Inferring the dynamical growth of structures from high-redshift cosmological data sets

Natàlia Porqueres i Rosa

München 2019

Inferring the dynamical growth of structures from high-redshift cosmological data sets

Natàlia Porqueres i Rosa

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

vorgelegt von
Natàlia Porqueres i Rosa
aus Barcelona

München, den 16.07.2019

Erstgutachter: PD Dr. Torsten Enßlin

Zweitgutachter: Prof. Dr. Jochen Weller

Tag der mündlichen Prüfung: 16.09.2019

Contents

List of Figures

List of Tables

Zusammenfassung

Diese Arbeit widmet sich dem Verständnis der Bildung kosmischer Großstrukturen mithilfe der Analyse kosmologischer Beobachtungen. Dabei konzentriert sich meine Arbeit insbesondere auf die Analyse der räumlichen Materieverteilung und ihrer Dynamik im weit entfernten und damit hochrotverschobenen Universum aus Beobachtungsdaten.

Diese räumliche Materieverteilung, wurde in der Entwicklungsgeschichte des Universums durch mehrere physikalische Prozesse geformt, und ihre Beobachtung hat demzufolge das Potenzial, offene und fundamentale Fragen der Grundlagenphysik zu beantworten. Insbesondere die Analyse des hochrotverschobenen Universums kann dabei unschätzbare Einblicke in die Entwicklung und Dynamik unseres Universums liefern und ermöglicht es die Parameter des kosmologischen Modells zu bestimmen. Jedoch können Beobachtungen und physikalische Theorien nicht trivial miteinander verbunden werden, und es bedarf neuer Methoden der Datenanalyse, um das Beobachtungsrauschen und die systematischen Unsicherheiten zukünftiger Daten zu behandeln. In dieser Arbeit habe ich den Ansatz der Bayes'schen physikalischen Vorwärtsmodellierung verfolgt, welcher es erlaubt Vorhersagen des Standardmodells der Kosmologie vollständig und selbstkonsistent mit Daten zu testen. Obwohl dieses, vor Kurzem etablierte Standardmodell der Kosmologie, die meisten der derzeit existierenden kosmologischen Beobachtungen mit grosser Genauigkeit erklären kann, verbleiben Spannungen zwischen Modellvorhersagen und Beobachtungen, welche trotz zunehmender Datenqualität bestehen bleiben. Einige dieser Spannungen bestehen bei Messungen der parameter H_0 und σ_8 und auch bei Messungen von $H(z)$ bei hohen Rotverschiebungen mittels Lyman-α Daten. Die Ursachen dieser Spannungen sind unbekannt und könnten sowohl auf systematische Effekte in den Daten zurückzuführen sein als auch die ersten Anzeichen neuer Physik in den Beobachtungen sein. Eine genauere Behandlung systematischer Effekte in Daten ist daher notwendig, um neue Einsichten in die Physik des Universums zu gewinnen. Aus diesem Grund habe ich ein robustes Datenmodell entwickelt, welches unempfindlich gegenüber Beobachtungssystematiken ist und deswegen, selbst bei unbekannter Beobachtungssystematik, genaue kosmologische Aussagen ermöglicht. Die detaillierte Untersuchung der Verteilung der kosmischen Materie kann neue Erkenntnisse in der Grundlagenphysik liefern. Insbesondere die Analyse kosmischer Großstrukturen hat das Potenzial, zwischen homogener Dunkler Energie und Modifikationen der Schwerkraft zu unterscheiden, Modelle der Dunklen Materie zu testen als auch Neutrinomassen zu bestimmen. Aus diesem Grund wendet sich die Kosmologie derzeit der Analyse der kosmischen Großstruktur zu. Während sich in naher Zukunft die Beobachtungen hauptsächlich auf die Untersuchung der Galaxienverteilung bei hoher Rotverschiebung konzentrieren werden, zeichnet der Lyman-α Wald in Quasarspektren das kosmische Netz mit höherer Auflösung nach, als dies mit räumlichen Stichprobenraten von Galaxien zu erreichen wäre. Die vom Lyman-α Wald gemessenen Strukturen auf kleinen Skalen enthalten wertvolle Information über Neutrinomassen und Modelle der Dunklen Materie. Im Gegensatz zu Galaxien, die sich in Hochdichteregionen ansammeln, ermöglichen Beobachtungen des Lyman-α Waldes die Untersuchung unterdichter Regionen des Universums, die empfindlich auf diffuse Bestandteile des Universums wie die Dunkle Energie reagieren. Um den Informationsgehalt des Lyman-α Waldes zu erfassen, entwickelte ich ein statistisches Analyseverfahren, um die Materieverteilung bei Rotverschiebungen $z > 2$ aus dem Lyman-α Wald in Quasarspektren zu rekonstruieren.

Die Natur der Quasare, und im Allgemeinen der aktiven Galaxienkerne (AGN), ist noch immer ein Rätsel. Mit dem Ziel eines besseren Verständnisses ihrer Entstehung und Entwicklung, habe ich den Zusammenhang zwischen den Eigenschaften der AGN und denen der sie umgebenden kosmischen Strukturen untersucht. Diese Analyse lieferte den Nachweis einer Entwicklungssequenz zwischen zwei Arten von AGN. Detaillierte Analysen der dreidimensionalen (3d) Materieverteilung, die dem Lyman-α Wald zugrunde liegt, erfordern eine genaue Behandlung der systematischen Effekte. Die filamentartige Struktur der kosmischen Materieverteilung entsteht durch die nichtlineare gravitative Anhäufung von Materie im Rahmen des kosmischen Strukturwachstums. Die Erfassung von signifikanter physikalischer Information, die in dieser Filamentstruktur enthalten ist, erfordert eine 3d Analyse der Materieverteilung. Um dieses Ziel zu erreichen stellt meine Arbeit einen vollständig Bayes'schen Ansatz vor, der es ermöglicht die 3d Materieverteilung und ihre Dynamik bei Rotverschiebungen $z > 2$ aus dem Lyman-α Wald zu extrahieren. Dieses Verfahren liefert das unverfälschte Dichtefeld der Dunklen Materie und sein Leistungsspektrum, als auch Massen- und Geschwindigkeitsprofile kosmischer Strukturen, wie Galaxienhaufen und grosse leere Regionen. Des Weiteren ermöglicht das Verfahren die Bestimmung der Eigenschaften des intergalaktischen Mediums (IGM), wie das Temperatur-Dichte-Verhältnis des neutralen Wasserstoffs und die Spektralform der ersten Lichtquellen des Universums. Die genaue Bestimmung des Temperatur-Dichte-Verhältnisses lindert nicht nur Probleme bei der Dateninterpretation, die durch die unbekannte Astrophysik des IGM entstehen, sie könnte auch zur aktuellen Debatte um die Temperatur des neutralen Wasserstoffs in Regionen mit hoher Dichte beitragen. Wie in dieser Arbeit beschrieben, ist eine detaillierte und physikalisch plausible Inferenz der 3d Großstrukturen bei hoher Rotverschiebung möglich geworden, welche nun neue Wege zur ultimativen Überprüfung des kosmologischen Standardmodells anhand kommender Daten eröffnet.

Abstract

This thesis is devoted to understanding the formation of cosmic large-scale structures underlying the cosmological observations. In particular, my work focuses on inferring the spatial matter distribution and its dynamics in the high-redshift Universe from data. Imprinted by several physical processes throughout cosmic history, the matter distribution has the potential to answer outstanding questions of fundamental physics. In particular, the high-redshift Universe can provide invaluable information about the evolution of cosmological parameters governing the dynamics of our Universe. However, connecting theory and observations is a challenging task, requiring novel data analysis methods to cope with noise and systematic effects of next-generation data. In this thesis, I extended upon a Bayesian physical forward modelling approach that permits to jointly and fully self-consistently test the standard model with data.

While the recently established standard model of cosmology fits most cosmological observations to extraordinary accuracy, there remain some tensions between the model and observations that seem to persists despite the increasing quality of data. Some of these tensions are H_0, σ_8, and the high-redshift tension of $H(z)$ reported by Lyman-α analyses. The cause of these tensions might be related to systematic effects in the data but can also be first signs of new physics indicated by the data. A more accurate treatment of the systematic effects is, thus, inevitable. For this reason, I developed a robust data model that is insensitive to survey systematics and therefore provides unbiased cosmological results even in light of unknown systematics.

The detailed study of the cosmic matter distribution can provide new insights into fundamental physics. In particular, the analysis of cosmic large-scale structures has the potential to discriminate between homogeneous dark energy and modifications of gravity, test dark matter models, and constrain neutrino masses. For this reason, currently, cosmology turns to analyse the cosmic large-scale structure. While next-generation surveys will focus mostly on the analysis of the galaxy distribution at high redshift, the Lyman-α forest in quasar spectra traces the cosmic web with higher resolution than can be achieved with galaxy sampling rates. The small scales probed by the Lyman-α forest are sensitive to neutrino masses and dark matter models. Complementary to galaxy clustering in high-density regions, the Lyman-α forest traces the under-dense regions of the Universe, which are sensitive to diffuse components of the Universe such as dark energy. To harvest the information content of the Lyman-α forest, I developed a statistical framework to infer the matter distribution at $z > 2$ from the Lyman-α forest in quasar spectra.

The nature of quasars and, more generally, active galactic nuclei (AGN) is still mysterious. To achieve a better understanding of their formation and evolution, I investigated the relation between AGN and their large-scale structure environment. This analysis provided evidence of an evolutionary sequence between two types of AGN.

Detailed analyses of the three-dimensional matter distribution underlying the Lyman-α forest require an accurate treatment of systematic effects. The filamentary structure of the cosmic web arises as a result of the non-linear gravitational clustering governing structure formation. Capturing information entailed in this filamentary structure requires a three-dimensional (3d) analysis of the matter distribution. For this reason, this thesis presents a fully Bayesian framework to infer the 3d matter distribution and its dynamics at $z > 2$ from the Lyman-α forest. This method provides the unbiased dark matter density field and its corresponding power spectrum, recovering mass and velocity profiles of cosmic structures such as clusters and voids. Further, the method constrains the properties of the intergalactic medium (IGM), more specifically, the temperature-density relation of the neutral hydrogen and the spectral shape of the first luminous sources of the Universe. Besides avoiding biases due to the unknown astrophysics of the IGM, constraints on the temperature-density relation could contribute to the current debate on whether the neutral hydrogen is hotter in overdense regions or vice-versa.

As demonstrated in this work, detailed and physically plausible inference of 3d large-scale structures at high redshift has become feasible, providing new paths towards ultimate tests of the cosmological standard model with next-generation data.

Chapter 1

Introduction

1.1 Motivation

Currently, cosmology is at an exciting juncture. The standard model of cosmology has been extremely successful in numerous ways. Some of its predictions have been confirmed with high accuracy by observations and laboratory experiments. In particular, the standard model of cosmology fits the observations of the cosmic microwave background with extraordinary accuracy (see e.g. Planck Collaboration et al., 2018, 2019). The predicted helium content in metal-poor gas is in agreement with observations (Krauss and Romanelli, 1990; Smith et al., 1993; Hata et al., 1996). The number of neutrino families has been confirmed in laboratory experiments from Z-boson decay (Décamp et al., 1990; Adriani et al., 1992). However, the standard model of cosmology still faces open questions. Two of the largest contributions to the energy content of the late Universe - dark energy and cold dark matter (CDM) - have a mysterious nature. Dark matter has evaded direct detection in laboratory experiments and its nature is still unknown. In addition, some discrepancies seem to appear between the predictions of CDM and observations at small scales (Flores and Primack, 1994; Moore, 1994; Navarro et al., 1997; Moore et al., 1999). The nature of dark energy is even more puzzling, with a yet unknown dynamical evolution (see e.g. Efstathiou, 1999; Huterer and Turner, 1999; Saini et al., 2000; Weller and Albrecht, 2001; Nakamura and Chiba, 2001; Weller and Albrecht, 2002; Shafieloo et al., 2006; Sahni and Starobinsky, 2006).

Besides these open questions, some tensions between the standard model and observations seem to persist and increase. Among those are the H_0 and σ_8 tensions (e.g. Planck Collaboration et al., 2016a; Riess et al., 2016; Köhlinger et al., 2017; Riess et al., 2018; Abbott et al., 2018; Rusu et al., 2019) or the discrepancy of Lyman-α correlations with Planck observations (see e.g. Delubac et al., 2015; du Mas des Bourboux et al., 2017). These tensions might be related to systematic effects but may also be a sign of new physics beyond the current standard model such as decaying dark matter, dynamical dark energy or modifications of gravity. The resolution of these tensions requires new data with increasing accuracy and better control of systematic effects.

Insights on these questions can come from the study of the spatial matter distribution and growth of cosmic structures. According to the current picture of cosmology, large-scale structures of the Universe have their origin in primordial quantum fluctuations (Guth and Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983). These fluctuations were generated during an era of exponential expansion, driven by a quantum field known as inflaton (Guth, 1981; Linde, 1982; Albrecht and Steinhardt, 1982). This rapid expansion magnified the quantum fluctuations to macroscopic perturbations, which led to today's structures (Baumann, 2009). Through 13 billions years of cosmic history, these initial perturbations were modified by several physical processes such as decoupling of matter and radiation, acoustic oscillations, recombination, reionization, and gravitational collapse (see e.g. Peebles, 1980). The large-scale matter distribution then has the potential to answer outstanding questions in fundamental physics such as: which is the nature of dark energy and dark matter? Which are the neutrino masses and their mass hierarchy? Is the current paradigm of structure formation correct? To gain insights on all these questions, current cosmology turns to analyse the large-scale matter distribution (LSST Science Collaboration et al., 2009; Racca et al., 2016).

Next-generation surveys such as the Large Synoptic Survey Telescope (LSST, LSST Science Collaboration et al., 2009) and the Euclid satellite mission (Laureijs et al., 2011; Racca et al., 2016) aim at mapping the galaxy distribution at high-redshifts ($z \approx 3$). The analyses of the cosmic large-scale structures at high redshifts have the potential to provide valuable information about the dynamical behaviour of dark energy, determining whether its equation of state is redshift-dependent. Observations at $z > 2$ can also provide new insights into the redshift evolution of the Hubble function $H(z)$ tension reported in Delubac et al. (2015); du Mas des Bourboux et al. (2017) as well as σ_8, which is directly connected to the growth of structures (see e.g. Dodelson, 2003). However, extracting valuable information from these surveys requires good control of systematic effects. While previous surveys were limited by statistical noise, the next generation of surveys will be limited by systematic uncertainties (Ivezic et al., 2008; Laureijs et al., 2011; Amendola et al., 2018; Racca et al., 2016). In the past, such effects have been addressed by generating templates for such contaminations (Leistedt and Peiris, 2014; Bovy et al., 2012; Jasche and Lavaux, 2017). However, all these methods rely on a more or less robust estimate of the expected foreground contamination. Since future surveys can be subject to yet unknown contaminations, this thesis presents a novel likelihood to effectively deal with spurious effects induced by unknown foreground and target contaminations. Tests with contaminated simulated data showed that, while the standard analysis presented spurious effects on the matter density, this likelihood recovers the underlying matter distribution and unbiased power spectrum.

While most of the research of next-generation surveys will focus on galaxy clustering and lensing, the Lyman-α (Ly-α) forest can provide complementary information. The Ly-α forest is a set of absorption features in quasar spectra. These absorption lines are generated due to the scattering of photons by neutral hydrogen. Since the neutral hydrogen is found at low densities (Peirani et al., 2014), the Ly-α forest is sensitive to the underdense regions of the Universe, providing complementary information to galaxy surveys.

Since the non-linear effect of gravity in underdense regions is mitigated, voids are sensitive to the diffuse components of the Universe such as dark energy. Therefore, voids constitute a powerful laboratory to test the expansion of the Universe. More specifically, the analysis of the cosmic expansion in voids can discriminate between homogeneous dark energy and modified gravity since only the latter would affect the non-linear structures by modifying the Poisson equation.

In addition, the Ly-α forest probes the matter distribution at $z > 2$ with higher spatial resolution than can be achieved by galaxy surveys (see e.g. Lee et al., 2013, 2018). By probing down to scales of a few Mpc, the Ly-α forest is sensitive to neutrino masses. Neutrinos become non-relativistic at small redshift and, therefore, they free-stream as relativistic particles during most of the cosmic history. The effect of this free-streaming is suppression of power on small scales (Lesgourgues and Pastor, 2006). Then, the matter distribution traced by the Ly-α forest can provide constraints on neutrino masses (Rossi, 2014; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Yèche et al., 2017). Besides, the Ly-α forest can also provide information to discriminate between dark matter models based on their small-scale structure predictions (Viel et al., 2013).

In line with the potential of the Ly-α forest to solve outstanding questions of cosmology, a large number of ongoing surveys has been started: eBOSS (Myers et al., 2015), DESI (Levi et al., 2013; DESI Collaboration et al., 2016), CLAMATO (Lee et al., 2018). Additionally, future surveys like MSE (Maunakea Spectrographic Explorer McConnachie et al., 2016) and LSST (LSST Science Collaboration et al., 2009) will increase the amount of available Ly-α forest observations by a factor of 10. Most of the current and previous analyses of the Ly-α forest focus only on the analysis of the matter power spectrum (e.g. Croft et al., 1998; Seljak et al., 2006; Viel et al., 2006; Bird et al., 2011; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Nasir et al., 2016; Yèche et al., 2017; Boera et al., 2019), limited to first- and second-order statistics. However, non-linear dynamics transport information to high order correlations, corresponding to the filamentary structure of the cosmic web. Capturing the full information content of the cosmic large-scale structure requires to infer the entire three-dimensional cosmic matter distribution. The previous approaches (Kitaura et al., 2012b; Cisewski et al., 2014; Stark et al., 2015b; Ozbek et al., 2016; Horowitz et al., 2019) attempting to recover the three-dimensional matter distribution present biases that require better treatment of systematic effects. Going beyond these approaches, this thesis presents a Bayesian and statistically rigorous approach to perform dynamical matter clustering analyses with Ly-α forest data while accounting for all uncertainties inherent to the observations. Tested with simulated data emulating the CLAMATO survey, this method provides the unbiased dark matter distribution at $z = 2.5$.

1.2 Structure of this thesis

Chapter 2 introduces the standard model of cosmology. First, it describes the homogeneous Universe. Second, it focuses on the inhomogeneous cosmic web and the growth of structures from their initial conditions to the presently observed matter distribution and its power-spectrum. Chapter 3 discusses the observables to test cosmological and structure formation models. In particular, this chapter describes the observational effects on galaxy surveys and Ly-α forest datasets.

Harvesting the information from cosmological datasets requires statistical techniques. Due to cosmic variance, noise, and systematic observational effects, a unique recovery of the matter distribution is not possible and large-scale structure analyses have to match the statistical properties of the matter density field. For this reason, Chapter 4 presents the basic concepts of Bayesian statistics and data analysis. Particularly, this chapter describes Markov Chain Monte Carlo methods as relevant for this thesis.

This thesis presents an extension of a Bayesian framework for large-scale structure inference, BORG (Bayesian Origin Reconstruction from Galaxies), which is introduced and described in Chapter 5. More specifically, the extension of BORG in this thesis corresponds to the likelihood for contaminated data sets and an extension to the Ly-α forest.

Chapter 6-8 present the results of the thesis. Chapter 6 presents a study of the nature of active galactic nuclei (AGN), which confirmed an evolutionary transition between two different types of AGN. Investigating the effect of the large-scale environment on the evolution and formation of AGN is necessary to use these objects as tracers of the matter distribution. Since quasars are a type of AGN, this analysis aims at achieving a better understanding of the nature of AGN.

Chapter 7 presents a novel likelihood to effectively deal with contaminated datasets. The next generation of surveys will be affected by yet unknown contaminations. For this reason, this chapter presents the derivation and implementation of a likelihood that can obtain unbiased results from datasets affected by foreground and target contamination. The numerical implementation of the likelihood into the BORG framework is discussed. Tests with simulated data affected by galactic dust contamination demonstrate that this likelihood recovers the underlying matter distribution and unbiased matter power spectrum.

Chapter 8 presents a Bayesian framework to infer the three-dimensional density field from the Ly-α forest. The derivation of a likelihood based on the fluctuating Gunn-Peterson approximation is presented as well as the numerical implementation of the framework. Tests with simulated data emulating the CLAMATO survey demonstrated that the algorithm is able to infer the matter distribution at high-redshift and recover the correct power-spectrum. The algorithm can interpolate the information between lines of sight and recover mass and velocity profiles of cosmic structures such as clusters and voids. These results show that the inference of three-dimensional density fields from Ly-α forest data has become feasible.

Finally, Chapter 9 summarises the main results of this thesis and discusses further applications and development of the method.

Chapter 2

Cosmic structure formation

2.1 Introduction

The current standard model of cosmology rests on profound observational grounds such as the cosmic microwave background (CMB, Planck Collaboration et al., 2019, 2018), galaxy spatial distribution (see e.g. Eisenstein et al., 2011), or supernova distance measurements (Riess et al., 1998; Perlmutter et al., 1999). Within this framework, the Universe is mainly composed of dark matter, which is necessary to explain observed gravitational interactions, and dark energy, responsible for the accelerated expansion. Visible matter, such as gas, stars, and galaxies, constitute only a small component of the energy density of the Universe (see e.g. Mo et al., 2010).

The Big Bang and inflation scenario provides a physical model for the initial conditions of the Universe (Guth and Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983). The success of the Big Bang model rests on three observations: the cosmic expansion (Riess et al., 1998; Perlmutter et al., 1999), the abundances of light elements which are predicted by Big Bang nucleosynthesis (Alpher et al., 1948) and the blackbody spectrum of the CMB measured for the first time by the FIRAS instrument in the COBE satellite (Mather et al., 1994; Wright et al., 1994). After the Big Bang, the Universe underwent a period of exponential expansion known as inflation (Guth, 1981; Linde, 1982; Albrecht and Steinhardt, 1982). The inflationary paradigm is observationally supported by statistical homogeneity and isotropy of the CMB (Planck Collaboration et al., 2019).

During the inflationary era, the Universe was dominated by a quantum scalar field with negative pressure, the inflaton (see e.g. Mukhanov, 2005). Quantum fluctuations of the inflaton field were amplified to macroscopic scales due to the accelerated expansion. Later, these initial perturbations were modified by several physical processes, such as decoupling of radiation and matter, recombination, free-streaming of neutrinos and acoustic oscillations. During the matter and dark-energy dominated epochs, the initial perturbations evolved by gravitational interaction and formed the presently observed non-linear structures.

This chapter presents a brief introduction to the standard model of cosmology and structure formation models.

2.1.1 Einstein's equation

The standard model of cosmology arises as a specific solution to Einstein's equation for a homogeneous and isotropic matter and energy distribution. In this framework, the geometry of space-time is described by general relativity (GR). GR states that the matter and energy distribution of the Universe determines its geometry via Einstein's equation:

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} \mathcal{R} = 8\pi G T_{\mu\nu}. \tag{2.1}$$

In this equation, the space-time geometry is encoded in the Einstein tensor $G_{\mu\nu}$, which depends on the metric components $g_{\mu\nu}$. The mass-energy distribution is given by the energy-momentum tensor $T_{\mu\nu}$. In Einstein's equation, $R_{\mu\nu}$ is known as the Ricci tensor, $\mathcal{R} = g^{\mu\nu} R_{\mu\nu}$ is the Ricci scalar and G is Newton's gravitational constant.

Supernova observations led to the discovery that the expansion of the Universe is accelerated (Riess et al., 1998; Perlmutter et al., 1999). This accelerated expansion requires to introduce a cosmological constant Λ into Einstein's equation:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{2.2}$$

where Λ is seen as an energy component with negative pressure $P_\Lambda = -\rho_\Lambda c^2$.

2.1.2 Cosmological principle

The specific solution of Einstein's equation corresponding to the standard model of cosmology is derived from the matter-energy distribution stated by the cosmological principle. The cosmological principle states that the Universe at large scales is isotropic and homogeneous (see e.g. Peacock, 1999). Although the cosmological principle was introduced as an assumption to solve Einstein's equation, it is currently well supported by observational data such as the rotational invariance of the CMB (Planck Collaboration et al., 2019). Let's see the implications of the cosmological principle:

- Isotropy implies that comoving observers cannot define a preferred direction in any of their observables.
- Homogeneity means that there is no preferred location for any observer with the mean motion of cosmic matter.

This implies that the comoving observers measure the same observables and no location can be distinguished in a fundamental way from any other. This is also known as the Copernican principle.

2.1.3 Robertson-Walker metric

The cosmological principle defines a specific class of fundamental observers: comoving observers to whom the Universe appears isotropic and homogeneous. This definition is necessary since two observers at the same point in relative motion cannot both see an isotropic Universe.

Fundamental observers define a reference frame with a metric of the form

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 + g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j, \tag{2.3}$$

where t is the proper time of the fundamental observer and g_{ij} are the spatial components of the metric. The cross-terms of the metric g_{0i} must vanish for an isotropic Universe. Otherwise, they would define a preferred direction.

In a homogeneous and isotropic Universe, the only allowed motion is a pure expansion or contraction. In particular, for an expanding Universe with a scale factor $a\left(t\right)$, the metric reads

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - a\left(t\right)^2\,\mathrm{d}q^2 \tag{2.4}$$

with $\mathrm{d}q$ being the line element. Due to isotropy, the matter distribution can only depend on the radial coordinate. Therefore, the line element $\mathrm{d}q$ can be written in spherical coordinates (r, θ, ϕ) (Misner et al., 1973)

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - a\left(t\right)^2\left[\mathrm{d}r^2 + f_K^2\left(r\right)\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right]. \tag{2.5}$$

This metric is known as the Robertson-Walker metric. The function $f_K\left(r\right)$ encodes the curvature of the Universe:

$$f_K\left(r\right) = \begin{cases} \frac{1}{K^{1/2}}\sin\left(\sqrt{K}r\right), & K > 0 \text{ closed} \\ r, & K = 0 \text{ flat} \\ \frac{1}{|K|^{1/2}}\sinh\left(\sqrt{|K|}r\right), & K < 0 \text{ open} \end{cases}$$

An important implication of the Robertson-Walker metric is that physical distances depend on the dynamical evolution of the Universe, which is encoded in the scale factor. Therefore, if two observers are separated by a comoving distance x_0, their physical distance at some time t will be $a\left(t\right)x_0$.

2.1.4 Deriving the Friedman equations: A model for cosmic dynamics

As discussed in the previous sections, the Robertson-Walker metric describes the geometry of an expanding, homogeneous, and isotropic Universe. However, it does not inform about the dynamical evolution of the Universe. Friedmann proposed a solution of Einstein's equation for an expanding homogeneous and isotropic Universe (Friedmann, 1922, 1924). These equations, which describe the dynamics of the Universe, are known as the Friedmann equations.

To derive the Friedman equations, we first need to evaluate the Ricci tensor (see e.g. Carroll, 2004) for the Robertson-Walker metric (eq. 2.5). Then, the components of the

Ricci tensor are given by

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a}, && (2.6)\\ R_{ij} &= \left(2\dot{a}^2 + a\ddot{a}\right)\delta^K_{ij}, && (2.7) \end{aligned}$$

where δ^K_{ij} is a Kronecker delta. The Ricci scalar then reads

$$\mathcal{R} = 6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2}\right). \qquad (2.8)$$

Secondly, we need the components of the energy-momentum tensor $T_{\mu\nu}$, which informs about the energy-like characteristics of a system (energy, pressure, stress, etc.). The cosmological principle implies that the energy-momentum tensor can be completely specified by two components: the energy density ρ and the isotropic pressure p. Due to isotropy, the energy-momentum tensor is diagonal and all the space-components must be equal $T_{11} = T_{22} = T_{33} = p$. The energy-momentum tensor is, then, given by

$$T_{\mu\nu} = \left(\rho + \frac{p}{c^2}\right) u_\mu u_\nu - p g_{\mu\nu}, \qquad (2.9)$$

where ρ is the total energy density, p is the pressure and u_μ is the velocity (Carroll, 2004). Therefore, T_{00} is the total mass-energy density $\rho_{\rm tot}$, which contains the mass-energy density from all the components of the Universe (matter, radiation and dark energy). The time component of Einstein's equation can be obtained from eq. (2.1) and (2.6):

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho_{\rm tot}. \qquad (2.10)$$

This solution is one of the Friedman equations.

With the energy momentum tensor in eq. (2.9), Einstein's equation (eq. 2.1) can be reduced to

$$\begin{aligned} \frac{\dot{a}}{a} &= \sqrt{\frac{8\pi G}{3}\rho_m - K\frac{c^2}{a^2} + \frac{\Lambda c^2}{3}}, && (2.11)\\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda c^2}{3}. && (2.12) \end{aligned}$$

These equations describe very successfully the dynamics of the Universe.

Often, the Friedman equations are written in terms of the Hubble function,

$$H\left(t\right) \equiv \frac{\dot{a}}{a}, \qquad (2.13)$$

which characterizes the expansion of the Universe. The Hubble function allows us to define a critical density, corresponding to the density at the critical point between an expanding and contracting Universe. This critical density $\rho_{\rm crit}$ is given by

$$\rho_{\rm crit} \equiv \frac{3H_0^2}{8\pi G} \qquad (2.14)$$

Parameter	Symbol	Value
Hubble parameter	h	0.674 ± 0.005
Total matter density	Ω_{m}	0.315 ± 0.007
Baryon density	Ω_{b}	0.0493 ± 0.0002
Cosmological constant	Ω_{Λ}	0.6889 ± 0.0056
Curvature parameter	Ω_{k}	0.001 ± 0.002

Table 2.1: Cosmological parameters (Planck Collaboration et al., 2018)

with $H_0 = H\,(z=0)$. Then, the first Friedman equation reads

$$H^2 = H_0^2 \left[\frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_c}{a^2} + \Omega_\Lambda\right], \tag{2.15}$$

with

$$\Omega_m = \frac{\rho}{\rho_{\mathrm{crit}}},\ \Omega_r = \frac{8\pi G p}{c^2 H_0^2},\ \Omega_\Lambda = \frac{\Lambda}{3H_0^2},\ \Omega_c = \frac{Kc^2}{H_0^2}. \tag{2.16}$$

These parameters Ω are known as the cosmological parameters and have been measured with high accuracy by the Planck satellite mission (see e.g. Planck Collaboration et al., 2018). However, observations of galaxies, the Ly-α forest and Cepheids show tensions with the CMB measurments. These tensions seem to persist and increase with the increasing accuracy of surveys. Some of these tensions are the H_0 and σ_8 tensions (e.g. Planck Collaboration et al., 2016a; Riess et al., 2016; Köhlinger et al., 2017; Riess et al., 2018; Abbott et al., 2018; Rusu et al., 2019) and the high-redshift tension of $H(z)$ (see e.g. Delubac et al., 2015; du Mas des Bourboux et al., 2017).

2.1.5 Redshift

Most of the cosmological observations are made through light. Therefore, understanding how photons propagate in the Universe is inevitable to interpret the observational data. In the framework of general relativity, massless particles like photons travel along null geodesics, $\mathrm{d}s^2 = 0$ (Carroll, 2004). Therefore, in an expanding Universe, the comoving distance between a photon source and the observer is given by the traveling time of photons,

$$d \equiv \int_0^d \mathrm{d}q = \int_{t_s}^{t_o} \mathrm{d}t\, \frac{c}{a\,(t)}. \tag{2.17}$$

where $\mathrm{d}q$ is the line element from eq. (2.4). By construction, the comoving distance between two observers at fixed comoving coordinates remains constant. However, different observers will measure different duration of a light signal due to cosmic expansion. To illustrate this, let's assume that a photon source sends a signal with wavelength λ_s. Assuming that the duration (Δt) of a signal corresponds to the period of the wave, $\Delta t = \lambda/c$, an observer will measure a wavelength of $\lambda_o = c\Delta t_o$,

$$\frac{\lambda_o}{\lambda_s} = \frac{\Delta t_o}{\Delta t_s} = \frac{a\,(t_o)}{a\,(t_s)} \equiv 1 + z, \tag{2.18}$$

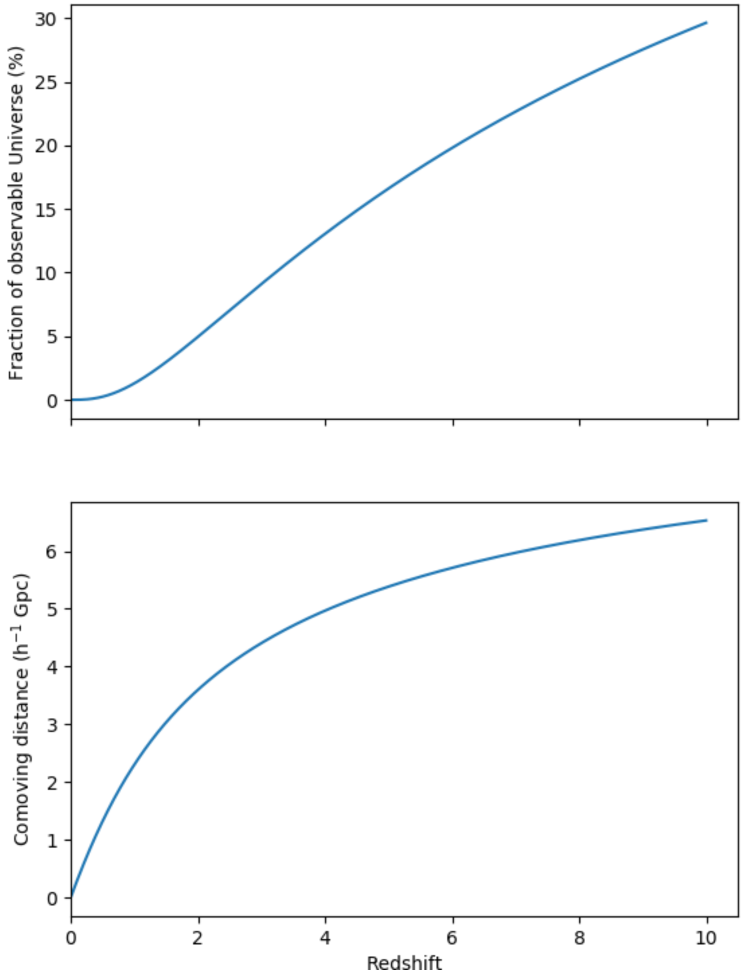


Figure 2.1: Lower panel: Comoving distance as a function of redshift for a flat universe. Upper panel: Fraction of the observable universe accessible at a redshift z for a flat universe. The next generation of surveys, mapping the matter distribution out to $z \approx 3$ will observe 8% of the observable universe.

where z is defined as the cosmological redshift.

