions

The results of ARGO applied to the CMASS catalogue are described in this section.

However, we first test the outcome of the Poisson and the NB likelihood model in ARGO with a snapshot of a mock galaxy catalogue, based on the BigMultiDark simulations[1]. This is not a light-cone, but a snapshot at fixed redshift of $z = 0.57$, which corresponds to the mean redshift of the the BOSS galaxy catalogue and also the light-cone mock that we are going to use in the following. We will use the terms *snapshot mock* and *light-cone mock* to avoid confusions. Using this snapshot mock catalogue, we can calibrate the bias parameters without the disturbance of light-cone effects. We find that the snapshot mock galaxy population is well described by a Poisson likelihood with negligible over-dispersion shown in Figure 6.1.

This can be explained due to the resolution of the mesh grid and as the mock galaxies are created here to mimic luminous red galaxies (LRGs), the most massive galaxies known. This means, in opposite to the mock galaxies we analyzed with the Bolshoi simulation, that in a Lagrangian picture the LRGs are located at the primordial density peaks that collapsed in a very early stage of the evolution of the Universe and thus, by hierarchical merging, accumulated most of the matter in its surroundings. This *halo exclusion* effect is well studied in literature (compare Section 2.8.2) and consequently can reduce over-dispersion (and even reverse it). This is exactly the case for the BOSS-type galaxies. The dependency on the mesh grid resolution is a very important component when analyzing the dispersion of the galaxy counts per cell. Larger cell sizes for a given galaxy population will reduce the over-dispersion, or as mentioned above even reverse it. These conditions have to be taken into account for a proper treatment of the dispersion of galaxy counts per cell.

In the following we firstly work out the auxiliary variables that we have introduced to overcome the physical effects a light-cone brings along in Section 6.3.1. Afterwards we will analyze the density field of the reconstruction in Section 6.3.2. The distribution function of the linear density field is investigated in Section 6.3.3. I show the different slices of the inferred density field in Section 6.3.4, where we also apply a Zel'dovich like transformation and finally summarize our results in Section 6.3.5.

6.3.1 Light-cone auxiliaries

We calculate the mean field $\mu(w(\boldsymbol{r}))$ depending on the completeness in ten bins starting with $w(r) = 0$ as described in Equation 6.5. We find that the mean field at $w(\boldsymbol{r}) = 0$ is comparable within 0.5% to $\mu_{\rm G} = -\sigma_{\rm G}^2/2 = 0.543$ that is expected from the Gaussian random field created from the initial

---

[1] We obtained this catalogue with a query from the CosmoSim database created by the German Astrophysical Virtual Observatory as a result of a collaboration between the Leibniz-Institute for Astrophysics Potsdam (AIP) and the Spanish MultiDark Consolider Project `https://www.cosmosim.org`.

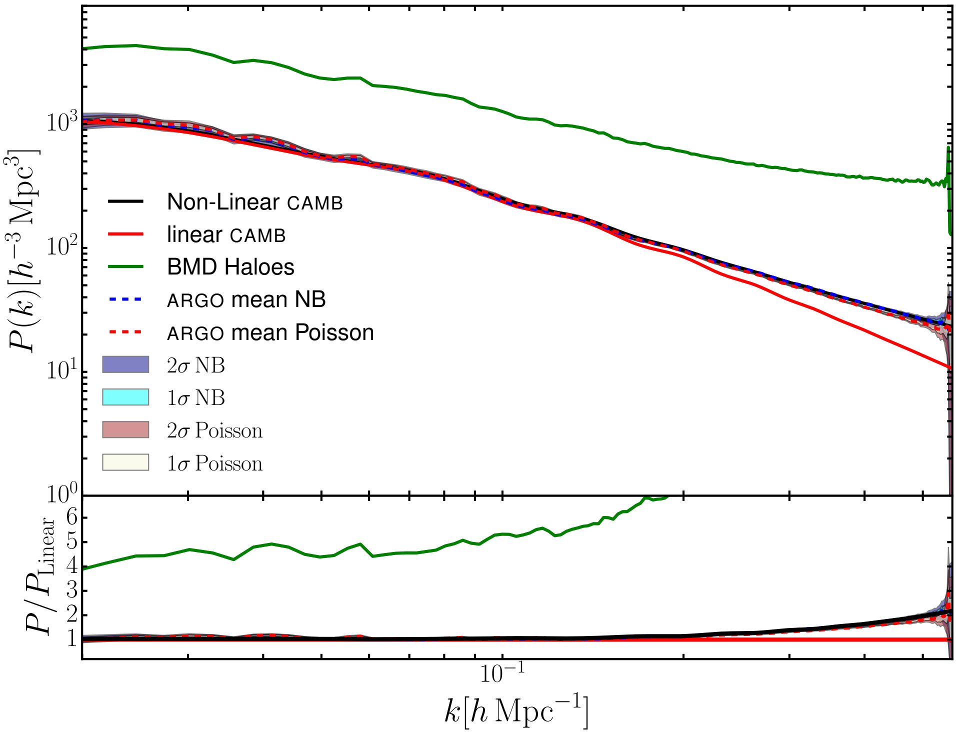


Figure 6.1: Comparison of small scale power spectrum for a snapshot mock catalogue at $z = 0.57$, between Poisson and NB likelihood reconstructions.

covariance matrix at a resolution of $6.25\,h^{-1}\,\mathrm{Mpc}$, shown in Figure 6.2 on the left. It can also be seen that $\boldsymbol{\mu}$ is strongly varying within the volume from $\mu_{\mathrm{min}} = -0.63$ and $\mu_{\mathrm{max}} = -0.46$. This is a deviation of up to 18% from μ_{G} which underlines the importance of a completeness dependent mean field sampling as shown here.

Also as discussed in Equation 6.13, we translate the constant linear bias of the CMASS galaxies in redshift-space at different redshifts to a real-space bias and weight all galaxies to a reference redshift $z_{\mathrm{ref}} = 0.43$, corresponding to the lower edge of the CMASS galaxy catalogue. The outcome of this mapping is described in Equation 6.13 and demands a correction factor computed in Section 6.1.4. The final bias values w.r.t. the position inside the reconstructed volume are shown in 6.2 on the right, where the positions of the galaxies are illustrated with the grey shaded area.

6.3.2 Density inference

In this Section I describe the outcome of the density sampling applied on both, the CMASS galaxy catalogue and also the light-cone mock catalogue. Although the focus here mainly lies on the density field reconstruction, I emphasize that these results are obtained in real-space and thus the velocity sampling has been applied simultaneously. Nevertheless I leave the discussion of the cosmic flows to the next Section (6.4). In concordance with the theoretical prediction found in Equation 6.20 we set the bias correction parameter f_{b} to a value between 0.66 ± 0.1 and 0.70 ± 0.05 within the theoretical uncertainty.

The resulting real-space power spectra and the consequent ratios with the theoretical linear power spectrum of the reconstructions can be appreciated in Figure 6.3. The top panel shows the reconstruction of the BOSS CMASS galaxy catalogue whereas the bottom panel shows the corresponding reconstruction

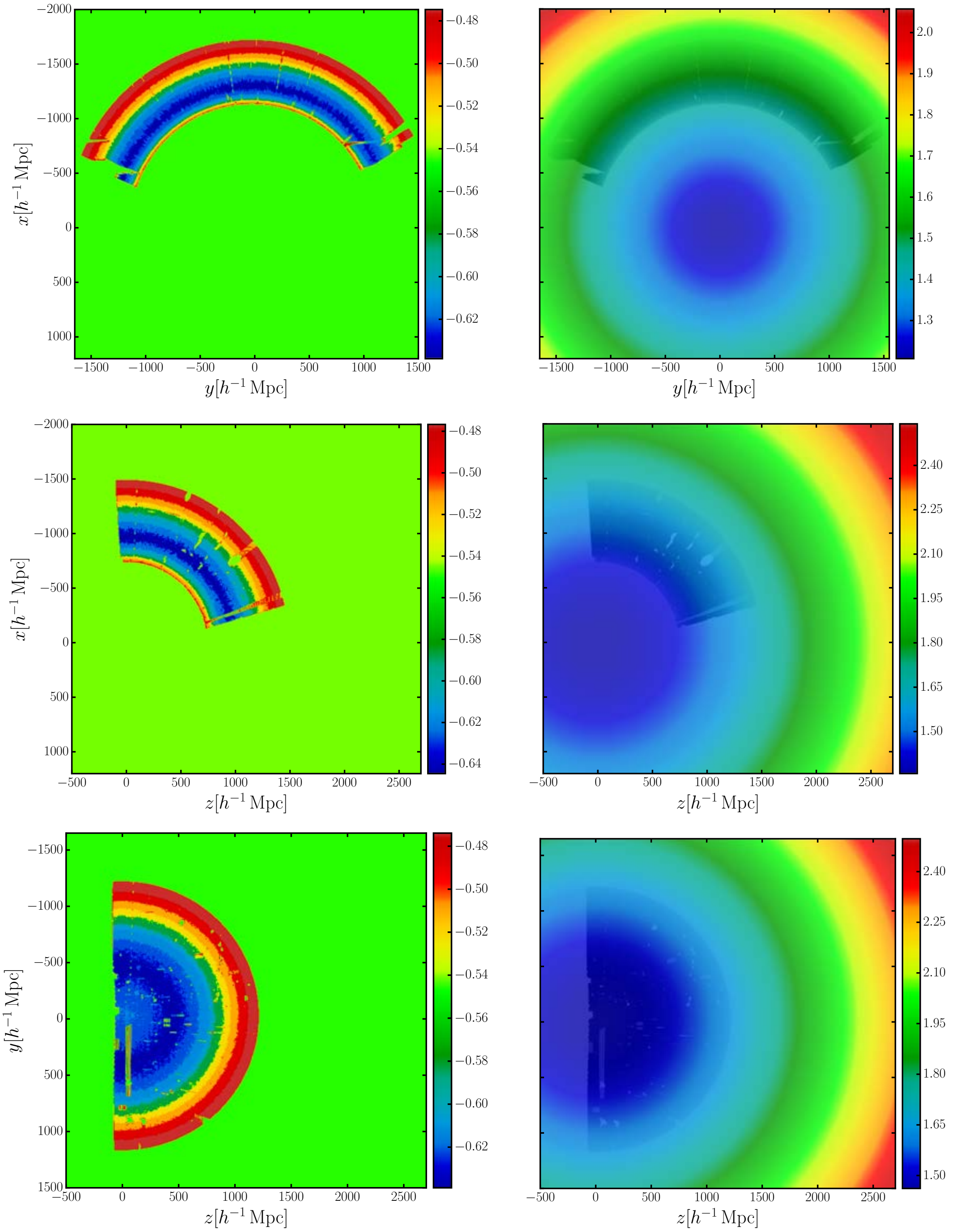


Figure 6.2: Slices of the completeness dependent mean field $\boldsymbol{\mu}$ shown on the right and position dependent bias field $\boldsymbol{b}$ on the left are shown in three different cuts. All slices are shown with a thickness of $6.25\,h^{-1}$ Mpc. The shaded area on the right hand side indicates where the data is located within the box.

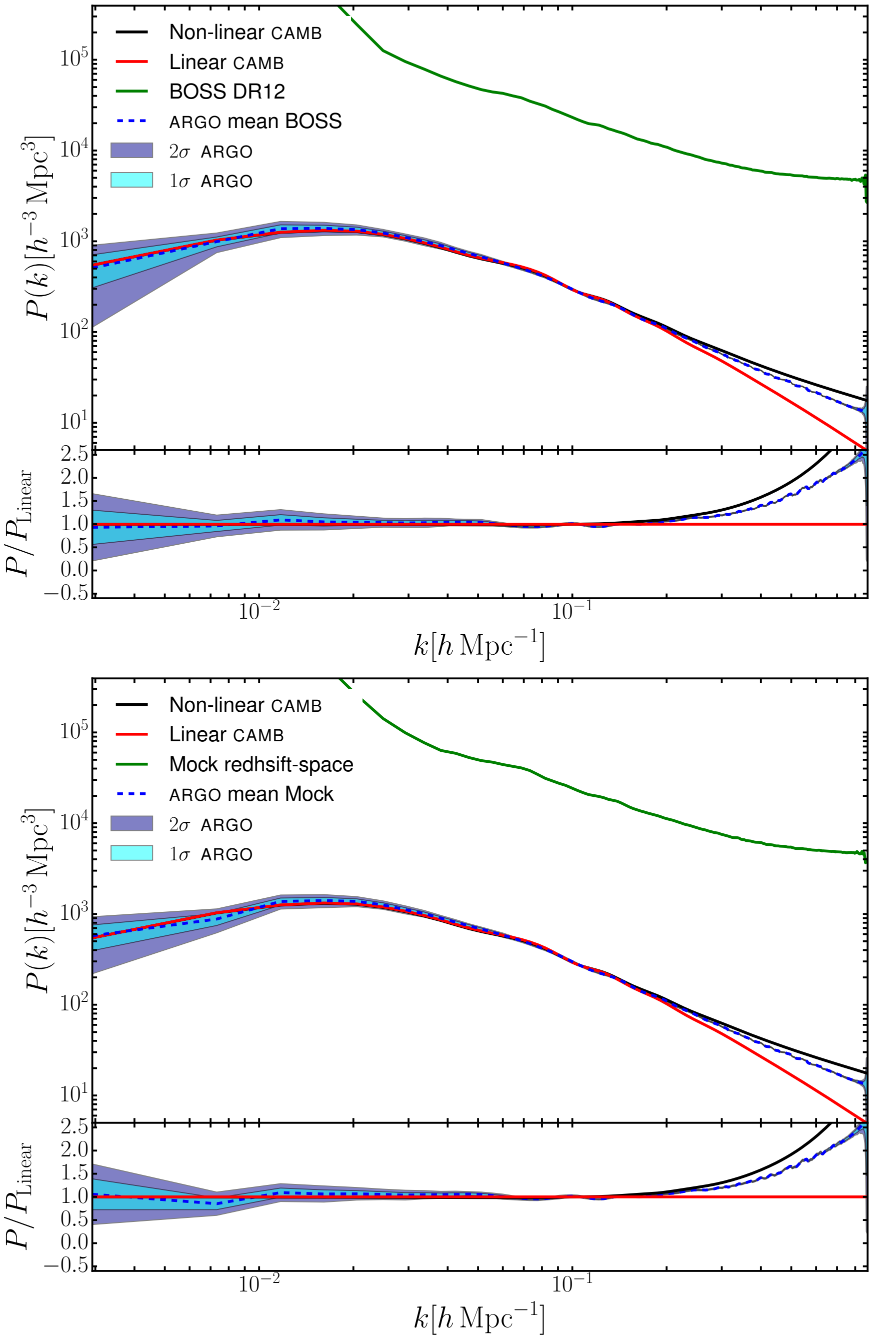


Figure 6.3: Power spectra of the reconstruction process are shown. the green solid line shows the raw galaxy input, the solid black and red lines show the theoretical predictions for the non-linear and linear density power spectrum, respectively, obtained from CAMB. **Top:** the reconstruction of the BOSS CMASS galaxy catalogue is shown, **Bottom:** the corresponding reconstruction of the light-cone mock catalogue is shown. The mean of each reconstruction is plotted as dashed blue line including the one and two σ variances. Both plots include a bottom panel, in which the ratio of the power spectra with the theoretical linear prediction is shown.

of the light-cone mock catalogue. Both chains have been run with $f_{\mathrm{b}} = 0.67$. We show the theoretical predictions from CAMB (Howlett et al., 2012; Lewis et al., 2000) (based on CMBFAST Seljak & Zaldarriaga (1996)) in solid black for the full non-linear power spectrum created with HALOFIT and the pure linear power spectrum in solid red. The solid green line represents the power spectrum of the raw data, not corrected for the window function. Finally the blue dashed line shows the mean of

6000 reconstructions obtained from ARGO after the burn-in phase. Additionally the one and two σ variance bands are shown in cyan and dark blue. We achieve unbiased reconstructions up to scales of about $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$. As expected, we encounter a higher variance on largest scales that are prone to cosmic variance. The negligible variance on small scales is due to the outweighing domination of the prior function as 90% of the reconstructed volume is empty. For scales below $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$ we can clearly see a lack of power. This is caused by the fact that in this analysis we did not apply any correction for the FOG that are predominant at these scales and therefore reduce the power. However, on scales relevant for BAO and LSS analyses, the reconstructions show excellent agreement with the theoretical predictions.

6.3.3 Density distribution

Now we can analyze the validity of the lognormal model. We do this by evaluating the distribution function of the linear density field $\boldsymbol{\delta}_{\mathrm{L}}$ by histogramming the matter overdensity at each cell, shown in Figure 6.4. To estimate the one σ variance band, we took the mean over 6000 iterations.

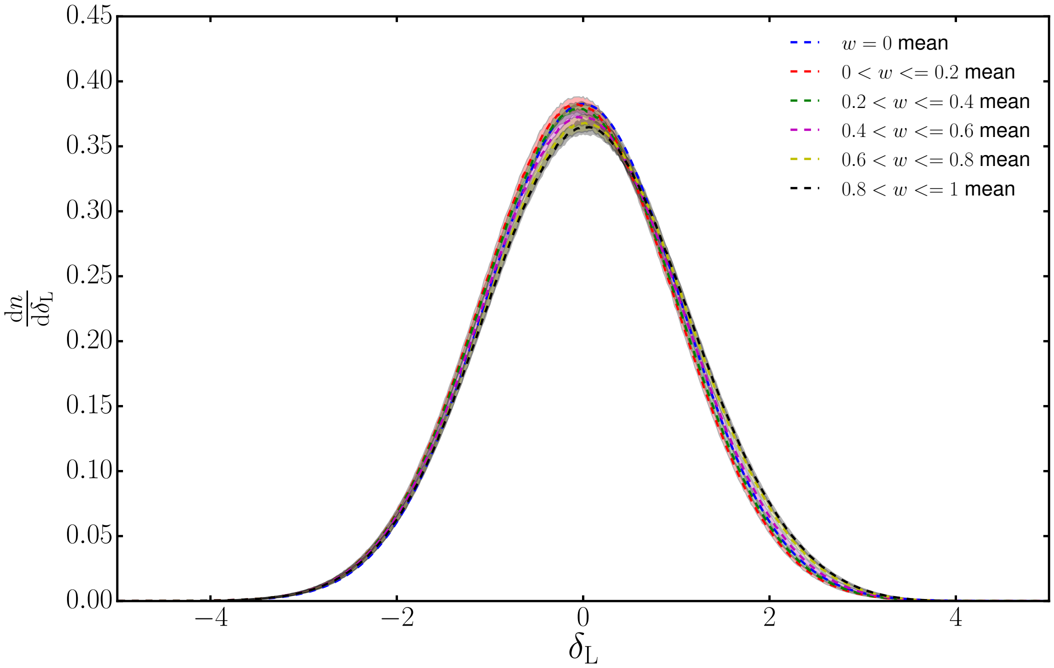


Figure 6.4: Distribution of the matter statistics for different completeness values from 6000 reconstructions on a mesh of 512^3 cells and resolution $d_{\mathrm{L}} = 6.25\,h^{-1}$ Mpc, shown for the linear component $\boldsymbol{\delta}_{\mathrm{L}}$ of the reconstructed density field. The corresponding skewness range is between -10^{-4} and -0.09 with means being always smaller than $|\langle\boldsymbol{\delta}_{\mathrm{L}}\rangle| < 0.13$. The skewness is thus reduced by two orders of magnitude, as compared to a skewness of $\sim$7 corresponding to the galaxy overdensity on a mesh with the same cell resolution.

We find a good agreement through six different completeness domains by reducing significantly the skewness of the galaxy distribution from 6.4 to less than 0.03. Also, the absolute average mean over the different completeness domains is smaller than $|\langle\boldsymbol{\delta}_{\mathrm{L}}\rangle| < 0.18$.