Given a cosmological model, the observed redshift allows measuring the distance between the observer and the source as

$$d_{\mathrm{com}}\left(z_0, z_s\right) = \int_{t(z_s)}^{t(z_0)} \mathrm{d}t = \int_{a(z_s)}^{a(z_0)} \mathrm{d}a \, \frac{c}{a^2 H\left(a\right)}. \tag{2.19}$$

Figure 2.1 shows the relation between the redshift and the comoving distance and the fraction of the observable universe accessed at a given redshift. This fraction has been computed as $V_{\mathrm{com}}/V_{\mathrm{observable}}$ where the comoving volume corresponds to $V_{\mathrm{com}} = (4\pi/3)\, d_{\mathrm{com}}^3$ and the observable universe radius in the ΛCDM model is given by $d_{\mathrm{observable}} = 3.24 c/H_0 =$ 14 Gpc (Ryden, 2003). The next generation of surveys (LSST and Euclid) will map the galaxy distribution out to $z \approx 3$, observing a significant fraction of the observable universe: 8%.

2.2 Cosmic structure growth

Previous sections describe the homogeneous and isotropic Universe. However, homogeneity cannot hold at small scales as indicated by the existence of galaxies and their distribution in the sky. This section, therefore, focuses on the inhomogeneities of the Universe, starting from the initial seed fluctuations to the evolved matter perturbations traced by the galaxy distribution.

After the Big Bang, the Universe underwent a phase of exponential expansion driven by a quantum field, the inflaton (Guth, 1981; Linde, 1982; Albrecht and Steinhardt, 1982). The quantum fluctuations of the inflaton were magnified to macroscopic cosmological perturbations due to the exponential expansion. This inflationary scenario predicts a statistically homogeneous and isotropic density field with nearly-Gaussian perturbations (Guth and Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983).

After the inflation, radiation and matter were coupled by electron-photon scattering. At this time, the energy density of the Universe was dominated by radiation and, therefore, the gravitational potential was determined by radiation perturbations. Due to radiation pressure, the growth of photons perturbations in this epoch was suppressed. If perturbations do not grow, the gravitational potential decays due to the cosmic expansion, which dilutes the energy density field.

Although the radiation perturbation did not grow, matter perturbations slowly collapsed and grew during the radiation-dominated era (Dodelson, 2003). The Universe transited from radiation to matter domination at the time of equality. Close to the time of equality, the radiation pressure became less important and the matter perturbations started growing faster. In the matter-dominated era, the cosmological perturbations were modified by self-gravity. At $T \approx 3000$ K, electrons and protons recombined to form hydrogen atoms, increasing the mean-free path of photons. At this point, radiation and matter decoupled and the last-scattering surface, corresponding to the cosmic microwave background, was established.

After recombination and decoupling of matter and radiation, hydrogen clouds started collapsing due to gravitational interaction. During this slow collapse, the only photons in the Universe were those corresponding to the cosmic microwave background and 21 cm radio emissions from spin-transitions in hydrogen atoms (Loeb and Furlanetto, 2013). This period is known as the dark ages. Later, the first stars and galaxies formed and large-scale structures appeared. Stars and galaxies emitted high-energy photons that led to the reionization of the Universe (Loeb and Furlanetto, 2013). Finally, the Universe transited to dark-energy domination, with an accelerated expansion (Riess et al., 1998; Perlmutter et al., 1999).

All these physical processes modified and imprinted the large-scale structures of the Universe. For this reason, the analysis of the matter distribution can provide significant information on fundamental physics.

2.2.1 Evidence for dark matter

In the standard model of cosmology, structures evolve from small seed perturbations by gravitational interaction: matter flows away from regions where the density is below the mean and falls towards high-density regions. However, gravitational clustering is a slow process. To be able to explain today's structures only by baryonic matter, the CMB fluctuations are required to be two orders of magnitudes larger than the observed fluctuations (Einasto, 2009). Therefore, another kind of matter was required. This observation led to the introduction of dark matter in cosmology, for which observational indications were already found in the velocities of the Coma cluster (Zwicky, 1933).

Dark matter does not interact with radiation. While baryon perturbations did not grow because they were tightly coupled with radiation perturbations, the growth of dark matter perturbations was not suppressed by radiation pressure. Therefore, dark matter perturbations collapsed earlier than baryonic matter, setting up the gravitational potential. After the decoupling, baryons were released from the relatively smooth density field and fell into the potential wells of dark matter (Dodelson, 2003). Introducing dark matter then solved the problem of explaining the presently observed structures: dark matter perturbations started growing earlier and determined the gravitational potential.

Later, further evidence of dark matter was found in the rotation curves of galaxies: they flatten at large radii instead of showing the expected Keplerian fall, indicating the presence of additional non-observable matter (Rubin and Ford, 1970). Weak lensing and the X-ray studies of bullet clusters also provided evidence for dark matter, indicating that the gravitational centre of the cluster does not always need to match that of the visible matter distribution (Clowe et al., 2004; Markevitch et al., 2004; Clowe et al., 2006).

Although the nature of dark matter is still unknown and its corresponding particle is elusive to laboratory experiments (see e.g. Liu et al., 2017), some properties of dark matter are constrained by large-scale structure observations. More specifically, to explain the matter distribution,

- Dark matter is required to become non-relativistic early on: very high velocities would have resulted in significant damping of the small scale structures (see e.g. Dodelson, 2003), which is incompatible with observations. Non-relativistic dark matter is often called Cold Dark Matter (CDM).

- Dark matter is assumed to be stable and only weakly self-interacting. This is necessary to account for its contribution to the critical density.

The identification of the dark matter particle is a current milestone of cosmology and particle physics. The analysis of the matter clustering can provide some insights into the nature of dark matter by testing dark matter models based on their predictions at small scales (see e.g. Flores and Primack, 1994; Moore, 1994; Navarro et al., 1997; Moore et al., 1999).

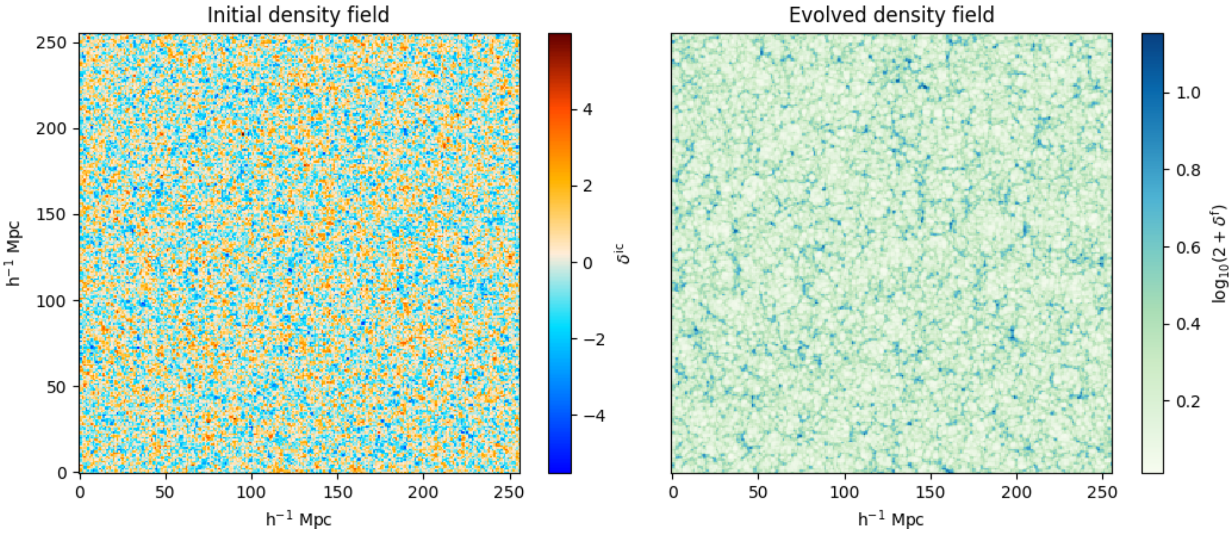


Figure 2.2: The initial density field is well-described by Gaussian statistics (left panel). However, the non-linear gravitational collapse introduces mode coupling and phase correlations, resulting in the filamentary structure of the evolved density field, which can be seen in the right panel.

2.2.2 Gaussian initial conditions

As described above, the Universe underwent an inflationary period that generated the initial conditions of large-scale structures. The inflationary model predicts the initial seed perturbations to be generated from the superposition of a high number of independent quantum fluctuations. According to the inflationary scenario, these seed perturbations are nearly Gaussian distributed. This was confirmed by the Planck satellite mission (Planck Collaboration et al., 2016b), which strongly constrained the deviation from Gaussianity. These observations of the CMB provide a well-supported model for the initial conditions of cosmic structures.

The initial conditions can be expressed in terms of fluctuations of the density around the cosmic mean density $\bar{\rho}$ by defining the density contrast δ as

$$\delta = \frac{\rho - \bar{\rho}}{\bar{\rho}}. \tag{2.20}$$

Since the density contrast is defined by subtracting the cosmic mean density, its Gaussian fluctuations are centered at zero. Therefore, the distribution of the initial density contrast can be written as

$$P\left(\delta^{\rm ic}\right) = \frac{1}{\sqrt{\det\left(2\pi\mathbf{S}\right)}} \exp\left[-\frac{1}{2}\int \mathrm{d}x \int \mathrm{d}y\, \delta^{\rm ic}_x S^{-1}_{xy} \delta^{\rm ic}_y\right], \tag{2.21}$$

where $S_{xy} = \langle \delta^{\rm ic}_x \delta^{\rm ic}_y \rangle$ is the covariance matrix. The Gaussian distribution is completely characterized by the mean and two-point statistics.

While the Gaussian approximation predicts the density amplitudes δ to be symmetrically distributed among positive and negative values, strong and weak energy conditions of general relativity (Carroll, 2004) require $\delta \geq -1$. This indicates that the Gaussian assumption is only valid at early times when the density fluctuations are small $|\delta| \ll 1$.

As a consequence of the isotropy of the Universe, the covariance matrix S_{xy} introduced above becomes diagonal in Fourier space. Therefore, we can write

$$\langle \bar{\delta}\left(\mathbf{k}\right) \bar{\delta}\left(\mathbf{k}'\right) \rangle = \left(2\pi\right)^3 \delta^D \left(\mathbf{k} - \mathbf{k}'\right) P_{\bar{\delta}(k)}, \tag{2.22}$$

with the Fourier transform being $\bar{\delta}\left(\mathbf{k}\right) = \int \mathrm{d}^3k \;\; \delta\left(\mathbf{x}\right) \exp\left(-i\mathbf{kx}\right)$, the density power spectrum $P_{\bar{\delta}(k)}$ and $\delta^D\left(x\right)$ being a Dirac delta. In a homogeneous and isotropic universe, the different Fourier modes are uncorrelated and their probability distribution is

$$P\left(\bar{\delta}\left(\mathbf{k}\right)\right) = \frac{1}{\sqrt{2\pi P_{\bar{\delta}(k)}}} \exp\left(-\frac{1}{2}\frac{|\bar{\delta}\left(\mathbf{k}\right)|^2}{P_{\bar{\delta}(k)}}\right). \tag{2.23}$$

Therefore, the initial density fluctuations are well described by Gaussian statistics and the power spectrum completely characterizes its properties. However, the non-linear gravitational collapse will amplify amplitudes of the density perturbations and introduce mode coupling and phase correlations that invalidate Gaussianity (Dodelson, 2003). As a consequence, the evolved density field traced by observed galaxies cannot be described by Gaussian statistics.

2.2.3 The shape of the matter power spectrum

As different processes in the cosmic history modify the density fluctuations, they imprinted the shape of the matter power spectrum. In particular, the evolution of cosmological perturbations depends on the moment they enter the horizon. Therefore, perturbations of different scales will have a different evolution (Dodelson, 2003). A perturbation crosses the horizon when its wavelength $\lambda = 2\pi/k$ is equal to $d_H\left(a\right) = c/H\left(a\right)$. This implies that large-scale perturbations enter the horizon at later times than small scale perturbation and, therefore, will be imprinted by different processes. This is reflected in the shape of the power spectrum.

While large-scale perturbations entered the horizon when the Universe was matter dominated, the small-scale perturbations entered during the radiation era. As a consequence, the growth of small scale perturbations is suppressed, resulting in lower amplitudes of the power spectrum. During the radiation-dominated era, the gravitational potential was determined by radiation perturbations. Since photons did not cluster, radiation perturbations did not grow and the gravitational potential decayed due to cosmic expansion. The expansion time scale t_{Hubble} is smaller than the collapse time of dark matter t_{dm}:

$$t_{\mathrm{Hubble}} \propto \frac{1}{\sqrt{G\rho_r}} < \frac{1}{\sqrt{G\rho_m}} \propto t_{\mathrm{dm}} \tag{2.24}$$

since $\rho_r > \rho_m$ in the radiation dominated era. Therefore, the growth of matter perturbations entering the horizon during the radiation era was suppressed.

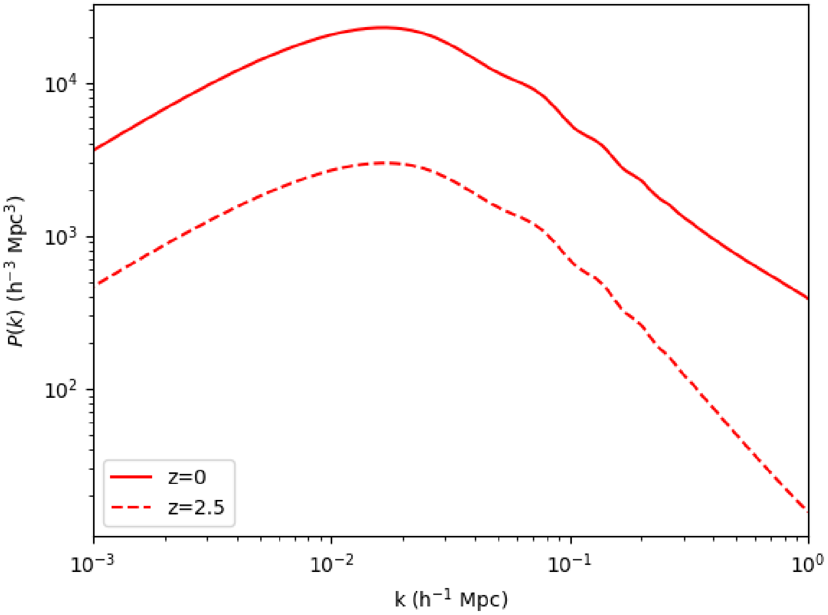


Figure 2.3: Matter power spectrum at redshift $z = 0$ and $z = 2.5$. The turn-over of the matter power spectrum corresponds to the size of the horizon at the time of equality, when the Universe transited from a radiation-dominated era to matter domination. This is due to the fact that small scale perturbations entered the horizon earlier, during radiation domination, and their growth was suppressed by radiation pressure. The wiggles of the power spectrum at $k \approx 0.04$ h Mpc^{-}1 are the baryonic acoustic oscillations (BAO), due to the photon-baryon coupling. The period of BAO depends on the sound speed and, therefore, contains information about the baryon content of the Universe.

In contrast, large-scale perturbations entered the horizon at later times, when the Universe was matter-dominated. The growth of the large-scale matter perturbations, therefore, was not suppressed. This difference between large- and small-scale perturbations produces a turn-over in the power spectrum, as shown in Fig. 2.3. The location of this turn-over corresponds to the transition from radiation-dominated to matter-domination, at the equality time: structures that entered before the equality time ($k > k_{\rm equality}$) have lower amplitudes due to growth suppression. Consequently, the mode $k_{\rm equality}$ of the turn-over is a measurement of the horizon size at the time of equality, which contains information on the dark matter content of the Universe.

Besides the turn-over, another relevant feature of the matter power spectrum is the wiggles (Eisenstein and Hu, 1998, 1999) corresponding to the baryonic acoustic oscillations (BAO). The BAO originated due to photon-matter interactions: when an overdense region attracted matter gravitationally, the coupling with photons generated pressure. These counteracting forces generated oscillations that are imprinted in the matter distribution. The period of these oscillations is determined by the sound speed, which depends on the baryon density. Therefore, the BAO contains information about the baryon component of the Universe.

2.2.4 The Boltzmann equation

As discussed above, the components of the Universe are affected by different interactions: photons are scattered by their interaction with baryons, electrons and protons are coupled, and gravity affects all the components of the Universe. All these interactions need to be taken into account to accurately describe the matter and photon distribution. This can be achieved with the Boltzmann equation, which can account for all these couplings (see e.g. Dodelson, 2003).

The Boltzmann equation states that the rate of change in the abundance of a particle type is the difference between its production and elimination rates,

$$\frac{\mathrm{d}f\left(\mathbf{r},\mathbf{p},t\right)}{\mathrm{d}t}=C[f], \tag{2.25}$$

where $f\left(\mathbf{r},\mathbf{p},t\right)$ is the number density in phase space and $C[f]$ contains all collision terms. For collisionless dark matter, $C[f]$ vanishes and therefore,

$$\frac{\mathrm{d}f}{\mathrm{d}t}=\frac{\partial f}{\partial t}+\frac{\partial f}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}t}+\frac{\partial f}{\partial p}\frac{\mathrm{d}p}{\mathrm{d}t}+\frac{\partial f}{\partial \hat{\mathbf{p}^i}}\frac{\mathrm{d}\hat{\mathbf{p}^i}}{dt}=0 \tag{2.26}$$

where p is the modulus of the momentum and $\hat{\mathbf{p}^i}$ are its unitary vectors. However, the last term does not contribute to first order perturbation theory (Dodelson, 2003). Equation (2.26) states that the number of particles in an element of the phase space does not change but the phase space element moves.

At scales smaller than the Hubble radius[1] $d \ll c/H_0$, relativistic effects, such as the curvature of space-time, are believed to be negligible (Dodelson, 2003). Therefore, the equations of motion of non-relativistic dark matter can be approximated by Newtonian gravity. In the Newtonian limit, $\mathrm{d}x^i/\mathrm{d}t = p/m$ and $\mathrm{d}p/\mathrm{d}t = m\nabla\Phi$, where Φ is the gravitational potential given by the Poisson equation

$$\nabla^2\Phi\left(\mathbf{r},t\right)=4\pi Gm\int f\left(\mathbf{r},\mathbf{p},t\right)\mathrm{d}^3\mathbf{p}. \tag{2.27}$$

Then equation (2.26) reads

$$\frac{\mathrm{d}f}{\mathrm{d}t}=\frac{\partial f}{\partial t}+\frac{\mathbf{p}}{m}\cdot\nabla f-m\nabla\Phi\cdot\frac{\partial f}{\partial \mathbf{p}}=0 \tag{2.28}$$

which is also known as the Vlasov equation. This differential equation is non-linear since the gravitational potential Φ depends on the distribution function. There is no general analytical solution of this equation for collissionless dark matter. Therefore, its behaviour is studied by two different approaches: N-body simulations that provide a sampled representation of $f\left(\mathbf{r},\mathbf{p},t\right)$ and approximated analytical approaches based on fluid dynamics.

[1] The Hubble radius is $d_H = c/H_0$.

Fluid dynamics approach

Although dark matter accounts for a large fraction of the cosmic matter density, dark matter particles are extremely light compared to the typical mass of galaxies and stars. For this reason, their number density is expected to be very high to account for their contribution to the total cosmic density. Under these conditions, discreteness effects are negligible (Bernardeau et al., 2002). At large scales, where multi-streaming effects are negligible, collisionless dark matter can be approximated as a fluid (Peebles, 1980).

In the fluid approach, the Vlasov equation (eq. 2.28) is approximated by taking momentum moments of the distribution function $f\left(\mathbf{r},\mathbf{p},t\right)$ (Bernardeau et al., 2002). The zeroth order moment connects the phase space density $f\left(\mathbf{r},\mathbf{p},t\right)$ with the local mass density $\rho\left(\mathbf{r},t\right)$:

$$\int \mathrm{d}^3\mathbf{p}\, m f\left(\mathbf{r},\mathbf{p},t\right) \equiv \rho\left(\mathbf{r},t\right) \tag{2.29}$$

and the first momentum moment defines the peculiar velocity flow $\mathbf{v}\left(\mathbf{r},t\right)$:

$$\int \mathrm{d}^3\mathbf{p}\, \mathbf{p} f\left(\mathbf{r},\mathbf{p},t\right) \equiv \rho\left(\mathbf{r},t\right)\mathbf{v}\left(\mathbf{r},t\right). \tag{2.30}$$

The equations describing the evolution of $\rho\left(\mathbf{r},t\right)$ and $\mathbf{v}\left(\mathbf{r},t\right)$ are derived from taking moments of the Vlasov equation (eq. 2.28). The zeroth order of the Vlasov equation gives the continuity equation:

$$\frac{\partial\rho\left(\mathbf{r},t\right)}{\partial t} + \nabla\left[\rho\left(\mathbf{r},t\right)\mathbf{v}\left(\mathbf{r},t\right)\right] = 0, \tag{2.31}$$

which describes conservation of mass. The first moment of the Vlasov equation leads to the Euler equation, which describes the conservation of momentum,

$$\frac{\partial\mathbf{v}\left(\mathbf{r},t\right)}{\partial t} + \left[\mathbf{v}\left(\mathbf{r},t\right)\cdot\nabla\right]\mathbf{v}\left(\mathbf{r},t\right) + \nabla\Phi\left(\mathbf{r},t\right) = 0. \tag{2.32}$$

The second moment is related to the stress tensor. However, the fluid approximation is usually truncated at the first order moment. This is only valid for the early stages of the structure formation, before velocity dispersions are generated.

The Poisson equation links the Newtonian potential with the mass density $\rho\left(\mathbf{r},t\right)$ as

$$\nabla^2\Phi\left(\mathbf{r},t\right) = 4\pi G\rho\left(\mathbf{r},t\right). \tag{2.33}$$

Although there is no general analytic solution for the fluid dynamics of CDM, there are several approximate solutions derived from perturbation theory, which are described in the following sections.

Eulerian linear perturbation theory

To study the linear growth of structures, the density and velocity can be expanded around the background density $\bar{\rho}\left(t\right) = \Omega_m \rho_{\mathrm{crit}} a\left(t\right)^{-3}$ and Hubble flow[2] $\mathbf{v}\left(t\right)$:

$$\rho\left(\mathbf{r}, t\right) = \bar{\rho}\left(t\right) + \delta\rho\left(\mathbf{r}, t\right), \tag{2.34}$$

$$\mathbf{v}\left(\mathbf{r}, t\right) = \bar{\mathbf{v}}\left(t\right) + \delta\mathbf{v}\left(\mathbf{r}, t\right), \tag{2.35}$$

$$\Phi\left(\mathbf{r}, t\right) = \bar{\Phi}\left(t\right) + \delta\Phi\left(\mathbf{r}, t\right). \tag{2.36}$$

By introducing comoving spatial coordinates $(\mathbf{x}, \mathbf{u})$, the background density $\bar{\rho}\left(t\right)$ becomes independent of time. The transformation to comoving coordinates reads

$$\mathbf{r}\left(t\right) = a\left(t\right)\mathbf{x}\left(t\right), \tag{2.37}$$

$$\delta\mathbf{v}\left(t\right) = a\left(t\right)\mathbf{u}\left(t\right). \tag{2.38}$$

A solution of these equations is the linear growth factor

$$D_+\left(a\right) = \frac{\delta\left(\mathbf{x}, a\right)}{\delta\left(\mathbf{x}, 1\right)}. \tag{2.39}$$

The perturbations can be defined in terms of the density contrast $\delta\left(\mathbf{r}, t\right) \equiv \delta\rho\left(\mathbf{r}, t\right)/\bar{\rho}\left(t\right)$. Then, the previous equations (2.31) and (2.32) are

$$\frac{\mathrm{d}\delta\left(\mathbf{x}, t\right)}{\mathrm{d}t} = -\nabla_x \mathbf{u}\left(\mathbf{x}, t\right), \tag{2.40}$$

$$\frac{\mathrm{d}\mathbf{u}\left(\mathbf{x}, t\right)}{\mathrm{d}t} + 2\frac{\dot{a}}{a}\mathbf{u}\left(\mathbf{x}, t\right) = 4\pi G\bar{\rho}\delta\left(\mathbf{x}, t\right). \tag{2.41}$$

Eliminating $\mathbf{u}\left(\mathbf{r}, t\right)$, the amplitude of the linear density perturbations is given by

$$\frac{\mathrm{d}^2\delta\left(\mathbf{x}, t\right)}{\mathrm{d}t^2} + 2\frac{\dot{a}}{a}\frac{\mathrm{d}\delta\left(\mathbf{x}, t\right)}{\mathrm{d}t} = 4\pi G\bar{\rho}\delta\left(\mathbf{x}, t\right). \tag{2.42}$$

In the linear regime, where $|\delta| \ll 1$, the growth of the perturbations depends on the component that dominates the dynamics. As long as $\Omega_m\left(a\right) \approx 1$, the density contrast evolves like $\delta\left(a\right) \propto a^{3\omega+1}$ with $\omega = 1/3$ for the radiation-dominated era and $\omega = 0$ for the matter-dominated era (see e.g. Peacock, 1999).

Lagrangian perturbation theory

In the fluid approximation described above, the motion is described in Eulerian coordinates. However, the Lagrangian coordinate system allows developing a non-linear perturbation theory, following the trajectories of individual particles instead of studying the dynamics of density and velocity fields.

[2] The Hubble flow is the motion due solely to the cosmic expansion.

In the Lagrangian description of fluids, we are not interested in the position of particles but the displacement field $\Psi(\mathbf{q})$. This field maps the initial comoving position of a particle $\mathbf{q}$ into its comoving Eulerian coordinate $\mathbf{x}$:

$$\mathbf{x}(\mathbf{q}, t) \equiv \mathbf{q} + \boldsymbol{\Psi}(\mathbf{q}, t). \tag{2.43}$$

The Jacobian of the transformation between Lagrangian and Eulerian coordinates is derived by requiring the Lagrangian mass element to be conserved:

$$\rho(\mathbf{x}, t)\mathrm{d}^3\mathbf{x} = \rho(\mathbf{q})\mathrm{d}^3\mathbf{q} \to \bar{\rho}(t)[1+\delta(\mathbf{x}, t)]\mathrm{d}^3\mathbf{x} = \bar{\rho}(t)\mathrm{d}^3\mathbf{q} \tag{2.44}$$

Therefore, the Jacobian is

$$J(\mathbf{q}, t) = \frac{1}{1+\delta(\mathbf{x}, t)}. \tag{2.45}$$

This result is valid as long as the particle trajectories do not cross: when trajectories cross, fluid elements with different initial positions $\mathbf{q}$ end up at the same Eulerian positions $\mathbf{x}$. If there particle trajectories cross, often referred to as shell-crossing, the Jacobian vanishes, indicating a collapse to infinite density. Therefore, the description of dynamics as a mapping does not hold when shell-crossing occurs.

With the mapping of coordinates defined by eq. (2.43), the equation of motion (eq. 2.41) can be written in terms of the displacement field $\Psi(\mathbf{q})$. For this, we take the divergence of the equation and make use of the Poisson equation to obtain

$$J(\mathbf{q}, t)\nabla_{\mathbf{x}} \cdot \left[\frac{\partial^2 \boldsymbol{\Psi}}{\partial t} + \frac{\dot{a}}{a}\frac{\partial \boldsymbol{\Psi}}{\partial t}\right] = \frac{3}{2}\Omega_m(t)\left(\frac{\dot{a}}{a}\right)^2 [J(\mathbf{q}, t) - 1]\,. \tag{2.46}$$

Therefore, the Lagrangian equation of motion requires gradients with respect to the Eulerian coordinates $\mathbf{x}$, which are related to the shear tensor of the displacement field $\partial\boldsymbol{\Psi}/\partial\mathbf{q}$. This equation is then a non-linear differential equation for $\boldsymbol{\Psi}(\mathbf{q}, t)$ and can be solved perturbatively.

The first order of Lagrangian perturbation theory (LPT) is the Zel'dovich approximation (Zel'dovich, 1970), which consists of a linear solution of eq. (2.46). At linear order, $J(\mathbf{q}, t)\nabla_{\mathbf{x}} \approx \nabla_{\mathbf{q}}$. By defining $\psi \equiv \nabla_{\mathbf{q}} \cdot \boldsymbol{\Psi}$,

$$\psi'' + \frac{\dot{a}}{a}\psi' = \frac{3}{2}\Omega_m(t)\left(\frac{\dot{a}}{a}\right)^2 \psi \tag{2.47}$$

where $\psi' = \partial\psi/\partial t$. Then, the linear solution of this equation is given by

$$\psi^{(1)}(\mathbf{q}, t) = -D_+(t)\delta(\mathbf{q}), \tag{2.48}$$

where D_+ was defined in eq. (2.39).

The Zel'dovich approximation does not account for the gravitational interaction between particles and defines the trajectory as straight inertial motion in the direction of the initial velocity vector. Therefore, the Zel'dovich approximation describes a one-dimensional collapse that forms two-dimensional structures, named ”sheets” or ”pancakes”. This one-dimensional collapse would lead to an unphysical infinite density of the sheets. Therefore, its validity is restricted to the linear regime[3] and breaks down when gravitationally bound objects start to form. Therefore, after the formation of pancakes, the Zel’dovich approximation is not valid: particles falling into the pancakes will oscillate rather than move along the directions of their initial velocities as predicted by the approximation. Nevertheless, the Zel’dovich approximation provides insights into the formation of the cosmic web at large scales.

To describe the departure of the large-scale matter distribution from the Gaussian initial conditions, we need to include the second-order terms of LPT. The second-order LPT provides an improvement over the Zel’dovich approximation by accounting for the non-local gravitational instability and introduces corrections due to gravitational tidal effects (Bernardeau et al., 2002). The correction introduced by second-order LPT reads

$$\mathbf{x}(\mathbf{q},t) = \mathbf{q} + \Psi^{(1)}(\mathbf{q},t) + \Psi^{(2)}(\mathbf{q},t) \tag{2.49}$$

with $\psi^{(1)}(\mathbf{q},t) = \nabla_{\mathbf{q}} \cdot \Psi^{(1)}(\mathbf{q},t) = -D_+(t)\delta(\mathbf{q})$ and the second-order solution is related to the tidal effects as

$$\psi^{(2)}(\mathbf{q},t) = \nabla_{\mathbf{q}} \cdot \Psi^{(2)}(\mathbf{q},t) = \frac{1}{2}\left(\frac{D_2}{D_+}\right)^2 \sum_{i\neq j}\left[\Psi^{(1)}_{i,j}\Psi^{(1)}_{j,j} - \Psi^{(1)}_{i,j}\Psi^{(1)}_{j,i}\right], \tag{2.50}$$

with $\Psi_{i,j} \equiv \partial\Psi_i/\partial\mathbf{q}_j$ and $D_2\,(a)$ is the second-order growth factor. For a flat universe with a cosmological constant,

$$D_2\,(a) \approx -\frac{3}{7}\left(D^+\,(a)\right)^2 \Omega_m^{-1/143}. \tag{2.51}$$

The second-order LPT recovers the filamentary structure of the cosmic web, accurately describing one- two- and three-point correlation functions of the matter distribution and representing the higher order statistics (Moutarde et al., 1991; Buchert et al., 1994; Bouchet et al., 1995; Scoccimarro, 2000; Leclercq et al., 2013).

Numerical simulations

In the course of structure formation, high-density objects such as galaxies and clusters are formed. In this regimes of $|\delta| \gg 1$, perturbation theory does not provide a good description of the dynamics. The non-linear dynamics introduce strong couplings between different modes $\delta\,(\mathbf{k})$ in Fourier space and produce non-Gaussian features due to phase correlations. To study the non-linear stages of structure formation, one relies on N-body simulations that solve the Vlasov equation.

In an N-body simulation, the density field is represented by the sum of a set of discrete particles. The equations of motion describing the trajectory of each particle are solved

[3] When a perturbation enters the non-linear regime, it detaches from the Hubble flow and starts to collapse.

considering the gravitational forces due to the interaction with the rest of the particles. This provides the new position and velocity of each particle after a small time-step.

Although there are different methods to perform N-body simulations (e.g. particle-particle, particle-mesh, nested grid particle-mesh), we focus on the particle-mesh method since this is the solution used in the BORG framework (Jasche and Lavaux, 2018). In this method, the gravitational potential and density field are computed on a regular grid. More specifically, the dark matter particles move according to the equation of motion for a small time-step. After each time-step, the gravitational potential is re-computed by assigning particles to the grid and the equations of motion are updated with the corresponding gravitational forces. In the BORG framework, the density and gravitational fields are computed using a cloud-in-cell scheme (Appendix B, Hockney and Eastwood, 1988) to assign the particles to the grid. More specifically, the initial density field is populated with dark matter particles that evolve according to the following equations (Jasche and Lavaux, 2018):

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}a} = \frac{\mathbf{p}}{\dot{a}a^2}, \tag{2.52}$$

where $\mathbf{x}$ and $\mathbf{p} = a^2\dot{\mathbf{x}}$ are the positions and momenta of the dark matter particles. Then, momenta are updated for the new distribution of particles according to

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}a} = -\frac{\nabla_{\mathbf{x}}\phi}{aH\left(a\right)}, \tag{2.53}$$

where the gravitational potential is given by the Poisson equation as

$$\nabla^2_{\mathbf{x}}\phi = \frac{3}{2}H_0^2\Omega_{m,0}\frac{\delta\left(\mathbf{x}\right)}{a}. \tag{2.54}$$

The numerical integration of the equations of motion in eq. (2.52) is done through a leap-frog integrator (see Appendix A).

Chapter 3

Tracers of the large-scale matter distribution

The previous chapters presented the standard cosmological model and structure-formation models. This chapter discusses suitable observables to test these models. Particularly, studying the large-scale structure requires observational probes of the cosmic density field. Most of the analyses of the matter distribution have been based on galaxies since they are abundant and luminous enough to be observed at large distances. The first galaxy survey was performed by the Center of Astrophysics in 1985 (Huchra et al., 1988), mapping the position of 1100 galaxies. Since then, the amount of available galaxy surveys has increased rapidly: the Sloan Digital Sky Survey (SDSS Eisenstein et al., 2011) provided the position of almost a 930000 galaxies in 2008, SDSS III (Alam et al., 2017) had observed 1.2 million of galaxies in 2016, and the Dark Energy Spectroscopic Instrument (DESI Levi et al., 2013; DESI Collaboration et al., 2016) will map 35 million galaxies. Besides the increase in the amount of data, the next generation of surveys (LSST and Euclid) aim at mapping the galaxy distribution out to $z \approx 3$ (LSST Science Collaboration et al., 2009; Racca et al., 2016). However, the analysis of galaxy surveys requires assumptions on the relation between galaxy and matter distributions, known as the bias model (Bardeen et al., 1986; Cole and Kaiser, 1989; Peacock and Smith, 2000; Seljak, 2000; Desjacques et al., 2018). This link is not yet well understood since it involves complex galaxy formation physics. In addition, different kinds of galaxies present different distributions, indicating that galaxies with different properties trace different density regimes. For this reason, other probes of the matter density have been used.

The distribution of peculiar velocities of galaxies has also been used to trace the matter density field (Bertschinger et al., 1990; Nusser and Dekel, 1992; Dekel et al., 1999; Frisch et al., 2002; Brenier et al., 2003; Mohayaee and Sobolevskiĭ, 2008; Lavaux, 2008; Kitaura et al., 2012a). Peculiar velocities are produced by gravitational instabilities and, therefore, they should be directly connected to the mass distribution. Spectra are used to measure the velocity of galaxies. However, the observed velocity is the combination of the Hubble flow and the peculiar velocity. Separating these two components requires to know the distance to the galaxy. This can be done by different distance indicators such as the Tully-Fisher relation (Tully and Fisher, 1977) or the diameter-velocity dispersion in elliptical galaxies (Gregg, 1995).

Another probe of the matter distribution is gravitational lensing (see e.g. Hoekstra, 2005; Simon et al., 2009; Alsing et al., 2016, 2017). Photon propagation is affected by the gravitational potential fluctuations along the line of sight, distorting the image of background sources. From the distortion pattern, one can infer the mass distribution along the line of sight. Applying this to a population of galaxies provides the cosmic shear, which is directly related to the underlying matter field (Mo et al., 2010). However, this requires to measure the distortion and magnification of the sources compared to their relative unlensed images. Since galaxies are not intrinsically spherical, the intrinsic ellipticity has to be taken into account when estimating the cosmic shear. By assuming that the intrinsic ellipticities are uncorrelated, one can average galaxy images to estimate the local shear (Mo et al., 2010).

All these tracers of the matter distribution rely on galaxy observations. While galaxies mostly trace the over-dense regions of the Universe, the low-density regions have the potential to provide relevant information about dark energy and neutrino masses (see e.g. Hamaus et al., 2016). The non-linear effects are mitigated in voids and, therefore, voids are sensitives to the diffuse components of the Universe, such as dark energy. The study of voids can discriminate between homogeneous dark energy and modifications of gravity that would have an impact on non-linear structures. A powerful probe of the under-dense regions of the Universe is the Lyman-α (Ly-α) forest, which consists of absorption lines in the spectra of distant quasars. These absorption lines are generated by the scattering of photons due to the neutral hydrogen in the intergalactic medium (IGM), which also follows the dynamical evolution of the cosmic web (Peirani et al., 2014; Sorini et al., 2016). Besides providing complementary information to galaxy surveys, the Ly-α forest is observed at high-redshifts, where the current galaxy surveys are too sparse to trace the matter distribution. Finally, the Ly-α forest is observed with a higher spatial distribution that can be achieved with galaxy sampling (see e.g. Lee et al., 2018), probing scales down to few megaparsecs, which are sensitive to neutrino masses and dark matter models (Viel et al., 2013; Rossi, 2014; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Yèche et al., 2017).

In this chapter, galaxy redshift surveys and Ly-α forest datasets are presented. More specifically, their observational effects that need to be modelled are discussed.

3.1 Galaxy redshift surveys

Galaxies form due to the condensation of baryonic matter in gravitational potential wells defined by the dark matter distribution. Although the relation between the galaxy distribution and underlying matter density is not completely understood yet (the bias problem, Bardeen et al., 1986; Cole and Kaiser, 1989; Peacock and Smith, 2000; Seljak, 2000; Desjacques et al., 2018), the study of galaxy clustering has been very successful at providing information about the cosmic mass distribution (Colombi et al., 1996; Huterer et al., 2015; Basilakos and Nesseris, 2016; Frenk and White, 2012; Boyle and Komatsu, 2018; Mishra-

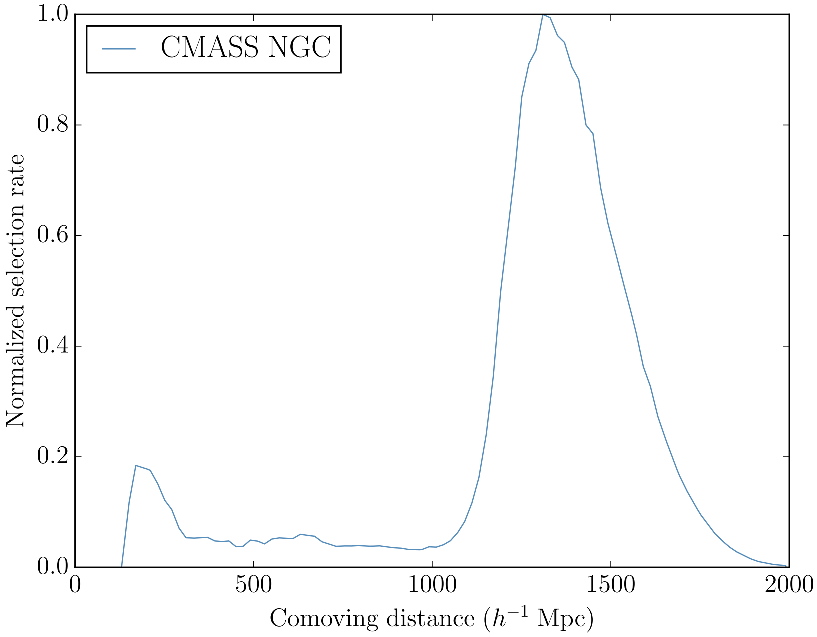


Figure 3.1: Radial selection function for the CMASS (north galactic cap) survey. The radial selection function encodes the probability that a galaxy at a redshift z is included in the survey. This function is defined by the observational criteria of the survey and needs to be accounted for in the data analysis.

Sharma et al., 2018).

The first galaxy surveys led to the discovery of cosmic structures such as filaments and voids (Gregory and Thompson, 1978; Gregory et al., 1981; Kirshner et al., 1981; Zeldovich et al., 1982). Consecutive galaxy surveys showed the hierarchical organization of galaxies, tracing the cosmic web (Geller, 1990; Schuecker and Ott, 1991; Shectman et al., 1996; Vettolani et al., 1997). Later, massive surveys of galaxies were performed in large areas of the sky and provided the redshifts of galaxies, measured from their spectra (2dF: Colless et al. 2003, SDSS: Eisenstein et al. 2011, 2MASS: Skrutskie et al. 2006, eBOSS: Dawson et al. 2016). In the upcoming years, photometric surveys aim at mapping the galaxy distribution at higher redshifts $z \approx 3$ (LSST, Ivezic et al. 2008; Euclid, Laureijs et al. 2011). To extract the information from these data, the treatment of several observational and systematic effects has to be included in the analysis.

First, the observational criteria of the survey need to be taken into account. Galaxy surveys provide a non-uniform sampling of the galaxy distribution in a volume defined by the boundaries of the survey. The survey strategy introduces selection effects such as cutoff in magnitude of distant galaxies. Modelling its effect requires a radial selection function, which accounts for the fact that the amount of observed galaxies with a given luminosity decreases with distance. The selection function, therefore, defines the probability that a random galaxy at a certain distance is included in the catalogue. Figure 3.1 shows an

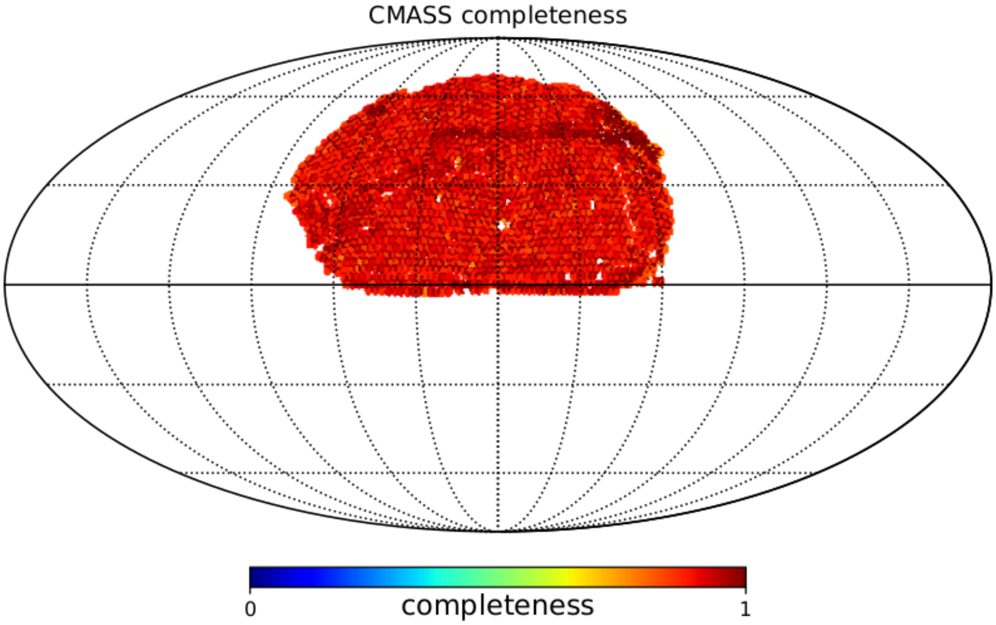


Figure 3.2: Completeness mask for the north galactic cap of the CMASS sample. The completeness mask is computed as the ratio of spectroscopic samples over the number of photometric objects, indicating the geometry of the survey and the completeness of the spectroscopic survey in each observed area of the sky. Therefore, the selection function informs about the observational criteria of the survey and needs to be included in the data analysis.

example of radial selection function for the CMASS sample. Besides the radial selection function, one needs to consider the angular selection function or geometry of the survey, which informs about the regions of the sky that have been observed. The angular selection function is combined with the completeness mask, which provides information on which areas of the sky have been observed in more detail. More specifically, the completeness mask is computed as the ratio between the spectroscopic observations and a previous photometric catalogue used to select the spectroscopic targets (see e.g. Anderson et al., 2012). An example of a completeness mask is shown in Fig. 3.2.

Besides the observational criteria of the survey, the data are affected by foreground and target contamination. Among those are the galactic dust contamination or stellar contamination. In the past, these effects have been modelled as templates. However, this approach requires a good understanding of the contamination. Future galaxy surveys (LSST and Euclid) will be affected by yet unknown contaminations. If not accounted for, these contaminations introduce spurious and erroneous effects in the matter distribution inferred from the data. For this reason, I developed a likelihood that effectively deals with contaminated datasets (see Chapter 7).

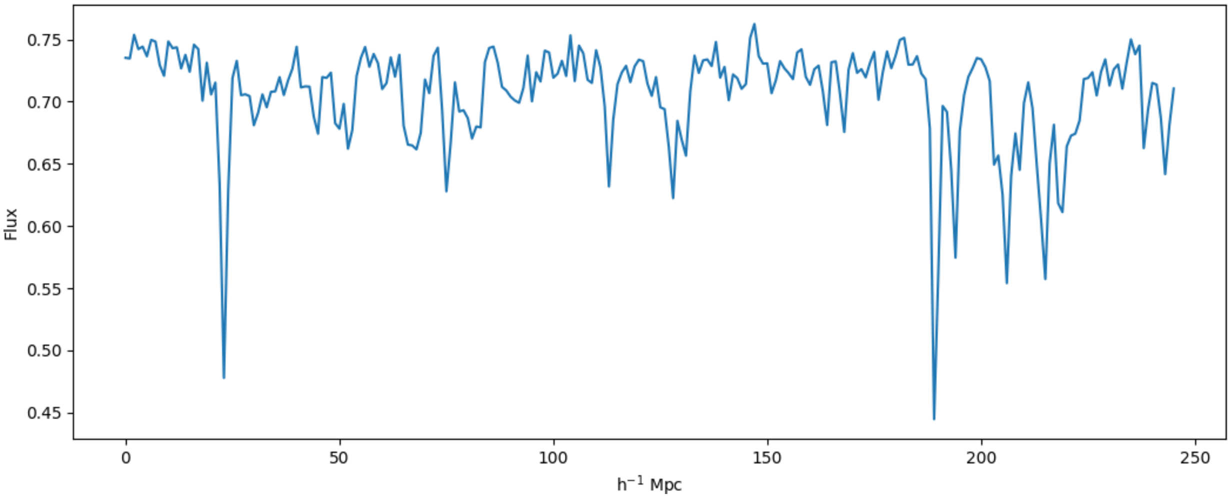


Figure 3.3: Example of Lyman-α forest spectrum from the CLAMATO mock survey. The flux F is the transmitted flux fraction ($F = F_{\text{transmitted}}/F_{\text{total}}$). The absorption lines trace the matter distribution along the line of sight: higher-density regions absorb the quasar flux, producing absorption lines, while low-density regions are transparent and transmit a large fraction of the flux.

3.2 The Lyman-α forest

As mentioned above, the Lyman-α (Ly-α) forest provides complementary information to galaxy surveys. The absorption lines that constitute the Ly-α forest are generated due to the scattering of quasar photons by neutral hydrogen (HI). Since hydrogen is neutral in low-density regions, the Ly-α forest traces the under-dense regions of the Universe (Peirani et al., 2014). More specifically, the Ly-α forest arises from regions where the density of matter is within a factor ten of the cosmic mean density (Peirani et al., 2014; Sorini et al., 2016).

The emission of quasars (or QSO) is characterized by a strong emission in the Ly-α line, corresponding to a wavelength of 1216 Å. The photons emitted by the quasar are redshifted due to the cosmic expansion. While traversing the Universe, the radiation that lies bluewards of the Ly-α line is shifted to 1216 Å at some redshift z. At that z, the photons at 1216 Å can be absorbed by the neutral hydrogen. Since HI clouds at different redshifts see the photons at different wavelengths, each cloud leaves a fingerprint as an absorption line. The wavelength and absorption of each line then trace the matter distribution along the line of sight.

The first studies of this feature in the QSO spectrum (Gunn and Peterson, 1965; Scheuer, 1965) focused on probing the existence of HI in the intergalactic medium (IGM). These tests, proposed by Gunn and Peterson (Gunn and Peterson, 1965), relied on the large cross-section of the Ly-α absorption by neutral hydrogen: if the Universe were filled with a diffuse distribution of neutral hydrogen, we would observe a significant trough of flux in the QSO spectrum at $\lambda < (1 + z_Q)\lambda_\alpha$. Although Gunn and Peterson detected the absorption trough in the QSO spectra, the light of the QSO was not completely absorbed. This led to the conclusion that the IGM is highly ionized, indicating that the Universe underwent a reionization phase after the recombination.

Observations of QSO with higher spectral resolution showed that the trough observed by Gunn and Peterson was actually a 'forest' of hundreds of absorption lines (Lynds, 1971). The reason for this forest is that the neutral hydrogen is not homogeneously distributed but forms discrete systems. Hydrodynamical simulations (Cen et al., 1994; Hernquist et al., 1996; Miralda-Escudé et al., 1995) showed that the HI distribution traces the dark matter distribution. On scales larger than 1 Mpc, the thermal pressure of the HI is negligible (Peeples et al., 2010) and the HI traces the dark matter distribution more closely than the galaxy distribution. The Ly-α forest is, therefore, a powerful tool to study the matter distribution.

The high resolution of these data requires an accurate treatment of the systematic effects. Most of the previous approaches to extract cosmological information from the Ly-α forest are based on the study of the power spectrum (Seljak et al., 2006; Viel et al., 2006; Slosar et al., 2011; Busca et al., 2013; Bautista et al., 2017; Blomqvist et al., 2019; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Yèche et al., 2017). However, harvesting the complete information content of the matter distribution requires high-order statistics to capture the filamentary structure of the cosmic web. To go beyond previous approaches, I developed a fully Bayesian framework to infer the three-dimensional large-scale structures from the Ly-α forest (see Chapter 8). This framework jointly infers the matter distribution and the astrophysical parameters that model the HI properties. While the value of these parameters is still a matter of debate, this is the first approach to jointly infer the parameters and the matter distribution.

Chapter 4

Statistical data analysis

Cosmology is an entirely observational science: unlike other branches of physics, we cannot prepare laboratory experiments in Earth-based laboratories, where one can control and abstract unwanted effects. In cosmology, we have to make sense of observations subject to noise and uncontrollable systematic effects such as foreground contaminations. Therefore, cosmology relies very heavily on statistics since it requires a rigorous statistical analysis with accurate treatment of noise and observational effects. For this reason, this chapter presents the key concepts of statistical data analysis as relevant for this thesis. Part of this chapter focuses on Markov Chain Monte Carlo methods since they are one of the core elements of the BORG algorithm, which I applied and extended in this work.

4.1 Probability theory

There are two approaches to statistical data analysis: frequentist and Bayesian. The difference between these two approaches is the way they interpret probability: probabilities can be described as frequencies of outcomes in experiments or as certainty in propositions that do not involve random variables. While frequentists make statements on the data based on a fixed model, Bayesians make statements on the model given a fixed data. In frequentist statistics, the measurement outcome is uncertain. Whereas for Bayesian statistics, the measurement outcome is certain since it is observed, but the knowledge about the state of the world is uncertain.

Summarising, in frequency theory, a probability represents the percentage of times that something has happened in a very large number of identical repeats of an experiment. In Bayesian statistics, a probability distribution encodes our knowledge about some model parameter or set of competing theories, based on our current state of information. Therefore, frequentist statistics formulates the problems in terms of building estimators for data summaries while Bayesian statistics informs about the certainty of the model given some data. Hence, Bayesian statistics addresses the problem of inferring the true value given all the available information, including the data.

Bayesian and frequentist statistics derive from probability calculus. One of the laws of probability calculus is the Bayes' theorem, which describes the probability that an event A occurs once event B already occurred:

$$P(A|B)P(B) = P(B|A)\,P(A) = P(A,B)\,, \tag{4.1}$$

where $P(A,B)$ denotes the probability of "A and B". If the two events are independent, we can write

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B)\,. \tag{4.2}$$

Therefore, in the case of several events B_j of which only one will occur, the probability for event $P(A)$ can be obtain by marginalization over all possible B_j:

$$P(A) = \sum_j P(A|B_j)\,P(B_j)\,. \tag{4.3}$$

Both approaches, frequentist and Bayesian, can be expressed mathematically using the Bayes' theorem.

4.2 Bayes' theorem for data analysis

Here we illustrate the application of Bayes' theorem to data analysis, in particular, to cosmological datasets. Since we are interested in inferring the 3d matter distribution, the density field will be the signal s. However, extracting this information from the data d requires modelling the noise and systematic effects. These systematic effects can be encoded in a set of parameters θ, corresponding to parameters from the bias model, the characterisation of the noise distribution or astrophysical quantities that affect the data. The Bayes' theorem then can be written as

$$P(s|\theta,d) = \frac{P(d|\theta,s)\,P(s)}{P(d)}. \tag{4.4}$$

$P(d|\theta,s)$ is the likelihood, which encodes the data model that informs on how the data were generated. $P(s)$ is the prior, encoding our understanding of the Universe. $P(d)$ is a normalization constant, often referred to as 'evidence'. Finally, we are interested in measuring the posterior distribution $P(s|\theta,d)$. The posterior distribution allows deriving the mean and variance of the inference results. The variance of the posterior distribution contains relevant information since it quantifies the uncertainty in the inference results. For this reason, the BORG algorithm employed in this thesis provides the full posterior distribution.

For non-extreme priors, the repeated application of the Bayes' theorem will converge to a unique posterior distribution (Bernstein-von Mises theorem). If the data is more informative than the prior, the posterior distribution will converge to the same result despite the prior choice. In the case that the data are not informative enough to overrule the prior, one needs to assess how much of the final inference depends on the prior choice. One way to perform this test is to generate simulated data from the posterior distribution and compare it to the observed data (posterior predictive test).

4.2.1 Chosing the prior

While Bayesian and frequentist statistics agree on the Bayes' theorem, they argue about the use of the prior. Although there is some guidance on selecting the prior, there is no prescription to specify it. However, the prior specification is not a limitation of Bayesian statistics and does not undermine objectivity. Specifying the prior exposes assumptions to criticism and scientific discussion; therefore, Bayesian statistics becomes a systematic way to quantify assumptions.

In cosmology, theory (Guth and Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983) and observational data (Planck Collaboration et al., 2018, 2019) allowed to establish a well-supported model, the ΛCDM model. Therefore, cosmology is an ideal case for Bayesian statistics since we have a well-motivated prior on the statistical distribution of primordial matter fluctuations from which today's structures developed. This prior knowledge is supported by theory as well as observations of the Planck satellite mission (Planck Collaboration et al., 2018, 2019). It is then well founded to incorporate this prior knowledge in our analyses. In particular, the BORG framework incorporates the Gaussian prior for initial conditions to infer the cosmic matter distribution (see Chapter 5 for more details).

4.3 Markov Chain Monte Carlo methods

High-dimensional problems pose a challenge to interpret the posterior distribution. Direct visualisation of the distribution is only possible for low-dimensional parameter spaces. Statistical summaries (mean, median) require integrating out the rest of the parameters, which is usually not possible analytically and numerically not feasible for high-dimensional parameter spaces. Therefore, the evaluation of the posterior distribution in high-dimensional parameter problems requires numerical approximations such as sampling.

The real posterior distribution $P\left(s|\theta,d\right)$ can be estimated by drawing samples from it. Therefore, the real posterior distribution can be written as a sum of Dirac deltas $\delta^D\left(s\right)$ (Andrieu et al., 2003) as

$$P\left(s|\theta,d\right) \approx P_N\left(s|\theta,d\right) = \frac{1}{N}\sum_{i=1}^{N}\delta^D\left(s-s_i\right). \tag{4.5}$$

In the way this is constructed, the posterior distribution is proportional to the local density of samples in the parameter space (Gregory, 2010). Each sample is a possible version of the truth and the variation between them characterizes the standard deviation of the posterior distribution, which allows to quantify the uncertainty in the results.

Additionally, the sampling scheme also simplifies the calculation of statistical summaries. Marginalization is achieved through histograms of the samples: it is sufficient to count the number of samples in different bins of a parameter, ignoring the rest of the parameters. As an example, the expectation value of a function of the parameter $f\left(\theta\right)$ is given by

$$\langle f\left(\theta\right)\rangle = \int f\left(\theta\right) P\left(\theta\right) d\sigma \approx \frac{1}{N}\sum_{i=1}^{N} f\left(\theta_i\right) \tag{4.6}$$

where i labels the samples. Obtaining means and variances, therefore, requires only discrete sums instead of integrals.

Now we need to define a method to sample the posterior distribution. A powerful sampling method is the Monte Carlo Markov Chain (MCMC Neal, 1993). A Markov chain is a process that generates a sequence of random variables in such a way that the probability distribution of the next step $i+1$ only depends on the current step i. Therefore, the probability distribution at one step is conditionally independent of all previous steps. This can be expressed mathematically by defining a transition probability $T\left(s^{(i+1)}|s^{(i)}\right)$, which encodes the probability of jumping from the current state i to the following state $i+1$. Then, the probability distribution of a Markov chain can be written as

$$P\left(s^{(i+1)}|s^{(0)}, s^{(1)}, ..., s^{(i)}\right) = P\left(s^{(i+1)}|s^{(i)}\right) = T\left(s^{(i+1)}|s^{(i)}\right). \tag{4.7}$$

By marginalizing, we can obtain the probability of a state $s^{(i+1)}$:

$$P\left(s^{(i+1)}\right) = \sum_{s^{(i)}} P\left(s^{(i)}\right) T\left(s^{(i+1)}|s^{(i)}\right). \tag{4.8}$$

When the transition probability does not depend on i, it is called stationary transition probability.

A Monte Carlo Markov Chain (MCMC) method is constructed to provide a sequence of points in the parameter space with a density proportional to the target distribution (Gregory, 2010). After a certain number of steps, known as the warm-up phase, the MCMC reaches a state where successive elements of the chain are drawn from the high-density regions of the target distribution. The number of steps required by the warm-up phase depends on the initialisation of the MCMC. Once the high-probability regions are reached, the MCMC reconstructs the probability distribution heuristically.

MCMC algorithms explore the parameter space in a random walk such that the probability for being in a certain region of the parameter space is proportional to the posterior density in that region (Andrieu et al., 2003). Therefore, the target probability can be estimated as

$$P\left(s\right) = \frac{1}{N}\sum_{i=1}^{N} \delta^D\left(s - s^{(i)}\right), \tag{4.9}$$

where $\delta^D\left(s\right)$ is the Dirac delta. Then statistical summary quantities, like the mean, can be approximated with sums

$$I_N\left(f\right) = \frac{1}{N}\sum_{i=1}^{N} f\left(s^{(i)}\right) \xrightarrow{N\to\infty} I\left(f\right) = \int f\left(s\right) P\left(s\right) \mathrm{d}s \tag{4.10}$$

indicating that the estimate $I_N(f)$ is unbiased and, therefore, it will converge to $I(f)$ by the law of large numbers (Andrieu et al., 2003).

The efficiency of MCMC methods stems from the distribution of samples: by providing a high density of samples in high-probability regions and sparse sampling in low probability regions, MCMC methods adapt to the geometry of the problem and focus on regions with a larger contribution. MCMC methods have to fulfill the following properties (Mackay, 2003):

- The target distribution $P(s)$ is an invariant distribution of the chain:

$$P\left(s^{(i+1)}\right) = \int \mathrm{d}^N s^{(i)}\, T\left(s^{(i+1)}|s^{(i)}\right) P\left(s^{(i)}\right). \tag{4.11}$$

- The chain must be ergodic: all the states can be reached from each other (not necessarily in one move).

- The chain does not have a periodic set: the chain cannot oscillate between different states in a regular periodic movement.

The MCMC approach seeks the stationary distribution to be the target distribution: after an initial warm-up phase, all the samples are obtained from the real posterior distribution. This can be achieved by requiring detailed balance: the transition from $s^{(i)}$ to $s^{(i+1)}$ must be equally likely as the transition from $\mathbf{s}$ to $\mathbf{s}'$

$$T\left(s^{(i+1)}|s^{(i)}\right) P\left(s^{(i)}\right) = T\left(s^{(i)}|s^{(i+1)}\right) P\left(s^{(i+1)}\right). \tag{4.12}$$

This implies that the transition from one state to another is reversible. The requirement of detailed balance implies invariance of $P(s)$ under the Markov chain and guarantees that the probability of the chain converges to $P(s)$ (Mackay, 2003). One of the major challenges for the MCMC approach is the construction of numerically efficient transition kernels $T\left(s^{(i+1)}|s^{(i)}\right)$.

4.4 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is an MCMC method developed by Metropolis et al. (1953) and later extended by Hastings (1970). One of the advantages of the Metropolis-Hastings method is that defining transition probability does not require knowing the shape of the target distribution. After obtaining a proposal for a new sample y, we need to decide whether to accept this new state. This is done on the basis of a ratio r called the Metropolis ratio,

$$r \equiv \frac{P\left(y|s^{(i)}\right) T\left(s^{(i)}|y\right)}{P\left(s^{(i)}|y\right) T\left(y|s^{(i)}\right)}. \tag{4.13}$$

If $r \geq 1$, we accept the sample and $s^{(i+1)} = y$. If $r < 1$, we accept it with a probability r. This is done by sampling a random variable U from a uniform distribution in the interval

0 to 1. If $U < r$, we accept the sample, i.e. $s^{(i+1)} = y$. Otherwise, the chain stays in the current state i.e. $s^{(i+1)} = s^{(i)}$. This is often summarized as an acceptance probability $\alpha\left(s^{(i)}, y\right)$:

$$\alpha\left(s^{(i)}, y\right) = \min\left(1, r\right) = \min\left(1, \frac{P\left(y|s^{(i)}\right) T\left(s^{(i)}|y\right)}{P\left(s^{(i)}|y\right) T\left(y|s^{(i)}\right)}\right). \tag{4.14}$$

where $\alpha\left(s^{(i)}, y\right)$ is the acceptance probability.

For the Metropolis-Hastings algorithm to converge to the target distribution $P\left(s^{(i)}|d,\theta\right)$, it has to be positive recurrent (Gregory, 2010). This means that there exists a stationary distribution $P\left(s\right)$ such that if an initial value is sampled from $P\left(s\right)$, then all subsequent iterates will be distributed according to $P\left(s\right)$. Following Gregory (2010), we will show that the stationary distribution of the Metropolis-Hastings algorithm is the target distribution. We start from a sample $s^{(i)}$ from the target distribution. The probability of drawing $s^{(i)}$ from the posterior distribution is $P\left(s^{(i)}|d,\theta\right)$. Then, by applying eq. (4.1), the joint probability $P\left(s^{(i)}, s^{(i+1)}|d,\theta\right)$ is given by

$$P\left(s^{(i)}, s^{(i+1)}|d,\theta\right) = P\left(s^{(i)}|d,\theta\right) P\left(s^{(i+1)}|d,\theta\right). \tag{4.15}$$

The probability of accepting a sample $s^{(i+1)}$ depends on the transition function $T\left(s^{(i+1)}|s^{(i)}\right)$,

$$P\left(s^{(i)}, s^{(i+1)}|d,\theta\right) = P\left(s^{(i)}|d,\theta\right) T\left(s^{(i+1)}|s^{(i)}\right) \alpha\left(s^{(i)}, s^{(i+1)}\right). \tag{4.16}$$

Using eq. (4.14),

$$\begin{aligned}
P\left(s^{(i)}, s^{(i+1)}\right) &= P\left(s^{(i)}|d,\theta\right) T\left(s^{(i+1)}|s^{(i)}\right) \\
&\times \min\left(1, \frac{P\left(s^{(i+1)}|s^{(i)}\right) T\left(s^{(i)}|s^{(i+1)}\right)}{P\left(s^{(i)}|s^{(i+1)}\right) T\left(s^{(i+1)}|s^{(i)}\right)}\right) && (4.17)\\
&= \min\left(P\left(s^{(i)}|d,\theta\right) T\left(s^{(i+1)}|s^{(i)}\right), P\left(s^{(i+1)}|s^{(i)}\right) T\left(s^{(i)}|s^{(i+1)}\right)\right) && (4.18)\\
&= P\left(s^{(i+1)}|d,\theta\right) T\left(s^{(i)}|s^{(i+1)}\right) \alpha\left(s^{(i+1)}, s^{(i)}\right) && (4.19)\\
&= P\left(s^{(i+1)}|d,\theta\right) P\left(s^{(i)}|d,\theta\right). && (4.20)
\end{aligned}$$

Therefore we have derived the detailed balance equation:

$$P\left(s^{(i)}|d,\theta\right) P\left(s^{(i+1)}|d,\theta\right) = P\left(s^{(i+1)}|d,\theta\right) P\left(s^{(i)}|d,\theta\right), \tag{4.21}$$

which means that the Markov chain generated by the Metropolis-Hastings algorithm converges to a stationary distribution. We will now show that there is a stationary distribution. This implies that once a sample from the stationary distribution is obtained, the following samples are drawn from the same distribution:

$$\begin{aligned}
\int P\left(s^{(i)}|d,\theta\right) P\left(s^{(i+1)}|s^{(i)}\right) \mathrm{d}s^{(i)} &= \int P\left(s^{(i+1)}|d,\theta\right) P\left(s^{(i)}|s^{(i+1)}\right) \mathrm{d}s^{(i)} && (4.22)\\
&= P\left(s^{(i+1)}|d,\theta\right) \int P\left(s^{(i)}|s^{(i+1)}\right) \mathrm{d}s^{(i+1)} && (4.23)\\
&= P\left(s^{(i+1)}|d,\theta\right). && (4.24)
\end{aligned}$$

Although the Metropolis-Hastings algorithm guarantees that the probability of the chain converges to the target distribution, it is not efficient in high-dimensional parameter spaces. As mention above, MCMC methods explore the parameter space in a random walk. If we choose large steps of the chain, it is likely to end in a state with low probability, compromising the progress by high rejection rates (Mackay, 2003). However, small steps in the random walk will require many iterations to produce effectively independent samples. Suppressing the random walk behaviour would significantly improve the efficiency of the sampler. This can be done with Hamiltonian Monte Carlo methods, which use gradient information to reduce the random walk.

4.5 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) methods are efficient MCMC algorithms in high-dimensional parameter spaces (Duane et al., 1987; Neal, 1993). For this reason, HMC is the MCMC method implemented in the BORG framework (Jasche and Wandelt, 2013). By using the information in the gradients, HMC algorithms suppress the random walk behaviour: the gradients indicate which direction the chain should move to find states with higher probability. Therefore, HMC algorithms explore the parameter space in a more efficient way than the standard Metropolis-Hastings algorithms.