6.3.4 Zel'dovich transformation of matter overdensities

So far we have been reconstructing the linear component of the density field in Eulerian space at a reference redshift within the lognormal approximation. We can, however, get an estimate of the non-linear cosmic web by performing structure formation within a comoving framework, i.e., without including the displacement of structures, as our reconstructed linear density fields already reside at the final Eulerian coordinates. It is possible to derive this kind of mapping from cosmological perturbation theory. We will rely here on the classical Zel'dovich (1970) framework. By demanding mass conservation from Lagrangian $\boldsymbol{q}$ to Eulerian space $\boldsymbol{r}$ with $\varrho(\boldsymbol{q})\mathrm{d}\boldsymbol{q} = \varrho(\boldsymbol{r})\mathrm{d}\boldsymbol{r}$, we get an equation for the cosmic evolved density field within comoving coordinates: $1 + \delta^{\mathrm{PT}}(\boldsymbol{q}) = \mathbf{J}^{-1}$ (with the supercript $^{\mathrm{PT}}$ for perturbation theory), where $\mathbf{J}$ is the Jacobian matrix often called the tensor of deformation

$$\mathscr{D}_{ij} \equiv \frac{\partial^2 \boldsymbol{\Psi}(\boldsymbol{q}, z)}{\partial q_i \partial q_j} \,. \tag{6.24}$$

By doing the proper diagonalization one finds that the comoving evolved density field can be written as

$$\delta^{\mathrm{PT}}(\boldsymbol{q}, z) + 1 = \frac{1}{(1 - D(z)\lambda_1(\boldsymbol{q}))(1 - D(z)\lambda_2(\boldsymbol{q}))(1 - D(z)\lambda_3(\boldsymbol{q}))} \,, \tag{6.25}$$

where λ_i are the eigenvalues of the deformation tensor with $\lambda_1 \geq \lambda_2 \geq \lambda_3$. This framework is helpful to gain insight over the formation of the cosmic web (see Hahn et al., 2007). In fact we could use the reconstructed velocity field to compute the shear tensor and study the cosmic web (Bond et al., 1996). We will however, focus on the largest eigenvalue denoting the direction of first collapse along the principal axis to form the filamentary cosmic web. We can Taylor expand the previous equation within the Eulerian framework yielding

$$\delta^{\mathrm{PT}}(\boldsymbol{r}, z) \simeq D(z)\lambda_1(\boldsymbol{r}) + \lambda^{+}(\boldsymbol{r}, z) \,, \tag{6.26}$$

with λ^{+} being the higher order contributions including the rest of eigenvalues, which can be approximated by $\lambda^{+}(\boldsymbol{r}, z) \simeq -\langle D(z)\lambda_1(\boldsymbol{r})\rangle$. This expression avoids the problem of formation of caustics, as present in Equation 6.25. We have tested other expansions including the rest of eigenvalues, however, with less success in describing the non-linear cosmic web. The operation of retaining the information of the largest eigenvalue can also be interpreted as filtering out the noisy part of the Gaussian field. This technique could potentially be useful to effectively enhance the cosmic web of a low resolution simulation for mock catalog production. Since this theory is based on the Gaussian density field, we will compute the eigenvalues based on the linear component of the density field $\boldsymbol{\delta}_{\mathrm{L}}$. In particular, we will compute them from the gravitational potential $\phi_{\mathrm{L}} \equiv \nabla^{-2}\delta_{\mathrm{L}}$, solving the Poisson equation with the inverse Laplacian operator in Fourier space, to obtain the corresponding tidal field tensor. By applying Equation 6.26 we thus get the linear component of the gravitationally evolved density field in Eulerian space, which we will denote as $\delta_{\mathrm{L}}^{\mathrm{PT}}(\boldsymbol{r})$. We now can compute the non-linear component by doing the transformation $\delta^{\mathrm{PT}}(\boldsymbol{r}) = \exp(\delta_{\mathrm{L}}^{\mathrm{PT}}(\boldsymbol{r}) + \mu(\delta_{\mathrm{L}}^{\mathrm{PT}}(\boldsymbol{r})) - 1$, having the physical meaningful property of yielding positive definite density fields. To ensure that this field shares the same power spectrum, as the lognormal reconstructed density field $\delta(\boldsymbol{r}) = \exp(\delta_{\mathrm{L}}(\boldsymbol{r}) + \mu(\delta_{\mathrm{L}}(\boldsymbol{r}))) - 1$, we apply in Fourier space

$$\hat{\delta}_{\mathrm{L}}^{\mathrm{PT,f}}(\boldsymbol{k}) = \sqrt{P^{\mathrm{trans}}(k)} \frac{\hat{\delta}_{\mathrm{L}}^{\mathrm{PT}}(\boldsymbol{k})}{\sqrt{\langle|\hat{\delta}_{\mathrm{L}}^{\mathrm{PT}}(\boldsymbol{k})|^2\rangle_{\Delta k}}} \,, \tag{6.27}$$

where the non-linear transformed power spectrum $P^{\mathrm{trans}}(k)$ is found iteratively. The ratio between the target power spectrum and the one obtained at a given iteration is multiplied to $P^{\mathrm{trans}}(k)$ from the

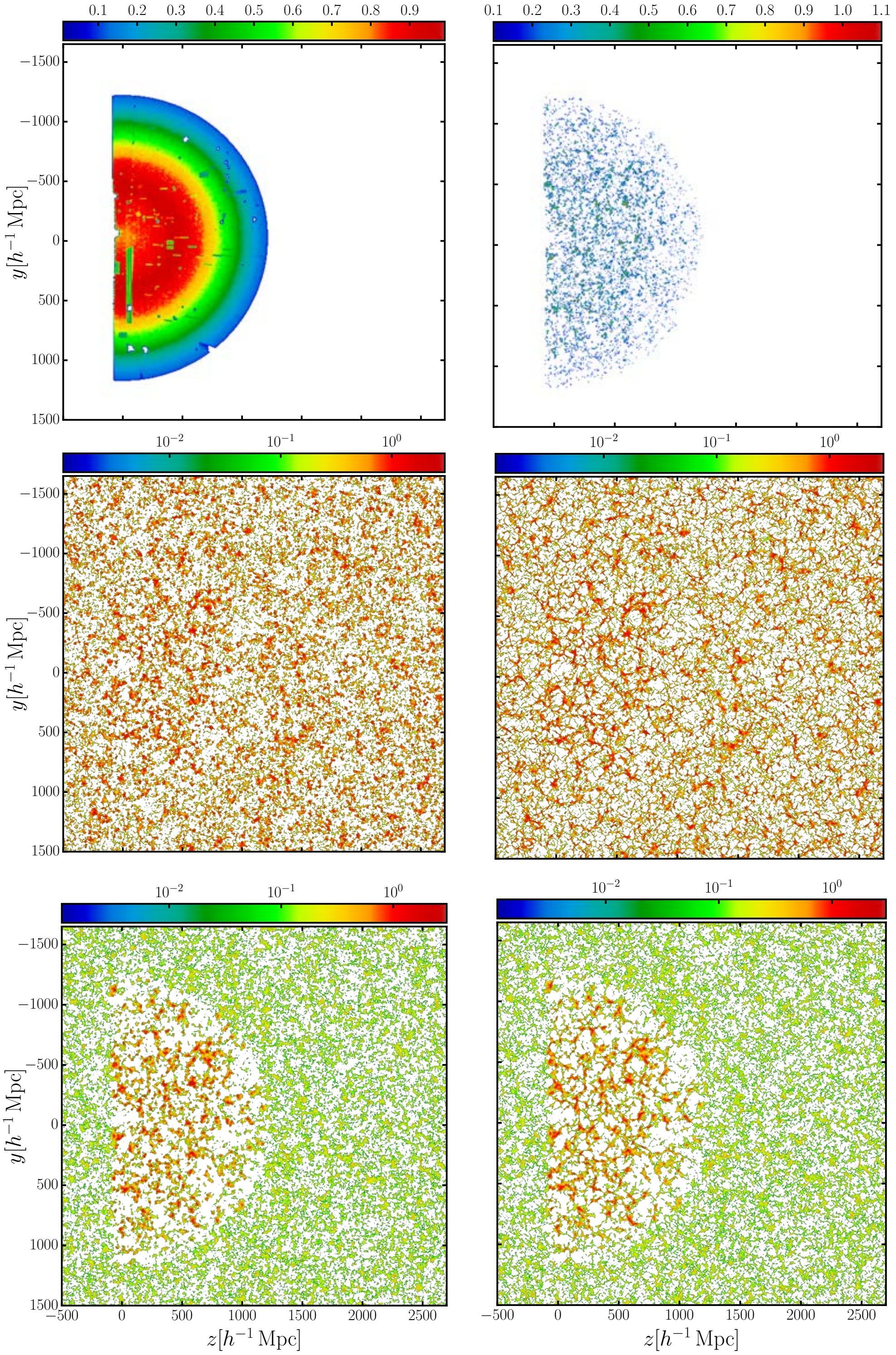
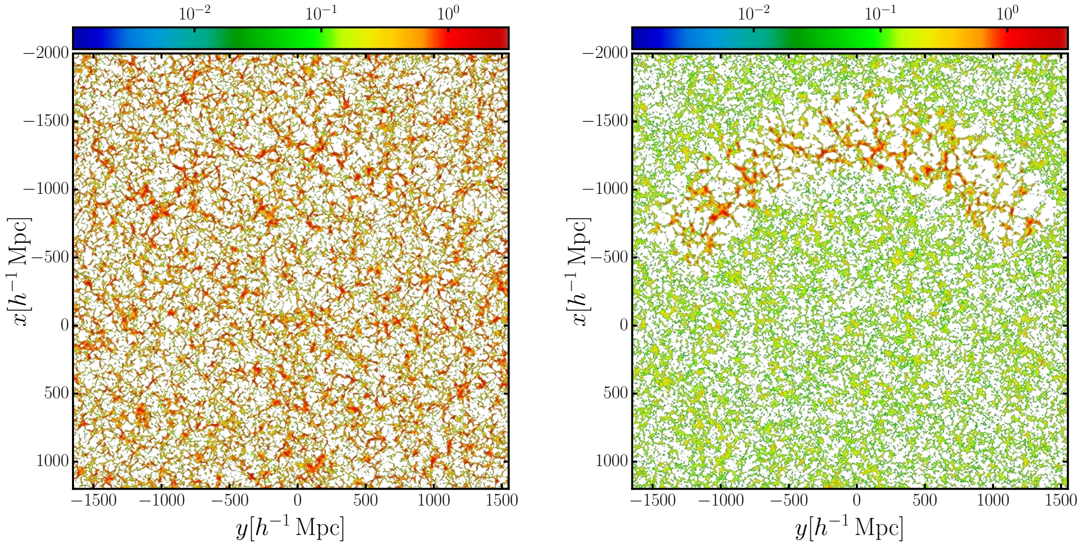


Figure 6.5: Based on a light-cone mock the different slices in the $z - y$ plane with $6.25\,h^{-1}$ Mpc resolution represent: **Top left:** Window function. **Top right:** Galaxy count density. **Middle left:** One reconstructed linear density field. **Middle right:** One reconstructed linear density field after Zel'dovich transformation. **Bottom left:** Mean over 6000 reconstructed linear density fields. **Bottom right:** Mean over 6000 reconstructed linear density fields after Zel'dovich transformation. Window is shown with $6.25\,h^{-1}$ Mpc , others with $\sim 30\,h^{-1}$ Mpc thickness.

previous iteration until the non-linear power spectra averaged in Δk-shells coincide $\langle|\hat{\delta}^{\rm PT,f}(\boldsymbol{k})|^2\rangle_{\Delta k} \simeq \langle|\hat{\delta}(\boldsymbol{k})|^2\rangle_{\Delta k}$ (i.e., the power spectrum from the non-linear transformed lognormal density field), in a given k-range within a given accuracy. As a starting guess of $P^{\rm trans}(k)$ we take $\langle|\hat{\delta}_{\rm L}(\boldsymbol{k})|^2\rangle_{\Delta k}$ (i.e., the power spectrum from the linear lognormal density field). In practice, less than 15 iterations are necessary to be accurate within better than 1% up to at least 70% of the Nyquist frequency using about 100 Δk-bins for meshes of 512^3 cells on cubical volumes of $3200\,h^{-1}$ Mpc side, requiring less than 20 min on six cores. This operation is justified, as we are dealing with the Gaussian component of the density field, permitting us to define a pseudo white noise $\hat{\delta}_{\rm L}^{\rm PT}(\boldsymbol{k})/\sqrt{\langle|\hat{\delta}_{\rm L}^{\rm PT}(\boldsymbol{k})|^2\rangle_{\Delta k}}$, which allows modifications of the two point statistics. In fact, the PDF of $\delta_{\rm L}^{\rm PT}$ to a high precision Gaussian. This calculation is parameter free, and does not require any further input than the lognormal field (and the window function to compute the completeness dependent renormalized mean fields). Effectively, these transformations retain the two-point statistics, while improving the three point statistics of the lognormal field, hereby extracting the cosmic web structure of the density field, which is diluted in the lognormal reconstructions. We note, that the distribution of peaks even prior to the non-linear tidal field transformation do not correspond to a random lognormal realisation, as they are based on the galaxy distribution within the posterior sampling analysis, which already suffered displacements due to the action of gravity. In Figure 6.5 we show exemplary the whole evolution of the analysis in the $z-y$ plane for the light-cone mock. Firstly we show the completeness (window function) on the top left with a thickness of a single cell. This is visually beneficial as one can see the polygons of the mask and also rejected areas. Secondly on the top right we present the raw galaxy count density in the same slice but now with a thickness of $\sim 30\,h^{-1}$ Mpc . The middle panel consists of a single reconstruction of the pure lognormal prior on the left and also after our Zel'dovich transformation method on the right with again $\sim 30\,h^{-1}$ Mpc thickness. One can easily see how the structures of the galaxy counts are visible in the dark matter density field and the empty parts are filled with a prior field. Also it is salient how the Zel'dovich transformation is promoting the filamentary structures of the cosmic web to yield a more realistic structure formation. At the bottom we show the corresponding averages over 6000 iterations for the above motioned slices. As expected the mean of the random prior fields vanish whereas the structures of the galaxy inputs are stable.

In the following we show samples of the Zel'dovich transformed BOSS DR12 CMASS reconstruction in different slices in Figure 6.6.



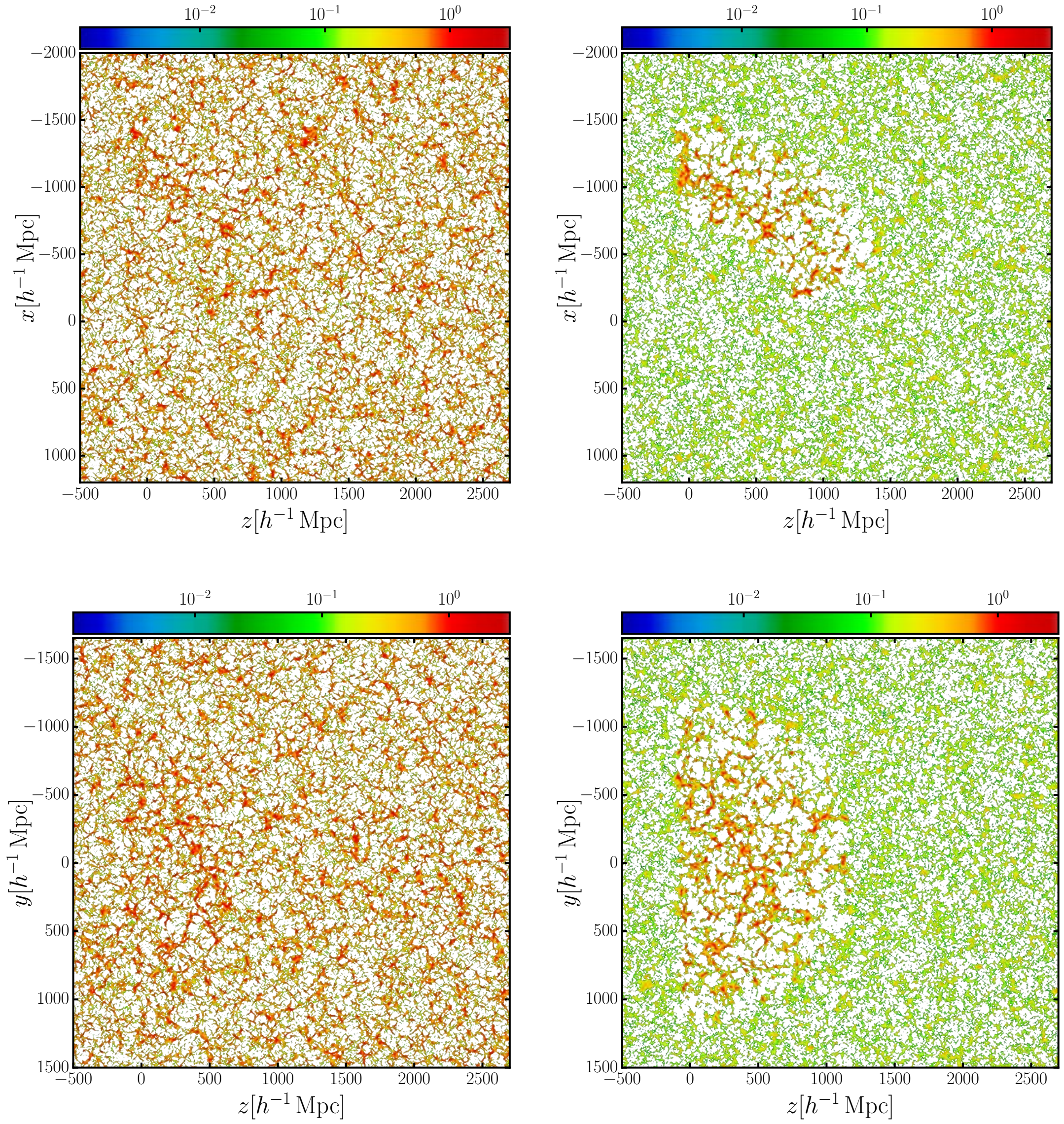


Figure 6.6: The **Left panel** shows slices of one reconstructed linear density field δ_{L} in different spatial coordinates, The **right panel** shows the same slice averaged over 6000 samples. All slices are shown for a thickness of $\sim 30\,h^{-1}$ Mpc and after Zel'dovich transformation was applied.

Within one reconstructed density field the data region is not distinguishable from the empty regions that are augmented with ARGO and postprocessingly undergo a Zel'dovich like transformation utilizing the first principal axis of collapse. In all spatial slice cuts, it can also be seen nicely how the reconstructed data regions are robust against the averaging, as shown in Figure 6.5. In Figure 6.7 we show the average of the reconstructions with the galaxies corrected to their real-space positions on top. We see a good agreement of the galaxy positions and the reconstructed density field.

Finally we investigate the robustness of our reconstructions also by comparing the mean and the standard deviation σ_i^2 of the density in each cell i. For this purpose we show in Figure 6.8 in the $y - x$ and $z - y$ slice the variance over 6000 iterations.

These plots demonstrate that the variance at low densities and unobserved densities is higher than in regions with high number density. In fact it can be disclosed that the variance depicts the inverse of

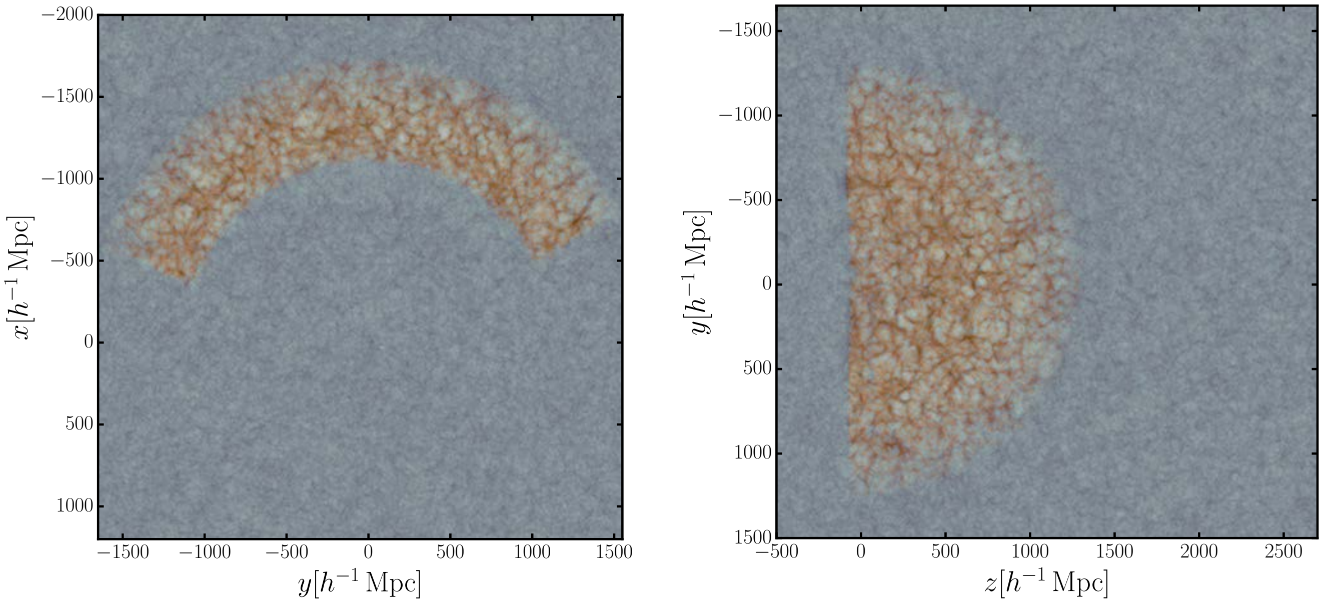


Figure 6.7: Average over 6000 iterations of the $y-x$ slice on the left and $z-y$ slice on the right with a thickness of $\sim 30\,h^{-1}$ Mpc , overplotted with the BOSS DR12 galaxies that have been mapped to real-space.

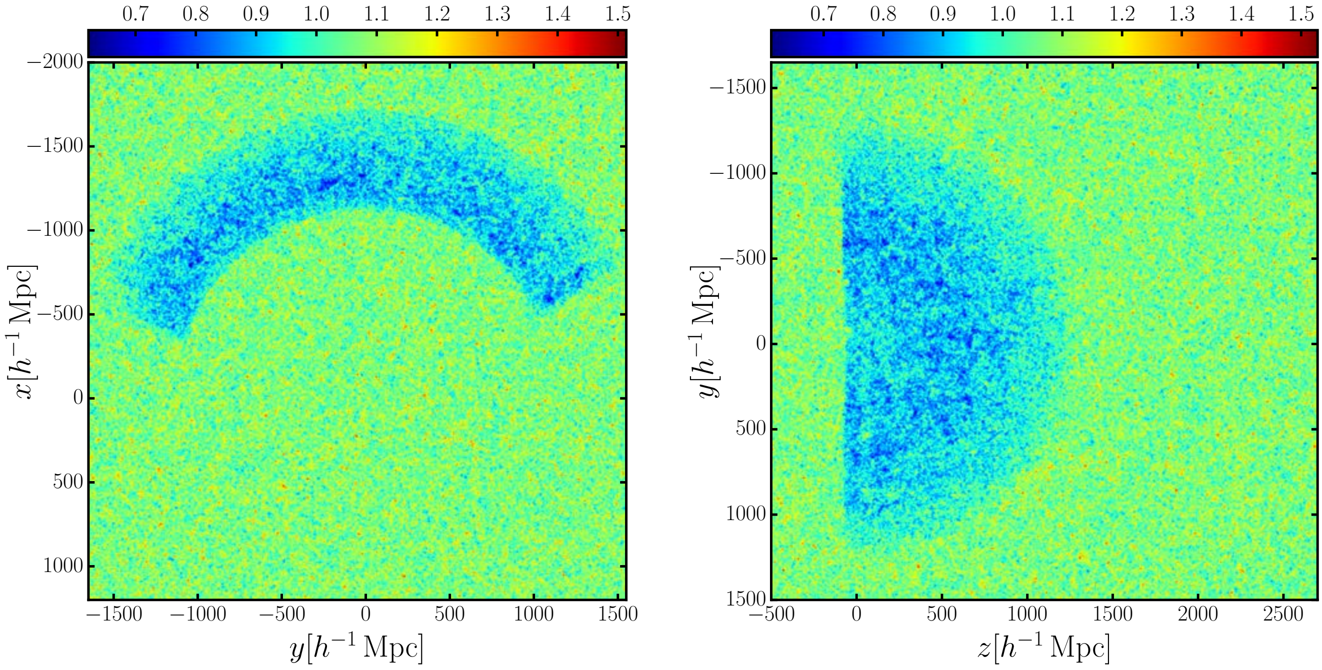


Figure 6.8: Variance over 6000 iterations of the $y-x$ slice on the left and $z-y$ slice on the right with a thickness of $\sim 30\,h^{-1}$ Mpc .

the cosmic web. It also can be seen, that the variance tends to increase at the edge of the light-cone where the selection function is rapidly decreasing towards zero.