To incorporate the gradient information, HMC makes use of Hamiltonian dynamics. The link between probability calculus and Hamiltonian dynamics is given by the canonical distribution. We can rewrite Bayes' theorem as

$$\begin{aligned} P\left(s|\theta,d\right) &= \frac{P(d|s,\theta)P(s)}{P(d)} && (4.25)\\ &= \frac{1}{P(d)}\exp\left[\ln\left(P(d|s,\theta)P(s)\right)\right]. && (4.26) \end{aligned}$$

This has the same form as the canonical distribution from statistical mechanics:

$$P\left(\mathbf{x},\mathbf{p}\right) = \frac{1}{Z_H}\exp\left[-\mathcal{H}\left(\mathbf{x},\mathbf{p}\right)\right]. \tag{4.27}$$

In the HMC approach, therefore, we interpret the negative logarithm of the target distribution as a physical potential $\Psi\left[P\left(\mathbf{x}\right)\right] = -\ln P\left(\mathbf{x}\right)$ and the evidence as the partition function $Z_H = P(d)$. A 'kinetic energy' $K\left(\mathbf{p}\right)$ is introduced in the Hamiltonian by defining momentum variables $\mathbf{p}$, which are then treated as nuisance parameters. The Hamiltonian then reads

$$\mathcal{H}\left(\mathbf{x},\mathbf{p}\right) = \Psi\left(\mathbf{x}\right) + K\left(\mathbf{p}\right). \tag{4.28}$$

More specifically, the kinetic energy is given by

$$K\left(\mathbf{p}\right) = \frac{1}{2}\mathbf{p}^{\dagger}\mathbf{M}\mathbf{p} \tag{4.29}$$

where $\mathbf{M}$ is called 'mass matrix'. This matrix characterizes the 'inertia' of the parameters when moving through the parameter space. Although it does not affect the accuracy of the sampler, it has an impact on its efficiency. Too large masses lead to slow exploration of the parameter space, while too low masses may result in large rejection rates (Neal, 1993) as the integration of particle trajectories becomes numerically unstable.

The method explores the augmented parameter space of $(\mathbf{x}, \mathbf{p})$ by alternating two types of proposals: first, a proposal of $\mathbf{p}$ randomizes the momentum variables without affecting $\mathbf{x}$, second, a proposal modifies $\mathbf{x}$ and $\mathbf{p}$. This second proposal is obtained from simulated dynamics by integrating the Hamiltonian equations

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}}, \qquad (4.30)$$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial \mathbf{x}}. \qquad (4.31)$$

The two alternated proposals of the HMC create samples of the joint density in eq. (4.27). Marginalizing over momentum variables, we obtain the desired distribution for $\mathbf{x}$.

Some important properties of the Hamiltonian dynamics are (Mackay, 2003)

- Hamiltonin dynamics are reversible in time.
- Hamiltonian dynamics do not violate detailed balance, meaning that they fulfill Liouville's theorem: the volume is conserved in phase space.
- The Hamiltonian is conserved in HMC methods:

  $$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial \mathbf{x}}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \frac{\partial \mathcal{H}}{\partial \mathbf{p}}\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial \mathbf{x}}\frac{\partial \mathcal{H}}{\partial \mathbf{p}} - \frac{\partial \mathcal{H}}{\partial \mathbf{p}}\frac{\partial \mathcal{H}}{\partial \mathbf{x}} = 0. \qquad (4.32)$$

  Therefore, the acceptance rate of the Metropolis-Hastings algorithm is always unity. Although, in practice, numerical errors can lead to a lower acceptance rate, HMC methods remain computationally much cheaper than standard Metropolis-Hastings algorithms by suppressing the random walk behaviour.

4.6 Slice sampler to sample the likelihood parameters

The BORG framework employs a HMC method to sample the matter distribution. However, modelling cosmological datasets requires to include other parameters related to the bias model, noise characterisation or astrophysical processes that imprint the data. Although these parameters can be considered mere nuisances for cosmological analysis, they need to be marginalised over in order to not bias the results. In the BORG framework, these model parameters are jointly sampled with the density distribution. While the HMC is used to infer the density field, these nuisance parameters are sampled via a slice sampler.

The slice sampler is an MCMC method that guarantees high acceptance rate (Neal, 2000) when the target distribution $P(x)$ can be evaluated at any point x. The slice sampler has unit acceptance rate and can perform global exploration for univariate distributions.

The slice sampler is based on the idea that one can sample from a distribution uniformly by looking only at the horizontal coordinates of the sampled points (Neal, 2000). Let's assume we want to sample a variable x with density function $p(x) \propto f(x)$. We can then obtain realizations of x by sampling uniformly from the region that lies under $f(x)$. This idea can be formalized by introducing a new variable y such that

$$p(x, y) = \begin{cases} 1/Z & \text{if } 0 < y < f(x) \\ 0 & \text{otherwise} \end{cases}$$

where $Z = \int f(x)\,\mathrm{d}x$. Marginalizing over y we obtain the distribution for x

$$p(x) = \int_0^{f(x)} \frac{1}{Z}\,\mathrm{d}y = \frac{f(x)}{Z}. \tag{4.33}$$

Therefore, the slice sampler provides a new sample x_1 from a previous x_0 in a three-step procedure (Neal, 2000), illustrated in Fig. 4.1:

- We choose a point x_0 and evaluate the function at that point, $f(x_0)$. Then, we draw a number y uniformly in $(0, f(x_0))$. This number y defines a horizontal slice (see Fig. 4.1). The regions of the slice y under the function $f(x)$ are identified (marked in bold) and the next step will only focus on these regions.

- An interval is located around x_0. One method to define the interval is the 'stepping-out' method (see (b) in Fig. 4.1): we place an interval randomly around x_0 with width w and expand it in steps of the same width until both ends of the interval are outside the slice (outside the bold areas). With this procedure, the slice sampler self-tunes the step size.

- A new proposal is drawn uniformly within this interval. If the proposal is accepted, the new value x_1 is used as a new sample and the next iteration starts. If the proposed value is rejected (it falls outside the bold areas), the interval is shrunk and the slice sampler draws a new proposal. Therefore, the rejected proposals are used to shrink the interval. This is an advantage of the slice sampler over the standard Metropolis-Hastings algorithm: while the efficiency of the Metropolis-Hastings depends on our choice of the step size, the slice sampler automatically shrinks the interval each time a proposal is rejected.

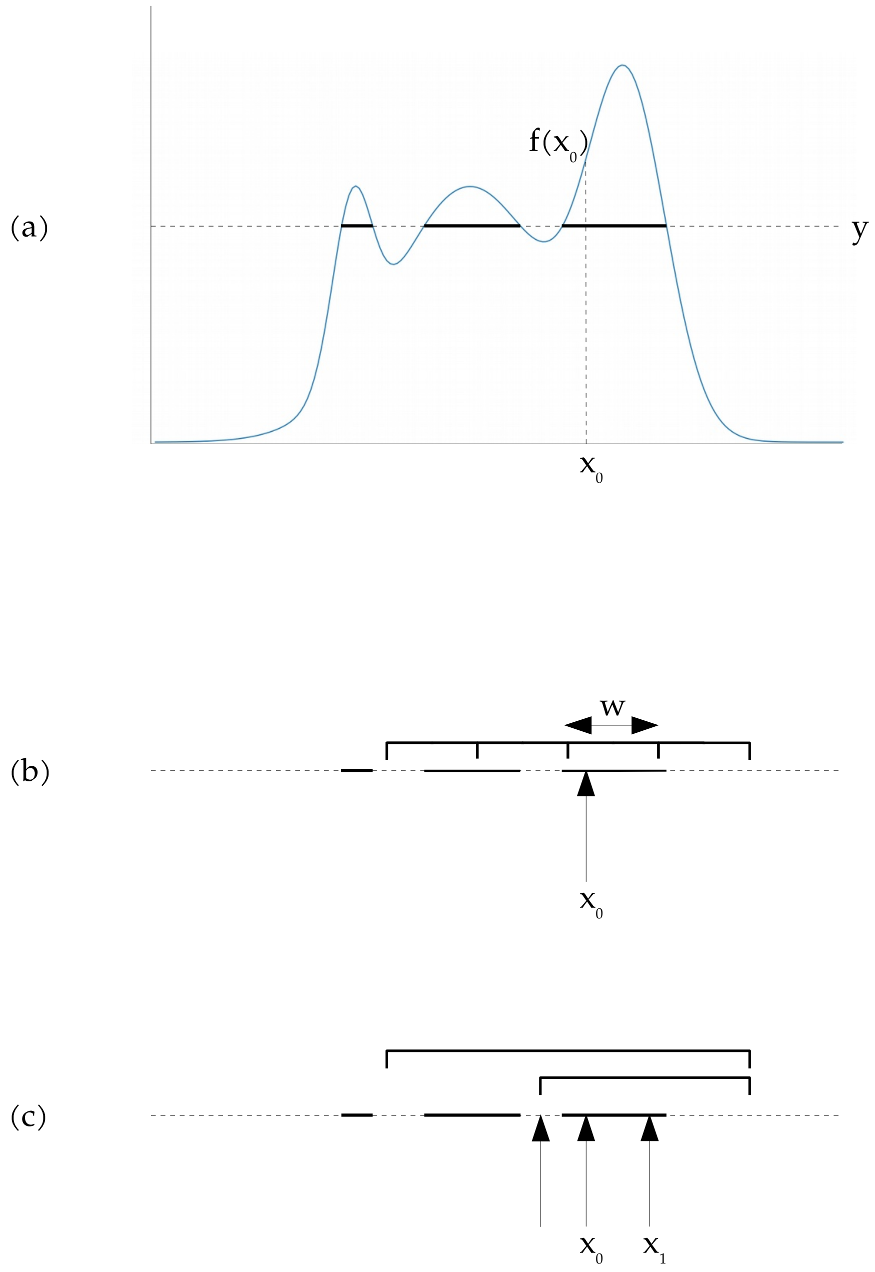


Figure 4.1: Single-variable slice sampler with the stepping-out and shrinkage method. (a) A vertical slice y is defined by drawing uniformally in $(0, f(x_0))$. This defines a horizontal slice indicated in bold. (b) An interval with width w is randomly placed around x_0. This interval is expanded in steps of the same width w until both ends are outside the bold slice. (c) A new point x_1 is found by drawing samples uniformly from the interval until the sample falls inside the slice. The samples outside the slice are used to shrink the interval. (Neal, 2000).

Chapter 5

Physical forward modelling - the BORG framework

The study of the matter distribution has the potential to provide invaluable information about fundamental physics. For this reason, current and future cosmological surveys aim at mapping the cosmic web (see e.g. LSST Science Collaboration et al., 2009; Refregier et al., 2010; Laureijs et al., 2011; Alam et al., 2016). However, most of the studies of the matter distribution focus only on the analysis of the power spectrum (Tegmark et al., 2002; Szalay et al., 2003; Tegmark et al., 2004; Pope et al., 2004), which only contains information about the one- and two-point statistics. Since the non-linear dynamics transport information to higher-order statistics, harvesting the full information content of the data requires to study the three-dimensional density field. In this context, Jasche and Wandelt (2013) presented the Bayesian Origin Reconstruction from Galaxies (BORG) algorithm. BORG is a fully Bayesian framework for large-scale structure inference from cosmological datasets, which can be equipped with different data models. In this thesis, I applied and extended the BORG algorithm. More specifically, I developed a likelihood to effectively deal with contaminated datasets (Chapter 7) and implemented it in BORG, making it more robust against foreground and target contaminations. Further, I extended the BORG algorithm to the analysis of the Ly-α forest, which allows studying the high-redshift Universe with higher accuracy than can be achieved with galaxy samples (see e.g. Lee et al., 2018). This chapter introduces the BORG algorithm, while the next chapters describe the extensions in more detail.

5.1 The conceptional idea of BORG

The original idea of the BORG algorithm originates from the problem of analysing the non-linear structure of the matter distribution underlying observed galaxies in cosmological surveys. As described in Chapter 2, present matter density fields exhibit a complex statistical structure due to non-Gaussian gravitational dynamics. It is clear that detailed modelling of the cosmic matter distribution requires to describe the high-order statistics corresponding to the filamentary structure of the cosmic web. At present, there exists no closed-form description of the non-linear density field in terms of a high-dimensional multivariate probability distribution. Although there are approximations to reproduce the statistical behaviour of the dark matter density field (e.g. log-normal distributions or multivariate Gaussians, Lahav et al. 1994; Zaroubi et al. 1999; Kitaura and Enßlin 2008; Kitaura et al. 2009; Jasche and Kitaura 2010), they only parametrise the one- and two-points statistics and fail to reproduce more complex structures such as filaments (Baugh et al., 1995; Peacock and Dodds, 1996; Smith et al., 2003). However, the initial density is well-described by a Gaussian distribution, as discussed in previous chapters. For this reason, Jasche and Wandelt (2013) proposed to translate the problem of inferring the matter distribution to the inference of the initial conditions. This is possible by implementing a dynamical model of structure formation that connects the initial and evolved density fields by deterministic gravitational evolution. This dynamical model naturally accounts for the high-order statistics of the density field, recovering the filamentary structure of the cosmic web (Jasche and Wandelt, 2013).

Inferring the matter distribution requires to infer the amplitudes of the primordial density field at each volume element of a regular grid, commonly between 256^3 and 512^3 volume elements. This implies 10^7 to 10^8 free parameters. Due to the high dimensionality of the problem, the BORG algorithm incorporates an HMC sampler to explore the parameter space efficiently. BORG is then a fully Bayesian algorithm since it provides the full posterior distribution instead of point estimates. From the posterior distribution, we can derive the non-linear and non-Gaussian uncertainties and propagate them to our inference results (see Chapter 6)

5.2 Inferring the origin of cosmic structure

Cosmology constitutes an ideal case for Bayesian statistics. The standard model of cosmology is well supported by theory and observations (Planck Collaboration et al., 2018, 2019). Not using this information as a prior would require very good arguments. Therefore, the BORG framework incorporates a Gaussian prior for the initial conditions. As mention in Chapter 2, the isotropy of the Universe requires a diagonal covariance matrix in Fourier space. Therefore, the prior for initial density perturbations $\delta^{\rm ic}$ in a finite box is given by

$$P\left(\hat{\delta}^{\rm ic}|\hat{\mathbf{S}}\right) = \frac{1}{\sqrt{|2\pi\hat{\mathbf{S}}|}} \exp\left(-\frac{1}{2}\sum_{k,k'}\hat{\delta}^{\rm ic}_k \hat{S}^{-1}_{kk'}\hat{\delta}^{\rm ic}_{k'}\right) \tag{5.1}$$

where the hat denotes Fourier-space and $\mathbf{S}$ is the covariance matrix. The diagonal coefficients can be written as a function of the initial power spectrum $P(k)$:

$$\hat{S}^{-1}_{kk} = \sqrt{\frac{P(k)}{(2\pi)^{3/2}}}. \tag{5.2}$$

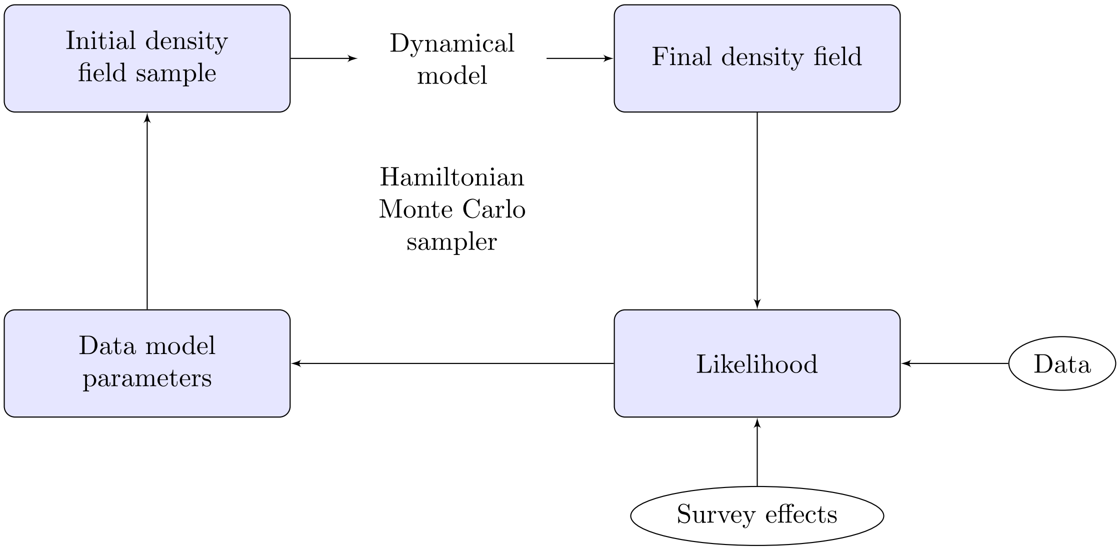


Figure 5.1: Flow chart depicting the iterative block sampling approach of the BORG inference framework. The BORG algorithm first infers three-dimensional initial and evolved density fields from the data. In the next step, the algorithm updates the parameters of the data model using the realisations of the density fields. The iteration of the algorithm using an MCMC method provides the joint posterior distribution.

The exact shape of the power spectrum is well described by theory and observations (Eisenstein and Hu, 1998, 1999; Dodelson, 2003). The BORG framework follows the power spectrum prescription of Eisenstein and Hu (1998), including BAO wiggles.

The initial condition prior selects physically reasonable density fields from the space of all possible states. However, the prior does not strictly limit the space of possible initial conditions that is explored to match the data. Despite the physical motivation, the Gaussian prior is a maximum entropy, thus, the least informative prior. If unlikely events are required to explain the data, the algorithm explores posterior regions that are unlikely.

After obtaining a realization of the initial density field, this is evolved by a dynamical structure formation model. By capturing the filamentary structure of the cosmic web, the dynamical structure formation model accounts for the high-order correlations. Using conditional probabilities, we can derive a prior for the final density field δ^{f} at a scale factor a:

$$P\left(\delta^{\mathrm{f}}\right) = \int P\left(\delta^{\mathrm{f}}, \delta^{\mathrm{ic}}\right) \mathrm{d}\delta^{\mathrm{ic}} \tag{5.3}$$

$$= \int P\left(\delta^{\mathrm{f}} | \delta^{\mathrm{ic}}\right) P\left(\delta^{\mathrm{ic}}\right) \mathrm{d}\delta^{\mathrm{ic}} \,. \tag{5.4}$$

Once the gravitational model $M\left(a, \delta_l^{\mathrm{ic}}\right)$ is defined, a prior distribution for the evolved density field can be obtained in two steps: first, a realization of the initial conditions is drawn from the Gaussian prior. It is then evolved forward in time by the structure formation model. This procedure provides samples from the joint prior distribution of the initial and evolved density fields:

$$P\left(\delta^{\mathrm{f}}, \delta^{\mathrm{ic}}\right) = P\left(\delta^{\mathrm{ic}}_l\right) \prod_l \delta^D \left(\delta^{\mathrm{f}}_l - M\left(a, \delta^{\mathrm{ic}}_l\right)\right). \tag{5.5}$$

5.3 The dynamical model

The BORG framework employs a dynamical model to evolve the initial conditions into the final density field. The dynamical model reproduces the filamentary structure of the cosmic web and, therefore, describes the high-order statistics. Sections 2.2.4 and 2.2.4 introduced some of the structure formation models: Lagrangian perturbation theory (LPT) and N-body simulations. While LPT is a valid approximation of the structure formation close to the linear regime (Moutarde et al., 1991; Buchert et al., 1994; Bouchet et al., 1995; Scoccimarro, 2000; Leclercq et al., 2013), it breaks down at $|\delta| \gg 1$. In these high-density regimes, where galaxies and clusters form, we need to rely on N-body simulations. For these reasons, the BORG framework incorporates different dynamical structure formation models: one can choose between LPT and N-body (particle-mesh) methods, depending on the density regimes of the problem.

More specifically, the BORG framework populates the initial density field with dark matter particles, which evolve according to the dynamical model. The equations of motion of these particles are integrated with a leapfrog integrator (see Appendix A). The final distribution of particles is then assigned to a grid using a cloud-in-cell scheme (Appendix B, Hockney and Eastwood, 1988), yielding the final density field. Note that the BORG algorithm infers the initial density field at Lagrangian coordinates, while final density fields are recovered at the corresponding final Eulerian coordinates. Therefore, the algorithm accounts for the displacement of matter in the course of structure formation (Jasche and Lavaux, 2018).

5.4 Connecting the Bayesian forward model with data

Inferring the matter distribution requires to connect the model with observations. To achieve this, we need to model the data taking the noise and systematic effects into account. Based on the data model, we can build a likelihood distribution, describing the statistical process by which the data was created given a specific model prediction. In this section, I will describe the standard Poissonian likelihood for galaxy surveys. I will also introduce the likelihood I developed for the Ly-α forest data. Appendix C contains the gradients required for the HMC scheme.

5.4.1 Modelling galaxy clustering as a point process

Galaxy surveys provide the position of each object, allowing the study of galaxy clustering. These positions of galaxies are then converted into number counts in each volume element of a 3d regular grid. To achieve this, the sky coordinates of galaxies (α, ξ) and their comoving distance $d_{\rm com}$ are transformed to the Cartesian grid coordinates as

$$\begin{aligned} x &= d_{\rm com}\left(z\right)\cos\left(\xi\right)\cos\left(\alpha\right), \qquad (5.6)\\ y &= d_{\rm com}\left(z\right)\cos\left(\xi\right)\sin\left(\alpha\right), \qquad (5.7)\\ z &= d_{\rm com}\left(z\right)\sin\left(\xi\right), \qquad (5.8)\end{aligned}$$

with ξ being the declination and α being the right ascension. The comoving distance $d_{\rm com}$ is calculated from the redshift given a fiducial cosmology as

$$d_{\rm com}(z) = \frac{c}{H_0}\int_0^z \frac{{\rm d}z'}{\sqrt{\Omega_r\left(1+z'\right)^4+\Omega_m\left(1+z'\right)^3+\Omega_k\left(1+z'\right)^2+\Omega_\Lambda}} \qquad (5.9)$$

where Ω_x are the cosmological parameters defined in eq. (2.16). Once the galaxies are located in the volume elements of a three-dimensional regular grid, the galaxy number-counts are N computed at each volume element.

To account for the luminosity-dependent effects, data are split into several magnitude bins, which are treated as independent datasets. This allows modelling the galaxy bias and selection function effects differently for fainter and brighter galaxies. Therefore, for each magnitude bin l, there is a likelihood distribution $P\left(N^l|\delta^{\rm ic}\right)$, where N^l are the number counts of observed galaxies (Jasche and Wandelt, 2013). The joint likelihood in all bins is obtained by

$$P\left(\{N^l\}|\delta^{\rm ic}\right) = \prod_l P\left(N^l|\delta^{\rm ic}\right), \qquad (5.10)$$

where $\{N^l\}$ contains the galaxy counts in all magnitude bins.

The statistical noise due to the discrete nature of galaxy surveys is often modeled as a Poisson process (Layzer, 1956; Peebles, 1980; Martínez and Saar, 2003).

$$P\left(N^l|\lambda\left(\delta^{\rm ic}\right)\right) = \prod_x \frac{1}{N_x^l!}\exp\left(-\lambda_x^l\left(\delta^{\rm ic}\right)\right)\left(\lambda_x^l\left(\delta^{\rm ic}\right)\right)^{N_x^l} \qquad (5.11)$$

with $\lambda_x^l\left(\delta^{\rm ic}\right)$ being the expected number of galaxies in the voxel x in the luminosity bin l for a given initial density field $\delta^{\rm ic}$.

The BORG framework also takes the galaxy bias and observational effects such as the completeness mask or the radial selection function (Jasche and Lavaux, 2017) into account. The galaxy bias B^l accounts for the non-local differences between matter and galaxy distributions. The geometry and completeness of the survey are encoded in the linear survey response R_x^l. Then, the mean number of galaxies is

$$\lambda_x^l\left(\delta^{\rm ic}\right) = R_x^l\bar{N}^l\left(1+B^l\left[M\left(a,\delta^{\rm ic}\right)\right]_x\right) \qquad (5.12)$$

where $\bar{N}^l$ is the expected number of galaxies in the l-th bin (Jasche and Wandelt, 2013).

Finally, the likelihood for the galaxy redshift surveys reads

$$\begin{aligned} P\left(\{N\}|\delta^{\rm ic},\bar{N}\right) &= \prod_{x,l}\frac{1}{N_x^l!}\exp\left(-R_x^l\bar{N}^l\left(1+B^l\left[M\left(a,\delta^{\rm ic}\right)\right]_x\right)\right) \\ &\times \left(R_x^l\bar{N}^l\left(1+B^l\left[M\left(a,\delta^{\rm ic}\right)\right]_x\right)\right)^{N_x^l}. \end{aligned} \tag{5.13}$$

Although that model has been successful at inferring the matter distribution (see e.g. Lavaux and Jasche, 2016), foreground and target contaminations may introduce artefacts on the density field. For this reason, Jasche and Lavaux (2017) developed a template method to account for foreground templates. However, this is not sufficient to deal with yet unknown systematic effects. For this reason, I developed a novel likelihood that effectively deals with unknown contaminations in cosmological datasets (Chapter 7).

5.4.2 A physical forward data model for the Lyman-α forest

While galaxies trace the matter distribution at high densities, the Ly-α forest provides a complementary probe. In this thesis, I extended the BORG algorithm to the analysis of the Ly-α forest, making the 3d analysis of the matter distribution at $z > 2$ feasible, where currently galaxy surveys are too sparse to trace the cosmic web.

The Ly-α forest is measured along lines of sight to background sources, generally quasars. Therefore, these data require a projector along the line of sight to identify the volume elements intersected by the spectra. To achieve that, I implemented a projector in the BORG framework: first, the coordinates of the quasar are located in the regular grid according to eq. (5.8); second, the volume elements intersected by the line of sight are calculated from the wavelengths in the spectrum. For that, the redshifts are obtained from the wavelengths as

$$z = \frac{\lambda - 1}{\lambda_0} \tag{5.14}$$

where $\lambda_0 = 1216$ Å corresponding to the wavelength of the Ly-α transition. The redshift is then used to compute the comoving distance from tabulated values of $d_{\rm com}(z)$ and finally, this is transformed into the regular grid coordinates.

For the analysis of the Ly-α forest, I developed a likelihood based on the fluctuating Gunn-Peterson Approximation (FGPA Gunn and Peterson, 1965; Lynds, 1971):

$$F(z,\hat{\mathbf{x}}) = \exp\left[-A(1+\delta(\mathbf{x}))^{\beta}\right], \tag{5.15}$$

where F is the transmitted flux fraction, z is the considered absorption redshift, $\mathbf{x}$ is the corresponding comoving distance, $\hat{\mathbf{x}}$ indicates the associated unit vector, δ is the density contrast and A and β are astrophysical parameters that are related to the physics of neutral hydrogen. A more detailed description of the fluctuating Gunn-Peterson approximation can be found in Chapter 8.

The likelihood is then derived by assuming Gaussian pixel-noise in the data (Bird et al., 2011; Cisewski et al., 2014; Ozbek et al., 2016; Lee et al., 2018; Horowitz et al., 2019). Therefore,

$$P\left(\delta^{\mathrm{ic}}, \delta^{\mathrm{f}} | F\right) = \prod_{n,x} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{\left(\left(F_n\right)_x - \exp\left[-A\left(1+\delta_x\right)^{\beta}\right]\right)^2}{2\sigma_n^2}\right] \tag{5.16}$$

where n labels the different lines of sight, x corresponds to the volume elements intersected by the n-th spectra and F is the observed transmitted flux fraction[1]. The astrophysical parameters A and β are related to the temperature-density relation of the intergalactic medium and the spectral shape of the background sources (Gunn and Peterson, 1965). However, the exact value of these parameters is still unknown (see e.g. Calura et al., 2012; Garzilli et al., 2012; Rudie et al., 2012; Bolton et al., 2014; Rorai et al., 2017). It is, therefore, necessary to marginalize over A and β to avoid introducing a bias in the results. For this reason, I modified the MCMC framework of the BORG algorithm to jointly sample the astrophysical parameters and the density field. To date, this is the first approach to sample these quantities jointly. Chapter 8 will show that the algorithm provides tight constrains of A and β, which could contribute to the debate of the temperature-density relation of the IGM (more detailes are provided in Section 8.7.6).

5.4.3 Choosing the optimal mass matrix

As mention above, the BORG framework employs an HMC algorithm to explore the high-dimensional parameter space. HMC methods involve a large number of tunable parameters that are contained in the mass matrix $\mathbf{M}$ (see Section 4.5). Although the choice of these parameters does not affect the results, they can strongly affect the efficiency of the algorithm (Neal, 1993). Jasche and Wandelt (2013) performed a stability analysis of the numerical integrator scheme to determine the mass matrix in BORG, fixed at values corresponding to the covariance of the initial conditions:

$$M_{ij} = S_{ij}^{-1} - \delta_{ij}^{K} D_1 \frac{\partial J_i\left(\delta^{\mathrm{ic}}\right)}{\partial \delta_i^{\mathrm{ic}}}\left(\psi_i\right), \tag{5.17}$$

where ψ_i is assumed to be the mean initial density after the warm-up phase, D^1 depends on the growth factor from Lagrangian Perturbation Theory and J depends on the dynamical model and the cloud-in-cell kernel (Hockney and Eastwood, 1988). For computational efficiency, the mass matrix is diagonalised.

[1] The transmitted flux fraction is defined as $F = F_{\mathrm{transmitted}}/F_{\mathrm{total}}$. Therefore, F has no units.

Chapter 6

Imprints of the large-scale structure on AGN formation and evolution

The material displayed in this chapter and Appendix D has been published to Astronomy & Astrophysics in Porqueres et al. (2018).

I led this research project as the principal investigator and authored the corresponding publication as the first author. My contributions consisted of defining the research question, performing the data analysis, and preparing the material for publication. The project was further done in collaboration with Jens Jasche (JJ), Torsten Enßlin and Guilhem Lavaux (GL) who contributed in scientific discussions of the results. JJ and GL provided the matter density field for the analysis. All authors provided feedback on the text and accepted the final manuscript.

6.1 Introduction

Active galactic nuclei (AGN) are believed to have a significant impact on galaxy evolution. Black hole masses correlate with several properties of the host galaxy, including for example the stellar mass and the velocity dispersion. This suggests that black holes and their host galaxies have an intertwined evolution and that AGN affect galaxy evolution.

The large-scale environment also plays a role in galaxy evolution. It has been shown that star formation rates are influenced by the large-scale environment (e.g. Balogh et al., 2004; Einasto et al., 2005; Lietzen et al., 2009; Chen et al., 2013). For example, colours and morphologies of galaxies depend on the density of their host cluster. It is shown that luminosity functions of elliptical galaxies are strongly affected by the environmental density amplitude (Einasto et al., 2008). Since the AGN phase may be a short period in the evolution of all massive galaxies, an interesting question is how the large-scale environment affects AGN, particularly, how the formation and properties of AGN depend on the environmental density.

The standard model of AGN (Antonucci, 1993) assumes that the energy they release is produced by the accretion of gas onto a central supermassive black hole. However, the coupling between black hole accretion and star formation in AGN remains unclear. In order to determine how these two processes happen and whether they synchronize, some works focused on the study of AGN in individual voids since more violent processes such as gas stripping are less likely to occur at low densities. Earliest works (Kirshner et al., 1981; Cruzen et al., 2002) were limited to individual voids and concluded that AGN properties are similar in voids and clusters. However, those studies could not exclude the possibility that the AGN observed in the studied void were not residing in filaments, which would mask the local environment in the void.

Statistically significant studies have emerged recently with the release of large surveys, in particular, the Sloan Digital Sky Survey (SDSS Strauss et al., 2002). It was shown that the occurrence of the AGN activity as a function of the environment depends on the properties of the host galaxy (Kauffmann et al., 2003; Constantin et al., 2008; Hwang et al., 2012; Silverman et al., 2008; Pimbblet et al., 2013). Strongly accreting AGN are found predominantly in low-density regions (Kauffmann et al., 2003; Constantin and Vogeley, 2006). According to Constantin et al. (2008) void AGN in massive hosts exhibit stronger accretion and younger stellar emission than their cluster counterparts. As found in Hwang et al. (2012), the morphology of the host galaxy strongly affects the dependence of the AGN fraction with the environment.

The study of the AGN clustering signal, defined as the cross-correlation of AGN with reference galaxies, shows that active and inactive galaxies have the same environment on large scales (Li et al., 2006; Jiang et al., 2016; Karhunen et al., 2014) while they show an overdensity of nearby neighbours (at distances < 1 h^{-1} Mpc). This indicates that halos hosting AGN and inactive galaxies have similar masses.

Constantin et al. (2008) have analysed the environmental dependence of the properties of each AGN spectral type. Although some differences between spectral types can be explained by the orientation angles with respect to the line of sight since the disk obscuration affects the AGN spectra, Tempel et al. (2009) and Constantin et al. (2008) suggested that the different spectral types may form a sequence in galaxy evolution. Since the dynamical evolution in low-density regions is expected to be slower, this evolutionary sequence might appear as a dependence of the spectral type with the environmental matter clustering. The study of this dependence also suggests an evolutionary sequence from quasars to radio-loud galaxies (Lietzen et al., 2011).

The goal of this paper is to study the relation between AGN and their global environmental density field, particularly the dependence of the incidence and properties of different spectral types with the density. Unlike most of the previous works (e.g. Constantin and Vogeley, 2006; Lietzen et al., 2011; Karhunen et al., 2014; Coziol et al., 2017), in which the density field is obtained from galaxy counts, we use a three-dimensional high-resolution density field obtained from a Bayesian reconstruction applied on the 2M++ sample (Lavaux and Hudson, 2011; Jasche and Wandelt, 2013; Lavaux and Jasche, 2016).

This paper is structured as follows: in Section 2, we present the method to reconstruct the large-scale density field. In Section 3, we detail the dataset and the spectral classification of AGN into Seyferts, LINERs, and Transition objects. In Section 4, we study the dependence of the properties and occurrence rates of the different types of AGN with the environment. We have investigated this dependence as a function of the density contrast as well as of a web-type classification based on the shear tensor, identifying cosmic web structures like voids, sheets, filaments, and clusters. We summarize our results in Section 5.

6.2 Methodology

This section provides a detailed overview of the methods used to study AGN properties in the cosmic large-scale structure (LSS). We also provide a brief overview of the Bayesian inference method providing the three-dimensional density field used in this work.

6.2.1 Bayesian large-scale structure inference

This work builds upon previous results of applying the Bayesian Origin Reconstruction from Galaxies (BORG) algorithm to the data of the 2M++ galaxy compilation (Lavaux and Hudson, 2011; Jasche and Wandelt, 2013; Jasche et al., 2015; Lavaux and Jasche, 2016).

The BORG algorithm is a fully probabilistic inference method aiming at reconstructing matter fields from galaxy observations. This algorithm incorporates a physical model for gravitational structure formation, which allows inferring the three-dimensional density field and the corresponding initial conditions at an earlier epoch from present observations.

Specifically, the algorithm explores a large-scale structure posterior distribution consisting of a Gaussian prior for the initial density field at an initial cosmic scale factor of $a = 10^{-3}$ linked to a Poissonian model of galaxy formation at a scale factor $a = 1$ via a second order Lagrangian perturbation theory (for details see Jasche and Wandelt 2012). The model accurately describes one-, two- and three-point functions and represents very well higher-order statistics, as it was calculated by Moutarde et al. (1991); Buchert et al. (1994); Bouchet et al. (1995); Scoccimarro (2000); Leclercq et al. (2013). Thus BORG naturally accounts for the filamentary structure of the cosmic web typically associated with higher-order statistics induced by nonlinear gravitational structure formation processes. The posterior distribution also accounts for systematic and stochastic uncertainties, such as survey geometries, selection effects and noise typically encountered in cosmological surveys.

We also note, that the BORG algorithm infers initial 3D density fields at their Lagrangian coordinates, while final density fields are recovered at corresponding final Eulerian coordinates. Therefore the algorithm accounts for the displacement of matter in the course of structure formation.

As mentioned above, in this work we have used inferred LSS properties previously obtained by applying the BORG algorithm to data of the 2M++ galaxy sample (Lavaux and Hudson, 2011). Three-dimensional matter density fields have been inferred on a cubic Cartesian grid of side length of 677.7 h^{-1} Mpc consisting of 256^3 equidistant voxels. This results in a grid resolution of 2.6 Mpc h^{-1}. Further we assumed a standard ΛCDM model

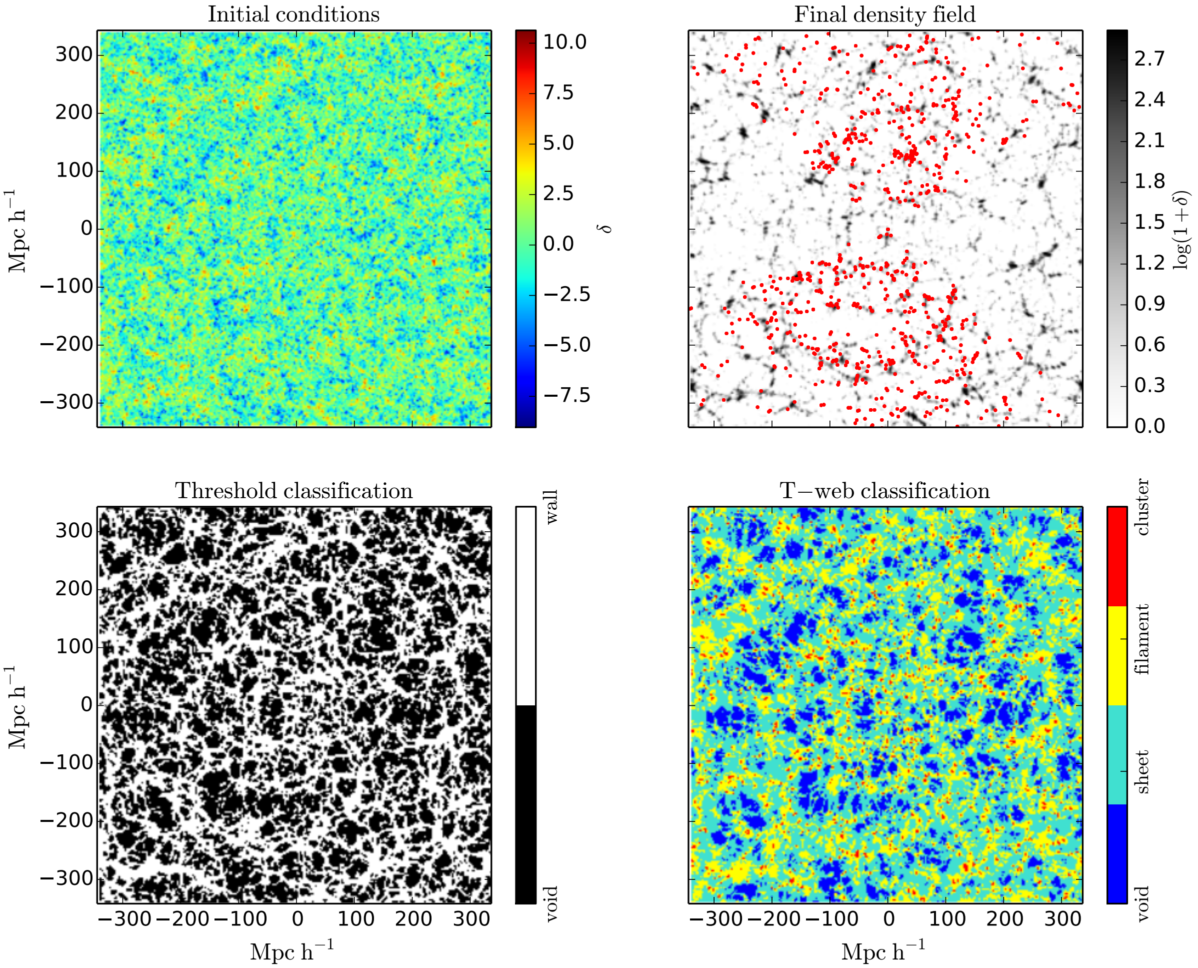


Figure 6.1: Slice of the three-dimensional density field. The upper panels show the initial and final density contrast. AGN are shown on top of the final density field (in red). The bottom left panel shows the different structures defined by a threshold in the density contrast $\delta = -0.6$ following the approach in Constantin et al. (2008). However, Constantin et al. (2008) defined the density field by smoothing the galaxy distribution, which might result in thinner structures. The bottom right panel shows the cosmic web structures according to a web-type classifier based on the tidal shear tensor.

with the following set of cosmological parameters: $\Omega_m = 0.3175$, $\Omega_\Lambda = 0.6825$, $\Omega_b = 0.049$, $h = 0.6711$, $\sigma_8 = 0.8344$, $n_s = 0.9624$ (Planck Collaboration et al., 2014). We assume $H = 100h$ km s^{-1} Mpc^{-1}. A slice of the final density field is shown in the upper panel in Fig. 6.1 with the AGN superimposed.

The observer is at the centre of the box and for the sake of this work, we concentrate on a spherical region with a radius of 120 Mpc h^{-1}. In Appendix D, we describe the alignment of observed AGN coordinates with the density field. Since AGN are affected by the redshift space distortions, we do the analysis in the redshift space.

6.2.2 T-web classification

A detailed web-type classification can be achieved via a T-web analysis (see e.g. Hahn et al., 2007; Forero-Romero et al., 2009; Leclercq et al., 2015). For a more complex classification of the large-scale environment, we use a dynamic web classifier that dissects the entire large-scale structure into different structure types: voids, sheets, filaments, and clusters. This classification is based on the tidal shear tensor,

$$T_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j}, \tag{6.1}$$

where Φ is the gravitational potential which can be obtained from the Poisson equation $\nabla^2\Phi(\mathbf{x}) = \delta(\mathbf{x})$. Different structures are classified according to the sign of the T_{ij} eigenvalues.

The interpretation of this classification is straightforward because the sign of the eigenvalue defines whether a structure is expanding (negative) or contracting (positive) in the direction of the eigenvector. Therefore, the signature of the tidal tensor T_{ij} defines the number of axes along which there is a plausible gravitational collapse or expansion. As summarized in Table 6.1, when the three eigenvalues are negative, the region is defined as a void. If there is only one positive eigenvalue, it is a sheet (two-dimensional structure), while it is classified as a filament when there is only one negative eigenvalue. Clusters are identified as three positive eigenvalues. The result of the web-type classification is shown in the bottom right panel in Fig. 6.1, in which the structures are defined at scales of 2.6 h^{-1} Mpc.

Web-type	Eigenvalues
Void	$\mu_1 < 0, \mu_2 < 0, \mu_3 < 0$
Sheet	$\mu_1 < 0, \mu_2 < 0, \mu_3 > 0$
Filament	$\mu_1 < 0, \mu_2 > 0, \mu_3 > 0$
Cluster	$\mu_1 > 0, \mu_2 > 0, \mu_3 > 0$

Table 6.1: Web-type classification according to the eigenvalues μ_i of the shear tensor.

6.2.3 Computation of the abundance and occurrence rate

The abundance is defined as the number density of objects as a function of the environmental density. Since the mean density regions are more extended than voids and clusters, we needed a volume correction which is computed as the fraction of volume with a given density $v(\delta) = \frac{V(\delta)}{V_{\mathrm{total}}}$. Then the logharithm of the number density is computed as

$$\ln\left(\frac{N}{v}\right) = \ln\left(\frac{N_{\mathrm{objs}}(\delta_{\mathrm{min}} < \delta < \delta_{\mathrm{max}})}{v(\delta_{\mathrm{min}} < \delta < \delta_{\mathrm{max}})}\right) \tag{6.2}$$

where δ_{min} and δ_{max} define the density bin.

The occurrence rate is defined as the number of objects of a spectral type in a given density divided by the total number of objects in this density bin, for example, $N_{\mathrm{Sy}}(\delta)/N_{\mathrm{AGN}}(\delta)$. These quantities are computed with the Blackwell-Rao estimator (see Section 6.2.4).

6.2.4 Blackwell-Rao estimator

The Markovian samples described in Section 6.2.1 permits us to provide an uncertainty quantification of our results. We employed the Blackwell-Rao estimators:

$$\langle x|d\rangle = \frac{1}{N}\sum_i x_i, \tag{6.3}$$

in which x is the quantity we want to study, for example the abundance, N is the number of Markovian samples and d the observations. This estimator was necessary because we were studying nonlinear functions of the density and $\langle f(\delta)\rangle \approx \langle f(\frac{1}{N}\sum_i \delta_i)\rangle$ only when the quantity of study $f(\delta)$ is linear. It also allowed us to calculate the uncertainty as

$$\langle (x - \langle x\rangle)^2|d\rangle = \frac{1}{N}\sum_i (x_i^2 + \sigma_i^2) - \langle x|d\rangle^2, \tag{6.4}$$

where σ_i^2 is the variance of the posterior distribution.

All the quantities in this study have been computed using the Blackwell-Rao estimators on 50 Markovian samples.

6.3 Data

In this section, we describe the datasets employed in this work and provide a description of derived quantities and data selections.

6.3.1 AGN catalog

The analysis presented in this work is based on the catalogue of the MPA/JHU collaboration (Kauffmann et al., 2003), which is a subset of the Sloan Digital Sky Survey Data

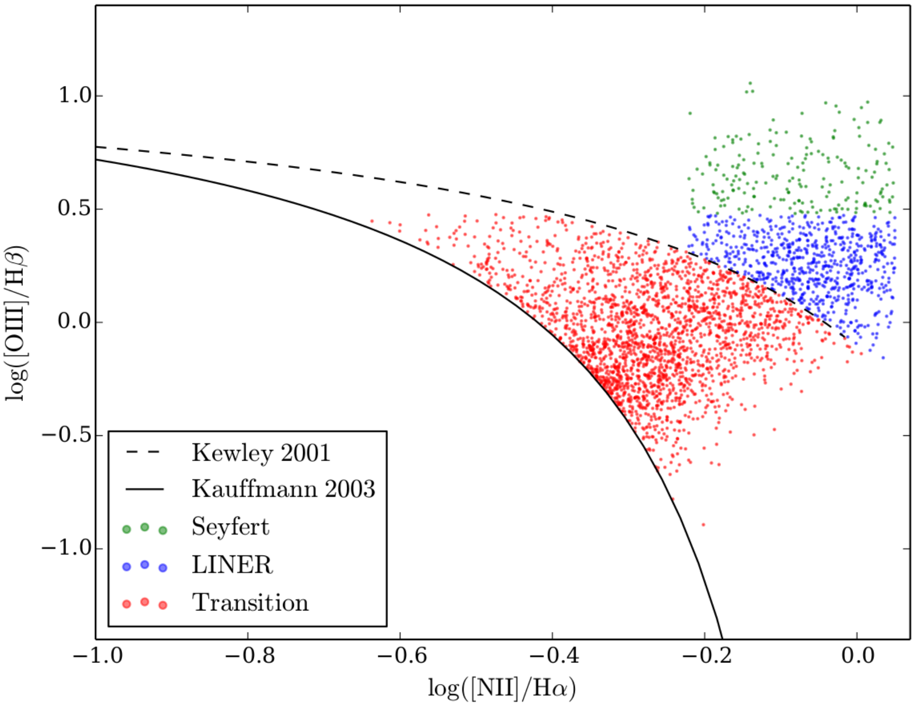


Figure 6.2: BPT diagram to classify AGN into Seyferts, LINERs, and Transition objects.

Release Seven (Strauss et al., 2002). This catalogue contains properties of the host galaxy such as the stellar mass and the dust attenuation strength as well as the velocity dispersion, which is proportional to the mass of the central black hole (Ferrarese and Merritt, 2000). It also contains the amplitude of the 4000 Å break and the strength of the Hδ absorption line which are indicators of the mean stellar age and recent starbursts. As a proxy of the morphology of the host galaxy, we will use the concentration index C and the effective stellar surface mass-density. Strateva et al. (2001) have shown that early-type galaxies have $C > 2.6$ while spiral and irregular galaxies have $C < 2.6$. The [OIII] line is produced by ionizing radiation that escapes along the polar axis of the dusty obscuring structure where it photo-ionizes and heats the medium (Heckman and Best, 2014). Therefore the [OIII] luminosity is related to the nuclear activity and so it is an indicator of the accretion rate (Kauffmann et al., 2003; Heckman et al., 2004; Heckman and Best, 2014). Following Constantin et al. (2008), we consider that AGN with $L_{[\mathrm{OIII}]} < 10^{39}$ erg s^{-1} are relatively weakly accreting objects.

According to the standard model (Antonucci, 1993), active galaxies are classified as type 1 AGN when the broad emission-line region is observed directly while type 2 are those for which it is obscured by the interstellar medium. In type 1, the continuum is dominated by nonthermal emission and thus it is difficult to estimate their host galaxy properties. For this reason, type 1 objects are excluded from the MPA/JHU catalogue (Kauffmann et al., 2003).

6.3.2 AGN classification

Baldwin et al. (1981) have shown that it is possible to distinguish type 2 AGN from normal star-forming galaxies by the intensity ratios of relatively strong emission lines. The MPA/JHU database contains the intensity ratios [OIII]λ5007/Hβ and [NII]λ6583/Hα (hereafter [OIII]/Hβ and [NII]/Hα) that allow us to classify AGN according to the Baldwin, Phillips & Terlevich (BPT) diagram (Fig. 6.2).

Spectral type	Definition
Transition object	Starburst galaxy
Seyfert	AGN with [OIII]/H$\beta > 3$
LINER	AGN with strong low ionization lines

Table 6.2: Spectral classification of type 2 AGN.

Spectral type	Total	Weakly accreting
Seyferts	197	33
LINERs	683	322
Transition objects	2176	1400

Table 6.3: Number of objects of each spectral type used in this study. We limit the study to $z < 0.041$ (120 h^{-1} Mpc).

The [OIII] line can be excited by AGN as well as by massive stars but it is known to be weak in metal-rich star-forming regions. However, for star-forming galaxies, the [OIII]/Hβ ratio increases while the [NII]/Hα ratio increases in high gas-phase metallicities (Charlot and Longhetti, 2001).

In the BPT diagram, AGN lie above the extreme starburst line defined by Kewley et al. (2001) and star-forming galaxies are below the Kauffmann et al. (2003) line. Objects between the lines host a mixture of star formation and nuclear activity and are classified as Transition objects. The objects in which the AGN component is dominant can be split into two groups: those with [OIII]/H$\beta > 3$ are Seyferts and the rest are classified as Low-Ionization Nuclear Emission-line Regions (LINER) (Kauffmann et al., 2003). LINERs typically have lower luminosities than Seyferts and their low ionization lines such as [OI] or [NII] are relatively strong. LINERs spectra could be produced in cooling flows, starburst-driven winds and shock-heated gas (Filippenko and Terlevich, 1992) and this opened a debate whether LINERs should be considered a low-luminosity extension of the AGN sequence (Ho et al., 2003). Table 6.3 shows the number of objects of each spectral type used in this study, up to redshift $z < 0.04$ (120 h^{-1} Mpc).

We followed Kauffmann et al. (2003) classification criteria, as shown in Fig. 6.2 and Table 6.2. The advantage of this criteria is that it is less sensitive to the projected aperture size of the fibres than other classification schemes based on lower ionization lines (Kauffmann et al., 2003).

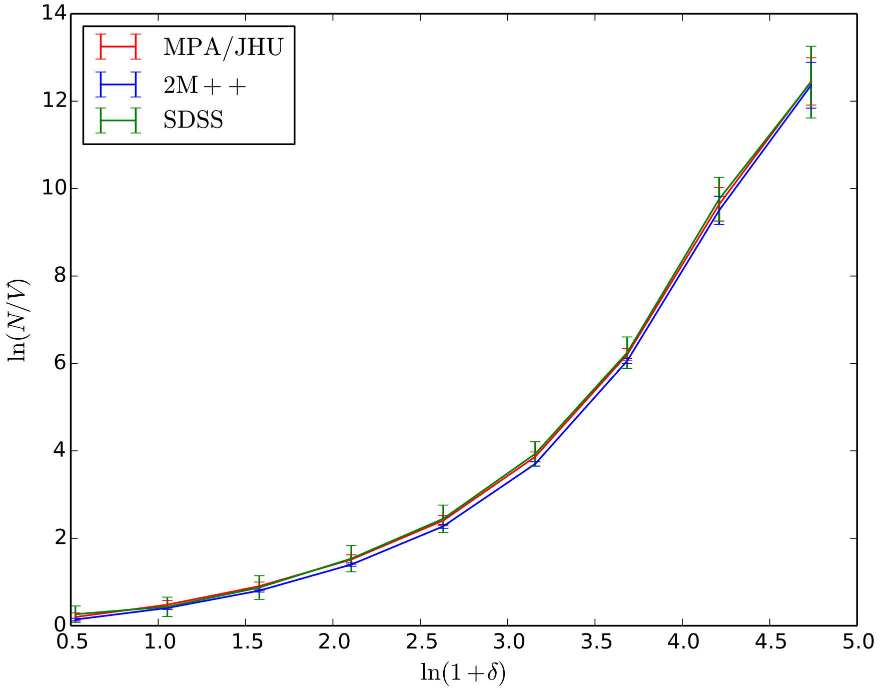


Figure 6.3: Abundance of AGN and galaxies. The curves are rescaled in order to compare the results for different catalogues. Galaxies and AGN show the same abundances which suggest that their halos have similar masses.

6.4 Results

6.4.1 AGN abundance

Since AGN can be observed at large distances, they are believed to be ideal tracers of the structure of the Universe at the largest scales. AGN played a key role in the early phases of cosmological studies but it was shown by source counts that AGN population evolve with cosmic epoch (Merloni and Heinz, 2013). Hence, the cosmological information, such as the geometry or the cosmological parameters, is masked by the evolution of AGN themselves. In this section, we study the abundances of AGN in different density environments and compare those to the corresponding galaxy abundances.

Figure 6.3 shows the abundances, the number density of objects as a function of the environmental density (Section 6.2.3), for the AGN catalogue (MPA/JHU) and two galaxy catalogues (2M++ and SDSS) limited in the same redshift range $z < 0.04$. In order to compare the results of the different catalogues, we rescaled the curves by dividing by the abundance in the first bin.

As can be seen, the abundances show the same trend for AGN and galaxies. This result is compatible with the findings of Jiang et al. (2016) and Li et al. (2006) who studied

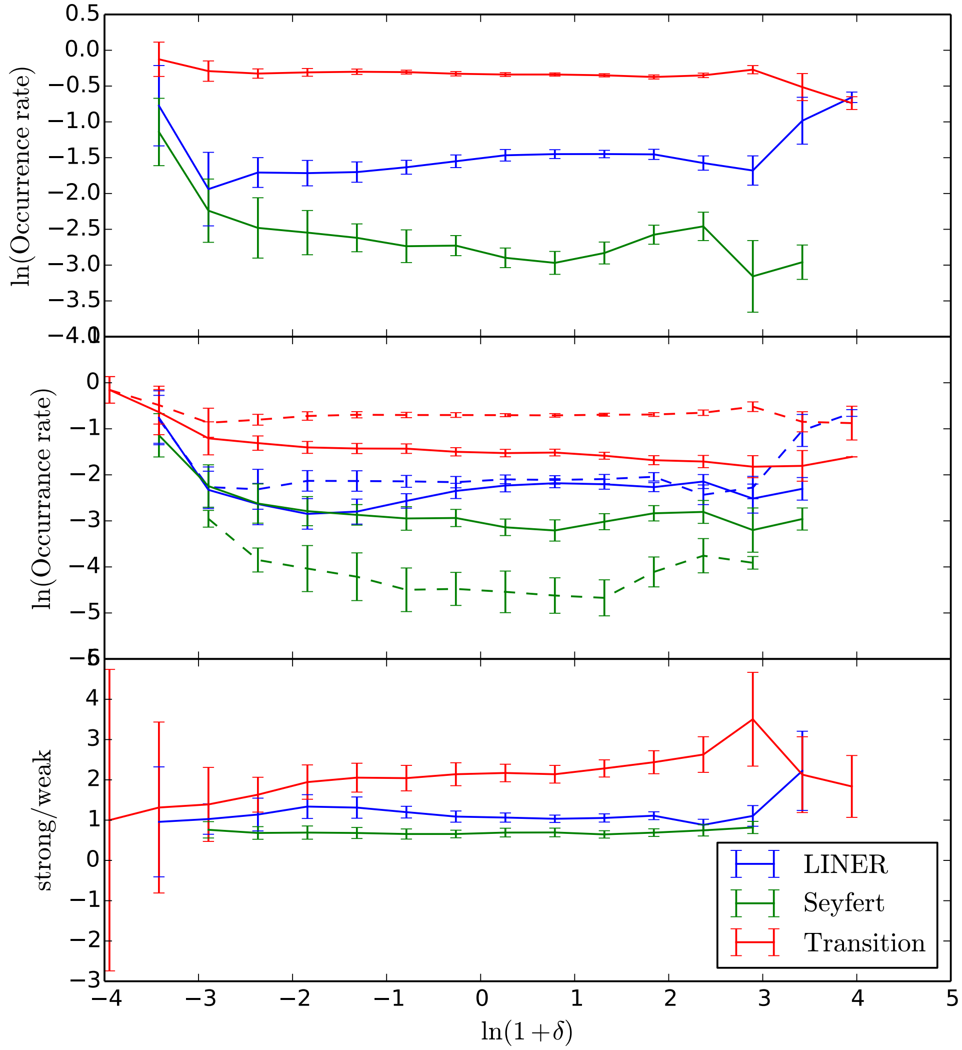


Figure 6.4: Occurrence rate for different spectral types as a function of the density contrast. The upper panel shows that Transition objects are less abundant in high densities, while LINERs show the opposite trend. The middle and bottom panels show that strongly (solid line) accreting Transition objects (defined as $L_{[\mathrm{OIII}]} > 10^{39}$ erg s^{-1}) are more sensitive to density contrast than weakly accreting (dashed line). This is compatible with the evolutionary sequence suggested by Constantin et al. (2008) with Transition objects transforming into LINERs

.

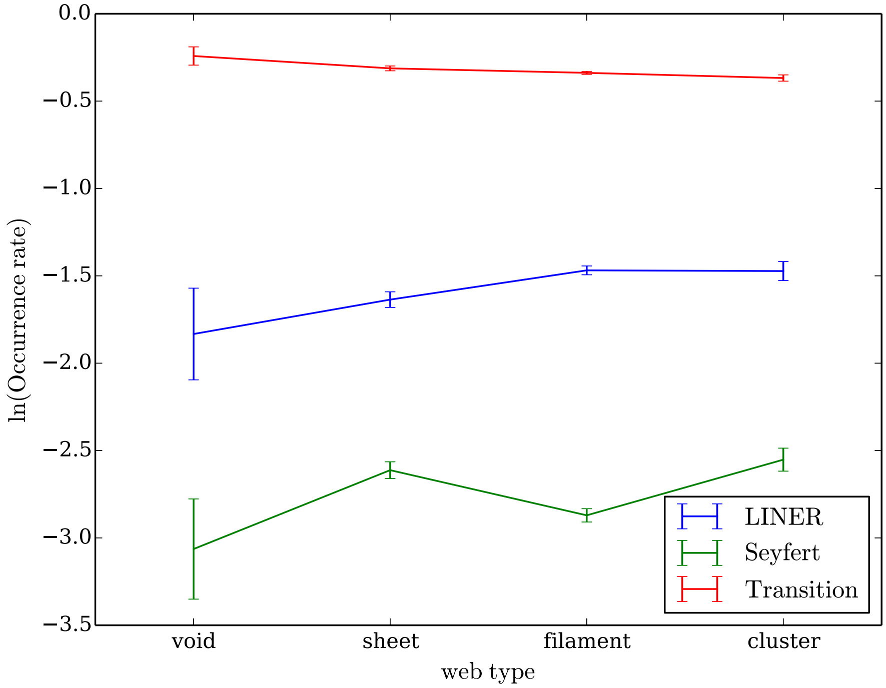


Figure 6.5: AGN ocurrence rate in different structures. Transition objects are more abundant in voids while LINERs are more abundant in clusters which is compatible with Transition objects evolving into LINERs.

the cross-correlation function of AGN and a reference galaxy sample and concluded that the AGN clustering signal is the same as that of galaxies at large scales (> 1 Mpc h^{-1}) indicating that the halos of active and inactive galaxies have similar masses. They found some differences in the number of neighbours of AGN in small scales (< 100 kpc h^{-1}) but our study is limited to scales of 2.6 Mpc h^{-1} which is the resolution of the reconstructed density field.

6.4.2 Occurrence rate

The evolutionary sequence suggested by Constantin et al. (2008) was derived from the study of the occurrence rate of different spectral types in different environments (defined in Section 6.2.3). Since the dynamical evolution in underdensities is slower, the fraction of different spectral types can provide some information on AGN evolution.

The upper panel in Fig. 6.4 shows that Seyferts and Transition objects have larger occurrence rates in underdensities than in clusters while LINERs are equally represented in under and overdensities. This might indicate an evolutionary sequence since Seyferts and Transition objects in very high-density regimes, with faster dynamical evolution, have already evolved into LINERs. This is also consistent with the increase of LINERs occurrence

δ threshold	Spectral Type	Wall	Void	Significance
-0.7	LINER	17 ± 1	22.2±0.9	Significant
	Seyfert	8± 1	6.35 ± 0.07	Not signif.
	Transition	73± 2	71.4± 0.1	Not signif.
-0.6	LINER	18 ± 1	22.5 ± 0.1	Significant
	Seyfert	8.1± 0.7	6.27 ± 0.09	Significant
	Transition	73 ± 1	71.3± 0.2	Not signif.
-0.5	LINER	18±1	22±2	Significant
	Seyfert	7.6 ± 0.5	6.26 ± 0.09	Significant
	Transition	74± 1	71.1 ± 0.2	Significant

Table 6.4: Occurrence rate (in %) for different density thresholds between void and wall overdensities. We can see that the significance of the results depends on the threshold.

rate in clusters.

Since the dynamical evolution can be faster for highly accreting objects, we study the effect of the accretion rate on the occurrence rates. Following Constantin et al. (2008), we consider that objects with $L_{[\mathrm{OIII}]} > 10^{39}$ erg s^{-1} are weakly accreting. The middle panel in Fig. 6.4 shows that weakly accreting Transition objects are equally represented in any density regime while their counterparts with high accretion rates become less frequent with density. This might indicate that Transition objects with high accretion rate make their transformation to another spectral type faster. We can see that weakly accreting Seyferts show stronger dependences with density than high accreting ones. LINERs do not show a significant difference between high and low accretion rates except at very high densities, where weakly accreting objects are more abundant. This higher occurrence of weak LINERs at very high densities might be related to their position in the cluster: the highest densities are found in the centre of clusters, where the interaction with late-type galaxies containing gas is less likely and hence the gas accretion might be reduced (Christlein and Zabludoff, 2004; Hwang et al., 2012).

In order to compare our results to Constantin et al. (2008), we study the AGN occurrence rates for voids and walls. Following their approach, we define voids according to a density threshold between wall and void overdensity. Table 6.4 shows occurrence rates of objects in voids and walls for different density thresholds. As can be seen, the ordering of the occurrence rate between AGN in wall and void regions do not depend on the density threshold. However, the difference between voids and walls could be a statistical fluctuation. For this reason, we include the significance of the results in Table 6.4. The difference between voids and walls is considered to be significant when

$$|\mathrm{Occurrence}_{\mathrm{wall}} - \mathrm{Occurrence}_{\mathrm{void}}| > 2\sqrt{\epsilon^2_{\mathrm{void}} + \epsilon^2_{\mathrm{wall}}}, \tag{6.5}$$

being ϵ the uncertainty in the occurrence rate. We can see that the significance of the results strongly depends on the threshold. Constantin et al. (2008) set the threshold at $\delta = -0.6$

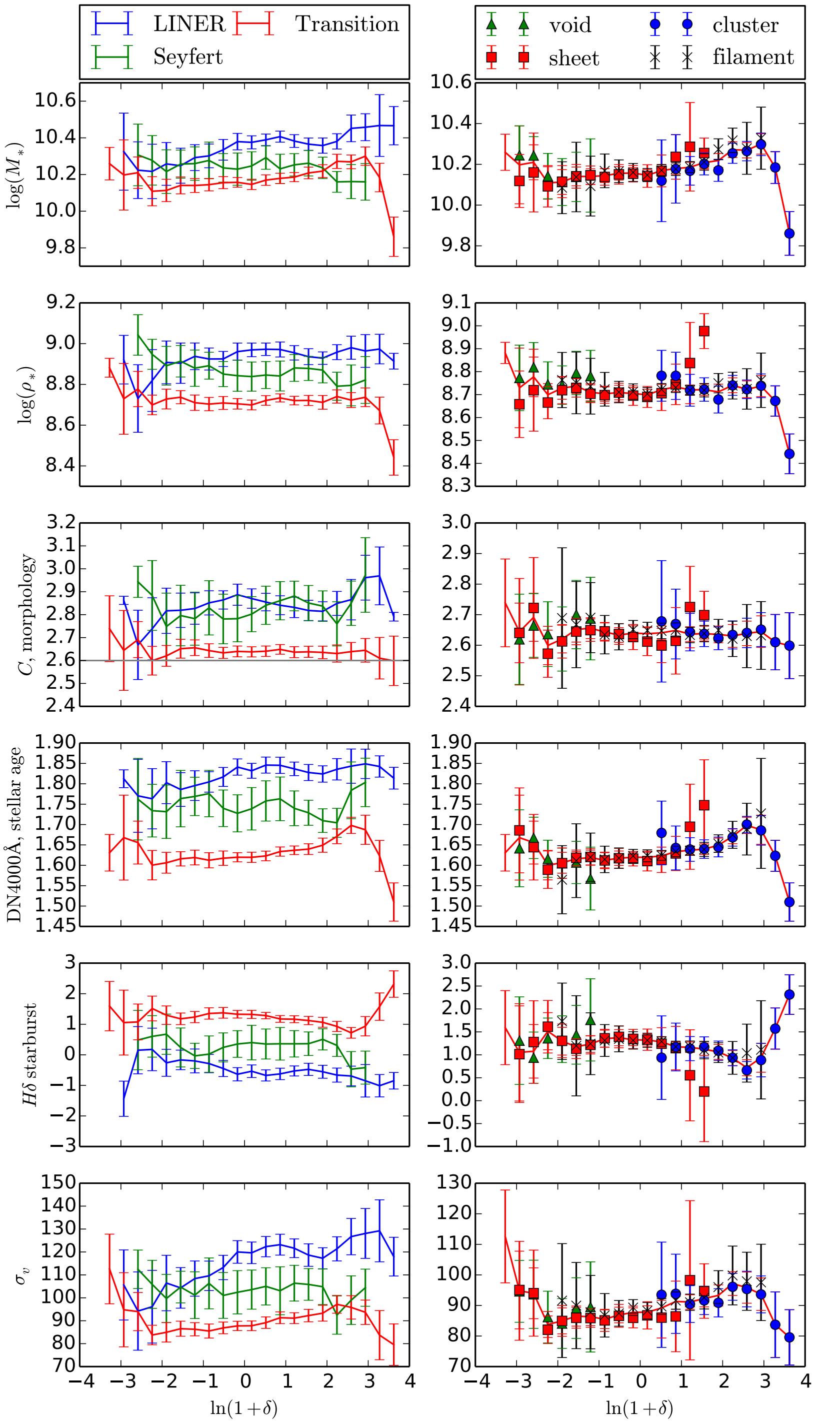


Figure 6.6: Properties of different spectral types of AGN as a function of density contrast. We can see that the stellar mass drops with the density due to gas stripping and high-density objects show a stronger starburst and younger stellar populations. The line in $C = 2.6$ separates the early- and late-type galaxies, showing that Seyferts and LINERs are found in early-type hosts. The left panels show that AGN properties are more dependent on the density than on the web-type.

and found that Seyferts are equally represented in voids and walls while Transition objects and LINERs show a preference for walls. We can only reproduce their result for Seyferts and LINERs when the threshold is $\delta = -0.7$, showing that the results obtained by density thresholding depend on the details of the definition of the density field.

6.4.3 Cosmic web analysis

In the previous section, we studied the occurrence rate as a function of density. In this section, we analyse whether the occurrence rate depends on the AGN location in the cosmic web. We classified different environments as voids, sheets, filaments and clusters as described in Section 6.2.2. Figure 6.5 shows that Transition objects are relatively more abundant in voids than in clusters while LINERs show opposite trends and Seyferts do not show a clear trend. This might also support that LINERs can be a later stage in the evolutionary sequence and hence they are more abundant where the dynamical evolution is faster due to interaction and merging.

6.4.4 Properties of AGN and LSS environment

We have studied how AGN formation depends on the environmental density. In this section, we focus on how the large-scale environment affects AGN properties. We studied the dependence of AGN and their galaxy host properties with the environmental density and the web-type structure. While the density is related to the amplitude of the large-scale gravitational potential, web-types provide information on the shape of the large-scale potential.