6.3.5 Remarks & Conclusions

We found it crucial to sample the mean field in a completeness dependent way. Premature attempts to a reconstruct a light-cone catalogue with widespread empty regions through the reconstructed volume showed that our sampling algorithm yielding a $\boldsymbol{\mu}$ that was in between the range shown in Figure 6.2. This value was neither right for the lognormal random field nor the data region, thus ARGO tries to compensate for this within the HMC causing an overall gradient in the reconstruction. The gradient then causes a severe excess of power at large scales. The renormalized lognormal prior is to some extent compensating the poor three-point description of the lognormal model. Also the Zel'dovich transformation, while retaining the two-point statistics, is improving the three-point statistics as information of the Hessian of the potential is used to estimate the filamentary structures.

The over-dispersion of the galaxy distribution and therefore the stochastic bias is strongly depending on the galaxy population, i.e., galaxy masses, luminosity etc. As we already showed in Section 5.2 more massive tracers tend to have less stochastic bias. This relation can even go beyond Poissonity for very massive objects into under-dispersion (Baldauf et al., 2012, 2013). However in our case the deviation from Poissonity has shown to be negligible for the S_2 Bolshoi subsample as well as for the LRG in the BOSS analysis and the mock galaxies. This has been verified in Figure 6.1 as snapshot analyzed at $z = 0.57$ yielded the same power on small scales whether a Poisson or NB likelihood (up to $\beta = 0.4$) model has been used for the inference analysis. The detailed understanding of each tracer's stochasticity is a crucial ingredient for a successful multi-tracer analysis.

The resconstructed density maps can be used for environmental studies (see, e.g., Nuza et al., 2014), or to study the warm hot inter-galactic medium in filaments, cross correlating them with temperature maps of the cosmic microwave background (see, e.g., Génova-Santos et al., 2015). They can further be used as a reference for future applications including reconstructions of the initial conditions (see e.g. Heß et al., 2013; Jasche & Wandelt, 2013; Wang et al., 2013, 2014). Another promising application is to use the large scale potential derived from the density fluctuations $\Phi \propto \nabla^{-2}\delta_{\mathrm{M}}$ to study the ISW effect.

6.4 Cosmic flows reconstructions

In this section I show in detail how ARGO maps iteratively the galaxies in redshift-space (raw data) into real-space for a galaxy catalogue from a galaxy redshift survey like BOSS or an equivalent light-cone mock. We use the light-cone mock to test the results of our velocity sampling within ARGO with the velocities given by the N-body simulation. As stated in Equation 4.76 we use linear theory including a dispersion term $\boldsymbol{v}_{\mathrm{disp}}$. This dispersion is sampled from a Gaussian of zero mean and a variance corresponding to $\sim 50\,\mathrm{s}^{-1}\,\mathrm{km}$, which has been confirmed by previous studies to yield good results on large-scale structures (Kitaura et al., 2012b),

$$\boldsymbol{v}_i \quad = \quad -f_{\Omega}(a_i)\, H(a_i)\, a_i\, \boldsymbol{\nabla}\nabla^{-2}\delta_i + \boldsymbol{v}_{\mathrm{disp}}\,. \tag{6.28}$$

We solve this equation in Fourier space so that $\boldsymbol{\nabla}\nabla^{-2} \rightarrow -\frac{\boldsymbol{k}}{|\boldsymbol{k}^2|}$. From the above equation we see that each galaxy requires the peculiar velocity to be computed at each individual redshift (scale factor) as the growth rate and the Hubble function are functions of redshift. In practise this is not feasible. We construct a number of 10-15 peculiar velocity fields defined on the same mesh but at different redshifts, i.e., from density fields multiplied with the corresponding growth factors and rates. Each galaxy will get assigned a peculiar velocity field, interpolated to its position within the cell taken from the peculiar velocity mesh at the corresponding redshift bin. In this way we avoid ~ 600000 FFTs on a 512^3 mesh grid.

As we discussed Section 5.3.1, dealing with quasi-virialized RSDs within our framework requires a complete web classification to identify the clusters (knots). In order to increase the power on small scales, a statistical mapping is done to assign a galaxy within a specific range to these identified knots. This is a computationally expansive operation already on a 128^3 mesh and only affects scales $k \geq 0.3\,h\,\mathrm{Mpc}^{-1}$. More precise ways of dealing with quasi-virialized RSD are shown in Heß et al. (2013); Jennings & Jennings (2015). In this project we focus on the coherent bulk flows.

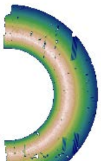
We analyze the reconstructed density that I described in Section 6.3.4 and the velocity inference using the monopole and quadrupole correlation functions in configuration space. With these methods we can effectively explore the influence of RSDs and the outcome of our reconstructions. According to Alam et al. (2016) we define the correlation functions in terms of Legendre polynomials as

$$\xi_l(\boldsymbol{s}) = \frac{2l+1}{2} \int\limits_{-1}^{1} L_l(\mu)\xi(\mu, \boldsymbol{s})\mathrm{d}\mu\,, \tag{6.29}$$

where l gives the order of the Legendre polynomial $L_l(\mu)$, μ is the cosine of the angle between the separation $\boldsymbol{s}$ and line-of-sight direction and $\xi(\mu, \boldsymbol{s})$ the two dimensional correlation function. $l = 0 \rightarrow L_0 = 1$, gives the monopole where as $l = 2 \rightarrow L_2 = (3\mu^2 - 1)/2$, gives the quadrupole correlation function.

In the following we will call the monopole ξ and the quadrupole Q. The results are shown in Figure 6.9. On the left panel we show the monopoles $\xi \cdot r^2$ and on the right the quadruples $Q \cdot r^2$. $Q = 0$ implies that the clustering is isotropic. We start on top with the results from the light-cone mock, and at the bottom panel the results of the BOSS DR12 catalogue. In each plot the red solid line represents the uncorrected galaxy positions in redshift-space. The green solid line is computed from the real-space positions of the mock catalogue. The blue dashed line with the corresponding one and two σ bands are taken from averaging 6000 reconstruction iterations and calculating their mean and variances. The corresponding ratios with the real-space mocks is shown below each plot. We note that showing the real-space mocks for the BOSS reconstruction has solely visual purposes and does not stand for real-space positions of the BOSS galaxy catalogue.

Let us firstly describe the findings for the monopole in the mock catalogue (top left panel). The redshift-space monopole shows higher clustering at distances up to $60\,h^{-1}\,\mathrm{Mpc}$ compared to the real-space one. On large distances beyond the BAO peak ($r \geq 130\ h^{-1}\,\mathrm{Mpc}$) the redshift-space correlation drops below the real-space function. Except for very close pairs where the green curve shows a little peak around $r \sim 5\ h^{-1}\,\mathrm{Mpc}$ the reconstruction is in very good agreement within the variance bands with the real-space mock over the entire distance range r. We also see, that the variance is increasing for larger distances due to decreasing number of pairs. On the top right we show the quadrupole of the mock catalogue. As stated above $Q = 0$ (consequently $Q \cdot r^2 = 0\,h^{-2}\,\mathrm{Mpc}^2$) indicates isotropy of the clustering. As we analyze one single mock catalogue only, we cannot estimate the variance of the anisotropy of the mock. Moreover, the real-space curve is up to small deviations close to $Q = 0$ at all distances r, so that we can rely on it. The redshift-space quadrupole is peaking at small distances and then drops below $Q \cdot r^2 = 50\,h^{-2}\,\mathrm{Mpc}^2$ at scales $r \geq 20\ h^{-1}\,\mathrm{Mpc}$. The anisotropy at small distances is manly driven by the fingers-of-god (FOG) effect whereas on all other distances the Kaiser effect is dominating. The mean of the ARGO reconstructions is well recovering the real-space positions beyond of the FOG dominated area larger than about $r \sim\ 40\ h^{-1}\,\mathrm{Mpc}$. The real-space positions are well within the one and two σ bands over the entire range (but a small dip appears at $\sim\ 140\ h^{-1}\,\mathrm{Mpc}$). We can see how the reconstruction is following closely the features in the real-space quadrupole, tracking the ups and downs. For this reason we are confident to yield robust real-space positions within our velocity sampling. We proceed with the correlation functions computed for the BOSS CMASS



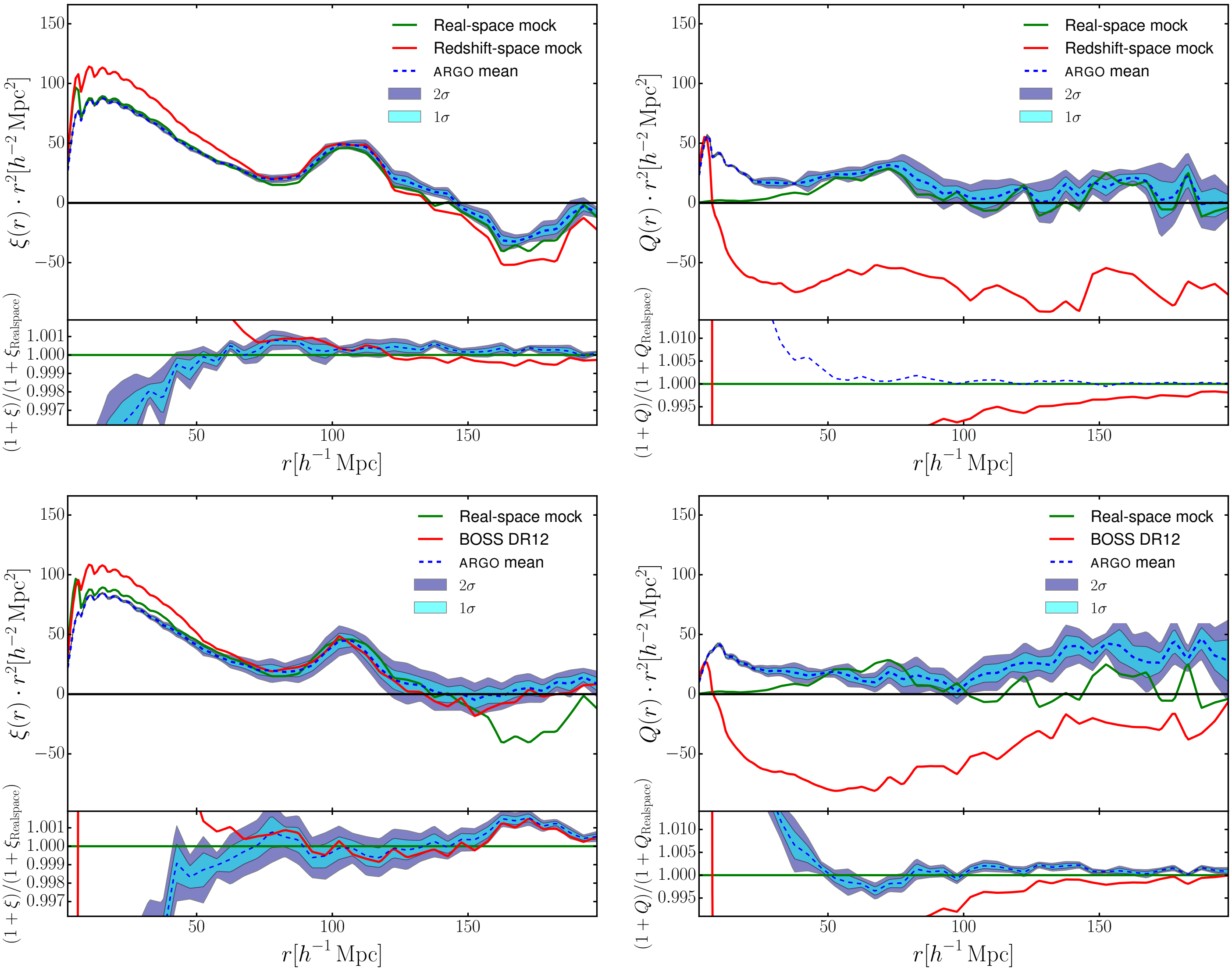


Figure 6.9: **Left:** Monopole correlation function. **Right:** Quadrupole correlation function. **Top:** Light-cone mock based on BMD simulations. **Bottom:** BOSS DR12 data. In red the raw input data in redshift-space curve is shown, in green the real-space curve of the light-cone mock. The dashed blue line represents the mean of 6000 ARGO samples with the corresponding variances. All correlation functions are divided by the real-space mock correlation function.

data sample at the bottom left panel. Different from the mock catalogue the BOSS data needs to be weighted in the scheme that we described in Section 4.5.5, Equation 4.53. We apply these weights within the NGP algorithm to estimate the galaxy number density on the mesh within ARGO and we also account for these weights when calculating the correlation functions. In Appendix A.4 we show the impact if the weights are not applied. The redshift-space monopole for the BOSS catalogue shows a similar shape compared to the mock catalogue at small distances. However at large distances beyond the BAO peak ($r \geq 130\ h^{-1}\,\mathrm{Mpc}$), the red curve does not drop as strongly as the counterpart in the mock catalogue. These are well known systematics of the BOSS catalogue and partly overcome by the applied weights. At $r \geq 150\,h^{-1}\,\mathrm{Mpc}$ the correlation function is nearly constant, confirmed also in Alam et al. (2016). Besides these systematic deviations the monopoles of mock and data are matchable (the real-space mock data are shown in green). The lower right panel shows the quadrupole for the BOSS data. Interestingly at very small distances the FOG seem to be causing less anisotropy in the data as compared to the mock. This is due to the FOG model in Rodríguez-Torres et al. (2016). Beginning at $r \sim 45\ h^{-1}$ Mpc the mean of our reconstructions' quadrupole is in conformity with the zero-line within the variance bands, concluding that the inferred real-space positions do isotropize the BOSS data.

The raw quadrupole of the BOSS data in redshift-space shows less anisotropy at large distances of $r \sim 120\,h^{-1}$ Mpc compared to the light-cone mock. This can be due to cosmic variance effects or systematics in the data. However, with only one realization of the mocks we can not make a concrete claim about this feature.

For visual reasons the real-space quadrupole of the mock catalogue is shown in the plot. The results show that we yield entirely satisfactory recovery of real-space positions relevant for analyses of the large-scale structures.

6.4.1 Comparison to mock velocities

In this section I will compare closely the inferred velocities of the mock light-cone catalogue to the velocities obtained from the actual N-body simulation. The results are shown in Figure 6.10. Each point in the two dimensional plot corresponds to the y-component of the velocity of a mock galaxy. The inferred velocity is placed on the y-axis and the mock velocities on the x-axis, additionally the color bar shows the density of galaxies. We use the correlation coefficient to estimate the degree of correlation, which is defined as the off-diagonal covariance normalized over the individual variances $r_{ij} = \frac{\textbf{COV}(i,j)}{\sigma_i^2\sigma_j^2}$. We start with a single reconstruction on the top left. Just by eye we can see the tendency of points along the diagonal line though the origin that we plot in black solid. We yield a correlation coefficient of $r = 0.56$, facing a small tilt also in the correlation. On the top right we show the correlation for the mean of 6000 reconstructions. Clearly the correlation is improved significantly to $r = 0.64$, removing the tilt. The majority of galaxies are located in the central reddish region along the black line. This result is obtained despite the presence of 10% of the CMASS galaxy being satellites that are dominated by viral motions around the central galaxy. As a proxy we consider two cases shown in the bottom panel, excluding galaxies for which the velocity difference between mock and reconstructed exceeds 650 and 400 s^{-1} km, respectively. The first one, shown at the bottom left panel, removes $\sim$3.5% of the galaxies, and the second one on the bottom right removes $\sim$10%. Since not all satellite galaxies will be outliers the true answer will be statistically closer to the first case, raising the statistical correlation coefficient to about $r = 0.75$, which is a considerable improvement with respect to other methods in literature.

We show the velocities reconstructed with ARGO on a 512^3 mesh, and compare these to the mock velocities on the same mesh grid. The mock velocity mesh is created by a modified NGP algorithm. We assign the positions of the galaxies to the grid cells and average the peculiar velocities of these galaxies in each dimension x, y, z. This method is sufficiently precise if the bulk flow is dominating, yet gets spoiled by viral motions. To have a fair comparison we apply a Gaussian smoothing of $r_\mathrm{S} = 4\,h^{-1}$ Mpc to mock velocities. The results are shown in Figure 6.11 with a thickness of $50\,h^{-1}$ Mpc in the $z - y$ slice.

We note that the general pattern is agreeing for the mocks and the reconstructions. However it is important to note that due to the averaging of the mock galaxy velocities within a grid cell, the amplitude of the mock slice is reduced as expected. We still see a correlation over the volume compared to the mean of 6000 iterations. What is also noticeable in the mean reconstruction, is that although most of the empty region is averaged to zero, empty regions immediately at the border to the data

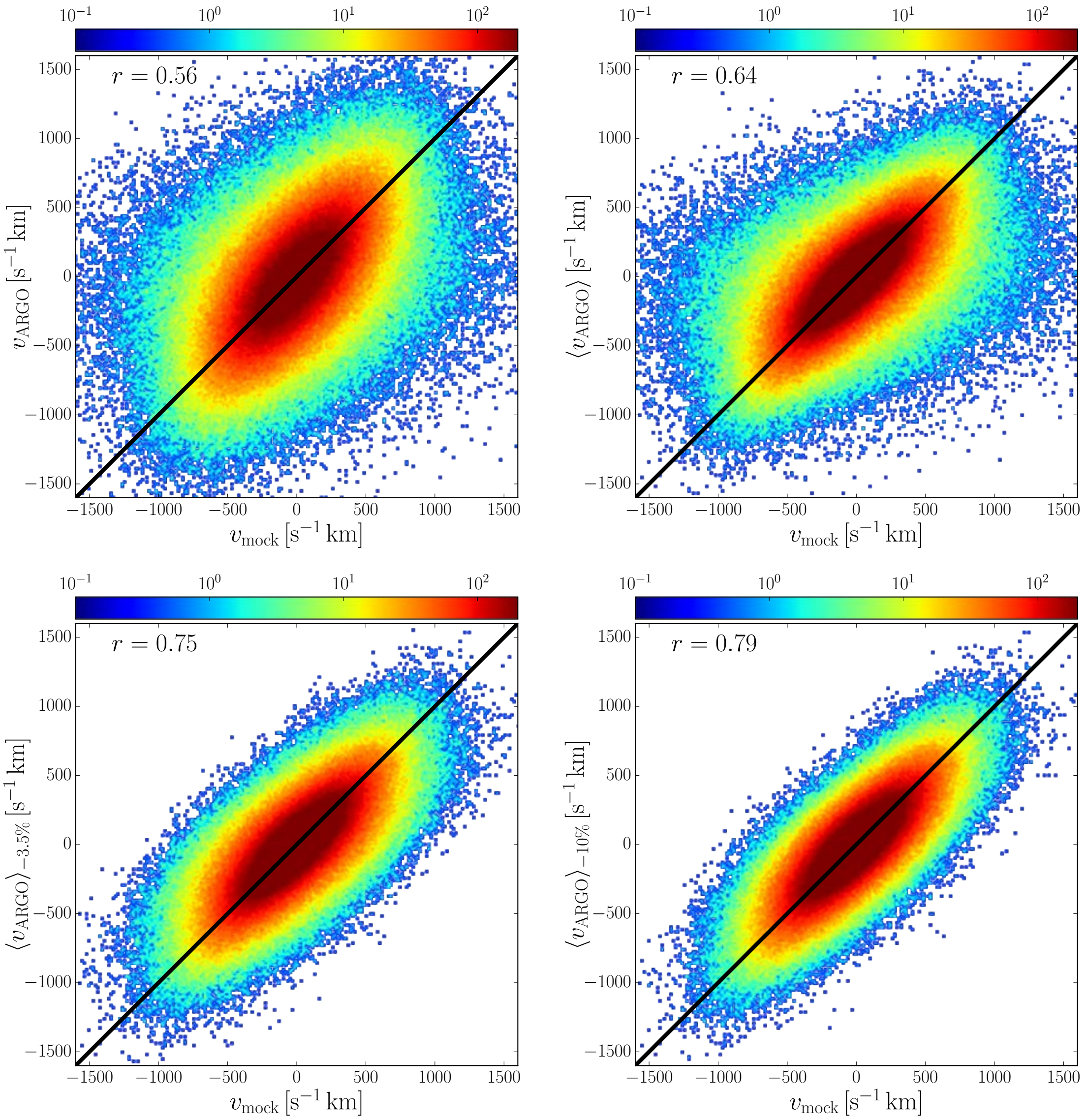


Figure 6.10: Correlation of the y-component of the velocity field of the light-cone mocks and for reconstructions based on the mocks with resolutions of $d_{\mathrm{L}} = 6.25\,h^{-1}$ Mpc with an additional Gaussian smoothing of $r_{\mathrm{S}} = 2\,h^{-1}$ Mpc. **Upper left panel:** for one reconstructed sample yielding a correlation coefficient of $r = 0.56$, **upper right panel:** for the mean over 6000 reconstructed samples with a correlation coefficient of $r = 0.64$, **lower left panel:** same as upper right panel, but excluding galaxies for which the difference in the velocity reconstruction exceeds $|\boldsymbol{v}| = 650\,s^{-1}$ km (i.e., excluding about 3.5% of the sample) yielding $r = 0.75$, and **lower right panel:** same as upper right panel, but excluding galaxies for which the difference in the velocity reconstruction exceeds $|\boldsymbol{v}| = 400\,s^{-1}$ km (i.e., excluding about 10% of the sample), yielding $r = 0.79$. The color denotes the number density of mock galaxies.

region seem to be more stable against averaging. This is in fact a reasonable result as in linear theory the velocity is determined by the potential flow, which means that the flow will be directed from low to high density regions. In turn, this means that the border of the empty regions will be constrained by the data. Thus we do not expect the closest boundary region to completely vanish due to averaging.