Comparing spectral types

The left panels in Fig. 6.6 show AGN properties as a function of environmental density for each spectral type. Transition objects show stronger dependency on the density contrast than LINERs and Seyferts. This can be related to their larger amount of gas since the availability of gas at merging can trigger star formation and affect the dynamics of galaxy interaction.

We can see in Fig. 6.6 that the stellar mass of the host galaxy, $\log(M_*)$, decreases at high densities for Transition objects and Seyferts. A possible explanation for this phenomena is the gas stripping due to interactions with other galaxies. LINERs do not show this behaviour because their gas content is smaller. We can also see that galaxy hosts in underdensities are more massive than in the mean density, especially for Transition objects. This indicates that void AGN are hosted in the most massive galaxies in voids. This result is consistent with Constantin et al. (2008), where they found that void AGN hosts are not dominated by the abundant less massive galaxies in the underdense regions.

The surface mass density $\log(\rho_*)$ and the concentration index C allow to study the galaxy morphology (Kauffmann et al., 2003). These two quantities can be used to separate early-type galaxies (Hubble types Sa, S0, and E) with $C > 2.6$ and surface mass density in

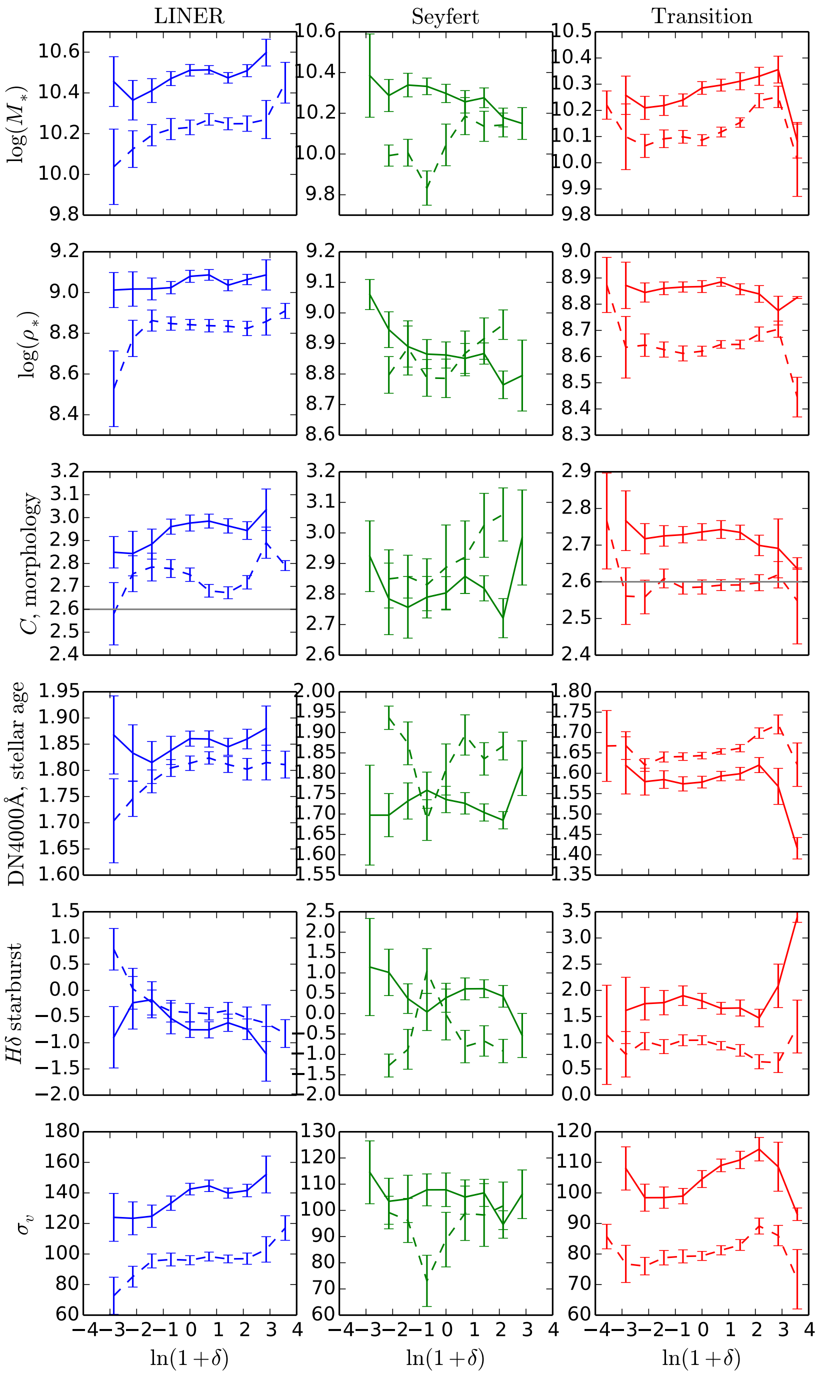


Figure 6.7: AGN properties for each spectral type and reported accretion rates. As can be seen, strongly (solid line) and weakly (dashed) accreting LINERs show a larger difference in underdensities. Transition objects with high accretion rate have more powerful starburst and smaller concentration indices at high densities (irregular galaxies). Seyferts show a transition at $\ln(1 + \delta) \approx -0.7$.

the range $3\times10^{8}-3\times10^{9}M_{\odot}\mathrm{kpc}^{-2}$, from late-type galaxies spirals and irregulars (Strateva et al., 2001). Figure 6.6 shows that LINERs and Seyferts occupy early-type galaxy hosts in any density regime. Transition objects also prefer early-type hosts in voids. However, the mean of their concentration index is $C\approx2.6$ for mean and large densities which indicates that the host galaxy of these objects can present different morphologies. We completed the analysis of their morphology in the next section, considering the differences due to accretion rates.

We also study the stellar population of the host galaxy. The 4000Å break is an indicator of the mean age of the stellar population which has smaller values for younger populations. Strong values of Hδ absorption arises due to recent bursts of star formation (Kauffmann et al., 2003). We can see that Transition object hosts are younger than Seyfert and LINER galaxy hosts. This is consistent with their larger amount of gas and the fact that Transition objects lie between the extreme starburst line and the pure star formation line, meaning that they can still produce new stars. The amplitude of the 4000Å break decreases at highest densities for Transition objects suggesting that strong interaction and merging in clusters trigger star formation in these objects. This is consistent with the Hδ trend since Transition objects show a larger increase of the Hδ absorption at high densities, indicating a more powerful starburst. These two quantities also show that Seyferts in underdensities have a slight preference for younger galaxy hosts while LINER hosts are similar in under and mean density regions but become older in high-density regimes, which may be related to the slower dynamical evolution in voids.

The stellar velocity dispersion σ_v is proportional to the black hole mass (Ferrarese and Merritt, 2000). We can see that velocity dispersions are larger in higher densities for LINERs. However, it decreases for high-density Transition objects. This could suggest that Transition objects at high densities host less massive black holes because they are still forming. However, Transition objects show a recent starburst in the same density regime that can affect the velocity dispersion. Since new stars are formed in gas clouds, they have the same velocity and this might affect the measurement of stellar velocity dispersions in the galaxy. Hence, velocity dispersion might become a bad estimator for the mass of the central black hole after a recent starburst.

In order to determine if AGN properties depend on the cosmic web or the density, the right panels in Fig. 6.6 show the properties at different web-type structures as a function of density. We can see that AGN properties do not depend on the web-type classification but on the density since different structures show the same values in the same density range. Since these properties are related to the star formation, it is expected that they depend on the density more than on the shape of the gravitational potential.

Comparing accretion rates

Since some authors (Constantin et al., 2008; Kauffmann et al., 2003) found that AGN properties depend on the accretion rate, we compared how the environment affects AGN properties in each spectral type for different accretion rates.

Figure 6.7 shows that strongly accreting Seyferts have less massive hosts at high densi-

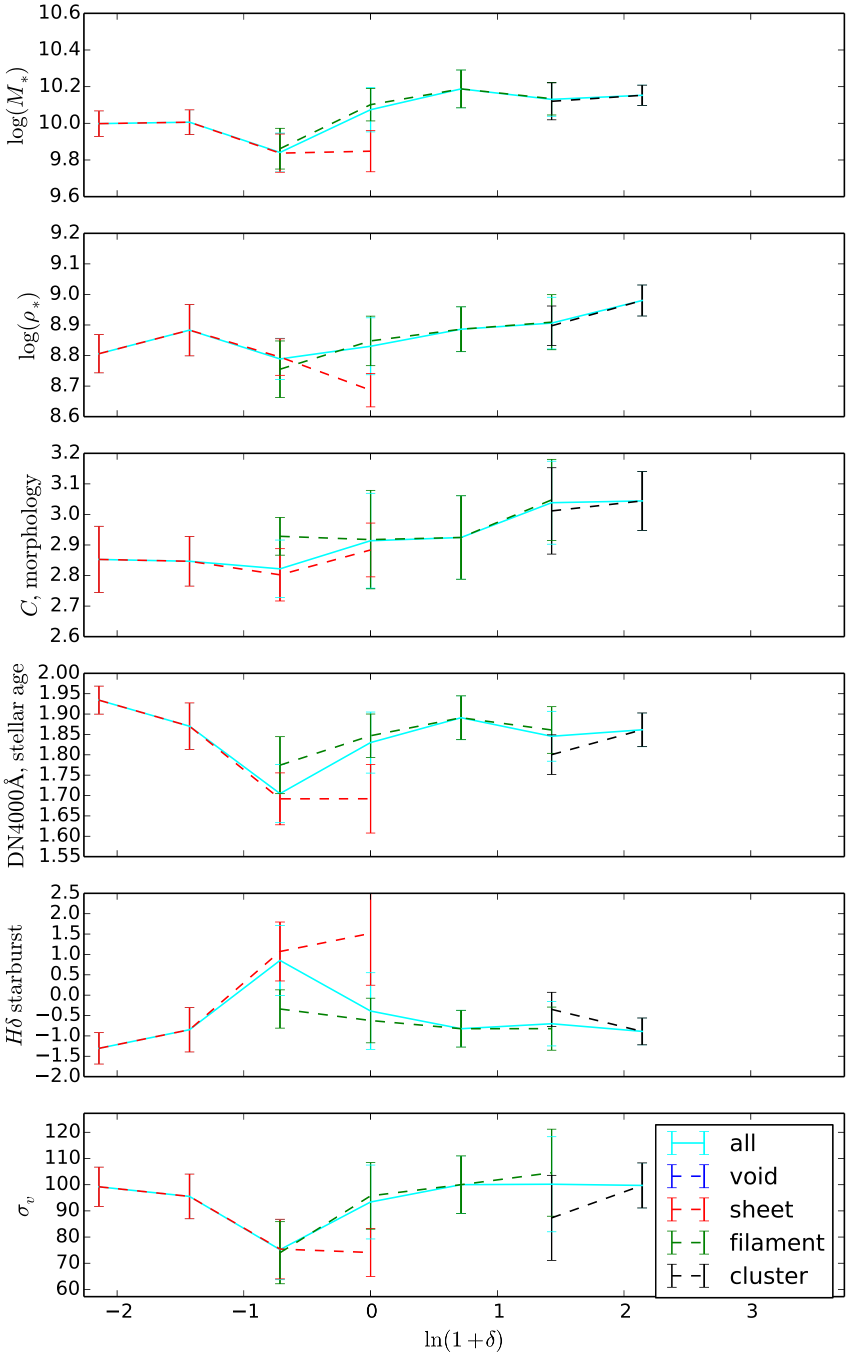


Figure 6.8: Properties of weakly accreting Seyferts. Around $\ln(1+\delta) \approx -0.7$ these exhibit a different behavior if located in sheets or in filaments. This difference in AGN formation times in these two environments might indicate a transition.

ties while Transition objects and LINERs do not show this trend. All spectral types have more massive hosts for high accretion rates. However, Transition objects have massive hosts in voids for any accretion rate.

In the previous section, we found that Seyferts and LINERs prefer early-type galaxies in any density regime while Transition objects can be found in different morphologies. In this section, we compare their concentration index and stellar mass density for different accretion rate. Figure 6.7 shows that Transition objects with low accretion rate can be found in late-type galaxies ($C < 2.6$) but they prefer early-type galaxies in underdensities. This can be due to a higher abundance of early-type galaxies in voids. We can also see that the concentration index decreases at high density for objects with high accretion rate. A possible explanation is that the interaction and merging in clusters lead to an irregular galaxy.

Transition objects with higher accretion rates show stronger starbursts ($H\delta$) than weakly accreting objects. This is consistent with the 4000Å break, indicating a younger population for Transition objects with a high accretion rate, especially at high densities. LINERs show different behaviour only in the low-density regime while at high densities the starburst and stellar age is equivalent to high and low accretion rate. Weakly accreting LINERs in voids show a young stellar population and stronger starburst. This might correlate with the concentration index: weakly accreting LINERs in voids are found in late-type galaxies, probably spiral galaxies with a larger amount of gas that can form new stars.

Weakly accreting Seyferts show a transition at $\ln(1+\delta) = -0.7$ in all their properties. In order to study this feature, Fig. 6.8 shows the properties for those objects in different web-type structures. Unlike the other spectral types, weakly accreting Seyferts show a different behaviour if located in sheets or in filaments of the same density. This may indicate different populations of Seyferts as a function of web-types of the same density. Specifically, objects in sheets are on average younger than in filaments. This may be a signature of recent AGN formation but further investigation with next-generation data and more detailed density reconstructions are required to confirm this feature in the future.

6.5 Conclusions

In this work, we have studied the effect of the large-scale environment on AGN: how the environment affects the formation and properties of AGN. We have used a 3D high-resolution (2.6 Mpc h^{-1}) density field obtained from a Bayesian reconstruction applied to the 2M++ galaxy catalogue. The study is based on the MPA/JHU AGN catalogue (Kauffmann et al., 2003), which contains only Type 2 AGN. We limit our study to objects within 120 Mpc h^{-1} ($z < 0.04$).

We confirm that the environment affects the formation and properties of AGN. Particularly, AGN properties and formation depend on the environmental density more than on the web-type. Hence, the amplitude of the large-scale gravitational potential affects AGN more than the shape of the large-scale potential. However, Seyferts with low accretion rate show some sensitivity to whether they reside within sheets or filaments.

The AGN abundance is the same as that of galaxies, indicating that halos are similar for active and inactive galaxies. When comparing spectral types and accretion rates, we have found differences in occurrence rates. Weakly accreting LINERs are more abundant in underdensities than in clusters while Transition objects show opposite trends. This is consistent with the AGN evolution of Seyferts/Transition objects into LINERs suggested by Constantin et al. (2008).

AGN properties are also affected by environmental density. The effect of the environment is stronger for Transition objects. This might be related to their larger amount of gas. It was found that the stellar mass of the host grows with the environmental density only for LINERs while Seyferts and Transition objects show a decrease of their stellar mass at very high densities. This might be a result of the gas stripping due to galaxy interaction and merging. It is also interesting that the AGN population in voids is not dominated by low-mass hosts but AGN are found in the most massive void galaxies. The starburst and age of the stellar population are also affected by the large-scale environment: younger populations and more powerful starbursts are found in clusters. This might be an indicator of the interactions and merging that trigger the star formation in AGN hosts.

We also find that the effect of the environment on AGN properties is different for weakly accreting objects. For instance, the morphology of Transition objects depend on their accretion rate, showing that weakly accreting objects are found in late-type galaxies. Seyferts with low accretion rate are younger in sheets than those in filaments in the same density regime, which might indicate different populations in these two environments.

These results indicate some particular properties of void AGN to be confirmed and studied by larger AGN samples of next-generation surveys.

Chapter 7

Explicit Bayesian treatment of unknown foreground contaminations in galaxy surveys

The material displayed in this chapter has been published to Astronomy & Astrophysics in Porqueres et al. (2019b).

I led this research project as the principal investigator and authored the corresponding publication as the first author. My contributions consisted of the method development, the implementation and testing of the method, generation of artificial mock data, and preparation of the material for the publication. The project was further done in collaboration with Doogesh Kodi Ramanah (DKR), Jens Jasche, and Guilhem Lavaux who contributed in discussions about the method and defining tests to validate the algorithm. DKR wrote 50 % of Sections 7.1 and 7.4. All authors provided feedback on the text and accepted the final manuscript.

7.1 Introduction

The next generation of galaxy surveys such as Large Synoptic Survey Telescope (LSST) (Ivezic et al., 2008) or Euclid (Laureijs et al., 2011; Racca et al., 2016; Amendola et al., 2018) will not be limited by noise but by systematic effects. In particular, deep photometric observations will be subject to several foreground and target contamination effects, such as dust extinction, stars, and seeing (e.g. Scranton et al., 2002; Ross et al., 2011; Ho et al., 2012; Huterer et al., 2013; Ho et al., 2015).

In the past, such effects have been addressed by generating templates for such contaminations and accounting for their overall template coefficients within a Bayesian framework. Leistedt and Peiris (2014), for example, compiled a total set of 220 foreground contaminations for the inference of the clustering signal of quasars in the Sloan Digital Sky Survey (SDSS-III) Baryon Oscillation Spectroscopic Survey (BOSS) (Bovy et al., 2012). Foreground contaminations are also dealt with in observations of the cosmic microwave background, where they are assumed to be an additive contribution to observed temperature fluctuations (e.g. Tegmark and Efstathiou, 1996; Tegmark et al., 1998; Hinshaw et al., 2007; Eriksen et al., 2008; Ho et al., 2015; Vansyngel et al., 2016; Sudevan et al., 2017; Elsner et al., 2017). In the context of large-scale structure analyses, Jasche and Lavaux (2017) presented a foreground sampling approach to account for multiplicative foreground effects which can affect the target and the number of observed objects across the sky.

All these methods rely on a sufficiently precise estimate of the map of expected foreground contaminants to be able to account for them in the statistical analysis. These approaches exploit the fact that the spatial and spectral dependence of the phenomena generating these foregrounds is well-known. But what if we are facing unknown foreground contaminations? Can we make progress in robustly recovering cosmological information from surveys subject to yet-unknown contaminations? In this work, we describe an attempt to address these questions and develop an optimal and robust likelihood to deal with such effects. The capability to account for 'unknown unknowns' is also the primary motivation behind the blind method for the visibility mask reconstruction recently proposed by Monaco et al. (2018).

The paper is organised as follows. We outline the underlying principles of our novel likelihood in Section 7.2, followed by a description of the numerical implementation in Section 7.3. We illustrate a specific problem in Section 4 and subsequently assess the performance of our proposed likelihood via a comparison with a standard Poissonian likelihood in Section 7.5. The key aspects of our findings are finally summarised in Section 7.6.

7.2 Robust likelihood

We describe the conceptual framework for the development of the robust likelihood which constitutes the crux of this work. The standard analysis of galaxy surveys assumes that the distribution of galaxies can be described as an inhomogeneous Poisson process (Layzer, 1956; Peebles, 1980; Martínez and Saar, 2003) given by

$$P(N|\lambda) = \prod_i \frac{e^{-\lambda_i}(\lambda_i)^{N_i}}{N_i}, \tag{7.1}$$

where N_i is the observed number of galaxies at a given position in the sky i and λ_i is the expected number of galaxies at that position. The expected number of galaxies is related to the underlying dark-matter density field ρ via

$$\lambda = S\bar{N}\rho^b \exp(-\rho_g \rho^{-\epsilon}), \tag{7.2}$$

where S encodes the selection function and geometry of the survey, $\bar{N}$ is the mean number of galaxies in the volume, and $\{b, \rho_g, \epsilon\}$ are the parameters of the non-linear bias model proposed by Neyrinck et al. (2014).

The key contribution of this work is to develop a more robust likelihood than the standard Poissonian likelihood by marginalizing over the unknown large-scale foreground

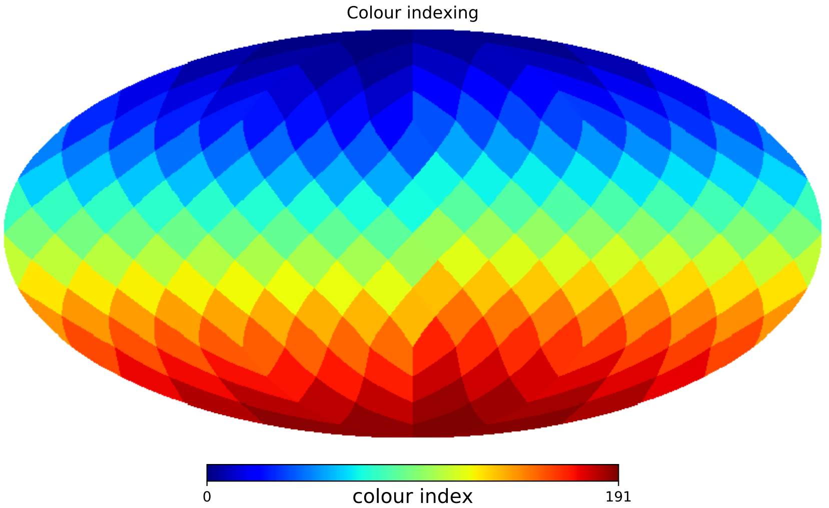


Figure 7.1: Schematic to illustrate the colour indexing of the survey elements. Colours are assigned to voxels according to patches of a given angular scale. Voxels of the same colour belong to the same patch, and this colour indexing is subsequently employed in the computation of the robust likelihood.

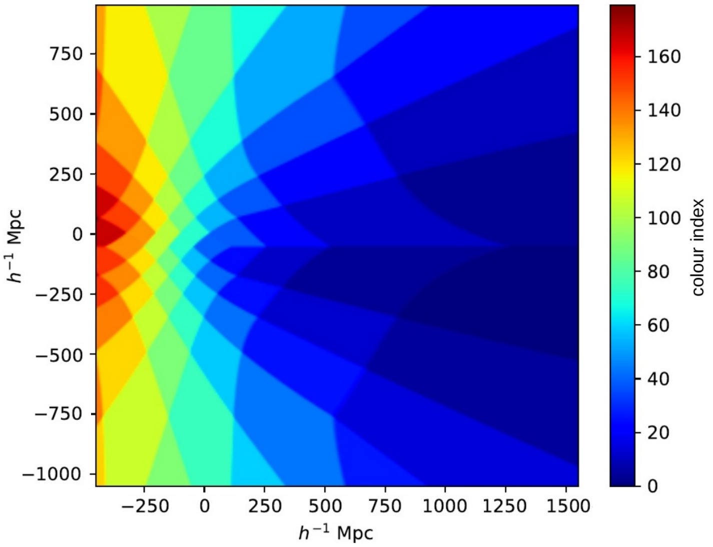


Figure 7.2: Slice through the three-dimensional (3d) coloured box. The extrusion of the colour indexing scheme (cf. Fig. 7.1) onto a 3d grid yields a collection of patches, denoted by a given colour, with a group of voxels belonging to a particular patch, to be employed in the computation of the robust likelihood. The axes indicate the comoving distances to the observer, who is located at the origin (0,0,0).

contamination amplitudes. We start with the assumption that there is a large-scale foreground modulation that can be considered to have a constant amplitude over a particular group of voxels. Assuming that A is the amplitude of this large-scale perturbation, we can write $\lambda_\alpha = A\bar{\lambda}_\alpha$, where the index α labels the voxels over which the perturbation is assumed to have constant amplitude. The likelihood consequently has the following form:

$$P(N|\bar{\lambda}, A) = \prod_\alpha \frac{e^{-A\bar{\lambda}_\alpha} A^{N_\alpha} (\bar{\lambda}_\alpha)^{N_\alpha}}{N_\alpha} \tag{7.3}$$

$$= e^{-A\sum_\alpha \bar{\lambda}_\alpha} A^{\sum_\alpha N_\alpha} \prod_\alpha \frac{(\bar{\lambda}_\alpha)^{N_\alpha}}{N_\alpha}. \tag{7.4}$$

We can marginalize over the unknown foreground amplitude A as follows:

$$P(N|\bar{\lambda}) = \int \mathrm{d}A \; P(N, A|\bar{\lambda}) \tag{7.5}$$

$$= \int \mathrm{d}A \; P(A|\bar{\lambda}) \; P(N|A, \bar{\lambda}) \tag{7.6}$$

$$= \int \mathrm{d}A \; P(A) \; P(N|A, \bar{\lambda}), \tag{7.7}$$

where, in the last step, we assumed conditional independence, $P(A|\bar{\lambda}) = P(A)$. This assumption is justified since the processes which generate the foregrounds are expected to be independent of the mechanisms involved in galaxy formation. As a result of this marginalization over the amplitude A, and using a power-law prior for A, $P(A) = \kappa A^{-\gamma}$ where γ is the power-law exponent and κ is an arbitrary constant, the likelihood simplifies to:

$$P(N|\bar{\lambda}) = \kappa \frac{\left(\sum_\alpha N_\alpha\right)!}{\left(\sum_\beta \bar{\lambda}_\beta\right)^{\sum_\alpha N_\alpha + 1 - \gamma}} \prod_\alpha \frac{(\bar{\lambda}_\alpha)^{N_\alpha}}{N_\alpha} \tag{7.8}$$

$$\propto \frac{1}{\left(\sum_\beta \bar{\lambda}_\beta\right)^{1-\gamma}} \prod_\alpha \left(\frac{\bar{\lambda}_\alpha}{\sum_\beta \bar{\lambda}_\beta}\right)^{N_\alpha}. \tag{7.9}$$

We employ a Jeffreys prior for the foreground amplitude A, which implies setting $\gamma = 1$. Jeffrey's prior is a solution to a measure invariant scale transformation (Jeffreys, 1946) and is, therefore, a scale-independent prior, such that different scales have the same probability and there is no preferred scale. This scale invariant prior is optimal for inference problems involving scale measurements as this does not introduce any bias on a logarithmic scale. Moreover, this is especially interesting because this allows for a total cancellation of unknown amplitudes in Eq. (7.9), resulting in the following simplified form of our augmented likelihood:

$$P(N|\bar{\lambda}) \propto \prod_\alpha \left(\frac{\bar{\lambda}_\alpha}{\sum_\beta \bar{\lambda}_\beta}\right)^{N_\alpha}. \tag{7.10}$$

7.3 Numerical implementation

We implement the robust likelihood in BORG (Bayesian Origin Reconstruction from Galaxies, Jasche and Wandelt, 2013), a hierarchical Bayesian inference framework for the non-linear inference of large-scale structures. It encodes a physical description for non-linear dynamics via Lagrangian Perturbation Theory (LPT), resulting in a highly non-trivial Bayesian inverse problem. At the core, it employs a Hamiltonian Monte Carlo (HMC) method for the efficient sampling of a high-dimensional and non-linear parameter space of possible initial conditions at an earlier epoch, with typically $\mathcal{O}(10^7)$ free parameters, corresponding to the discretised volume elements of the observed domain. The HMC implementation is detailed in Jasche and Kitaura (2010) and Jasche and Wandelt (2013). The essence of BORG is that it incorporates the joint inference of initial conditions, and consequently the corresponding non-linearly evolved density fields and associated velocity fields, from incomplete observations. An augmented variant, BORG-PM, employing a particle mesh model for gravitational structure formation, has recently been presented (Jasche and Lavaux, 2018). An extension to BORG has also been developed to constrain cosmological parameters via a novel application of the Alcock-Paczyński test (Ramanah et al., 2019).

For the implementation of the robust likelihood, the HMC method that constitutes the basis of the joint sampling framework requires the negative log-likelihood and its adjoint gradient, which are given by

$$\begin{aligned}\Psi &\equiv -\log P(N|\bar{\lambda}) \\ &= \sum_{\alpha} N_{\alpha} \log\Big(\sum_{\beta} \bar{\lambda}_{\beta}\Big) - \sum_{\alpha} N_{\alpha} \log \bar{\lambda}_{\alpha},\end{aligned}$$

and

$$\frac{\partial \Psi}{\partial \bar{\lambda}_{\gamma}} \frac{\partial \bar{\lambda}_{\gamma}}{\partial \rho} = \frac{\bar{\lambda}_{\gamma}}{\rho}\Big(b + \epsilon \rho_g \rho^{-\epsilon}\Big)\left[\frac{\sum_{\alpha} N_{\alpha}}{\sum_{\beta} \bar{\lambda}_{\beta}} - \frac{N_{\gamma}}{\bar{\lambda}_{\gamma}}\right]. \tag{7.11}$$

The labelling of voxels with the same foreground modulation is encoded via a colour indexing scheme that groups the voxels into a collection of angular patches. This requires the construction of a sky map which is divided into regions of a given angular scale, where each region is identified by a specific colour and is stored in `HEALPix` format (Górski et al., 2005), as illustrated in Fig. 7.1. An extrusion of the sky map onto a three-dimensional (3d) grid subsequently yields a 3d distribution of patches, with a particular slice of this 3d coloured grid displayed in Fig. 7.2. The collection of voxels belonging to a particular patch is employed in the computation of the robust likelihood given by eq. (7.11), where α corresponds to the colour index.

This is a maximally ignorant approach to deal with unknown systematic errors where we enforce that every modulation above a given angular scale is not known. Since the colouring scheme does not depend on any foreground information, the numerical implementation of the likelihood is therefore generic. Moreover, another advantage of our approach is that the other components in our forward modelling scheme do not require any adjustments

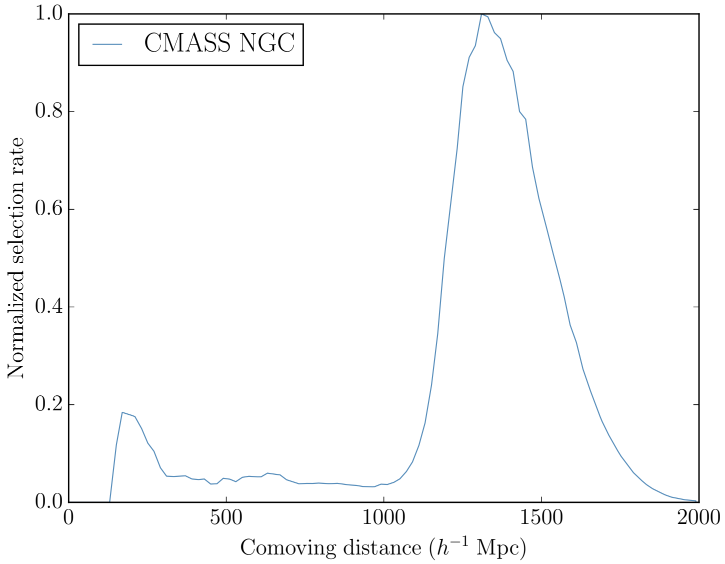


Figure 7.3: Radial selection function for the CMASS (north galactic cap) survey which is used to generate the mock data to emulate features of the actual SDSS-III BOSS data.

to encode this data model. However, we have not considered additive contaminations typically emanating from stars. We defer the extension of our data model to account for such additive contaminants to a future investigation.

7.4 Mock generation

We provide a brief description of the generation of the mock data set used to test the effectiveness of our novel likelihood, essentially based on the procedure adopted in Jasche and Kitaura (2010) and Jasche and Wandelt (2013). We first generate a realisation for the initial density contrast δ^{ic} from a zero-mean normal distribution with covariance corresponding to the cosmological power spectrum, such that we have a 3d Gaussian initial density field in a cubic equidistant grid with $N_{\mathrm{side}} = 256$, consisting of 256^3 voxels, where each voxel corresponds to a discretised volume element, and comoving box length of 2000 h^{-1} Mpc. This 3d distribution of initial conditions must then be scaled to a cosmological scale factor of $a_{\mathrm{init}} = 0.001$ using a cosmological growth factor $D^+(a_{\mathrm{init}})$.

The underlying cosmological power spectrum, including baryonic acoustic oscillations, for the matter distribution is computed using the prescription described in Eisenstein and Hu (1998, 1999). We assume a standard Λ cold dark matter (ΛCDM) cosmology with the set of cosmological parameters ($\Omega_{\mathrm{m}} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_{\mathrm{b}} = 0.0486$, $h = 0.6774$, $\sigma_8 = 0.8159$, $n_{\mathrm{s}} = 0.9667$) from Planck Collaboration et al. (2016b). We then employ LPT to transform the initial conditions into a non-linearly evolved field δ^{f} at redshift $z = 0$, which is subsequently constructed from the resulting particle distribution via the cloud-in-cell (CIC) method (e.g. Hockney and Eastwood, 1988).

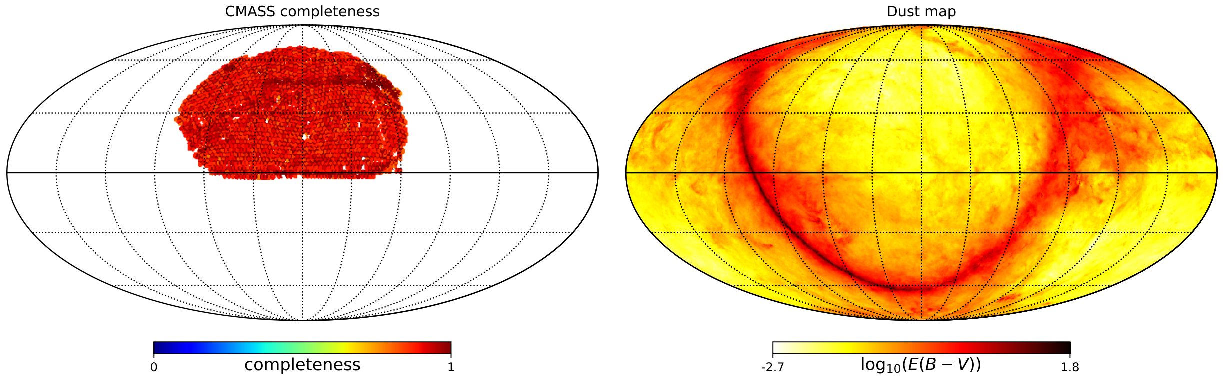


Figure 7.4: Observed sky completeness (*left panel*) of the CMASS component of the SDSS-III survey for the north galactic cap and dust extinction map (*right panel*) used to generate the large-scale contamination. This reddening map has been generated from the SFD maps (Schlegel et al., 1998).

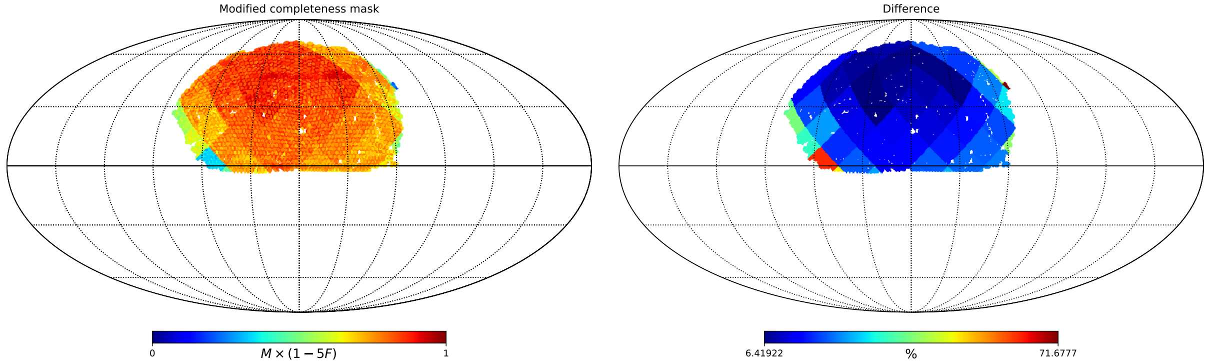


Figure 7.5: Contaminated completeness mask (*left panel*) and percentage difference compared to the original completeness mask (*right panel*). The contamination is introduced by multiplying the original mask by a factor of $(1-5F)$ where F is a foreground template, in this case, the dust extinction map downgraded to the angular resolution of the colour indexing map depicted in Fig. 7.1. The factor $\alpha = 5$ is chosen such that the mean contamination is 15%, an arbitrary choice to ensure that the contaminations are significant in the completeness mask. The difference between the original and contaminated masks shows that the effect is stronger on the edges of the survey.

Given the final density field δ^{f}, we generate a mock galaxy redshift catalogue subject to foreground contamination. For the test case considered in this work, we generate a data set that emulates the characteristics of the SDSS-III survey, in particular, the highly structured survey geometry and selection effects. We use a numerical estimate of the radial selection function of the CMASS component of the SDSS-III survey, shown in Fig. 7.3, obtained by binning the corresponding distribution of tracers $N(d_{\mathrm{com}})$ in the CMASS sample (e.g. Ross et al., 2017), where d_{com} is the comoving distance from the observer. The CMASS radial selection function is therefore estimated from a histogram of galaxy distribution over redshift. The procedure to construct the CMASS sky completeness is less trivial, however. We derive this CMASS mask, depicted in the left panel of Fig. 7.4, from the SDSS-III BOSS Data Release 12 (Alam et al., 2015) database by taking the ratio of spectroscopically confirmed galaxies to the target galaxies in each polygon from the mask.

In order to emulate large-scale foreground contamination, we construct a reddening map that describes dust extinction, illustrated in the right panel of Fig. 7.4. This dust template is derived from the data provided by Schlegel et al. (1998) via straightforward interpolation, rendered in HEALPix format (Górski et al., 2005)[1]. The contamination is produced by multiplying the completeness mask of CMASS, shown in the left panel of Fig. 7.4, by a factor of $(1 - \eta F)$, where F is the foreground template rescaled to the angular resolution of the colour indexing scheme, and η controls the amplitude of this contamination. To obtain a mean contamination of 15% in the completeness, we arbitrarily chose $\eta = 5$ to ensure that the foreground contaminations are significant. This mean value corresponds to the average contamination per element of the sky completeness. Figure 7.5 shows the contaminated sky completeness and the percentage difference, with the edges of the survey being more affected by the contamination due to their proximity to the galactic plane where the dust is more abundant. The mock catalogue is produced by drawing random samples from the inhomogeneous Poissonian distribution described by eq. (7.1) and using the modified completeness.

7.5 Results and discussion

In this section, we discuss results obtained by applying the BORG algorithm with the robust likelihood to contaminated mock data. We also compare the performance of our novel likelihood with that of the standard Poissonian likelihood typically employed in large-scale structure analyses. In order to test the effectiveness of our likelihood against unknown systematic errors and foreground contaminations, the algorithm is agnostic about the contamination and assumes the CMASS sky completeness depicted in the left panel of Fig. 7.4.

We first study the impact of the large-scale contamination on the inferred non-linearly evolved density field. To this end, we compare the ensemble mean density fields and corresponding standard deviations for the two Markov chains with the Poissonian and novel likelihoods, respectively, illustrated in the top and bottom panels of Fig. 7.6, for

[1] The construction of this template is described in more depth in Section 3 of Jasche and Lavaux (2017).

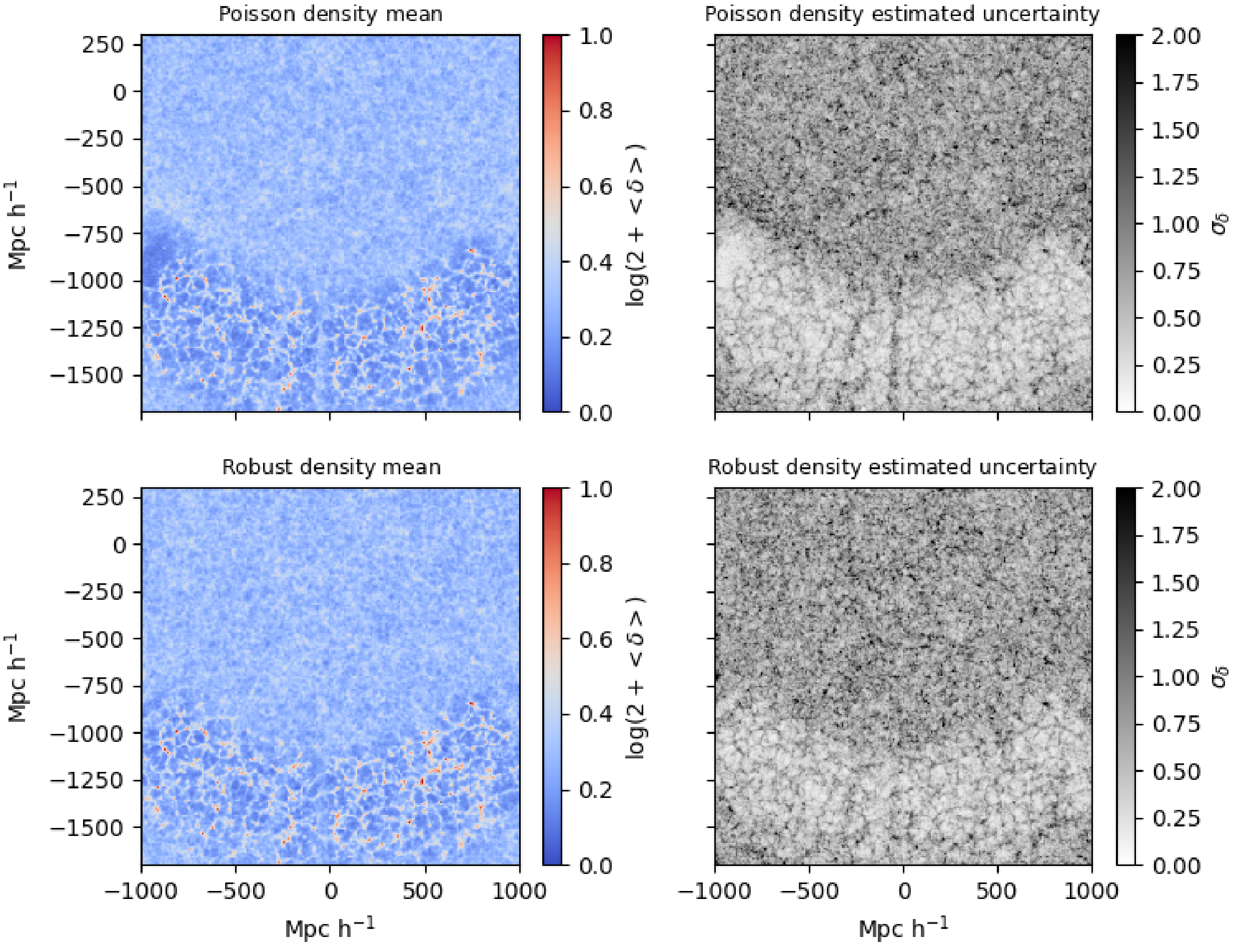
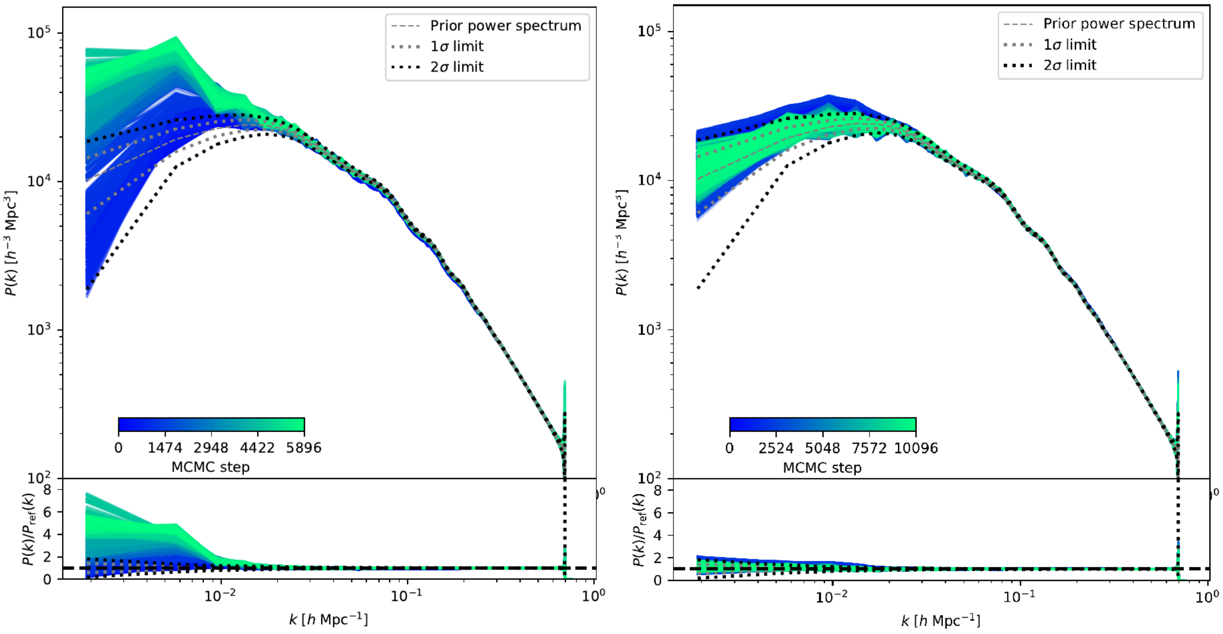


Figure 7.6: Mean and standard deviation of the inferred non-linearly evolved density fields, computed from the MCMC realisations, with the same slice through the 3d fields being depicted above for both the Poissonian (upper panels) and augmented (lower panels) likelihoods. The filamentary nature of the non-linearly evolved density field can be observed in the regions constrained by the data, with the unobserved or masked regions displaying larger uncertainty, as expected. Unlike our robust data model, the standard Poissonian analysis yields some artefacts in the reconstructed density field, particularly near the edges of the survey, where the foreground contamination is stronger.



(a) Standard Poissonian likelihood

(b) Robust likelihood

Figure 7.7: Reconstructed power spectra from the inferred initial conditions from a BORG analysis with unknown foreground contamination for the robust likelihood (left panel) and the Poissonian likelihood (right panel) over the full range of Fourier modes considered in this work. The σ limit corresponds to the cosmic variance $\sigma = \sqrt{1/k}$. The colour scale shows the evolution of the power spectrum with the sample number. The power spectra of the individual realisations, after the initial burn-in phase, from the robust likelihood analysis, possess the correct power across all scales considered, demonstrating that the foregrounds have been properly accounted for. In contrast, the standard Poissonian analysis exhibits spurious power artefacts due to the unknown foreground contaminations, yielding excessive power on these scales.

a particular slice of the 3d density field. As can be deduced from the top-left panel of Fig. 7.6, the standard Poissonian analysis results in spurious effects in the density field, particularly close to the boundaries of the survey since these are the regions that are the most affected by the dust contamination. In contrast, our novel likelihood analysis yields a homogeneous density distribution through the entire observed domain, with the filamentary nature of the present-day density field clearly seen. While we can recover well-defined structures in the observed regions, the ensemble mean density field tends towards the cosmic mean density in the masked or poorly observed regions, with the corresponding standard deviation being higher to reflect the larger uncertainty in these regions. From this visual comparison, it is evident that our novel likelihood is more robust against unknown large-scale contaminations.

From the realisations of our inferred 3d initial density field, we can reconstruct the

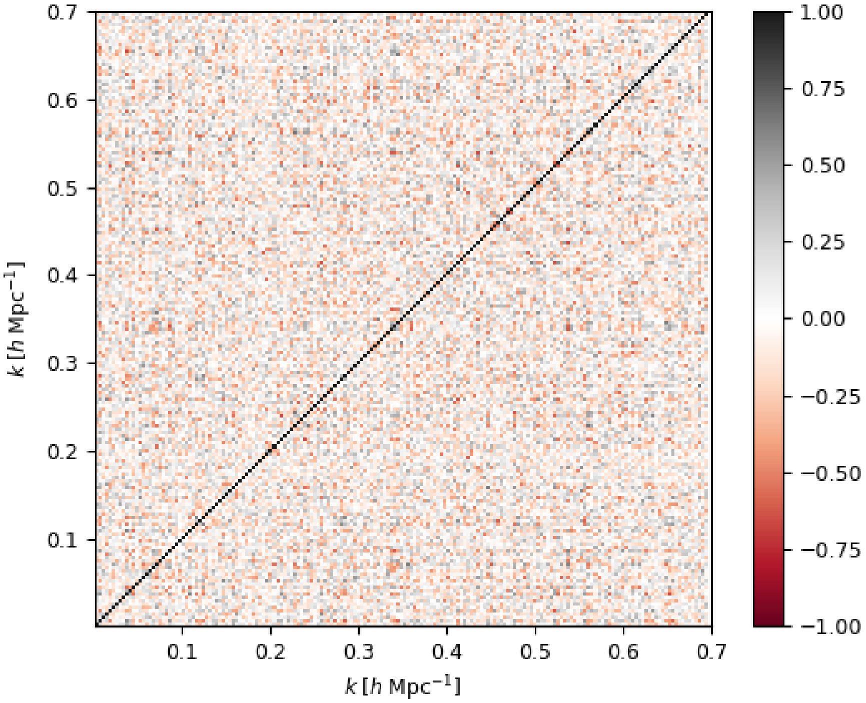


Figure 7.8: Correlation matrix of power spectrum amplitudes with respect to the mean value for the robust likelihood normalised using the variance of amplitudes of the power spectrum modes. The correlation matrix shows that our augmented data model does not introduce any spurious correlation artefacts, thereby implying that it has properly accounted for the selection and foreground effects.

corresponding matter power spectra and compare them to the prior cosmological power spectrum adopted for the mock generation. The top panel of Fig. 7.7 illustrates the inferred power spectra for both likelihood analyses, with the bottom panel displaying the ratio of the *a posteriori* power spectra to the prior power spectrum. While the standard Poissonian analysis yields excessive power on the large scales due to the artefacts in the inferred density field, the analysis with our novel likelihood allows us to recover an unbiased power spectrum across the full range of Fourier modes.

In addition, we tested the combined effects of the foreground and unknown noise amplitudes by estimating the covariance matrix of the Fourier amplitudes of the reconstructed power spectra. As depicted in Fig. 7.8, our novel likelihood exhibits uncorrelated amplitudes of the Fourier modes, as expected from ΛCDM cosmology. The strong diagonal shape of the correlation matrix indicates that our proposed data model correctly accounted for any mode coupling introduced by survey geometry and foreground effects.

The above results clearly demonstrate the efficacy of our proposed likelihood in robustly dealing with unknown foreground contaminations for the inference of non-linearly evolved dark matter density fields and the underlying cosmological power spectra from deep galaxy redshift surveys. This method can be inverted to constrain foreground properties of the contamination. The inferred dark matter density allows for galaxy catalogues to be built without contaminations. These can be compared to the observed number counts to reconstruct the foreground properties as the mismatch between the two catalogues.

7.6 Summary and conclusions

The increasing requirement to control systematic and stochastic effects to high precision in next-generation deep galaxy surveys is one of the major challenges for the coming decade of surveys. If not accounted for, unknown foreground effects and target contaminations will yield significant erroneous artefacts and bias cosmological conclusions drawn from galaxy observations. A common spurious effect is an erroneous modulation of galaxy number counts across the sky, hindering the inference of 3d density fields and associated matter power spectra.

To address this issue, we propose a novel likelihood to implicitly and efficiently account for unknown foreground and target contaminations in surveys. We described its implementation in a framework of non-linear Bayesian inference of large-scale structures. Our proposed data model is conceptually straightforward and easy to implement. We illustrated the application of our robust likelihood to a mock data set with significant foreground contaminations and evaluated its performance via a comparison with an analysis employing a standard Poissonian likelihood to showcase the contrasting physical constraints obtained with and without the treatment of foreground contamination. We have shown that foregrounds, when unaccounted for, lead to spurious and erroneous large-scale artefacts in density fields and corresponding matter power spectra. In contrast, our novel likelihood allows us to marginalise over unknown large-angle contamination amplitudes, resulting in a homogeneous inferred density field, thereby recovering the fiducial power spectrum amplitudes.

We are convinced that our approach will contribute to optimising the scientific returns of current and coming galaxy redshift surveys. We have demonstrated the effectiveness of our robust likelihood in the context of large-scale structure analysis. Our augmented data model remains nevertheless relevant for more general applications with other cosmological probes, with applications potentially extending even beyond the cosmological context.

Chapter 8

Inferring high-redshift large-scale structure dynamics from the Lyman-α forest

The material displayed in this chapter has been submitted for publication to Astronomy & Astrophysics (Porqueres et al., 2019a).

I led this research project as the principal investigator and authored the corresponding publication as the first author. My contributions consisted of the method development, the implementation and testing of the method, generation of artificial mock data, and preparation of the material for the publication. The project was further done in collaboration with Jens Jasche (JJ), Guilhem Lavaux (GL), and Torsten Enßlin who contributed in discussions about the method and its implementation. GL and JJ provided support in implementing the method and improving the numerical scalability of the algorithm. All authors provided feedback on the text and accepted the final manuscript.

8.1 Introduction

Currently, cosmology is at the crossroads. While the standard model of cosmology ΛCDM fits the bulk of cosmological observations to extraordinary accuracy, some tensions between the model and observations seem to persist and increase (Planck Collaboration et al., 2016b, 2019). Amongst those are the H_0 and σ_8 tensions (e.g. Planck Collaboration et al., 2016a; Riess et al., 2016; Köhlinger et al., 2017; Riess et al., 2018; Abbott et al., 2018; Rusu et al., 2019). Additionally, Lyman-α auto-correlation and cross-correlations with quasar data reported a 2.3σ tension with the flat ΛCDM prediction obtained from Planck observations (see e.g. Delubac et al., 2015; du Mas des Bourboux et al., 2017).

The resolution of these tensions may encompass systematic effects but may also be the first signs of new physics indicated by novel cosmological data of increasing quality. New observations and better control on systematic effects in data analyses are inevitable to gain new insights into the physical processes driving the evolution of the universe.

For this reason, in this work, we present a novel and statistically rigorous approach to extract cosmologically relevant and significant information from high-redshift Lyman-α forest observations tracing the dynamic evolution of cosmic structures.

Detailed analyses of the spatial distribution of matter and growth of cosmic structures can provide significant information to discriminate between models of homogeneous dark energy and modifications of gravity, determine neutrino masses and their mass hierarchy and investigate the dynamical clustering behaviour of warm or cold dark matter (e.g. Colombi et al., 1996; Huterer et al., 2015; Basilakos and Nesseris, 2016; Frenk and White, 2012; Boyle and Komatsu, 2018; Mishra-Sharma et al., 2018).

To harvest this information, cosmology now turns to analyze the inhomogeneous matter distribution with next-generation galaxy surveys such as the Large Synoptic Survey Telescope (LSST) and the Euclid satellite mission probing the galaxy distribution out to redshifts $z \sim 3$ and beyond (LSST Science Collaboration et al., 2009; Refregier et al., 2010; Laureijs et al., 2011; Alam et al., 2016).

While most of this research focuses on studying the cosmic large-scale structure with galaxy clustering and weak lensing observations, the Lyman-α (Ly-α) forest has the potential to provide important complementary information. Firstly, the Ly-α forest probes the matter distribution at higher redshift with higher spatial resolution than can be achieved with galaxy sampling rates. By probing scales down to 1 Mpc, the Ly-α forest is sensitive to neutrino masses and dark matter models (Viel et al., 2013; Rossi, 2014; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Yèche et al., 2017). Secondly, while galaxy clustering probes high-density regions, the Ly-α forest is particularly sensitive to under-densities in the matter distribution (Peirani et al., 2014; Sorini et al., 2016). In addition, the Ly-α at $z > 2$ is redshifted to an optical band for which the atmosphere is transparent. The Ly-α emission becomes then one of the more powerful probes at redshifts that are otherwise challenging to access with ground-based galaxy surveys (Lee, 2016).

Specifically, auto-correlations of the Ly-α forest along the line of sight have been used to infer cosmological parameters which agree by and large with CMB observations (e.g Seljak et al., 2006; Viel et al., 2006; Slosar et al., 2011; Busca et al., 2013; Bautista et al., 2017; Blomqvist et al., 2019). Ly-α forest data has also been used to constrain neutrino masses (Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Yèche et al., 2017), test warm dark matter models (Viel et al., 2013) and study the thermal history of the intergalactic medium (Nasir et al., 2016; Boera et al., 2019).

In line with these promises, a large number of ongoing surveys mapping the Ly-α forest at higher redshifts has been started. The eBOSS observations in the SDSS DR14 are expected to provide the spectra of 435 000 quasars over 7500 deg^2 and re-observe the lines of sight from the previous data release with low signal-to-noise (SNR$<$ 3, Myers et al., 2015). DESI (Dark Energy Spectroscopic Instrument, Levi et al., 2013; DESI Collaboration et al., 2016) will map the Ly-α forest over a larger fraction of the sky (14 000 deg^2), targeting 50 high-redshift quasars per deg^2. Increasing the density of tracers, the ongoing CLAMATO survey (COSMOS Lyman Alpha Mapping and Tomography, Lee et al., 2018) maps the Ly-α forest over a small part of the sky (600 arcmin^2) with a separation between lines of sight of 2.4 h^{-1} Mpc, 20 times smaller than the BOSS survey (Eisenstein et al., 2011; Dawson

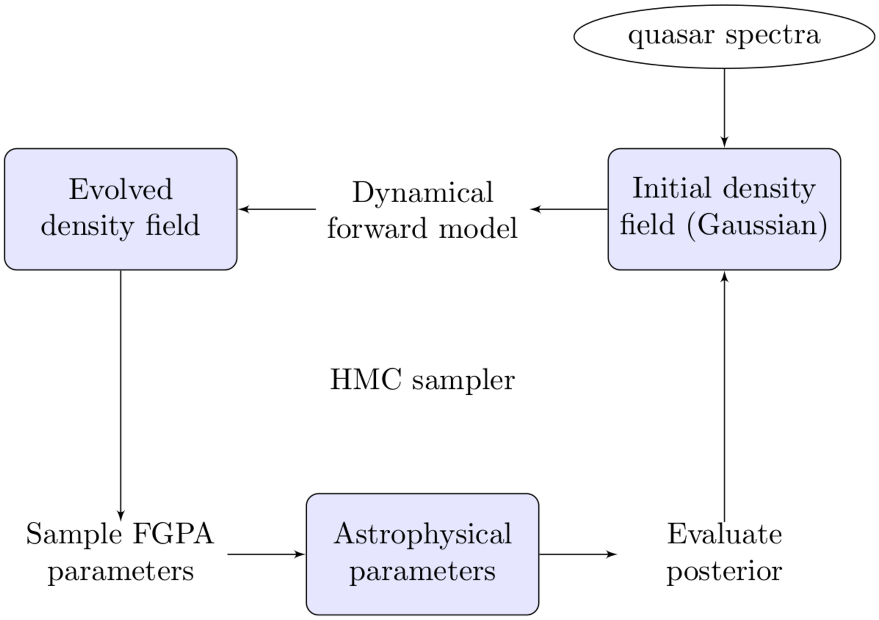


Figure 8.1: Flow chart depicting the iterative block sampling approach of the BORG inference framework for the Ly-α forest. As can be seen, the BORG algorithm first infers three-dimensional initial and evolved density fields from quasar spectra using assumed astrophysical parameters of the FGPA model. Then, using realizations of the evolved density field, the algorithm will update these astrophysical parameters and a new iteration of the process begins. Iterating this procedure results in a valid Markov Monte Carlo Chain that will correctly explore the joint posterior distribution of the three-dimensional matter distribution and astrophysical properties underlying Ly-α forest observations.

et al., 2013) and smaller than the expected separation in DESI (3.7h^{-1} Mpc, Horowitz et al., 2019). Future surveys like MSE (Maunakea Spectrographic Explorer McConnachie et al., 2016) and LSST (LSST Science Collaboration et al., 2009) will increase the amount of available Ly-α forest observations by a factor of 10.

At present, major analyses of the Ly-α forest focus only on the analysis of the matter power spectrum (e.g. Croft et al., 1998; Seljak et al., 2006; Viel et al., 2006; Bird et al., 2011; Palanque-Delabrouille et al., 2015; Rossi et al., 2015; Nasir et al., 2016; Yèche et al., 2017; Boera et al., 2019). However, these approaches ignore significant amounts of information contained in the higher-order statistics of the matter density field as generated by non-linear gravitational dynamics in the late time universe (He et al., 2018).

Capturing the full information content of the cosmic large-scale structure requires a field-based approach to infer the entire three-dimensional cosmic large-scale structure from observations. This poses a particular challenge for the analyses of Ly-α forest observations, which provide sparse inherently one-dimensional information along the lines of sight. Various approaches to perform three-dimensional density reconstructions from one-dimensional Ly-α forests have been proposed in the literature (e.g. Kitaura et al., 2012b; Cisewski et al., 2014; Stark et al., 2015b; Ozbek et al., 2016; Horowitz et al., 2019). Gallerani et al. (2011) and Kitaura et al. (2012b) proposed a Gibbs sampling scheme to jointly infer density and velocity fields and corresponding power-spectra. However, these approaches assume matter density amplitudes to be log-normally distributed. The lognormal distribution reproduces one- and two-point statistics but fails to reproduce higher-order statistics associated with the filamentary dark matter distribution. In an attempt to extrapolate information from one-dimensional quasar spectra into the three-dimensional volume, Cisewski et al. (2014) applied a local polynomial smoothing method. Ozbek et al. (2016) and Stark et al. (2015b) employed a Wiener filtering approach to reconstruct the three-dimensional density field between lines of sight of Ly-α forest data. In order to reproduce higher order statistics, Horowitz et al. (2019) recently used a large-scale optimization approach to fit a gravitational structure growth model to Ly-α data, showing that this approach allows recovering the more filamentary structure of the cosmic web. Although the approach improves over linear and isotropic Wiener filtering approaches, it shows systematic deviations of reconstructed matter power-spectra and underestimates density amplitudes at scales corresponding to the mean separation between lines of sight (Horowitz et al., 2019).

To go beyond previous approaches, in this work, we present a fully Bayesian and statistically rigorous approach to perform dynamical matter clustering analyses with high redshift Ly-α forest data while accounting for all uncertainties inherent to the observations. The aim of this work is to provide a novel and fully Bayesian approach to infer physically plausible three-dimensional density and velocity fields from Ly-α forest data. Our approach builds upon the algorithm for Bayesian Origin Reconstruction from Galaxies (BORG, Jasche and Wandelt, 2013; Jasche and Lavaux, 2018), which employs physical models of structure formation and sophisticated Markov Chain Monte Carlo techniques to optimally extract large-scale structure information from data and quantify corresponding uncertainties.

To make such inferences feasible, we developed a likelihood based on the fluctuating Gunn-Peterson approximation (FGPA, Gunn and Peterson, 1965) and jointly constrained the astrophysical properties of the intergalactic medium. We tested and validated our approach with simulated data emulating the CLAMATO survey, showing that the algorithm recovers the unbiased dark-matter field.

The paper is organized as follows. Section 8.4 provides a brief overview of our Bayesian inference framework, BORG, as required for this work. In Section 8.2, we discuss the physics of the Ly-α forest underlying the data model described in Section 8.3. Section 8.5 describes the generation of simulated data employed to test and validate the algorithm. The performance of the algorithm is tested in Section 8.6 and the results are shown in Section 8.7, where we also discuss possible applications of the method to scientifically exploit the Ly-α forest. Finally, Section 8.8 summarizes the results and discusses further extensions of the algorithm.

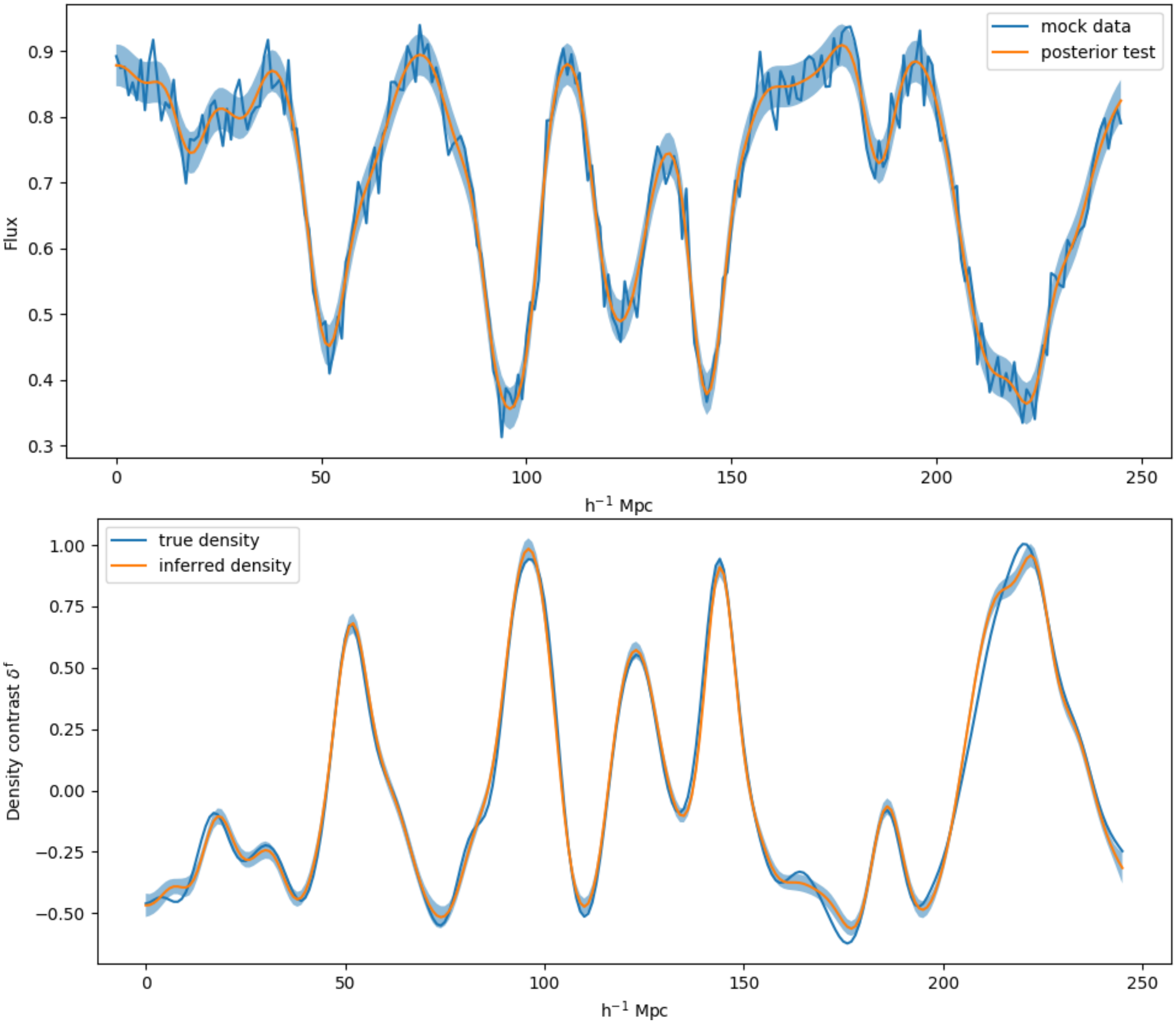


Figure 8.2: Top panel: Posterior prediction to test the model. These tests check whether the data model can accurately account for the observations. Any significant mismatch would immediately indicate a breakdown of the applicability of the data model or error of the inference framework. Our method recovers the transmitted flux fraction correctly, confirming that the data model can accurately account for the observations. Bottom panel: Comparison of the inferred ensemble mean density field along the line of sight to the ground truth. It can be seen that high-density regions yield a suppression of transmitted flux, while underdense regions transmit the quasar signal, in agreement with the FGPA model.

8.2 The physics of the Lyman-α forest

Developing a Bayesian framework to infer the matter distribution from the Ly-α forest requires to incorporate the corresponding light absorption physics into the data model.

When traversing the universe, photons emitted from quasars will be scattered by neutral hydrogen (HI) gas residing inside cosmic large-scale structures. Scattering of quasar light results in observed spectra covered with absorption features referred to as the Ly-α forest. While high-density regions obscure the light and attenuate quasar fluxes, under-dense regions are almost transparent. Therefore, the signal comes from the under-dense regions. Due to the cosmological redshift, different HI regions absorb photons from different wavelengths in the quasar spectrum (see e.g. Mo et al., 2010). Consequently, every HI absorber leaves an absorption feature to the spectrum, permitting to trace the distance and density of HI regions along the line of sight. The Ly-α forest, therefore, provides a formidable probe of the cosmic large-scale structure along the observers past light cone.