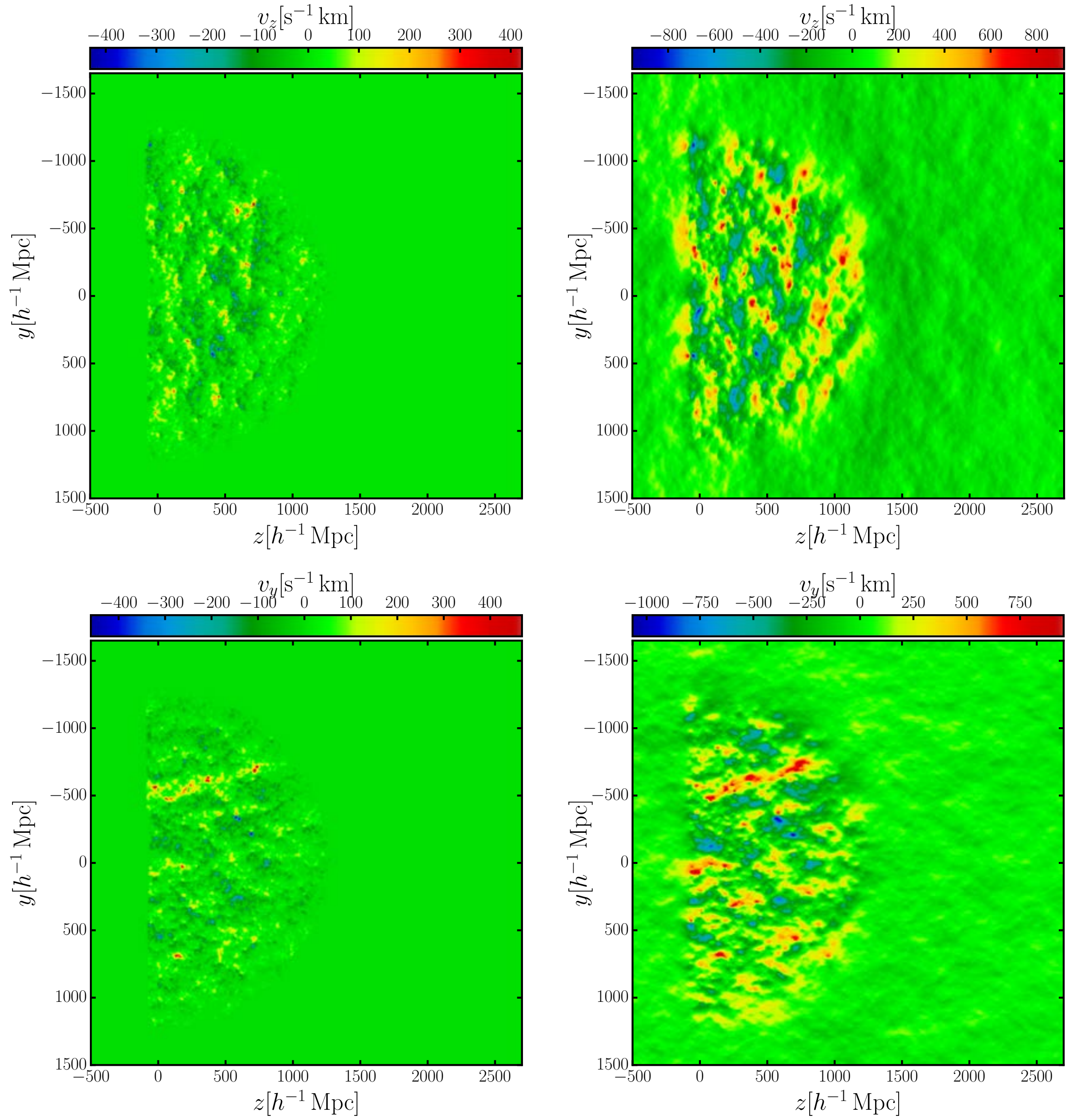


Figure 6.11: Slices shown in the $z - y$ plane of the y and z-component of the velocity field with a thickness of $50\,h^{-1}$ Mpc , shown for the light-cone mock and for ARGO reconstructions based on the mock, with resolutions of $d_\mathrm{L} = 6.25\,h^{-1}$ Mpc and an additional Gaussian smoothing of $r_\mathrm{S} = 4\,h^{-1}$ Mpc. **Upper left panel:** v_z of the mock, **upper right panel:** v_z for the mean of 6000 ARGO reconstructions, **lower left panel:** v_y of the mock, and **lower right panel:** v_y for the mean over 6000 reconstructions. The different magnitude for the mock velocities is due to the averaging off all galaxies' velocity within one grid cell, whereas the velocity of the ARGO reconstructions are inferred on the mesh grid itself.

The color code represents the magnitude of the corresponding velocity, v_z or v_y respectively.

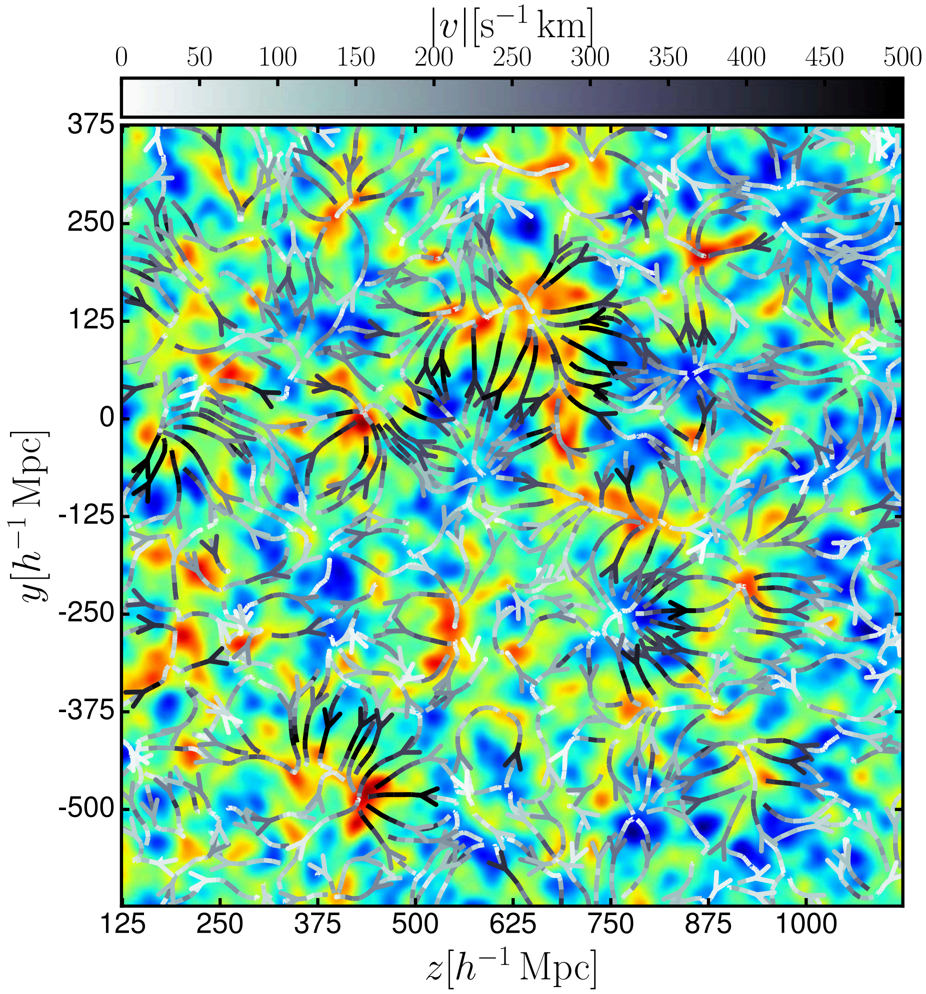


Figure 6.12: Slice of a reconstructed density field from BOSS CMASS DR12 data in the $z-y$ plane, smoothed with a Gaussian kernel of radius $r_{\mathrm{S}} = 10\,h^{-1}$ Mpc. The two dimensional $v_z + v_y$ velocity is overplotted on top of the density field. The color bar marks the magnitude of the velocity vector $|\boldsymbol{v}|$. Red color denotes high density areas, and blue color low densities.

In Figure 6.12 we show a streamline visualization of the bulk flows on top of a smoothed density field reconstructed from BOSS data. For visual reasons we just show a patch of $1250\,h^{-1}$ Mpc side length in the $z-y$ plane with a thickness of $30\,h^{-1}$ Mpc . The density was smoothed additionally with a radius of $r_{\mathrm{S}} = 10\,h^{-1}$ Mpc. In this figure, the density of streamlines show the strength of the field, i.e. the gradient of the potential, whereas the grey color code indicates the particular magnitude of the velocity vector $|\boldsymbol{v}|$ in units of s^{-1} km. It can be nicely seen how the flow is pointing from underdense areas, shown in blue, to overdense regions, shown in red. This is in good agreement with the previous findings. Figure 6.12 visualizes impressively the findings of the cosmic web classification, which we mathematically expressed with the tidal field tensor. Let us illustrate this point with three example structures that we find in the figure:

1. At $z \approx 875\ h^{-1}$ Mpc , $y \approx 75\ h^{-1}$ Mpc .

   This is a highly underdense region that we classified as a void. The eigenvalues of the tidal field tensor T_{ij} of this structure are all negative, which can be seen as in all directions the velocity field is streaming out of the void into the more dens structures around the void.

2. At $z \approx 760\ h^{-1}\,\mathrm{Mpc}$, $y \approx -125\ h^{-1}\,\mathrm{Mpc}$.
   This is a filament, meaning that along the elongated axis the velocity is flowing through this structure, whereas the filament itself is fed by the in-streaming flow perpendicular to this elongated axis. This means, that the filament is fed from two perpendicular axes and flows along the axis, that is not collapsed, in accordance to the two eigenvalues of the tidal field tensor larger than one, and one eigenvalue less than one.

3. At $z \approx 900\ h^{-1}\,\mathrm{Mpc}$, $y \approx -200\ h^{-1}\,\mathrm{Mpc}$.
   This is an isolated high density area, called knot. We can see that the surrounding densities are lower, as shown by the blue color. Characteristically we see how the knot is fed by its surroundings as the flows from all directions are pointed to the knot, corresponding to three positive eigenvalues of the tidal field tensor.

6.4.2 Velocity correlation

Another useful tool to analyze the goodness of the reconstructed velocity field is the power spectrum of the scaled divergence of velocity field $P_{\theta\theta}$ with

$$\theta \equiv -\frac{1}{f_\Omega H a}\nabla \cdot \boldsymbol{v}\,. \tag{6.30}$$

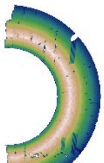
The outcome of the velocity field reconstruction with different smoothing scales is shown in Figure 6.13. As the smoothing scale could be considered as a free parameter within our model, we use the velocity divergence power spectrum to estimate the optimal smoothing length.

In particular, one expects $P_{\theta\theta}$ to converge towards the linear power spectrum in the transition to the non-linear regime at about $k \sim 0.15 - 0.2\,h\ \mathrm{Mpc}^{-1}$ (Hahn et al., 2015; Jennings, 2012).

This is expected as the velocity divergence is closer to the Gaussian field than the gravitationally evolved density field (see, e.g., Kitaura et al., 2012b). In fact while the density is enhanced in the potential wells, virialization prevents galaxies from getting larger and larger velocities. As a consequence, the power spectrum of the velocity divergence is close to the linear density field in the quasi-linear regime, eventually being even more suppressed at high k values. Figure 6.13 shows that such an agreement down to scales of $k \sim 0.2\,h\ \mathrm{Mpc}^{-1}$ is indeed achieved for smoothing scales of about $r_{\mathrm{S}} = 2\,h^{-1}$ Mpc. In fact for a smoothing scale r_{S} between 1 and 2 h^{-1} Mpc one can potentially obtain unbiased results beyond $k = 0.5\,h\ \mathrm{Mpc}^{-1}$. Therefore a smoothing scale for our reconstructions with 512^3 mesh grid were run using $r_{\mathrm{S}} = 2\,h^{-1}$ Mpc.

6.4.3 Radial selection function sampling

As RSDs affect the distances in line-of-sight direction, the radial selection function $f(r)$ shown on the left in Figure 3.2 will be affected by redshift-space positions on the galaxies. This is called Kaiser-Rocket effect. Ideally the radial selection function should be evaluated in real-space. We address this problem within ARGO and sample the radial selection function as described in Section 4.8. We therefore create once the *raw* window function without any radial dependence from the angular mask at the beginning of the Monte-Carlo chain. Then while sampling the velocities we can trivially multiply the newly found radial selection function with the *raw* window iteratively. The outcome of this sampling is shown in Figure 6.14.



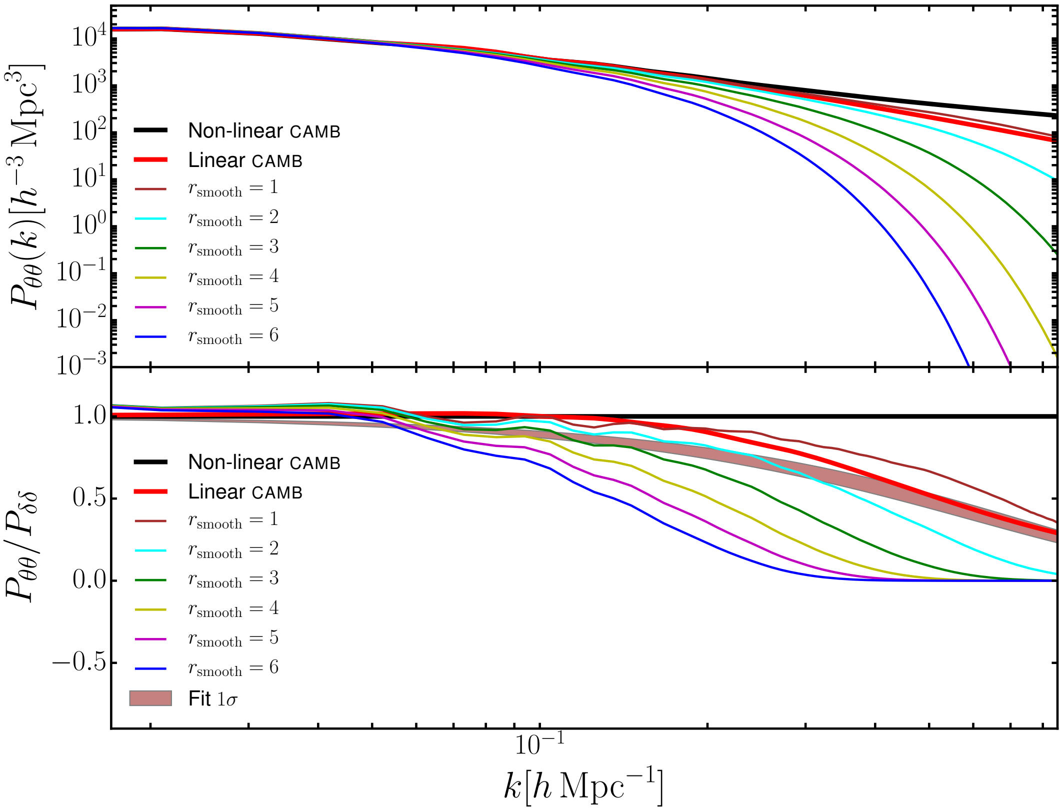


Figure 6.13: **Upper panel:** power spectrum of the scaled divergence of the peculiar velocity field θ, for different smoothing scales (shown in units of $h^{-1}\,\mathrm{Mpc}$) for a typical realisation on a mesh of 512^3 with resolution $d_{\mathrm{L}} = 6.25\,h^{-1}$ Mpc. **Lower panel:** ratio wr.t. the non-linear power spectrum from Heitmann et al. (2010). The shaded region represents the theoretical fit for the velocity divergence bias $b_v = \mathrm{e}^{-(k/a)^b}$ by Hahn et al. (2015) with the σ region being computed based on the largest uncertainty found with the parameters a and b. The wiggles are due to the more pronounced baryon acoustic oscillations in the mean theoretical power spectrum than in the particular realisation used in this plot.

It can be seen that the mean of the inferred radial selection functions coincides with $f(r)$ in redshift-space within the variances down to $r = 1150\ h^{-1}$ Mpc in consensus with the findings of Rodríguez-Torres et al. (2016). Therefore, we see very little dependence whether the selection function is sampled or not. However this may not represent the general case and might be exceptional to the CMASS survey. This method can be still applied to any survey and may proof to be important for higher redshifts.

6.4.4 Remarks & Conclusion

I showed the velocity sampling method in ARGO for a light-cone galaxy catalogue and a mock light-cone. This method relies entirely on linear theory with a dispersion term included, and is able to isotropize the galaxy positions down to distances $r \sim 40\,h^{-1}$ Mpc . The relevant physical quantities are calculated for 10-15 redshifts bins. Afterwards the galaxy positions are interpolated to these bins. After the velocities are computed, the galaxy positions are corrected for the distortion due to peculiar velocities. As linear theory tends to overestimate the coherent bulk flow, we apply a smoothing of the overdensity field of which the velocities are calculated from. We estimate the optimal smoothing within our reconstruction procedure from the power spectrum of the velocity divergence. Although we are using

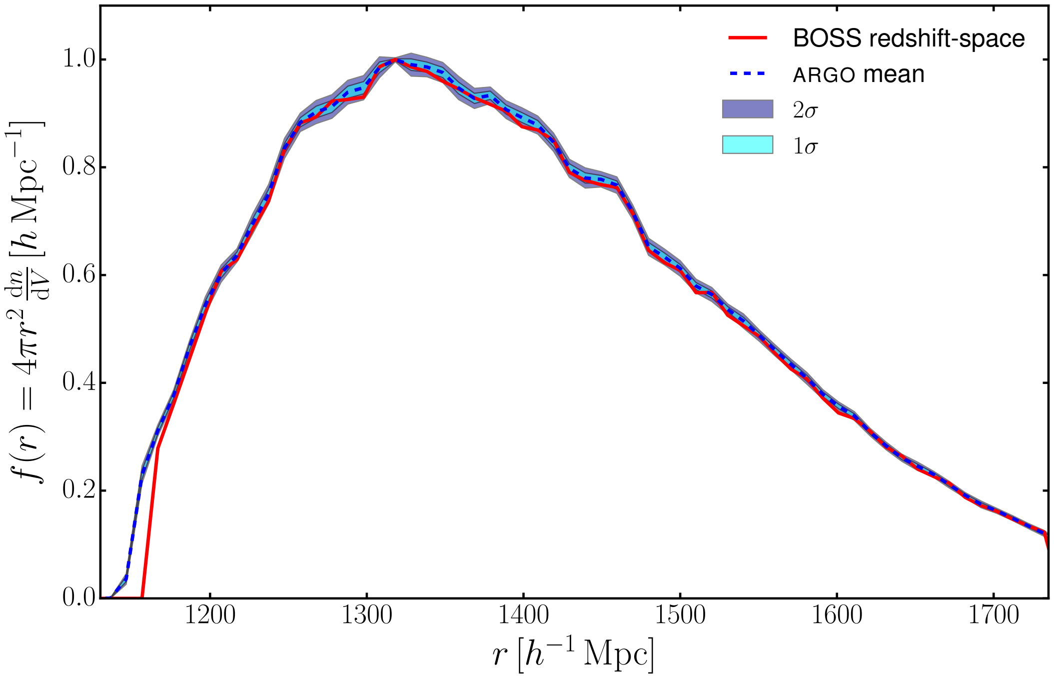


Figure 6.14: Radial selection function $f(r)$ for the BOSS CMASS galaxy survey, normalized to unity before (red solid line) and after RSDs corrections with ARGO (blue dashed line) with corresponding variance bands. The mean is calculated by calculating $f(r)$ for 6000 reconstructions.

only linear theory here, our method includes a couple of ingredients which can explain the improvement we achieved compared to previous methods, such as being a self-consistent iterative method, yielding linearized density fields, for which the pixel window function has been exactly solved (the counts in cells, i.e., the nearest grid point). This is done as we reconstruct the density given the galaxy count per cell exactly so that the reconstructed power spectrum is not suffering from shot noise effects. Also the non-linear bias description with a power law is an important element yielding robust velocity fields. Tidal field corrections could be included in the model. One could try to get improved velocity reconstructions from the linear component rather than from the non-linear one, like it is done in Falck et al. (2012); Kitaura & Angulo (2012). Nevertheless, there is a (nearly constant) bias from the lognormal transformation present in the linear density field, which we intent to avoid to reduce the number of parameters (see Neyrinck et al., 2009). The next aim of our reconstruction method is to have less reliance on cosmological parameters. Therefore it is desirable to have a joint growth rate sampling. A possible idea for this would be to use the two dimensional power spectrum or correlation function and try to derive isotropy self-consistently. However degeneracies of growth rate, bias and the mass fluctuation on the scale of 8 h^{-1} Mpc, σ_8, impede this undertaking. An interesting outcome from this analysis is that, before the redshift dependent bias was included we have not been able to receive a quadrupole correlation function that was as much compatible with zero for the complete r range, as we just showed. This demonstrates that without redshift dependent bias the density field is involving an intrinsic gradient that induces an imbalance while reconstructing the velocity field.

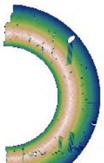
7

Summary and Outlook

The presented thesis is a contribution to study the physical properties of large-scale structure formation within the ΛCDM cosmology model. The main aim of my work was to build a reliable algorithm that is able to jointly reconstruct cosmic density and velocity fields.

Cosmological reconstructions offer a unique environment to combine observational data taken from galaxy surveys and models that describe physical phenomena and interactions leading to structure formation at large scales. On the other hand, within reconstructions the accurate statistical and numerical modelling of large data sets and their interpretation in a probabilistic context is a growing challenge in the era of precision cosmology today. In this thesis I addressed all these issues mentioned above within the realized projects, putting emphasis to the reconstruction method ARGO. I recapitulated the basics of the ΛCDM cosmology. Initially, I introduced the concept of a homogeneous and isotropic universe which then was followed by analytic approaches to describe density and velocity perturbations. Also, the most recent bias and structure formation models that we studied with our methods are comprised. Subsequently, I provided a basic summary of both, N-body simulations with the associated process of halo finding, and galaxy redshift surveys, especially the Sloan Digital Sky Survey. Further, in a general introduction to Monte-Carlo methods I presented background knowledge necessary to follow the functioning of our approach. I put strong emphasis on sampling methods, the historical development of random walk solvers and then more sophisticated methods like the Hamiltonian Monte Carlo method, to avoid random walks and route the direction of the sampled phase-space using Hamiltonian equations of motion. Finally, I explain our reconstruction method ARGO in detail. This algorithm relies on Bayesian inference that samples the density of dark matter perturbation and the corresponding velocity field on the mesh grid given a set of tracers. ARGO intrinsically utilizes accounts for observational masking effects, a complex biasing scheme that may or may not incorporate light-cone effects such as growth functions, a prior model for structure formation, and also jointly infers the velocity field.

In this work I focused on two main aspects, which are totally new to reconstruction methods, firstly modelling a non-linear, stochastic galaxy bias; and secondly including redshift-space distortions within the reconstruction process. I demonstrated that these ingredients are essential to obtain accurate dark matter field reconstructions from galaxy redshift surveys by comparing them to *state-of-the-art* procedures.



The work comprised in this thesis led to three publications.

- The first project showed for the first time the importance of modelling the non-linear bias with a power law relation, and the deviation from Poissonity to account for the discreteness of the galaxy distribution. Also a threshold bias scheme has been applied in a reconstruction for the first time. This work was based on a high resolution dark matter simulation.
- The second project presented the method to correct for redshift space distortions in a realistic mock galaxy catalog of luminous red galaxies.
- The third and main project of my thesis joined all the knowledge gained in those previous works to thoroughly verify the accuracy in the reconstruction method from galaxy redshift surveys suffering additionally observational systematic uncertainties, such as, a radial selection function, survey geometry, or cosmic evolution, and was applied to both accurate mock catalogs and observational data. Moreover, I have shown that the method presented here yields accurate reconstructions of the cosmic flows and the cosmic web on large scales.

A detailed summary of the conclusions extracted in these publications and improvements we achieved, is presented below.

Non-Poisson likelihoods and stochastic bias

As more and more dense galaxy redshift surveys are about to scan the sky, a better understanding and precise modelling of the power spectrum at a large range of scales is needed. However, the crude assumption of galaxy counts per volume following a Poisson distribution fails especially at small scales. We found that a reconstruction based upon a Poisson likelihood lacks power as compared to theoretical predictions at already $k = 0.2\ h\,\mathrm{Mpc}^{-1}$.

Therefore we developed a Bayesian reconstruction algorithm, able to produce unbiased samples of the underlying dark matter field from non-linear stochastic biased tracers up to scales of $k \sim 1\,h\,\mathrm{Mpc}^{-1}$. We divide the bias into a deterministic and a stochastic part. The deterministic bias is modelled with an analytic relation linking galaxy and dark matter density, whereas the stochastic bias must be modelled statistically. This is because stochastic bias alters the dispersion of the galaxy counts per cell given an underlying dark matter density. Incomplete models only consider a linear relation of galaxy and dark matter density, which is a very good approximation on largest scales, but fails on smaller scales. We implemented a power law relation with additional threshold for the deterministic bias, supported by the outcome of studies with simulations (most recently in Neyrinck et al., 2014) and also theoretical predictions (Bardeen et al., 1986; Cen & Ostriker, 1992).

For the stochastic bias we focused on the negative binomial (NB) distribution function and discussed its parameter to change the dispersion, showing that the disperion of the NB distribution is larger or equal to the dispersion of a Poisson distribution.

We have also introduced an iterative sampling scheme to deal with strongly biased objects tracing the high density peaks and thus are modelled with a threshold bias.

In particular, we have tested our algorithm with the Bolshoi N-body simulation, inferring the underlying dark matter density field from a subsample of the corresponding halo catalogue. We found that a Poisson likelihood yields reconstructions with power spectra deviating more than 10% at $k = 0.2\,h\,\mathrm{Mpc}^{-1}$. Our method shows that we can draw nearly unbiased ralizations of the underlying dark matter density field (compatible within $1\,\sigma$) from the posterior distribution up to scales of about $k \sim 1\,h\,\mathrm{Mpc}^{-1}$ in terms of power-spectra and cell-to-cell correlations with the negative binomial probability distribution function.

We have furthermore analytically shown that our method is able to perform a multi-tracer analysis, meaning that within our statistical approach we can account for different sets of tracers, e.g. galaxy populations with different biases, combining them in the likelihood model of ARGO.

Bayesian redshift-space distortions correction

In this project we have presented a Bayesian technique to correct for both coherent and virialized redshift-space distortions inherent in galaxy catalogues by estimating the distance to the individual galaxies. We have demonstrated that this technique is accurate at least up to $k \lesssim 0.3\,h\,\mathrm{Mpc}^{-1}$ in the isotropization of the 2D power spectrum based on precise galaxy mock catalogues describing the *Constant Mass* (CMASS) galaxy sample of the *Baryon Oscillation Spectroscopic Survey* (BOSS) of the Sloan Digital Sky Survey (SDSS). However, the method is general enough to be precise down to far smaller scales, as indicated by the recovered power spectra.

While traditional redshift-space distortions measurements focus on the growth rate, our approach is complementary and more general with the advantage that it deals with non-linear structure formation, non-linear and stochastic galaxy bias, yielding also, as a by-product the real-space positions of the individual galaxies. This technique is promising for a broad number of applications, such as correcting for photo-metric redshift-space distortions including the cosmic web information, or to make precise environmental studies, as demonstrated in Nuza et al. (2014) with a similar forward method recovering the corresponding primordial fluctuations. We have demonstrated in particular that it is a potentially interesting technique for the estimation of the growth rate of the matter density by the isotropization of the 2D power spectrum.

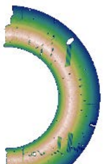
The Clustering of Galaxies in the Completed SDSS-III Baryon Oscillation Spectroscopic Survey: Phase-space Reconstructions of Cosmic Flows and Cosmic Web from Luminous Red Galaxies

In this work, we have presented a Bayesian phase-space (density and velocity) reconstruction of the cosmic large-scale matter density and velocity field from the SDSS-III BOSS DR12 CMASS galaxy clustering catalogue. We have demonstrated that our models can yield accurate results on scales larger than $k \sim 0.2\,h^{-1}\,\mathrm{Mpc}$.

In particular we have used a set of simple assumptions. Let us list them here:

- the statistical posterior distribution of galaxies is described by the lognormal-negative binomial model,
- linear theory relates the peculiar velocity field to the density field,
- the volume is a fair sample, i.e. ensemble averages are equal to volume averages,
- cosmic evolution is modelled within linear theory with redshift dependent growth factors, growth rates, and bias,
- a power law bias, based on the linear bias multiplied by a correction factor, which can be derived from renormalized perturbation theory, relates the expected galaxy number counts to the underlying density field.

This has permitted us to derive the three bias parameters, namely amplitude, dispersion and threshold, consistently from the data, with a given smoothing scale and a particular ΛCDM cosmological parameter set.



We have included a number of novel aspects in the ARGO code extending it to account for cosmic evolution in the linear regime. These improvements are implemented within the statistical approach of ARGO:

- the density fields with a lognormal-NB model,
- the mean fields of the lognormal renormalized priors for different completeness values,
- the number density normalization at different redshift bins,
- the real-space positions of galaxies from the reconstructed peculiar velocity fields,
- and the real-space radial selection function from the reconstructed real-space positions of galaxies (accounting for the "Kaiser-rocket" effect).

Our results show that we can get unbiased dark matter power spectra up to $k \sim 0.2\,h$ Mpc^{-1}, and unbiased isotropic quadrupoles down to scales of about 20 h^{-1} Mpc, being far superior to redshift space distortion corrections based on traditional BAO reconstruction techniques, which start to deviate at scales below 60 h^{-1} Mpc.

As a test case study we also analyze deviations of Poissonity in the likelihood, showing that the power in the monopole and the scatter in the quadrupoles increases towards smaller scales.

The agreement between the reconstructions with mocks and BOSS data is remarkable. In fact, the identical algorithm with the same set-up and parameters were used for both mocks and observations. This confirms that the cosmological parameters used in this study are already close to the true ones, the systematics are well under control, and gives further support to ΛCDM at least on scales of about $0.01 \lesssim k \lesssim 0.2\,h\,\mathrm{Mpc}^{-1}$. Our method to use the first eigenvalue of the displacement field in Eulerian coordinated to reconstruct the filamentary large-scale structure is working remarkably well.

We also found that the reconstructed velocities have a statistical correlation coefficient compared to the true velocities of each individual light-cone mock galaxy of $r \sim 0.7$ including about 10% of satellite galaxies with virial motions. The power spectra of the velocity divergence agree well with theoretical predictions up to $k \sim 0.2\,h\,\mathrm{Mpc}^{-1}$. This is far superior to the results obtained from simple linear reconstructions of the peculiar velocities directly applied on the smoothed galaxy field for which statistical correlation coefficients of the order of 0.5 are obtained (Planck Collaboration et al., 2016). Also, while linear theory tends to overestimate the peculiar velocity field, the chosen grid resolution with the additional smoothing compensates for this yielding unbiased reconstructed peculiar motions. We have seen that for a given resolution the additional Gaussian smoothing radius (and the cell resolution) can be derived from the velocity divergence power spectrum to match the linear power spectrum in the quasi-linear regime ($0.1 \lesssim k \lesssim 0.5\,h\,\mathrm{Mpc}^{-1}$). We demonstrated that the reconstructed linear component reduces the skewness by two orders of magnitude with respect to the density directly derived from smoothing the galaxy field on the same scale.

We have furthermore demonstrated how to compute the Zel'dovich density field from the lognormal reconstructed density fields based on the tidal field tensor in a parameter free way. The recovered filamentary network remarkably connects the discrete distribution of galaxies. The real space density fields obtained in this work could be used to recover the initial conditions with techniques which rely on knowing the dark matter field at the final stage (see e.g. Wang et al., 2014).

Summarizing, we found that Bayesian reconstruction methods are particularly suited to test bias and structure formation models as the numerous reconstructions from one MCMC run are giving a reliable estimate of the mean reconstruction and also a precise estimate of the variance within one run chain. In this way, we tested the outcome of a data model over a large range of iteration steps.

Future work based this thesis

My thesis work can be beneficial in two ways, firstly numerous projects may benefit from the outcome of the reconstructions, and secondly promising extensions to ARGO can be implemented to broaden the range of sampled quantities.

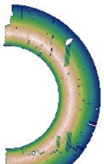
The reconstructed density and peculiar velocity fields obtained here can be used for a number of studies. Among those, the cosmological effect of CMB photons scattered by electrons in the gas of galaxy clustered (named after Sunyaev-Zel'dovich (SZ)) is promising to study with reconstructions. It is an excellent tool for measuring the peculiar motions of clusters of galaxies and hence studying the evolution of structure in the Universe. Commonly, the SZ is divided into the the *thermal effect*, arising due to the electron temperature in the inter-cluster gas, and the *kinetic effect*, where electrons scatter CMB photons due to their high bulk motion. A precise reconstruction of the large-scale velocity fields therefore, are crucial to estimate the contribution of the energy of the CMB photons due to the primodial last scattering surface or scattering with electrons at later times. A related analysis is the integrated Sachs-Wolfe (ISW) effect, which is caused by the change of the temperature due to the gravitational potential along its path to the observer. Therefore, cross correlating the potential of the matter perturbations in the large-scale structure projected in line of sight, and the anisotropies of the CMB is a promising project, for which our reconstructions are highly beneficial.

Baryon acoustic oscillations (BAO) measurements benefit from our reconstructions too, As we are able to statistically infer the real-space positions of the analyzed galaxies. BAO scale being a *standard ruler* is a key observation to constrain the Hubble parameter $H(z)$, as we measure the angular diameter distance $d_{\mathrm{A}}(z)$ of the BAO without anisoptropies. This would be complementary to the standard BAO analyses, such as done in Alam et al. (2016), in which the line-of-side and transverse angular diameter distance is measured separately.

In addition, our presented method may be used for improved environmental studies, as web classification methods can be directly applied on the dark matter density field.

Also, we seek to extend the usage of our reconstructions to study the influence of particle interactions and properties to the formation of large-scale structures. A promising analysis with upcoming denser and deeper surveys therefore is the probe of neutrino masses and their hierarchies (Lesgourgues et al., 2004) or decaying dark matter sensitivity (Poulin et al., 2016). If these models are included to the ΛCDM theory, we expect free streaming path lengths of particles to change the amplitude of matter perturbations at small scales, due to interactions. These scales need to be precisely understood and thus, the biasing relation we introduced to stochasticity is a vital ingredient for these scales.

The improvements I have shown for ARGO and the successful application to a large redshift survey are vital in order to be prepared to apply ARGO to upcoming denser and deeper surveys. This will allow us to probe ΛCDM cosmology on a broader range of scales with higher statistics, stressing the importance of an improved biasing relation.



A Additional

A.1 Inverse tests

Inverse tests are tools to analyze the outcome of our reconstructions if the target distribution is known. To validate the various refinements within ARGO, we ran several tests and mock scenarios, which I will list here.

- **General window handling:**
  We use the snapshot BMD mock at $z = 0.57$ and dismiss half of the volume at the z-coordinate. We account for the resulting empty volume by also weightingthe window function with $w = 0$. Thus, we create an artificial mock catalogue however, with a perfectly matched window. This analysis is ought to compare the prior lognormal field, augmented by ARGO in the empty area, and the N-body density field constrained by the data, in order to test the influence of a renormalized mean field $\boldsymbol{\mu}$ and the godness of the prior assumtion.

- **Light-cone without cosmic growth:**
  This test was performed in order to quantify the implementation of the growth factor in ARGO. We used the snapshot mock catalogue and mimic the geometry of the BOSS survey with the corresponding window function. The significant difference now is however, that all structures are at the same redshift and thus evolved with the same cosmic growth. Comparing this reconstruction with a light-cone sample from real data gives information about the correct growth model we applied to ARGO.

- **Lognormal-NB with light-cone geometry:**
  To check whether a single mean field $\boldsymbol{\mu}$ is sufficient to describe the data and the prior dominated regions, respectively, we create a lognormal random field of which we sample a random galaxy mock catalogue with a NB distribution. We use the same geometry as in our light-cone galaxy sample. In this test, we know that our lognormal-NB reconstruction model should be able to perfectly reconstruct the initial power spectrum of the random field, as the mean field of the random lognormal field is known by construction.

Figure A.1 shows the procedure to create a mock galaxy catalogue from a lognormal-NB field after the window function has been applied. Firstly, the window is taken from the BOSS light-cone sample. Then, we create a lognormal random field. Finally we apply the window to the lognormal random field and sampling galaxies with a NB probability according to the window. In this way we know the exact mean field $\boldsymbol{\mu}$ by construction. We also use the BMD snapshot mock and mimic a light-cone mock, however, without cosmic growth. The results can be seen in Figure A.2. We ran the reconstruction with one $\boldsymbol{\mu}$ for the whole box and the same NB likelihood and bias parameters. In red color the reconstruction of the lognormal-NB mock is shown, yielding a smooth power spectrum very close to

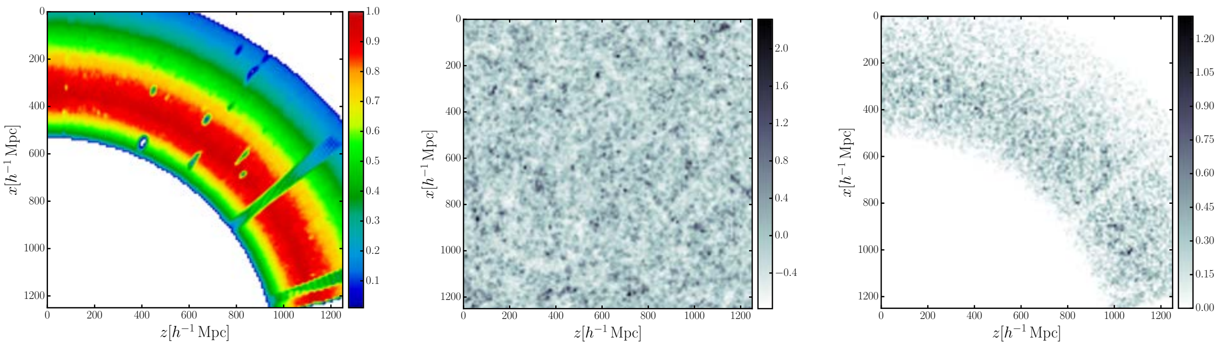


Figure A.1: Slices of the inverse tests. Left, the window function, central the lognormal random field, on the right mock galaxies sampled with a NB probability from the lognormal field according to the window function.

the theoretical predictions of CAMB. The blue curve is the reconstruction of the BMD mock, which shows a clear mismatch at large scales. This test proves that $\boldsymbol{\mu}$ must be sampled according the the completeness, as it will differ for the data dominated area and the predictions from lognormal random field.

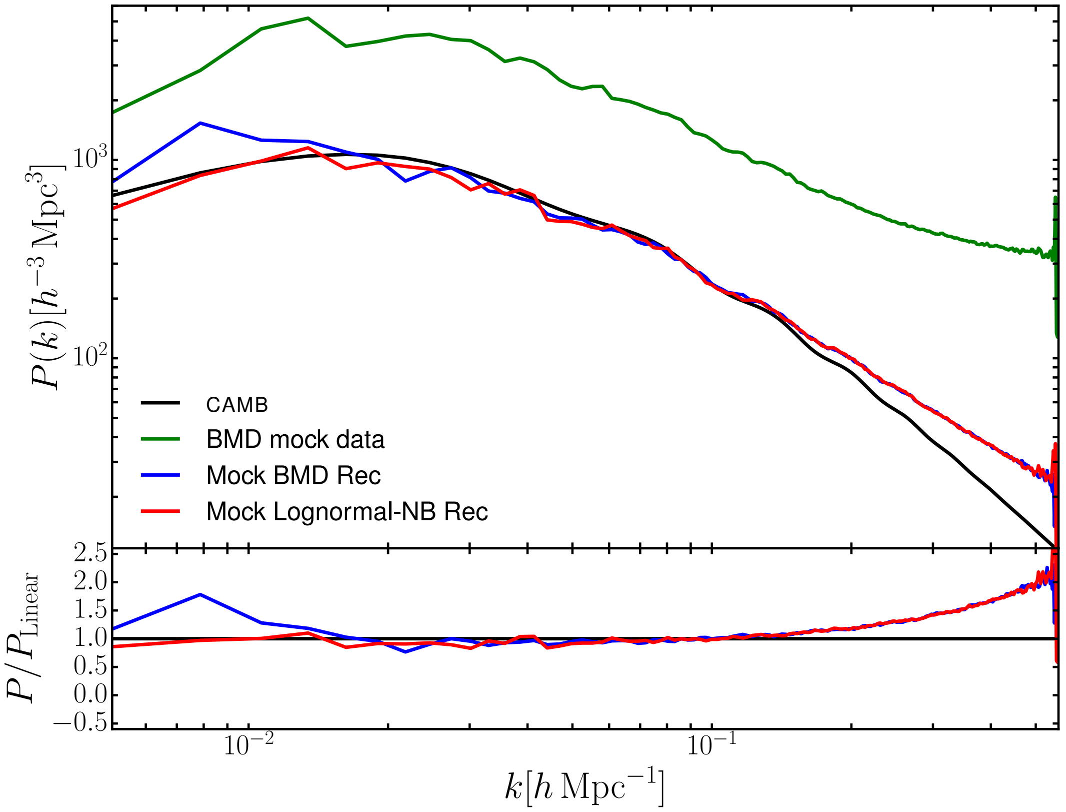


Figure A.2: Power spectra of the different reconstruction tests after the artificial window function was applied, showing the input power spectrum in green of the BMD mock galaxies of the full snapshot, in blue the reconstruction of the BMD mock with window correction. In red color the reconstruction of the lognormal-NB mock is shown, also with the artificial window function that was applied.

A.2 Potential Scale Reduction factor

We assume N_{chains} chains of length N_{length}. The output of the chain is denoted as $x_{c,s}$, with $c \in \{1, 2, ..., N_{\text{chains}}\}$ and $s \in \{1, 2, ..., N_{\text{length}}\}$. In our case x is the multidimensional ensemble $\{\delta_i\}$ of the overdensity of cell i in our reconstructed volume. For simplicity we show the calculations for a one dimensional observable x. Starting from an identical proposal distribution we calculate as follows

1. Calculate each chain's mean value
$$\bar{x}_c = \frac{1}{N_{\text{length}}} \sum_s x_{c,s} \,.$$

2. Calculate each chain's variance
$$\sigma_c^2 = \frac{1}{N_{\text{chains}} - 1} \sum_s (x_{c,s} - \bar{x}_c)^2 \,.$$

3. Calculate all chains' mean
$$\bar{x} = \frac{1}{N_{\text{chains}}} \sum_c \frac{1}{N_{\text{length}}} \sum_s x_{c,s} = \frac{1}{N_{\text{chains}}} \sum_c \bar{x}_c \,.$$

4. Calculate the weighted mean of each chain's variance
$$B = \frac{N_{\text{length}}}{N_{\text{chains}} - 1} \sum_c (\bar{x}_c - \bar{x})^2 \,.$$

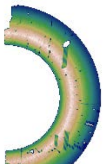
5. Calculate the average variance within one chain
$$W = \frac{1}{N_{\text{chains}}} \sum_c \sigma_c^2 \,.$$

6. The potential scale reduction factor (R) then is defined as
$$R = \sqrt{\frac{N_{\text{length}} - 1}{N_{\text{length}}} + \frac{N_{\text{chains}} + 1}{N_{\text{length}} N_{\text{chains}}} \frac{B}{W}} \,.$$

If all chains converge to the same target distribution, we expect the variance within one chain to be close to the variance between the N_{chains} chains, so that the R is close to one.

A.3 Single reconstructions sample of the Bolshoi simulation

In Section 5.2.4 we showed the averaged slice plots for the ensemble mean only. In Figure A.3 two different slices of a NB reconstruction of the Bolshoi simulation is shown. The structures look less smooth as expected and are also more noisy compared to the average plots.



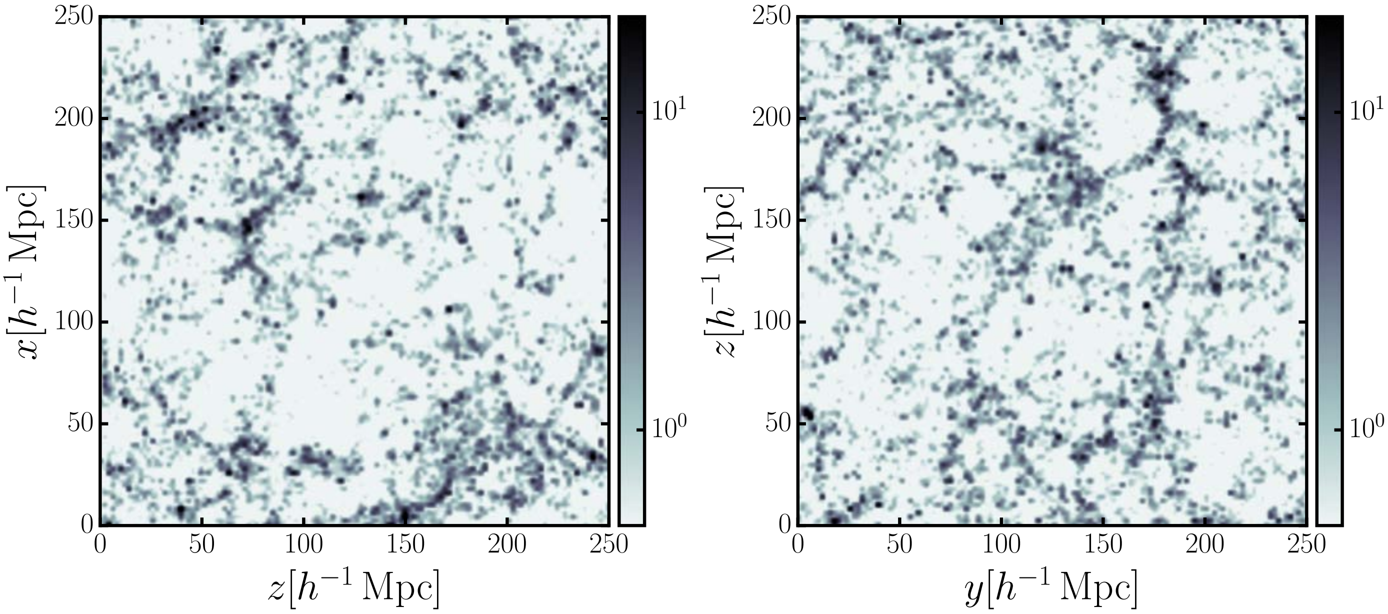


Figure A.3: Slices of a single reconstruction sample with NB likelihood and a thickness of $10\,h^{-1}$ Mpc.

A.4 Importance of weights for BOSS correlation function

In Figure 6.9 on the bottom left panel the monopole correlation function for the reconstructed real-space positions of the BOSS DR12 galaxies were shown. As we mentioned, the data needs to be reweighted to coup with various systematic effects. Here we show the huge impact if the weights are not applied. This can be seen in Figure A.4. The correlation function at distanced of $r \geq 140\ h^{-1}$ Mpc is rising, which is contradicting the theoretical expectations. This is not the case if the weights are properly applied. The weights also improve the scatter behaviour of the correlation function.

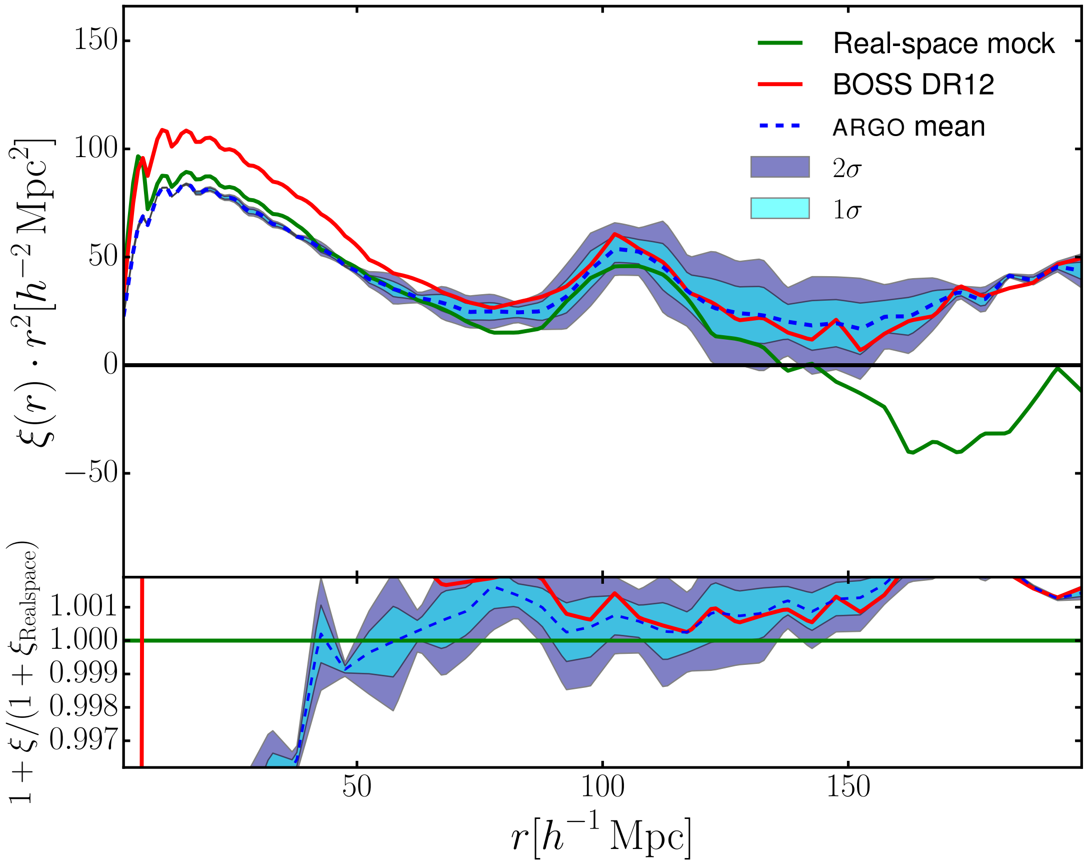


Figure A.4: Monopole correlation function for BOSS galaxies without applying weights.

Bibliography

Alam S., Albareti F. D., Allende Prieto C., Anders F., Anderson S. F., Anderton T., Andrews B. H., Armengaud E., Aubourg É., Bailey S., et al. 2015, *The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III*, [`arXiv:1501.00963`], ApJS, 219, 12

Alam S., Ata M., Bailey S., Beutler F., Bizyaev D., Blazek J. A., Bolton A. S., Brownstein J. R., Burden A., Chuang C.-H., Comparat J., Cuesta A. J., Dawson K. S., Eisenstein D. J., Escoffier S., Gil-Marín H., Grieb J. N., Hand N., Ho S., Kinemuchi K., Kirkby D., Kitaura F., Malanushenko E., Malanushenko V., Maraston C., McBride C. K., Nichol R. C., Olmstead M. D., Oravetz D., Padmanabhan N., Palanque-Delabrouille N., Pan K., Pellejero-Ibanez M., et al., 2016, *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample*, preprint, [`arXiv:1607.03155`]

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

Alsing J., Heavens A., Jaffe A. H., Kiessling A., Wandelt B., Hoffmann T., 2016, *Hierarchical cosmic shear power spectrum inference*, [`arXiv:1505.07840`], MNRAS, 455, 4452

Anderson L., Aubourg É., Bailey S., Beutler F., Bhardwaj V., Blanton M., Bolton A. S., Brinkmann J., Brownstein J. R., Burden A., Chuang C.-H., Cuesta A. J., Dawson K. S., Eisenstein D. J., Escoffier S., Gunn J. E., Guo H., Ho S., Honscheid K., Howlett C., Kirkby D., Lupton R. H., Manera M., Maraston C., McBride C. K., Mena O., Montesano F., Nichol R. C., Nuza S. E., Olmstead M. D., Padmanabhan N., Palanque-Delabrouille N., Parejko J., Percival W. J., Petitjean P., Prada F., Price-Whelan A. M., Reid B., Roe N. A., Ross A. J., Ross N. P., Sabiu C. G., Saito S., Samushia L., Sánchez A. G., Schlegel D. J., Schneider D. P., Scoccola C. G., Seo H.-J., Skibba R. A., Strauss M. A., Swanson M. E. C., Thomas D., Tinker J. L., Tojeiro R., Magaña M. V., Verde L., Wake D. A., Weaver B. A., Weinberg D. H., White M., Xu X., Yèche C., Zehavi I., Zhao G.-B., 2014, *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples*, [`arXiv:1312.4877`], MNRAS, 441, 24

Appel A. W., 1985, *An Efficient Program for Many-Body Simulation*, SIAM Journal on Scientific and Statistical Computing, vol. 6, no. 1, January 1985, p. 85-103., 6, 85

Aragon-Calvo M. A., 2012, *The MIP Ensemble Simulation: Local Ensemble Statistics in the Cosmic Web*, preprint, [`arXiv:1210.7871`]

Ata M., Kitaura F.-S., Müller V., 2014, *Non-Gaussian inference from non-linear and non-Poisson biased distributed data*, in Heavens A., Starck J.-L., Krone-Martins A., eds, IAU Symposium Vol. 306, Statistical Challenges in 21st Century Cosmology. pp 258–261, [`arXiv:1406.7796`], doi:10.1017/S1743921314010904

Ata M., Kitaura F.-S., Müller V., 2015, *Bayesian inference of cosmic density fields from non-linear, scale-dependent, and stochastic biased tracers*, [`arXiv:1408.2566`], MNRAS, 446, 4250

Ata M., Kitaura F.-S., Chuang C.-H., Rodríguez-Torres S., Angulo R. E., Ferraro S., Gil-Marín H., McDonald P., Hernández Monteagudo C., Müller V., Yepes G., Autefage M., Baumgarten F., Beutler F., Brownstein J. R., Burden A., Eisenstein D. J., Guo H., Ho S., McBride C., Neyrinck M., Olmstead M. D., Padmanabhan N., Percival W. J., Prada F., Rossi G., Sánchez A. G., Schlegel D., Schneider D. P., Seo H.-J., Streblyanska A., Tinker J., Tojeiro R., Vargas-Magana M., 2017, *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmic flows and cosmic web from luminous red galaxies*, [`arXiv:1605.09745`], MNRAS, 467, 3993

Athanassoula E., 1993, *N-body Techniques and their Impact on Galactic Stellar Dynamics*, in Combes F., Athanassoula E., eds, N-body Problems and Gravitational Dynamics. pp 116–130

Baldauf T., Seljak U., Desjacques V., McDonald P., 2012, *Evidence for quadratic tidal tensor bias from the halo bispectrum*, [`arXiv:1201.4827`], Phys. Rev. D, 86, 083540

Baldauf T., Seljak U., Smith R. E., Hamaus N., Desjacques V., 2013, *Halo stochasticity from exclusion and nonlinear clustering*, [`arXiv:1305.2917`], Phys. Rev. D, 88, 083507

Ballinger W. E., Peacock J. A., Heavens A. F., 1996, *Measuring the cosmological constant with redshift surveys*, [`arXiv:astro-ph/9605017`], MNRAS, 282, 877

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, *The statistics of peaks of Gaussian random fields*, ApJ, 304, 15

Barnes J., Hut P., 1986, *A hierarchical O(N log N) force-calculation algorithm*, Nature, 324, 446

Behroozi P. S., Conroy C., Wechsler R. H., 2010, *A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for* $0 < z < 4$, [`arXiv:1001.0015`], ApJ, 717, 379

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

Berlind A. A., Narayanan V. K., Weinberg D. H., 2001, *Biased Galaxy Formation and Measurements of* β, [`arXiv:astro-ph/0008305`], ApJ, 549, 688

Bernardeau F., 1992, *The quasi-Gaussian density-velocity relationship*, ApJ, 390, L61

Bernardeau F., Schaeffer R., 1992, *Galaxy correlations, matter correlations and biasing*, A&A, 255, 1

Bernardeau F., Schaeffer R., 1999, *Halo correlations in nonlinear cosmic density fields*, [`arXiv:astro-ph/9903387`], A&A, 349, 697

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

Beutler F., Saito S., Seo H.-J., Brinkmann J., Dawson K. S., Eisenstein D. J., Font-Ribera A., Ho S., McBride C. K., Montesano F., Percival W. J., Ross A. J., Ross N. P., Samushia L., Schlegel D. J., Sánchez A. G., Tinker J. L., Weaver B. A., 2014, *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: testing gravity with redshift space distortions using the power spectrum multipoles*, [`arXiv:1312.4611`], MNRAS, 443, 1065

Blake C., Baldry I. K., Bland-Hawthorn J., Christodoulou L., Colless M., Conselice C., Driver S. P., Hopkins A. M., Liske J., Loveday J., Norberg P., Peacock J. A., Poole G. B., Robotham A. S. G., 2013, *Galaxy And Mass Assembly (GAMA): improved cosmic growth measurements using multiple tracers of large-scale structure*, [`arXiv:1309.5556`], MNRAS, 436, 3089

Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, *Formation of galaxies and large-scale structure with cold dark matter*, Nature, 311, 517

Bolton A. S., Schlegel D. J., Aubourg É., Bailey S., Bhardwaj V., Brownstein J. R., Burles S., Chen Y.-M., Dawson K., Eisenstein D. J., Gunn J. E., Knapp G. R., Loomis C. P., Lupton R. H., Maraston C., Muna D., Myers A. D., Olmstead M. D., Padmanabhan N., Pâris I., Percival W. J., Petitjean P., Rockosi C. M., Ross N. P., Schneider D. P., Shu Y., Strauss M. A., Thomas D., Tremonti C. A., Wake D. A., Weaver B. A., Wood-Vasey W. M., 2012, *Spectral Classification and Redshift Measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey*, [`arXiv:1207.7326`], AJ, 144, 144

Bond J. R., Kofman L., Pogosyan D., 1996, *How filaments of galaxies are woven into the cosmic web*, [`arXiv:astro-ph/9512141`], Nature, 380, 603

Branchini E., Eldar A., Nusser A., 2002, *Peculiar velocity reconstruction with the fast action method: tests on mock redshift surveys*, [`arXiv:astro-ph/0110618`], MNRAS, 335, 53

Branchini E., Davis M., Nusser A., 2012, *The linear velocity field of 2MASS Redshift Survey, K_s = 11.75 galaxies: constraints on β and bulk flow from the luminosity function*, [`arXiv:1202.5206`], MNRAS, 424, 472

Bull P., 2016, *A Galaxy-Halo Model for Multiple Cosmological Tracers*, preprint, [`arXiv:1610.08948`]

Bull P., Akrami Y., Adamek J., Baker T., Bellini E., Beltrán Jiménez J., Bentivegna E., Camera S., Clesse S., Davis J. H., Di Dio E., Enander J., Heavens A., Heisenberg L., Hu B., Llinares C., Maartens R., Mörtsell E., Nadathur S., Noller J., Pasechnik R., Pawlowski M. S., Pereira T. S., Quartin M., Ricciardone A., Riemer-Sørensen S., Rinaldi M., Sakstein J., Saltas I. D., Salzano V., Sawicki I., Solomon A. R., Spolyar D., Starkman G. D., Steer D., Tereno I., Verde L., Villaescusa-Navarro F., von Strauss M., Winther H. A., 2016, *Beyond Λ CDM: Problems, solutions, and the road ahead*, [`arXiv:1512.05356`], Physics of the Dark Universe, 12, 56

Bunn E. F., Hoffman Y., Silk J., 1996, *The Wiener-filtered COBE DMR Data and Predictions for the Tenerife Experiment*, [`arXiv:astro-ph/9509045`], ApJ, 464, 1

Carroll S. M., 2004, *Spacetime and geometry. An introduction to general relativity*. Addison Wesley

Carron J., Szapudi I., 2014a, *Optimal observables in galaxy surveys*, in Heavens A., Starck J.-L., Krone-Martins A., eds, IAU Symposium Vol. 306, Statistical Challenges in 21st Century Cosmology. pp 235–238, [`arXiv:1407.0245`], doi:10.1017/S1743921314010783

Carron J., Szapudi I., 2014b, *Sufficient observables for large-scale structure in galaxy surveys*, [`arXiv:1310.6038`], MNRAS, 439, L11

Cen R., Ostriker J. P., 1992, *Galaxy formation and physical bias*, ApJ, 399, L113

Cen R., Ostriker J. P., 1993, *Cold Dark Matter Cosmogony with Hydrodynamics and Galaxy Formation: Galaxy Properties at Redshift Zero*, ApJ, 417, 415

Chodorowski M. J., Lokas E. L., Pollo A., Nusser A., 1998, *Recovery of the cosmological peculiar velocity from the density field in the weakly non-linear regime*, [`arXiv:astro-ph/9802050`], MNRAS, 300, 1027

Chuang C.-H., Wang Y., 2013, *Modelling the anisotropic two-point galaxy correlation function on small scales and single-probe measurements of H(z), $D_A(z)$ and f(z)σ_8(z) from the Sloan Digital Sky Survey DR7 luminous red galaxies*, [`arXiv:1209.0210`], MNRAS, 435, 255

Chuang C.-H., Zhao C., Prada F., Munari E., Avila S., Izard A., Kitaura F.-S., Manera M., et al. 2015, *nIFTy cosmology: Galaxy/halo mock catalogue comparison project on clustering statistics*, [`arXiv:1412.7729`], MNRAS, 452, 686

Clowe D., Randall S. W., Markevitch M., 2007, *Catching a bullet: direct evidence for the existence of dark matter*, [`arXiv:astro-ph/0611496`], Nuclear Physics B Proceedings Supplements, 173, 28

Cole S., Kaiser N., 1989, *Biased clustering in the cold dark matter cosmogony*, MNRAS, 237, 1127

Cole S., Fisher K. B., Weinberg D. H., 1995, *Constraints on Omega from the IRAS redshift surveys*, [`arXiv:astro-ph/9412062`], MNRAS, 275, 515

Cole S., Percival W. J., Peacock J. A., Norberg P., Baugh C. M., Frenk C. S., Baldry I., Bland-Hawthorn J., Bridges T., Cannon R., Colless M., Collins C., Couch W., Cross N. J. G., Dalton G., Eke V. R., De Propris R., Driver S. P., Efstathiou G., Ellis R. S., Glazebrook K., Jackson C., Jenkins A., Lahav O., Lewis I., Lumsden S., Maddox S., Madgwick D., Peterson B. A., Sutherland W., Taylor K., 2005, *The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications*, [`arXiv:astro-ph/0501174`], MNRAS, 362, 505

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

Coles P., Lucchin F., 2002, *Cosmology: The Origin and Evolution of Cosmic Structure*. Wiley

Conroy C., Wechsler R. H., Kravtsov A. V., 2006, *Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time*, [`arXiv:astro-ph/0512234`], ApJ, 647, 201

Courtois H. M., Hoffman Y., Tully R. B., Gottlöber S., 2012, *Three-dimensional Velocity and Density Reconstructions of the Local Universe with Cosmicflows-1*, [`arXiv:1109.3856`], ApJ, 744, 43

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

Davis M., Huchra J., 1982, *A survey of galaxy redshifts. III - The density field and the induced gravity field*, ApJ, 254, 437

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

Dawson K. S., Schlegel D. J., Ahn C. P., Anderson S. F., Aubourg É., Bailey S., Barkhouser R. H., Bautista J. E., Beifiori A., Berlind A. A., Bhardwaj V., Bizyaev D., Blake C. H., Blanton M. R., et al. 2013, *The Baryon Oscillation Spectroscopic Survey of SDSS-III*, [`arXiv:1208.0022`], AJ, 145, 10

Dekel A., Lahav O., 1999, *Stochastic Nonlinear Galaxy Biasing*, [`arXiv:astro-ph/9806193`], ApJ, 520, 24

Dodelson S., 2003, *Modern cosmology*. Academic Press

Doré O., Teyssier R., Bouchet F. R., Vibert D., Prunet S., 2001, *MAPCUMBA: A fast iterative multi-grid map-making algorithm for CMB experiments*, [`arXiv:astro-ph/0101112`], A&A, 374, 358

Duane S., Kennedy A., Pendleton B. J., Roweth D., 1987, *Hybrid monte carlo*, Physics Letters B, 195, 216

Eisenstein D. J., Hu W., 1998, *Baryonic Features in the Matter Transfer Function*, [arXiv:astro-ph/9709112], ApJ, 496, 605

Eisenstein D. J., Zehavi I., Hogg D. W., Scoccimarro R., Blanton M. R., Nichol R. C., Scranton R., Seo H.-J., Tegmark M., Zheng Z., Anderson S. F., Annis J., Bahcall N., Brinkmann J., Burles S., et al., 2005, *Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies*, [arXiv:astro-ph/0501171], ApJ, 633, 560

Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, *Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak*, [arXiv:astro-ph/0604362], ApJ, 664, 675

Eisenstein D. J., Weinberg D. H., Agol E., Aihara H., Allende Prieto C., Anderson S. F., Arns J. A., Aubourg É., Bailey S., Balbinot E., et al. 2011, *SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems*, [arXiv:1101.1529], AJ, 142, 72

Erdoğdu P., Lahav O., Zaroubi S., Efstathiou G., Moody S., Peacock J. A., Colless M., Baldry I. K., et al. 2004, *The 2dF Galaxy Redshift Survey: Wiener reconstruction of the cosmic web*, [arXiv:astro-ph/0312546], MNRAS, 352, 939

Falck B. L., Neyrinck M. C., Aragon-Calvo M. A., Lavaux G., Szalay A. S., 2012, *Straightening the Density-Displacement Relation with a Logarithmic Transform*, [arXiv:1111.4466], ApJ, 745, 17

Falck B., McCullagh N., Neyrinck M. C., Wang J., Szalay A. S., 2016, *The Effect of Corner Modes in the Initial Conditions of Cosmological Simulations*, preprint, [arXiv:1610.04862]

Feldman H. A., Kaiser N., Peacock J. A., 1994, *Power-spectrum analysis of three-dimensional redshift surveys*, [arXiv:astro-ph/9304022], ApJ, 426, 23

Fisher K. B., Lahav O., Hoffman Y., Lynden-Bell D., Zaroubi S., 1995, *Wiener reconstruction of density, velocity and potential fields from all-sky galaxy redshift su rveys*, [arXiv:astro-ph/9406009], MNRAS, 272, 885

Forero-Romero J. E., Hoffman Y., Gottlöber S., Klypin A., Yepes G., 2009, *A dynamical classification of the cosmic web*, [arXiv:0809.4135], MNRAS, 396, 1815

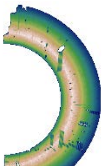
Frusciante N., Sheth R. K., 2012, *Lagrangian bias in the local bias model*, [arXiv:1208.0229], J. Cosmology Astropart. Phys., 11, 016

Fry J. N., 1984, *The Galaxy correlation hierarchy in perturbation theory*, ApJ, 279, 499

Fry J. N., Gaztanaga E., 1993, *Biasing and hierarchical statistics in large-scale structure*, [arXiv:astro-ph/9302009], ApJ, 413, 447

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, *The Sloan Digital Sky Survey Photometric System*, AJ, 111, 1748

Gamow G., 1946, *Expanding universe and the origin of elements*, Phys. Rev., 70, 572



Gelfand A. E., Smith A. F. M., 1990, *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association, 85, 398

Geller M. J., Huchra J. P., 1989, *Mapping the universe*, Science, 246, 897

Gelman A., 2010, *Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics*, preprint, [`arXiv:1001.2968`]

Gelman A., Rubin D., 1992, *Inference from iterative simulation using multiple sequences*, Statistical Science, 7, 457

Gelman A., Carlin J., Stern H., Dunson D., Vehtari A., Rubin D., 2013, *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC

Geman S., Geman D., 1984, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-6, 721

Génova-Santos R., Atrio-Barandela F., Kitaura F.-S., Mücket J. P., 2015, *Constraining the Baryon Fraction in the Warm Hot Intergalactic Medium at Low Redshifts with Planck Data*, [`arXiv:1501.01445`], ApJ, 806, 113

Gilks W., Richardson S., Spiegelhalter D., 1995, *Markov Chain Monte Carlo in Practice* . Chapman and Hall/CRC

Goldstein H., Poole Jr. C. P., Safko J. L., 2001, *Classical Mechanics (3rd Edition)*. Pearson

Goroff M. H., Grinstein B., Rey S.-J., Wise M. B., 1986, *Coupling of modes of cosmological mass density fluctuations*, ApJ, 311, 6

Gottloeber S., Klypin A., 2008, *The ART of Cosmological Simulations*, preprint, [`arXiv:0803.4343`]

Gramann M., 1993, *Second-order solution for determining density and velocity fields of galaxies*, ApJ, 405, L47

Granett B. R., Neyrinck M. C., Szapudi I., 2008, *An Imprint of Superstructures on the Microwave Background due to the Integrated Sachs-Wolfe Effect*, [`arXiv:0805.3695`], ApJ, 683, L99

Granett B. R., Branchini E., Guzzo L., Abbas U., Adami C., Arnouts S., Bel J., Bolzonella M., Bottini D., Cappi A., Coupon J., Cucciati O., Davidzon I., De Lucia G., de la Torre S., Fritz A., Franzetti P., Fumana M., Garilli B., Ilbert O., Iovino A., Krywult J., Le Brun V., Le Fèvre O., Maccagni D., Małek K., Marulli F., McCracken H. J., Polletta M., Pollo A., Scodeggio M., Tasca L. A. M., Tojeiro R., Vergani D., Zanichelli A., Burden A., Di Porto C., Marchetti A., Marinoni C., Mellier Y., Moutard T., Moscardini L., Nichol R. C., Peacock J. A., Percival W. J., Zamorani G., 2015, *The VIMOS Public Extragalactic Redshift Survey. Reconstruction of the redshift-space galaxy density field*, [`arXiv:1505.06337`], A&A, 583, A61

Grøn Ø., Hervik S., 2007, *Einstein's General Theory of Relativity with Modern Applications in Cosmology*. Springer Science & Business Media

Gunn J. E., Carr M., Rockosi C., Sekiguchi M., Berry K., Elms B., de Haas E., Ivezić Ž., Knapp G., Lupton R., Pauls G., Simcoe R., Hirsch R., Sanford D., Wang S., York D., Harris F., Annis J., Bartozek L., Boroski W., Bakken J., Haldeman M., Kent S., Holm S., Holmgren D., Petravick D., Prosapio A., Rechenmacher R., Doi M., Fukugita M., Shimasaku K., Okada N., Hull C., Siegmund

W., Mannery E., Blouke M., Heidtman D., Schneider D., Lucinio R., Brinkman J., 1998, *The Sloan Digital Sky Survey Photometric Camera*, [`arXiv:astro-ph/9809085`], AJ, 116, 3040

Guo Q., White S., Li C., Boylan-Kolchin M., 2010, *How do galaxies populate dark matter haloes?*, [`arXiv:0909.4305`], MNRAS, 404, 1111

Guth A. H., 1984, *The new inflationary universe*, Annals of the New York Academy of Sciences, 422, 1

Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, *Properties of dark matter haloes in clusters, filaments, sheets and voids*, [`arXiv:astro-ph/0610280`], MNRAS, 375, 489

Hahn O., Angulo R. E., Abel T., 2015, *The properties of cosmic velocity fields*, [`arXiv:1404.2280`], MNRAS, 454, 3920

Hamilton A. J. S., 1998, *Linear Redshift Distortions: a Review*, in Hamilton D., ed., Astrophysics and Space Science Library Vol. 231, The Evolving Universe. p. 185, [`arXiv:astro-ph/9708102`], doi:10.1007/978-94-011-4960-0˙17

Hamilton A. J. S., Tegmark M., 2004, *A scheme to deal accurately and efficiently with complex angular masks in galaxy surveys*, [`arXiv:astro-ph/0306324`], MNRAS, 349, 115

Hand L. N., Finch J. D., 1998, *Analytical Mechanics*. Cambridge University Press, doi:10.1119/1.19451

Harrison E. R., 1970, *Fluctuations at the Threshold of Classical Cosmology*, Phys. Rev. D, 1, 2726

Hastings W. K., 1970, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57, 97

Heavens A., 2009, *Statistical techniques in cosmology*, preprint, [`arXiv:0906.0664`]

Heavens A., Alsing J., Jaffe A., Hoffmann T., Kiessling A., Wandelt B., 2016, *Bayesian hierarchical modelling of weak lensing - the golden goal*, preprint, [`arXiv:1602.05345`]

Heitmann K., White M., Wagner C., Habib S., Higdon D., 2010, *The Coyote Universe. I. Precision Determination of the Nonlinear Matter Power Spectrum*, [`arXiv:0812.1052`], ApJ, 715, 104

Hernández-Monteagudo C., Ma Y.-Z., Kitaura F. S., Wang W., Génova-Santos R., Macías-Pérez J., Herranz D., 2015, *Evidence of the Missing Baryons from the Kinematic Sunyaev-Zeldovich Effect in Planck Data*, [`arXiv:1504.04011`], Physical Review Letters, 115, 191301

Heß S., Kitaura F.-S., 2016, *Cosmic flows and the expansion of the local Universe from non-linear phase-space reconstructions*, [`arXiv:1412.7310`], MNRAS, 456, 4247

Heß S., Kitaura F.-S., Gottlöber S., 2013, *Simulating structure formation of the Local Universe*, [`arXiv:1304.6565`], MNRAS, 435, 2065

Howlett C., Lewis A., Hall A., Challinor A., 2012, *CMB power spectrum parameter degeneracies in the era of precision cosmology*, [`arXiv:1201.3654`], JCAP, 1204, 027

Hubble E., 1929, *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*, Proceedings of the National Academy of Science, 15, 168

Hubble E., 1934, *The Distribution of Extra-Galactic Nebulae*, ApJ, 79, 8

Irvine W. M., 1961, PhD thesis, Harvard University

Jackson J. C., 1972, *A critique of Rees's theory of primordial gravitational radiation*, [arXiv:0810.3908], MNRAS, 156, 1P

Jain B., Bertschinger E., 1994, *Second-order power spectrum and nonlinear evolution at high redshift*, [arXiv:astro-ph/9311070], ApJ, 431, 495

Jasche J., Kitaura F.-S., 2010, *Fast Hamiltonian sampling for large-scale structure inference*, [arXiv:0911.2496], MNRAS, 407, 29

Jasche J., Wandelt B. D., 2013, *Bayesian physical reconstruction of initial conditions from large-scale structure surveys*, [arXiv:1203.3639], MNRAS, 432, 894

Jasche J., Kitaura F.-S., Li C., Enßlin T. A., 2010, *Bayesian non-linear large-scale structure inference of the Sloan Digital Sky Survey Data Release 7*, [arXiv:0911.2498], MNRAS, pp 1638–+

Jasche J., Leclercq F., Wandelt B. D., 2015, *Past and present cosmic structure in the SDSS DR7 main sample*, [arXiv:1409.6308], J. Cosmology Astropart. Phys., 1, 036

Jennings E., 2012, *An improved model for the non-linear velocity power spectrum*, [arXiv:1207.1439], MNRAS, 427, L25

Jennings E., Jennings D., 2015, *Non-linear stochastic growth rates and redshift space distortions*, [arXiv:1502.02052], MNRAS, 449, 3407

Jennings E., Baugh C. M., Pascoli S., 2011, *Modelling redshift space distortions in hierarchical cosmologies*, [arXiv:1003.4282], MNRAS, 410, 2081

Jewell J., Levin S., Anderson C. H., 2004, *Application of Monte Carlo Algorithms to the Bayesian Analysis of the Cosmic Microwave Background*, [arXiv:astro-ph/0209560], ApJ, 609, 1

Jing Y. P., 2005, *Correcting for the Alias Effect When Measuring the Power Spectrum Using a Fast Fourier Transform*, [arXiv:astro-ph/0409240], ApJ, 620, 559

Joachimi B., Taylor A. N., Kiessling A., 2011, *Cosmological information in Gaussianized weak lensing signals*, [arXiv:1104.1399], MNRAS, 418, 145

Kaiser N., 1984, *On the spatial correlations of Abell clusters*, ApJ, 284, L9

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

Kang H., Ryu D., Cen R., Song D., 2005, *Shock-heated Gas in the Large-Scale Structure of the Universe*, [arXiv:astro-ph/0410477], ApJ, 620, 21

Kauffmann G., White S. D. M., Guiderdoni B., 1993, *The Formation and Evolution of Galaxies Within Merging Dark Matter Haloes*, MNRAS, 264, 201

Kauffmann G., Heckman T. M., White S. D. M., Charlot S., et al., 2003, *Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey*, [arXiv:astro-ph/0204055], MNRAS, 341, 33

Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, *The environmental dependence of the relations between stellar mass, structure, star formation and nuclear activity in galaxies*, [arXiv:astro-ph/0402030], MNRAS, 353, 713

Kim J., Park C., Choi Y.-Y., 2008, *A Subhalo-Galaxy Correspondence Model of Galaxy Biasing*, [arXiv:0801.3169], ApJ, 683, 123

Kitaura F.-S., 2013, *The initial conditions of the Universe from constrained simulations*, [arXiv:1203.4184], MNRAS, 429, L84

Kitaura F.-S., Angulo R. E., 2012, *Linearization with cosmological perturbation theory*, [arXiv:1111.6617], MNRAS, 425, 2443

Kitaura F. S., Enßlin T. A., 2008, *Bayesian reconstruction of the cosmological large-scale structure: methodology, inverse algorithms and numerical optimization*, [arXiv:0705.0429], MNRAS, 389, 497

Kitaura F. S., Jasche J., Li C., Enßlin T. A., Metcalf R. B., Wandelt B. D., Lemson G., White S. D. M., 2009, *Cosmic cartography of the large-scale structure with Sloan Digital Sky Survey data release 6*, [arXiv:0906.3978], MNRAS, 400, 183

Kitaura F.-S., Jasche J., Metcalf R. B., 2010, *Recovering the non-linear density field from the galaxy distribution with a Poisson-lognormal filter*, [arXiv:0911.1407], MNRAS, 403, 589

Kitaura F.-S., Gallerani S., Ferrara A., 2012a, *Multiscale inference of matter fields and baryon acoustic oscillations from the Lyα forest*, [arXiv:1011.6233], MNRAS, 420, 61

Kitaura F.-S., Angulo R. E., Hoffman Y., Gottlöber S., 2012b, *Estimating cosmic velocity fields from density fields and tidal tensors*, [arXiv:1111.6629], MNRAS, 425, 2422

Kitaura F.-S., Erdoğdu P., Nuza S. E., Khalatyan A., Angulo R. E., Hoffman Y., Gottlöber S., 2012c, *Cosmic structure and dynamics of the local Universe*, [arXiv:1205.5560], MNRAS, 427, L35

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

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

Kitaura F.-S., Ata M., Angulo R. E., et al., 2016, *Bayesian redshift-space distortions correction from galaxy redshift surveys*, [arXiv:1507.08724], MNRAS, 457, L113

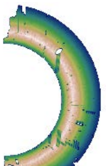
Kitching T. D., Heavens A. F., Das S., 2015, *3D weak gravitational lensing of the CMB and galaxies*, [arXiv:1408.7052], MNRAS, 449, 2205

Klypin A., Holtzman J., 1997, *Particle-Mesh code for cosmological simulations*, [arXiv:astro-ph/9712217], ArXiv Astrophysics e-prints

Klypin A. A., Trujillo-Gomez S., Primack J., 2011, *Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation*, [arXiv:1002.3660], ApJ, 740, 102

Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, *MultiDark simulations: the story of dark matter halo concentrations and density profiles*, [arXiv:1411.4001], MNRAS, 457, 4340

Komatsu E., Dunkley J., Nolta M. R., Bennett C. L., Gold B., Hinshaw G., Jarosik N., Larson D., Limon M., Page L., Spergel D. N., Halpern M., Hill R. S., Kogut A., Meyer S. S., Tucker G. S., Weiland J. L., Wollack E., Wright E. L., 2009, *Five-Year Wilkinson Microwave Anisotropy Probe Observations: Cosmological Interpretation*, [arXiv:0803.0547], ApJS, 180, 330



Kravtsov A. V., Klypin A. A., Khokhlov A. M., 1997, *Adaptive Refinement Tree: A New High-Resolution N-Body Code for Cosmological Simulations*, [`arXiv:astro-ph/9701195`], ApJS, 111, 73

Kwan J., Lewis G. F., Linder E. V., 2012, *Mapping Growth and Gravity with Robust Redshift Space Distortions*, [`arXiv:1105.1194`], ApJ, 748, 78

LSST Dark Energy Science Collaboration 2012, *Large Synoptic Survey Telescope: Dark Energy Science Collaboration*, preprint, [`arXiv:1211.0310`]

Lavaux G., Mohayaee R., Colombi S., Tully R. B., Bernardeau F., Silk J., 2008, *Observational biases in Lagrangian reconstructions of cosmic velocity fields*, [`arXiv:0707.3483`], MNRAS, 383, 1292

Lavaux G., Tully R. B., Mohayaee R., Colombi S., 2010, *Cosmic Flow From Two Micron All-Sky Redshift Survey: the Origin of Cosmic Microwave Background Dipole and Implications for* Λ*CDM Cosmology*, [`arXiv:0810.3658`], ApJ, 709, 483

Layzer D., 1963, *A Preface to Cosmogony. I. The Energy Equation and the Virial Theorem for Cosmic Distributions.*, ApJ, 138, 174

Leauthaud A., Tinker J., Behroozi P. S., Busha M. T., Wechsler R. H., 2011, *A Theoretical Framework for Combining Techniques that Probe the Link Between Galaxies and Dark Matter*, [`arXiv:1103.2077`], ApJ, 738, 45

Lee P. M., 2012, *Bayesian Statistics: An Introduction*. Wiley Press

Lesgourgues J., Pastor S., Perotto L., 2004, *Probing neutrino masses with future galaxy redshift surveys*, [`arXiv:hep-ph/0403296`], Phys. Rev. D, 70, 045016

Lewis A., Challinor A., Lasenby A., 2000, *Efficient computation of CMB anisotropies in closed FRW models*, [`arXiv:astro-ph/9911177`], Astrophys. J., 538, 473

Li X.-D., Park C., Sabiu C. G., Kim J., 2015, *Cosmological constraints from the redshift dependence of the Alcock-Paczynski test and volume effect: galaxy two-point correlation function*, [`arXiv:1504.00740`], MNRAS, 450, 807

Liddle A. R., Lyth D. H., 2000, *Cosmological Inflation and Large-Scale Structure*. Cambridge University Press

Loredo T. J., 2012, *Bayesian astrostatistics: a backward look to the future*, preprint, [`arXiv:1208.3036`]

Mann R. G., Peacock J. A., Heavens A. F., 1998, *Eulerian bias and the galaxy density field*, [`arXiv:astro-ph/9708031`], MNRAS, 293, 209

Martínez V. J., Saar E., Martínez-González E., Pons-Bordería M.-J., eds, 2009, Data Analysis in Cosmology Lecture Notes in Physics, Berlin Springer Verlag Vol. 665, doi:10.1007/978-3-540-44767-2.

Matarrese S., Verde L., Heavens A. F., 1997, *Large-scale bias in the Universe: bispectrum method*, [`arXiv:astro-ph/9706059`], MNRAS, 290, 651

Matsubara T., 2008, *Resumming cosmological perturbations via the Lagrangian picture: One-loop results in real space and in redshift space*, [`arXiv:0711.2521`], Phys. Rev. D, 77, 063530

Matsubara T., Suto Y., 1996, *Cosmological Redshift Distortion of Correlation Functions as a Probe of the Density Parameter and the Cosmological Constant*, [`arXiv:astro-ph/9604142`], ApJ, 470, L1

McCullagh N., Neyrinck M., Norberg P., Cole S., 2016, *Recovering dark-matter clustering from galaxies with Gaussianization*, [`arXiv:1511.02034`], MNRAS, 457, 3652

McDonald P., Roy A., 2009, *Clustering of dark matter tracers: generalizing bias for the coming era of precision LSS*, [`arXiv:0902.0991`], J. Cosmology Astropart. Phys., 8, 20

McDonald P., Seljak U., 2009, *How to evade the sample variance limit on measurements of redshift-space distortions*, [`arXiv:0810.0323`], J. Cosmology Astropart. Phys., 10, 7

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, *Equation of State Calculations by Fast Computing Machines*, The Journal of Chemical Physics, 21, 1087

Mo H. J., White S. D. M., 1996, *An analytic model for the spatial clustering of dark matter haloes*, [`arXiv:astro-ph/9512127`], MNRAS, 282, 347

Mo H. J., Jing Y. P., White S. D. M., 1996, *The correlation function of clusters of galaxies and the amplitude of mass fluctuations in the Universe*, [`arXiv:astro-ph/9602052`], MNRAS, 282, 1096

Mo H. J., Jing Y. P., White S. D. M., 1997, *High-order correlations of peaks and haloes: a step towards understanding galaxy biasing*, [`arXiv:astro-ph/9603039`], MNRAS, 284, 189

Mo H., van den Bosch F. C., White S., 2010, *Galaxy Formation and Evolution*. Cambridge University Press

Mohammed I., Seljak U., 2014, *Analytic model for the matter power spectrum, its covariance matrix and baryonic effects*, [`arXiv:1407.0060`], MNRAS, 445, 3382

Mohayaee R., Tully R. B., 2005, *The Cosmological Mean Density and Its Local Variations Probed by Peculiar Velocities*, [`arXiv:astro-ph/0509313`], ApJ, 635, L113

Mukhanov V., 2005, *Physical Foundations of Cosmology*. Cambridge University Press, doi:10.2277/0521563984

Murray I., Ghahramani Z., MacKay D., 2012, *MCMC for doubly-intractable distributions*, preprint, [`arXiv:1206.6848`]

Nadkarni-Ghosh S., Singhal A., 2016, *Phase space dynamics of triaxial collapse: joint density-velocity evolution*, [`arXiv:1407.1945`], MNRAS, 457, 2773

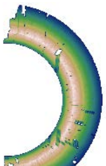
Neal R. M., 1993, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, in Technical Report CRG-TR-93-1. Dept. of Computer Science, University of Toronto

Neal R. M., 2012, *MCMC using Hamiltonian dynamics*, preprint, [`arXiv:1206.1901`]

Nesseris S., Perivolaropoulos L., 2008, *Testing* Λ*CDM with the growth function* $\delta(a)$*: Current constraints*, [`arXiv:0710.1092`], Phys. Rev. D, 77, 023504

Newman M. E. J., Barkema G. T., 1999, *Monte Carlo methods in statistical physics*. Clarendon Press

Neyrinck M. C., 2011, *Rejuvenating the Matter Power Spectrum. III. The Cosmology Sensitivity of Gaussianized Power Spectra*, [`arXiv:1105.2955`], ApJ, 742, 91



Neyrinck M. C., Szapudi I., Szalay A. S., 2009, *Rejuvenating the Matter Power Spectrum: Restoring Information with a Logarithmic Density Mapping*, [`arXiv:0903.4693`], ApJ, 698, L90

Neyrinck M. C., Szapudi I., Szalay A. S., 2011, *Rejuvenating Power Spectra. II. The Gaussianized Galaxy Density Field*, [`arXiv:1009.5680`], ApJ, 731, 116

Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, *A halo bias function measured deeply into voids without stochasticity*, [`arXiv:1309.6641`], MNRAS, 441, 646

Nusser A., Dekel A., Bertschinger E., Blumenthal G. R., 1991, *Cosmological velocity-density relation in the quasi-linear regime*, ApJ, 379, 6

Nuza S. E., Sánchez A. G., Prada F., Klypin A., Schlegel D. J., Gottlöber S., Montero-Dorta A. D., et al 2013, *The clustering of galaxies at $z \sim 0.5$ in the SDSS-III Data Release 9 BOSS-CMASS sample: a test for the ΛCDM cosmology*, [`arXiv:1202.6057`], MNRAS, 432, 743

Nuza S. E., Kitaura F.-S., Heß S., Libeskind N. I., Müller V., 2014, *The cosmic web of the Local Universe: cosmic variance, matter content and its relation to galaxy morphology*, [`arXiv:1406.1004`], MNRAS, 445, 988

Nußbaum M., 1975, *De Finetti, B.: Theory of probability*, Biometrische Zeitschrift, 17, 126

Okumura T., Matsubara T., Eisenstein D. J., Kayo I., Hikage C., Szalay A. S., Schneider D. P., 2008, *Large-Scale Anisotropic Correlation Function of SDSS Luminous Red Galaxies*, [`arXiv:0711.3640`], ApJ, 676, 889

Okumura T., Seljak U., Desjacques V., 2012, *Distribution function approach to redshift space distortions. Part III: halos and galaxies*, [`arXiv:1206.4070`], J. Cosmology Astropart. Phys., 11, 14

Padmanabhan N., Schlegel D. J., Finkbeiner D. P., Barentine J. C., Blanton M. R., Brewington H. J., Gunn J. E., Harvanek M., Hogg D. W., Ivezić Ž., Johnston D., Kent S. M., Kleinman S. J., Knapp G. R., Krzesinski J., Long D., et al. N., 2008, *An Improved Photometric Calibration of the Sloan Digital Sky Survey Imaging Data*, [`arXiv:astro-ph/0703454`], ApJ, 674, 1217

Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, *A 2 per cent distance to $z = 0.35$ by reconstructing baryon acoustic oscillations - I. Methods and application to the Sloan Digital Sky Survey*, [`arXiv:1202.0090`], MNRAS, 427, 2132

Peacock J. A., 1999, *Cosmological Physics*. Cambridge University Press

Peacock J. A., Dodds S. J., 1994, *Reconstructing the Linear Power Spectrum of Cosmological Mass Fluctuations*, [`arXiv:astro-ph/9311057`], MNRAS, 267, 1020

Peacock J. A., Heavens A. F., 1985, *The statistics of maxima in primordial density perturbations*, MNRAS, 217, 805

Peacock J. A., Cole S., Norberg P., Baugh C. M., Bland-Hawthorn J., Bridges T., Cannon R. D., Colless M., Collins C., Couch W., Dalton G., Deeley K., De Propris R., Driver S. P., Efstathiou G., et al., 2001, *A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey*, [`arXiv:astro-ph/0103143`], Nature, 410, 169

Pearson D. W., Samushia L., 2016, *Estimating the power spectrum covariance matrix with fewer mock samples*, [`arXiv:1509.00064`], MNRAS, 457, 993

Peebles P. J. E., 1973, *Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory*, ApJ, 185, 413

Peebles P. J. E., 1980, *The large-scale structure of the universe*. Princeton University Press

Peebles P. J. E., 1993, *Principles of Physical Cosmology*. Princeton University Press

Percival W. J., White M., 2009, *Testing cosmological structure formation using redshift-space distortions*, [`arXiv:0808.0003`], MNRAS, 393, 297

Percival W. J., Burkey D., Heavens A., Taylor A., Cole S., Peacock J. A., Baugh C. M., Bland-Hawthorn J., Bridges T., Cannon R., Colless M., Collins C., Couch W., Dalton G., De Propris R., Driver S. P., et al., 2004, *The 2dF Galaxy Redshift Survey: spherical harmonics analysis of fluctuations in the final catalogue*, [`arXiv:astro-ph/0406513`], MNRAS, 353, 1201

Perlmutter S., Aldering G., Goldhaber G., Knop R. A., Nugent P., Castro P. G., Deustua S., Fabbro S., Goobar A., Groom D. E., Hook I. M., Kim A. G., Kim M. Y., Lee J. C., Nunes N. J., Pain R., Pennypacker C. R., Quimby R., Lidman C., Ellis R. S., Irwin M., McMahon R. G., Ruiz-Lapuente P., Walton N., Schaefer B., Boyle B. J., Filippenko A. V., Matheson T., Fruchter A. S., Panagia N., Newberg H. J. M., Couch W. J., Project T. S. C., 1999, *Measurements of Ω and Λ from 42 High-Redshift Supernovae*, [`arXiv:astro-ph/9812133`], ApJ, 517, 565

Planck Collaboration 2015, *Planck 2015 results. XIII. Cosmological parameters*, preprint, [`arXiv:1502.01589`]

Planck Collaboration Ade P. A. R., Aghanim N., Arnaud M., Ashdown M., Aubourg E., Aumont J., Baccigalupi C., Banday A. J., Barreiro R. B., Bartolo N., Battaner E., Benabed K., Benoit-Lévy A., Bersanelli M., Bielewicz P., Bock J. J., Bonaldi A., Bonavera L., Bond J. R., Borrill J., Bouchet F. R., Burigana C., Calabrese E., Cardoso J.-F., Catalano A., Chamballu A., Chiang H. C., Christensen P. R., Clements D. L., Colombo L. P. L., Combet C., Crill B. P., Curto A., Cuttaia F., Danese L., Davies R. D., Davis R. J., de Bernardis P., de Zotti G., Delabrouille J., Dickinson C., Diego J. M., Dolag K., Donzelli S., Doré O., Douspis M., Ducout A., Dupac X., Efstathiou G., Elsner F., Enßlin T. A., Eriksen H. K., Finelli F., Forni O., Frailis M., Fraisse A. A., Franceschi E., Frejsel A., Galeotta S., Galli S., Ganga K., Génova-Santos R. T., Giard M., Gjerløw E., González-Nuevo J., Górski K. M., Gregorio A., Gruppuso A., Hansen F. K., Harrison D. L., Henrot-Versillé S., Hernández-Monteagudo C., Herranz D., Hildebrandt S. R., Hivon E., Hobson M., Hornstrup A., Huffenberger K. M., Hurier G., Jaffe A. H., Jaffe T. R., Jones W. C., Juvela M., Keihänen E., Keskitalo R., et al., 2016, *Planck intermediate results. XXXVII. Evidence of unbound gas from the kinetic Sunyaev-Zeldovich effect*, [`arXiv:1504.03339`], A&A, 586, A140

Plebanski J., Krasinski A., 2006, *An Introduction to General Relativity and Cosmology*. Cambridge University Press

Poulin V., Serpico P. D., Lesgourgues J., 2016, *A fresh look at linear cosmological constraints on a decaying Dark Matter component*, [`arXiv:1606.02073`], J. Cosmology Astropart. Phys., 8, 036

Press W. H., Schechter P., 1974, *Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation*, ApJ, 187, 425

Reid B. A., Samushia L., White M., Percival W. J., Manera M., Padmanabhan N., Ross A. J., Sánchez A. G., Bailey S., Bizyaev D., Bolton A. S., Brewington H., Brinkmann J., Brownstein J. R., Cuesta

A. J., Eisenstein D. J., Gunn J. E., Honscheid K., Malanushenko E., Malanushenko V., Maraston C., McBride C. K., Muna D., Nichol R. C., Oravetz D., Pan K., de Putter R., Roe N. A., Ross N. P., Schlegel D. J., Schneider D. P., Seo H.-J., Shelden A., Sheldon E. S., Simmons A., Skibba R. A., Snedden S., Swanson M. E. C., Thomas D., Tinker J., Tojeiro R., Verde L., Wake D. A., Weaver B. A., Weinberg D. H., Zehavi I., Zhao G.-B., 2012, *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measurements of the growth of structure and expansion rate at $z = 0.57$ from anisotropic clustering*, [`arXiv:1203.6641`], MNRAS, 426, 2719

Reid B. A., Seo H.-J., Leauthaud A., Tinker J. L., White M., 2014, *A 2.5 per cent measurement of the growth rate from small-scale redshift space clustering of SDSS-III CMASS galaxies*, [`arXiv:1404.3742`], MNRAS, 444, 476

Reid B., Ho S., Padmanabhan N., Percival W. J., Tinker J., Tojeiro R., White M., Eisenstein D. J., Maraston C., Ross A. J., Sánchez A. G., Schlegel D., Sheldon E., Strauss M. A., Thomas D., Wake D., Beutler F., Bizyaev D., Bolton A. S., Brownstein J. R., Chuang C.-H., Dawson K., Harding P., Kitaura F.-S., Leauthaud A., Masters K., McBride C. K., More S., Olmstead M. D., Oravetz D., Nuza S. E., Pan K., Parejko J., Pforr J., Prada F., Rodríguez-Torres S., Salazar-Albornoz S., Samushia L., Schneider D. P., Scóccola C. G., Simmons A., Vargas-Magana M., 2016, *SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues*, [`arXiv:1509.06529`], MNRAS, 455, 1553

Riess A. G., Filippenko A. V., Challis P., Clocchiatti A., Diercks A., Garnavich P. M., Gilliland R. L., Hogan C. J., Jha S., et al., 1998, *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*, [`arXiv:astro-ph/9805201`], AJ, 116, 1009

Robert C. P., Casella G., 2005, *Monte Carlo Statistical Methods*. Springer-Verlag New York, doi:10.1007/978-1-4757-4145-2

Rodríguez-Torres S. A., Chuang C.-H., Prada F., Guo H., Klypin A., Behroozi P., Hahn C. H., Comparat J., Yepes G., Montero-Dorta A. D., Brownstein J. R., Maraston C., McBride C. K., Tinker J., et al., 2016, *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modelling the clustering and halo occupation distribution of BOSS CMASS galaxies in the Final Data Release*, [`arXiv:1509.06404`], MNRAS, 460, 1173

Ross et al. A. J., 2015, *The clustering of the SDSS DR7 main Galaxy sample - I. A 4 per cent distance measure at $z = 0.15$*, [`arXiv:1409.3242`], MNRAS, 449, 835

Rubin V. C., Ford W. K. J., . Thonnard N., 1980, *Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 $R = 4$kpc to UGC 2885 $R = 122$ kpc*, ApJ, 238, 471

SDSS Collaboration 2000, SDSS Telescope image at Apache Point, New Mexico, `http://www.aip.de/highlight_archive/sdss2/sdss_telescope_hires.jpg`

Sachs R. K., Wolfe A. M., 1967, *Perturbations of a Cosmological Model and Angular Variations of the Microwave Background*, ApJ, 147, 73

Saito S., Baldauf T., Vlah Z., Seljak U., Okumura T., McDonald P., 2014, *Understanding higher-order nonlocal halo bias at large scales by combining the power spectrum with the bispectrum*, [`arXiv:1405.1447`], Phys. Rev. D, 90, 123522

Samushia L., Percival W. J., Raccanelli A., 2012, *Interpreting large-scale redshift-space distortion measurements*, [`arXiv:1102.1014`], MNRAS, 420, 2102

Samushia L., Reid B. A., White M., Percival W. J., Cuesta A. J., Lombriser L., Manera M., Nichol R. C., Schneider D. P., Bizyaev D., Brewington H., Malanushenko E., Malanushenko V., Oravetz D., et al., 2013, *The clustering of galaxies in the SDSS-III DR9 Baryon Oscillation Spectroscopic Survey: testing deviations from Λ and general relativity using anisotropic clustering of galaxies*, [`arXiv:1206.5309`], MNRAS, 429, 1514

Saslaw W. C., Hamilton A. J. S., 1984, *Thermodynamics and galaxy clustering - Nonlinear theory of high order correlations*, ApJ, 276, 13

Schaan E., Ferraro S., Vargas-Magaña M., Smith K. M., Ho S., Aiola S., Battaglia N., Bond J. R., De Bernardis F., Calabrese E., Cho H.-M., Devlin M. J., Dunkley J., Gallardo P. A., Hasselfield M., Henderson S., Hill J. C., Hincks A. D., Hlozek R., Hubmayr J., Hughes J. P., Irwin K. D., Koopman B., Kosowsky A., Li D., Louis T., Lungu M., Madhavacheril M., Maurin L., McMahon J. J., Moodley K., Naess S., Nati F., Newburgh L., Niemack M. D., Page L. A., Pappas C. G., Partridge B., Schmitt B. L., Sehgal N., Sherwin B. D., Sievers J. L., Spergel D. N., Staggs S. T., van Engelen A., Wollack E. J., 2015, *Evidence for the kinematic Sunyaev-Zel'dovich effect with ACTPol and velocity reconstruction from BOSS*, preprint, [`arXiv:1510.06442`]

Schaye J., Crain R. A., Bower R. G., Furlong M., Schaller M., Theuns T., Dalla Vecchia C., Frenk C. S., McCarthy I. G., Helly J. C., Jenkins A., Rosas-Guevara Y. M., White S. D. M., Baes M., Booth C. M., Camps P., Navarro J. F., Qu Y., 2015, *The EAGLE project: simulating the evolution and assembly of galaxies and their environments*, [`arXiv:1407.7040`], MNRAS, 446, 521

Scoccimarro R., 2000, *The Bispectrum: From Theory to Observations*, [`arXiv:astro-ph/0004086`], ApJ, 544, 597

Scoccimarro R., 2004, *Redshift-space distortions, pairwise velocities, and nonlinearities*, [`arXiv:astro-ph/0407214`], Phys. Rev. D, 70, 083007

Seljak U., 2012, *Bias, redshift space distortions and primordial nongaussianity of nonlinear transformations: application to Ly-α forest*, [`arXiv:1201.0594`], J. Cosmology Astropart. Phys., 3, 004

Seljak U., Zaldarriaga M., 1996, *A Line-of-Sight Integration Approach to Cosmic Microwave Background Anisotropies*, [`arXiv:astro-ph/9603033`], ApJ, 469, 437

Sheth R. K., Tormen G., 1999, *Large-scale bias and the peak background split*, [`arXiv:astro-ph/9901122`], MNRAS, 308, 119

Sheth R. K., Mo H. J., Tormen G., 2001, *Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes*, [`arXiv:astro-ph/9907024`], MNRAS, 323, 1

Smee S. A., Gunn J. E., Uomoto A., Roe N., Schlegel D., Rockosi C. M., Carr M. A., Leger F., et al. 2013, *The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey*, [`arXiv:1208.2233`], AJ, 146, 32

Smoot G. F., Bennett C. L., Kogut A., Wright E. L., Aymon J., Boggess N. W., Cheng E. S., de Amici G., Gulkis S., Hauser M. G., Hinshaw G., Jackson P. D., et al., 1992, *Structure in the COBE differential microwave radiometer first-year maps*, ApJ, 396, L1

Somerville R. S., Primack J. R., 1999, *Semi-analytic modelling of galaxy formation: the local Universe*, [`arXiv:astro-ph/9802268`], MNRAS, 310, 1087

Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, *Non-linear stochastic galaxy biasing in cosmological simulations*, [arXiv:astro-ph/9912073], MNRAS, 320, 289

Song Y.-S., Sabiu C. G., Kayo I., Nichol R. C., 2011, *Measuring coherent motions in the universe*, [arXiv:1006.4630], J. Cosmology Astropart. Phys., 5, 20

Springel V., 2005, *The cosmological simulation code GADGET-2*, [arXiv:astro-ph/0505010], MNRAS, 364, 1105

Sunyaev R. A., Zeldovich Y. B., 1970, *Small-Scale Fluctuations of Relic Radiation*, Ap&SS, 7, 3

Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, *Methods for rapidly processing angular masks of next-generation galaxy surveys*, [arXiv:0711.4352], MNRAS, 387, 1391

Tasitsiomi A., Kravtsov A. V., Gottlöber S., Klypin A. A., 2004, *Density Profiles of ΛCDM Clusters*, [arXiv:astro-ph/0311062], ApJ, 607, 125

Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, *The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests*, [arXiv:1001.3162], ApJ, 724, 878

Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, *Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering*, [arXiv:1005.1289], ApJ, 742, 16

Vale A., Ostriker J. P., 2004, *Linking halo mass to galaxy luminosity*, [arXiv:astro-ph/0402500], MNRAS, 353, 189

Verde L., 2007, *A practical guide to Basic Statistical Techniques for Data Analysis in Cosmology*, preprint, [arXiv:0712.3028]

Verde L., 2010, *Statistical Methods in Cosmology*, in Wolschin G., ed., Lecture Notes in Physics, Berlin Springer Verlag Vol. 800, Lecture Notes in Physics, Berlin Springer Verlag. pp 147–177, [arXiv:0911.3105], doi:10.1007/978-3-642-10598-2˙4

Vogelsberger M., Genel S., Springel V., Torrey P., Sijacki D., Xu D., Snyder G., Nelson D., Hernquist L., 2014, *Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe*, [arXiv:1405.2921], MNRAS, 444, 1518

Wandelt B. D., Larson D. L., Lakshminarayanan A., 2004, *Global, exact cosmic microwave background data analysis using Gibbs sampling*, [arXiv:astro-ph/0310080], Phys. Rev. D, 70, 083511

Wang Y., 2014, *Model-independent measurements of cosmic expansion and growth at $z = 0.57$ using the anisotropic clustering of CMASS galaxies from the Sloan Digital Sky Survey Data Release 9*, [arXiv:1404.5589], MNRAS, 443, 2950

Wang H., Mo H. J., Yang X., van den Bosch F. C., 2012, *Reconstructing the cosmic velocity and tidal fields with galaxy groups selected from the Sloan Digital Sky Survey*, [arXiv:1108.1008], MNRAS, 420, 1809

Wang H., Mo H. J., Yang X., van den Bosch F. C., 2013, *Reconstructing the Initial Density Field of the Local Universe: Methods and Tests with Mock Catalogs*, [arXiv:1301.1348], ApJ, 772, 63

Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, *ELUCID Exploring the Local Universe with the Reconstructed Initial Density Field. I. Hamiltonian Markov Chain Monte Carlo Method with Particle Mesh Dynamics*, [`arXiv:1407.3451`], ApJ, 794, 94

Watkins R., Feldman H. A., Hudson M. J., 2009, *Consistently large cosmic flows on scales of 100*h^{-1}*Mpc: a challenge for the standard* Λ*CDM cosmology*, [`arXiv:0809.4041`], MNRAS, 392, 743

Weinberg S., 2008, *Cosmology*. Oxford University Press

White M., 2001, *The mass of a halo*, [`arXiv:astro-ph/0011495`], A&A, 367, 27

White S. D. M., Frenk C. S., 1991, *Galaxy formation through hierarchical clustering*, ApJ, 379, 52

White M., Song Y.-S., Percival W. J., 2009, *Forecasting cosmological constraints from redshift surveys*, [`arXiv:0810.1518`], MNRAS, 397, 1348

White M., Tinker J. L., McBride C. K., 2014, *Mock galaxy catalogues using the quick particle mesh method*, [`arXiv:1309.5532`], MNRAS, 437, 2594

Yahil A., Strauss M. A., Davis M., Huchra J. P., 1991, *A redshift survey of IRAS galaxies. II - Methods for determining self-consistent velocity and density fields*, ApJ, 372, 380

Yang A., Saslaw W. C., 2011, *The Galaxy Counts-in-cells Distribution from the Sloan Digital Sky Survey*, [`arXiv:1009.0013`], ApJ, 729, 123

Yu Y., Zhang P., Lin W., Cui W., Fry J. N., 2011, *Gaussianizing the non-Gaussian lensing convergence field: The performance of the Gaussianization*, [`arXiv:1103.2858`], Phys. Rev. D, 84, 023523

Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, *Wiener Reconstruction of the Large-Scale Structure*, [`arXiv:astro-ph/9410080`], ApJ, 449, 446

Zehavi I., Zheng Z., Weinberg D. H., Blanton M. R., Bahcall N. A., Berlind A. A., Brinkmann J., Frieman J. A., Gunn J. E., Lupton R. H., Nichol R. C., Percival W. J., Schneider D. P., Skibba R. A., Strauss M. A., Tegmark M., York D. G., 2011, *Galaxy clustering in the completed sdss redshift survey: The dependence on color and luminosity*, The Astrophysical Journal, 736, 59

Zel'dovich Y. B., 1970, *Gravitational instability: an approximate theory for large density perturbations.*, A&A, 5, 84

Zeldovich Y. B., 1972, *A hypothesis, unifying the structure and the entropy of the Universe*, MNRAS, 160, 1P

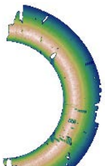
Zhang T.-J., Yu H.-R., Harnois-Déraps J., MacDonald I., Pen U.-L., 2011, *Increasing the Fisher Information Content in the Matter Power Spectrum by Nonlinear Wavelet Wiener Filtering*, [`arXiv:1008.3506`], ApJ, 728, 35

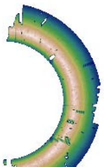
Zwicky F., 1933, *Die Rotverschiebung von extragalaktischen Nebeln*, Helvetica Physica Acta, 6, 110

da Ângela J., Shanks T., Croom S. M., Weilbacher P., Brunner R. J., Couch W. J., Miller L., Myers A. D., Nichol R. C., Pimbblet K. A., de Propris R., Richards G. T., Ross N. P., Schneider D. P., Wake D., 2008, *The 2dF-SDSS LRG and QSO survey: QSO clustering and the L-z degeneracy*, [`arXiv:astro-ph/0612401`], MNRAS, 383, 565 de la Torre S., Peacock J. A., 2013, *Reconstructing the distribution of haloes and mock galaxies below the resolution limit in cosmological simulations*, [arXiv:1212.3615], MNRAS, 435, 743

de la Torre S., Guzzo L., Peacock J. A., Branchini E., Iovino A., Granett B. R., Abbas U., Adami C., Arnouts S., Bel J., Bolzonella M., Bottini D., Cappi A., Coupon J., Cucciati O., Davidzon I., De Lucia G., Fritz A., Franzetti P., Fumana M., Garilli B., Ilbert O., Krywult J., Le Brun V., Le Fèvre O., Maccagni D., Małek K., Marulli F., McCracken H. J., Moscardini L., Paioro L., Percival W. J., Polletta M., Pollo A., Schlagenhaufer H., Scodeggio M., Tasca L. A. M., Tojeiro R., Vergani D., Zanichelli A., Burden A., Di Porto C., Marchetti A., Marinoni C., Mellier Y., Monaco P., Nichol R. C., Phleps S., Wolk M., Zamorani G., 2013, *The VIMOS Public Extragalactic Redshift Survey (VIPERS) . Galaxy clustering and redshift-space distortions at $z = 0.8$ in the first data release*, [arXiv:1303.2622], A&A, 557, A54

List of Figures





List of Tables

Trivia

A flicker book can be appreciated in the lower right corner on each odd page, representing the walk though a computed window function, perpendicular to the shown plane.