Since the Ly-α forest generated at $z > 2$ is redshifted to the optical atmospheric window, it can be observed by ground-based telescopes, becoming one of the more relevant probes at redshifts that are otherwise challenging to access with galaxy surveys. Observations of the Ly-α forest at lower redshift, from $z = 1.5$ down to $z = 0$ (Davé et al., 1999; Davé, 2001; Williger et al., 2010), require UV space-based spectrographs as Faint Object Spectrograph in the Hubble Space Telescope (Bahcall et al., 1993, 1996; Weymann et al., 1998).

In contrast to galaxy surveys, requiring assumptions on galaxy formation to model galaxy biases, modelling the Ly-α forest does not involve complicated models. At $z > 2$, HI gas, producing the Ly-α forest, is in radiative equilibrium of photoionization due to ultraviolet (UV) background and adiabatic cooling due to the cosmic expansion (Hui and Gnedin, 1997). This yields a tight relation between the temperature of the IGM and the dark matter density ρ

$$T = T_0 \left(\frac{\rho}{\bar{\rho}} \right)^{\gamma - 1}, \tag{8.1}$$

where T_0 and γ are constants that depend on the reionization history and the spectral shape of the UV background sources (Hui and Gnedin, 1997).

In thermal equilibrium, HI number density can be expressed as (Hui and Gnedin, 1997)

$$n_{HI}(\mathbf{x}) = \frac{\alpha(T(\mathbf{x}))}{\Gamma} \left[n_0(1+z)^3(1+\delta(\mathbf{x})) \right]^2 \propto (1+\delta(\mathbf{x}))^{\beta}, \tag{8.2}$$

where $\mathbf{x}$ is the comoving position, $\alpha(T) \propto T^{-0.7}$ is the radiative recombination rate of HI, Γ is the photoionization rate of neutral hydrogen, $n_0(1+z)^3$ is the mean baryon number density at redshift z and δ is the dark matter density contrast. On scales larger than Jeans' scale (100 kpc), thermal pressure in the gas is negligible and the IGM traces the dark matter distribution with high accuracy (Peirani et al., 2014).

Combining equations (8.1) and (8.2) leads to the fluctuating Gunn-Peterson approximation (FGPA Gunn and Peterson, 1965) for the transmitted flux:

$$F(z, \hat{\mathbf{x}}) = \exp\left[-A(1+\delta(\mathbf{x}))^{\beta} \right], \tag{8.3}$$

with z the considered absorption redshift, $\mathbf{x}$ the corresponding comoving distance, $\hat{\mathbf{x}}$ the associated unit vector, $\beta = 2 - 0.7(\gamma - 1)$ and $A \propto (1+z)^6 T_0^{-0.7} \Gamma^{-1}$. Therefore, the astrophysical properties of HI gas are encoded in the parameters A and β of the FGPA model.

8.3 A data model for the Lyman-α forest

To incorporate Ly-α forest observations into the Bayesian framework, we will now build a data model using the light absorption physics described in Sect. 8.2. Specifically, we assume the FGPA transmission model and Gaussian pixel noise for measured quasar spectra. A corresponding likelihood distribution can then be expressed as:

$$P(\delta^{\mathrm{f}}|F) = \prod_{n,x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left((F_n)_x - \exp\left[-A(1+\delta^{\mathrm{f}}_x)^{\beta} \right] \right)^2}{2\sigma^2} \right], \tag{8.4}$$

where n labels respective quasar spectra, x indexes different volume elements intersected by the n-th quasar line of sight and δ^{f} is the non-linear final density contrast evaluated at $z = 2.5$.

This likelihood is then implemented into the large-scale structure sampler of the BORG framework. The corresponding physical forward modelling approach will then proceed as follows. Using realizations of the three-dimensional field of primordial fluctuations, the dynamical structure formation model will evaluate non-linear realizations of the dark matter distribution at $z = 2.5$. Using these matter field realizations and the FGPA data model, BORG will predict quasar spectra that are compared to actual observations via the Gaussian likelihood of equation (8.4). More details on the sampling procedure are given in Chapter 5. To efficiently implement this non-linear and non-Gaussian forward modelling approach into an MCMC framework, we employed a Hamiltonian Monte Carlo (HMC) method, as described in Chapter 4.

In this work, we focus on the inference of the spatial dark matter distribution and its dynamics. For this reason, we will treat the parameters A and β mostly as nuisance parameters. Following similar approaches described in previous works (Jasche and Wandelt, 2013; Jasche and Lavaux, 2017; Ramanah et al., 2019; Jasche and Lavaux, 2018), we will use efficient slice sampling techniques to sample these parameters. We note that these parameters can carry valuable information on the astrophysical properties of the IGM. As illustrated in Section 8.7.6, our inference can be used to also make statements on the astrophysics of the IGM.

The entire iterative sampling approach is depicted in Fig. 8.1. As can be seen, the method iteratively infers three-dimensional matter density fields and parameters of the FGPA model, resulting in a valid MCMC approach to jointly explore density fields and astrophysical parameters.

8.4 Method

As mentioned above, this work extends the previously developed BORG algorithm in order to analyze the spatial matter distribution underlying Ly-α forest observations. Here we will provide only a brief summary of the algorithm and corresponding concepts, but interested readers will find more detailed descriptions in our previous works (Jasche and Wandelt, 2013; Jasche et al., 2015; Lavaux and Jasche, 2016; Jasche and Lavaux, 2017, 2018).

The BORG algorithm is a large-scale Bayesian inference framework aiming at inferring the non-linear spatial dark matter distribution and its dynamics from cosmological data sets. The underlying idea is to fit full dynamical gravitational structure formation models to observations tracing the spatial distribution of matter. Using non-linear structure growth models, the BORG algorithm can exploit the full statistical power of higher-order statistics imprinted to the matter distribution by gravitational clustering. In fitting dynamical structure growth models to data, the task of inferring non-linear matter density fields turns into a statistical initial conditions problem aiming at inferring the spatial distribution of primordial matter fluctuations from which present structures formed via gravitational structure growth. As such the BORG algorithm naturally links primordial initial conditions to late time observations and permits to infer dynamical properties as well as the structure formation history from present galaxy observations (Jasche and Wandelt, 2013; Jasche et al., 2015; Lavaux and Jasche, 2016; Jasche and Lavaux, 2017, 2018).

The BORG algorithm employs a large-scale structure posterior distribution based on the well-developed prior understanding of almost Gaussian primordial density fluctuations and gravitational structure formation to predict physically plausible realizations of present matter density fields. More specifically BORG encodes a Gaussian prior for the initial density contrast at an initial cosmic scale factor of $a = 10^{-3}$. Initial and evolved density fields are linked by deterministic gravitational evolution mediated by various physics models of structure growth. Specifically, BORG incorporates physical models based on Lagrangian Perturbation Theory (LPT) but also fully non-linear particle mesh (PM) models.

Our Bayesian inference approach has been previously shown to perform accurate dynamic mass estimation and provide mass measurements that agree well with complementary standard weak lensing and X-ray observations (Jasche and Lavaux, 2018). More recently, we have shown, that BORG can exploit the geometric shapes of the cosmic large-scale structure to significantly improve constraints on cosmological parameters via the Alcock-Paczynski test (Ramanah et al., 2019). Additional projects, conducted with the BORG algorithm, can be found at our web-page of the Aquila consortium[1].

In this work, we will rely on an LPT approach to structure formation since the Ly-α forest mostly arises from under-dense regions that can be conveniently modelled by perturbation theory (Peirani et al., 2014; Sorini et al., 2016). The dynamical model permits to recover the non-linear higher-order statistics associated with the filamentary matter distribution of the cosmic large-scale structure. Our approach also immediately provides detailed inferences of large-scale velocity fields and structure growth information at high

[1] `https://aquila-consortium.org`

redshifts as will be demonstrated in the remainder of this work.

A feature of particular relevance to this work is that BORG employs a modular statistical programming engine that executes a statistically rigorous Markov Chain to marginalize out any nuisance parameters associated to the data model, such as unknown biases or systematic effects. This statistical programming engine permits us to easily and straightforwardly implement complex data models. In particular, in this work, we will perform joint analyses of the cosmological large-scale structure and unknown astrophysical properties of the IGM medium described by the parameters of the FGPA model. The corresponding iterative block sampling inference approach is illustrated in Fig. 8.1.

8.5 Generating artificial Ly-α forest observations

To test our Ly-α forest inference framework we will generate artificial mock observations emulating the CLAMATO survey (Stark et al., 2015a; Lee et al., 2018).

Mock data is constructed by first generating Gaussian initial conditions on a cubic Cartesian grid of side length of $256h^{-1}$ Mpc with a resolution of $1h^{-1}$ Mpc. To generate primordial Gaussian density fluctuations we use a cosmological matter power-spectrum including the Baryonic wiggles calculated according to the prescription provided by (Eisenstein and Hu, 1998, 1999). We further assume a standard ΛCDM cosmology with the following set of parameters: $\Omega_m = 0.31$, $\Omega_\Lambda = 0.69$, $\Omega_b = 0.022$, $h = 0.6777$, $\sigma_8 = 0.83$, $n_s = 0.9611$ (Planck Collaboration et al., 2016b). We assumed $H = 100h$ km s^{-1} Mpc^{-1}.

To generate realizations of the non-linear density field, we evolve these Gaussian initial conditions via Lagrangian Perturbation Theory. This involves simulating displacements for 512^3 particles in the LPT simulation. Final density fields are constructed by estimating densities via the cloud-in-cell scheme from simulated particles on a Cartesian equidistant grid with 256^3 volume elements. We further apply a Gaussian smoothing kernel of $\sigma = 0.5h^{-1}$ Mpc to the fields to simulate the difference between dark matter and gas density fields (see e.g. Peirani et al., 2014; Stark et al., 2015a). A three dimensional quasar flux field is generated by applying the FGPA model, given in eq. (8.3), to the final density field and assuming constant parameters $A = 0.35$ and $\beta = 1.56$ at $z = 2.5$, corresponding to the values in Stark et al. (2015a).

From this three-dimensional quasar flux field, we generate individually observed skewers by tracing lines of sight through the volume. Specifically, we generate a total of 1024 lines of sight parallel to the z-axis of the box. The separation between lines of sight is the most limiting factor of Ly-α surveys. Although the CLAMATO survey achieves a separation of 2.4 h^{-1} Mpc, the lines of sight in this work are separated by 8 h^{-1} Mpc. This gives us the opportunity to explore the potential of our method to reconstruct the three-dimensional density field between lines of sight. Finally, we added Gaussian pixel-noise to the flux with a signal-to-noise ratio (SNR) of ≈ 5, which corresponds to a value in the SNR range of the CLAMATO survey ($\mathrm{SNR}_{\mathrm{min}} \approx 1.4, \mathrm{SNR}_{\mathrm{min}} \approx 10$ Lee et al., 2018). An example for one of these 1024 mock quasar spectra is illustrated in Fig. 8.2.

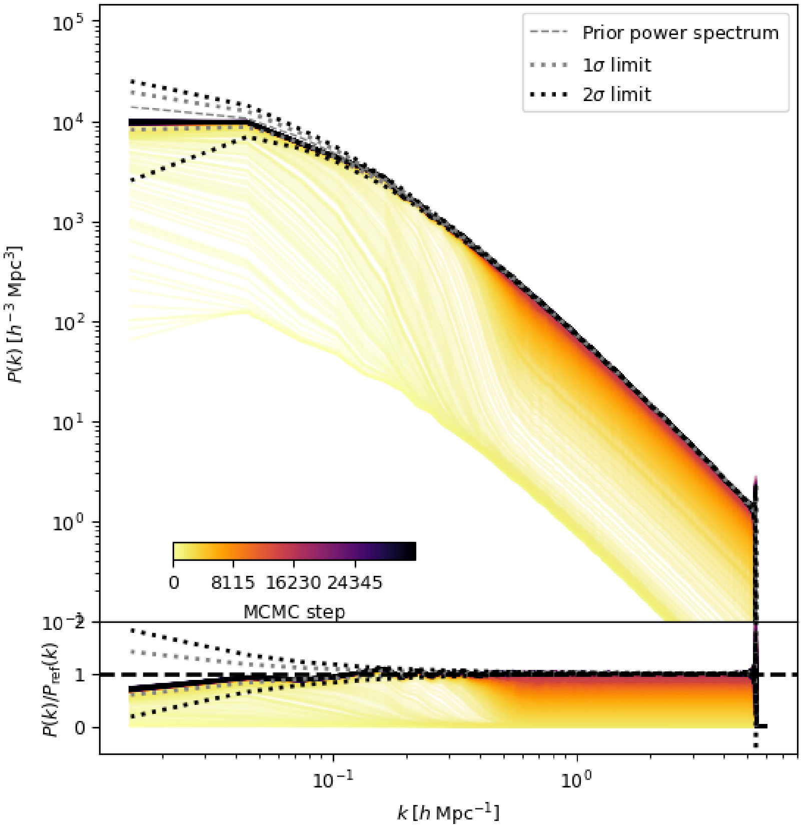


Figure 8.3: Burn-in of the posterior power spectra from the inferred initial conditions from a BORG analysis. The colour scale shows the evolution of the power spectrum with the sample number. The dashed lines indicate the underlying power spectrum and the 1- and 2-σ uncertainty limits. After the warm-up phase, the algorithm recovers the true matter power spectrum in all range of Fourier modes.

8.6 Testing sampler performance

In this Section, we present a series of tests to evaluate the performance of our method. In particular, we focus on the convergence (Section 8.6.1) and efficiency (Section 8.6.2) to infer the underlying dark matter density field. We provide tests of the posterior power-spectra and perform posterior predictive tests for quasar spectra to check that the inferred model agrees with the data.

8.6.1 The warm-up phase of the sampler

In the large sample limit, any properly set up Markov chain is guaranteed to approach a stationary distribution that provides an unbiased estimate of the target distribution. While Markov chains are typically started from a place remote from the target distribution after a finite amount of transition steps, the chain acquires a stationary state. Once the chain

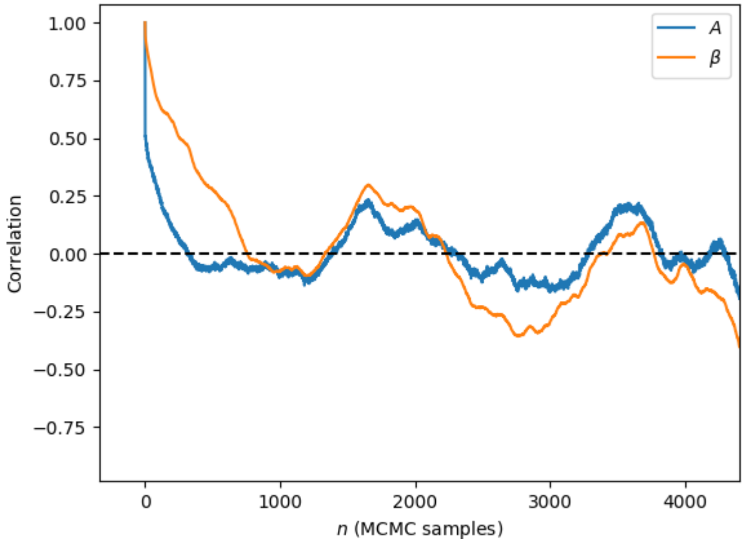


Figure 8.4: Autocorrelation of the parameters as a function of the sample lag in the Markov chain. The correlation length of the sampler can be estimated by determining the point when correlations drop below 0.1 for the first time. These parameters have a correlation length of 830 samples for β and 640 for A.

is in a stationary state, we may start recording samples to perform statistical analyses of the inference problem. To test when the Markov sampler has passed its initial warm-up phase, we follow a similar approach as described in previous works (Jasche and Wandelt, 2013; Jasche and Lavaux, 2017; Ramanah et al., 2019; Jasche and Lavaux, 2018; Porqueres et al., 2019b), by initializing the Markov chain with an over-dispersed state and trace the systematic drift of inferred quantities towards their preferred regions in parameter space. Specifically, we initialized the Markov chain with a random Gaussian cosmological density field scaled by a factor 10^{-3} and monitored the drift of corresponding posterior power-spectra during the initial warm-up phase. The results of this exercise are presented in Fig. 8.3. As can be seen, successive measurements of the posterior power-spectrum during the initial warm-up phase show a systematic drift of power-spectrum amplitudes towards their fiducial values. The sampler, therefore, correctly recovers the power of the initial density field and moves the chain towards regions of high-probability in the parameter space. By the end of the warm-up phase, the sampler has found an unbiased representation of the matter distribution at all Fourier modes considered in this work. Starting the sampler from an over-dispersed state, therefore, provides us with an important diagnostics to test the validity of the sampling algorithm.

8.6.2 Statistical efficiency of the sampler

By design, subsequent samples in Markov chains are correlated. The statistical efficiency of an MCMC algorithm is determined by the number of independent samples that can be drawn from a chain of a given length. To estimate the statistical efficiency of the sampler,

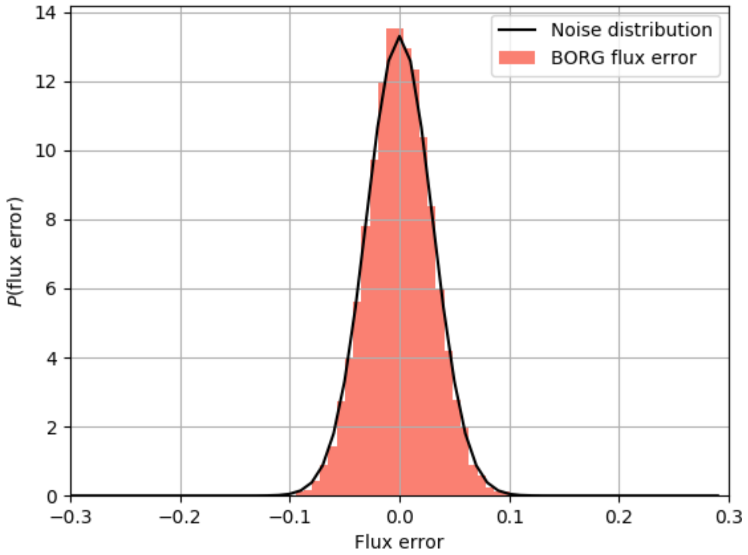


Figure 8.5: Histogram of the error in the fractional transmitted flux, which is computed as the difference between posterior-predicted fluxes and input spectra. The distribution of flux error matches the distribution of pixel noise in the data, indicating that the method is close to the theoretical optimum. This distribution of error flux can be compared to the one obtained in previous works, demonstrating that our method recovers the fluxes with significantly higher accuracy.

we estimate the correlation length of the astrophysical parameters A and β along the Markov chain. For a parameter θ the auto-correlation for samples with a given lag in the chain can be estimated as

$$C_n(\theta) = \frac{1}{N-n} \sum_{i=0}^{N-n} \frac{(\theta^i - \langle\theta\rangle)(\theta^{i+n} - \langle\theta\rangle)}{\mathrm{Var}(\theta)} \tag{8.5}$$

where n is the lag in MCMC samples, $\langle\theta\rangle$ is the mean and $\mathrm{Var}(\theta)$ is the variance. We typically determine the correlation length by estimating the lag n_C at which the auto-correlation C_n dropped below 0.1. The number n_C therefore present the number of transitions required to produce one independent sample. The results of this test are presented in Fig. 8.4. As can be seen, correlation lengths for parameters A and β amount to 640 and 830 samples, respectively. To significantly improve statistical efficiency, in the future, we will use similar strategies to obtain a faster mixing of the chain as described in Ramanah et al. (2019).

8.6.3 Posterior predictions for quasar spectra

To test whether inferred density fields provide accurate explanations for the observations, we perform a simple posterior predictive test (see e.g. Gelman et al., 2004). Generally, posterior predictive tests provide good diagnostics about the adequacy of data models in explaining observations and to identify possible systematic problems with the inference.

Here we will use density fields and the astrophysical parameters A and β, inferred by BORG, to predict expected quasar fluxes. If posterior predictions agree with actual observations within the uncertainty bounds, then the data model can be considered to be sufficient to analyze the data. In contrast, any misfits or systematic deviations would indicate a problem of the data model or the inference process.

Specifically, we applied the FGPA model given in eq. (8.3) to inferred density fields:

$$(F_{pp})_x = \frac{1}{N}\sum_{i=0}^{N} \exp\Big[- A_i\big(1 + (\delta^{\mathrm{f}}_i)_x\big)^{\beta_i}\Big] \tag{8.6}$$

where F_{pp} is the posterior-predicted flux, i labels the samples and x labels the volume elements.

The result of this test is presented in Fig. 8.2. As can be seen, the posterior predicted quasar spectrum nicely traces the data input within the observational 1σ uncertainty region. This demonstrates that the method correctly locates absorber positions and corresponding amplitudes of the underlying density field.

While Fig. 8.2 shows results only for a single line of sight, more generally, we also explored the flux errors for all lines of sight. The corresponding distribution of flux errors for posterior predictions is presented in Figure 8.5. The distribution of flux errors corresponds to the distribution of pixel-noise in the spectra, demonstrating that our method is close to the theoretical optimum. We note that this plot can also be compared to Fig. 14a in Horowitz et al. (2019). This comparison indicates that our method exhibits better control of the data model, including the handling of nuisance parameters and uncertainties.

8.6.4 Isotropy of the velocity field

One-dimensional Ly-α forest data introduces a particular geometry of the data into the inference problem. In particular, data is only available along one-dimensional lines of sight. In this work, we are interested in recovering the three-dimensional density field from Ly-α forest data by fitting a dynamical model. This model describes the formation of structures by displacing matter from its initial conditions to their final Eulerian positions. As such, we have found, that the large-scale velocity field is particularly sensitive to systematic effects introduced by the data, in particular, the survey geometry. Since the distribution of lines of sight defines a preferred axis, we thus test the isotropy of the velocity field to ensure that the algorithm correctly recovered a physically plausible three-dimensional velocity and corresponding density fields. The results of this test are shown in Fig. 8.6, showing that there is no preferred component of the three-dimensional velocity field. All velocity components have a similar distribution indicating that the algorithm correctly accounted and corrected for the geometry of the survey. This result is further supported by the recovery of an isotropic primordial power-spectrum as described in Section 8.7.1.

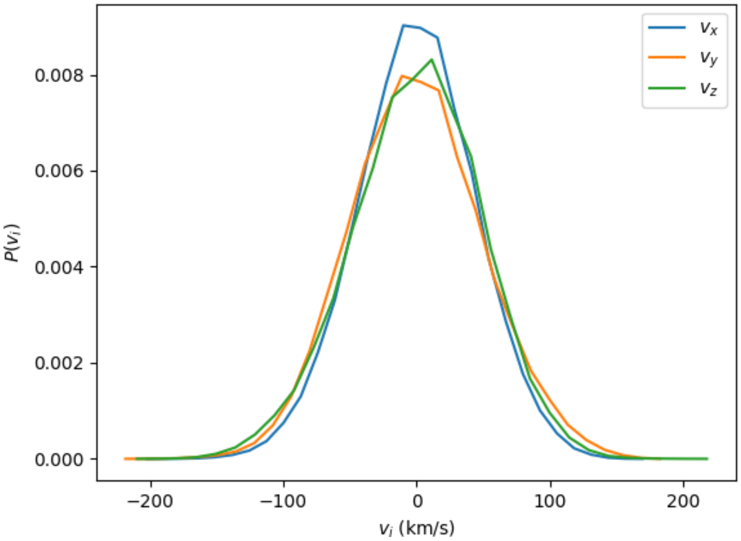


Figure 8.6: Distribution of the three Cartesian components of the velocity. The three distributions have the same mean and variance, indicating that the density field is isotropic as expected.

8.7 Analyzing the LSS in Ly-α forest data

In this Section, we present the results of applying our algorithm to simulated Ly-α forest data. We show that our method infers physically plausible and unbiased density fields and corresponding power-spectra at all scales considered in this work. We further demonstrate that the algorithm accurately extrapolates information into unobserved regions between lines of sight. We will further illustrate the performance of the algorithm by recovering mass and velocity profiles of clusters as well as voids.

8.7.1 Inference of matter density fields at high-redshift

As discussed above, the BORG algorithm performs a full-scale Bayesian analysis of the cosmological large-scale structure in Ly-α forest data. This is achieved by using the physical forward modelling approach to fit a dynamical model of structure growth to the data. As a result, we simultaneously obtain inferences of the primordial field of fluctuations from which structures formed, the non-linear spatial matter distribution at a redshift of $z = 2.5$ and corresponding velocity fields. More specifically, the BORG algorithm provides us with an ensemble of realizations of these fields drawn from the corresponding large-scale structure posterior distribution discussed in Chapter 5. This ensemble of realizations from the Markov Chain enables us to estimate any desired statistical summary and quantify corresponding uncertainties.

As an example, in Fig. 8.7, we illustrate ensemble mean and variances for primordial and final density fields. Specifically, Fig. 8.7 shows slices through the true density, and the ensemble mean and variances of inferred three-dimensional density fields, computed

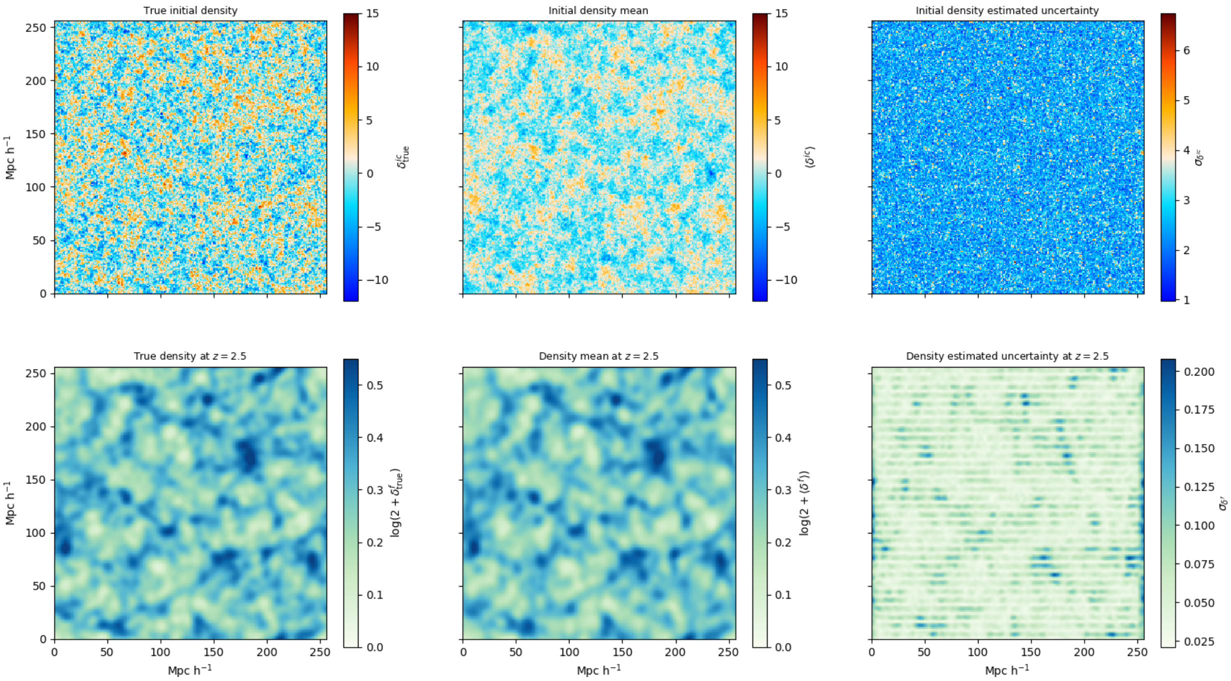


Figure 8.7: Slices through ground truth initial (left upper panel), final density field (left lower panel), inferred ensemble mean initial (middle upper panel) and ensemble mean final (middle lower panel) density field computed from 12600 MCMC samples. Comparison between these panels shows that the method recovers the structure of the true density field with high accuracy. Note that the algorithm infers the correct amplitudes of the density field. Right panels show standard deviations of inferred amplitudes of initial (upper right panel) and final density fields (lower right panel). The standard deviation of the final density field shows lower values at the position of the lines of sight. Note that the uncertainty of $\delta^{\rm f}$ presents a structure that correlates with the density field. Particularly, the variance is higher in high-density regions due to the saturation of the absorbed flux. In contrast, the standard deviation of the initial conditions are homogeneous and show no correlation with the initial density field due to the propagation of information from the final to the initial density field via the dynamical model.

from 12600 samples. A first visual comparison between ground truth and the inferred ensemble mean final density fields in the lower panels of Fig. 8.7 illustrates that the algorithm correctly recovered the three-dimensional large-scale structure from Ly-α forest data. On first sight, both fields are visually almost indistinguishable, indicating the high quality of the inference. The lower right panel of Fig. 8.7 shows the corresponding standard deviations for density amplitudes for respective volume elements as estimated from the Markov Chain. It can be seen that the estimated density standard deviations correlate with the inferred density field. This is expected for a non-linear data model, which couples signal and noise. In particular, one can observe that uncertainties are lowest for under-dense regions where Ly-α forest data provides high signal-to-noise, as quasar light is simply transmitted by structures. In contrast, one can observe the highest uncertainties for over-dense structures. Since over-dense structures absorb quasar light, this decreases the signal in these regions: the absorption is saturated at high-densities. Even more, if structures are sufficiently dense, then they will absorb all quasar light and the absorption becomes saturated. Once light absorption is saturated, the data provides no further information about the actual density amplitude of the absorber and data will only provide information on a minimally lowest density threshold required to explain the observations. Figure 8.7 therefore, clearly demonstrates that our method correctly accounts for and quantifies uncertainties inherent to Ly-α forest observations. It is interesting to remark, that while observations of galaxy clustering are most informative in high-density regions, the Ly-α forest is particularly informative for under-dense regions and therefore provides complementary information to galaxy surveys. Figure 8.7 demonstrates that, besides inferring the correct filamentary structure, our method clearly infers the correct amplitudes of the density field. Our results can also be compared to Fig. 2 in Horowitz et al. (2019), which presents a systematic underestimation of the density amplitude.

8.7.2 Analyzing posterior power-spectra

As demonstrated in the previous Section, the algorithm provides accurate reconstructions of final density fields. To provide more quantitative tests and to estimate whether inferred density fields are physically plausible, we perform analyses of posterior power-spectra measured from inferred density fields.

Reconstructing three-dimensional density fields from one-dimensional Ly-α data is technically challenging. For example, Kitaura et al. (2012b) used a Gibbs sampling approach to sample the large- and small-scales of the density field separately with a lognormal prior for the evolved density field. This approach inferred correct power-spectrum amplitudes at large scales $k < 0.1$ h Mpc^{-1} but obtained erroneous excess power at smaller scales, which was attributed to the inadequacy of the lognormal approximation or complex correlations in the data (Kitaura et al., 2012b). Horowitz et al. (2019) used an optimization approach to fit a dynamical forward model to the data, but the method obtains power-spectra that severely underestimate the power of density amplitudes.

As we aim to optimally extract cosmologically relevant information from Ly-α forest data, we are particularly interested in recovering physically plausible density fields from

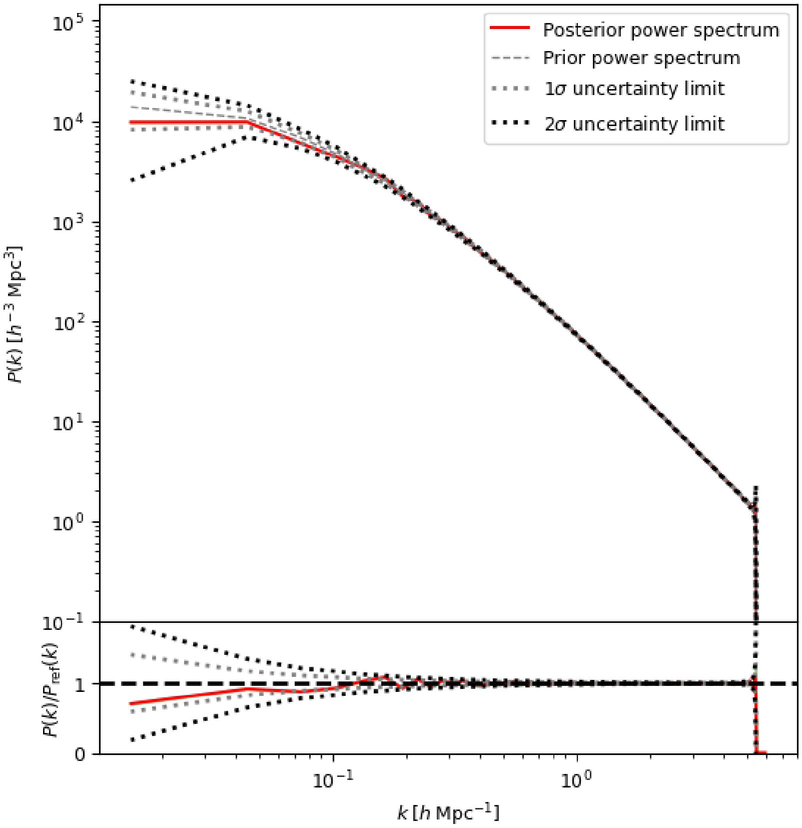


Figure 8.8: Mean posterior matter power-spectrum. Although the standard deviation is plotted, it is too small to be seen, showing the stability of the posterior power-spectrum. The dashed line indicates the underlying power spectrum and the 1- and 2-σ uncertainty limit. The mean and standard deviation are computed from 12600 samples of the Markov chain. The algorithm recovers the power-spectrum amplitudes within the $1-\sigma$ cosmic variance uncertainty limit throughout the entire range of Fourier modes considered in this work.

observations. Erroneous power in inferred power-spectra is often a sign of untreated or unknown survey systematics and indicates a break down of the assumptions underlying the data model or the inference approach. In this work, we use a sophisticated MCMC framework to marginalize out nuisance parameters and quantify uncertainties inherent to the observations. As already indicated in Section 8.6.1, our approach is able to identify the correct power distribution of density amplitudes from observations. To further quantify this result, we estimate the mean and variance of posterior power-spectra measured from the ensemble of Markov samples. The results are presented in Fig. 8.8. As can be seen, our method recovers the correct fiducial cosmological power-spectrum within the 1-σ cosmic variance uncertainty at all Fourier modes considered in this work. The unbiased recovery of the power-spectrum clearly indicates that the method correctly accounted for uncertainties and systematic effects of the data.

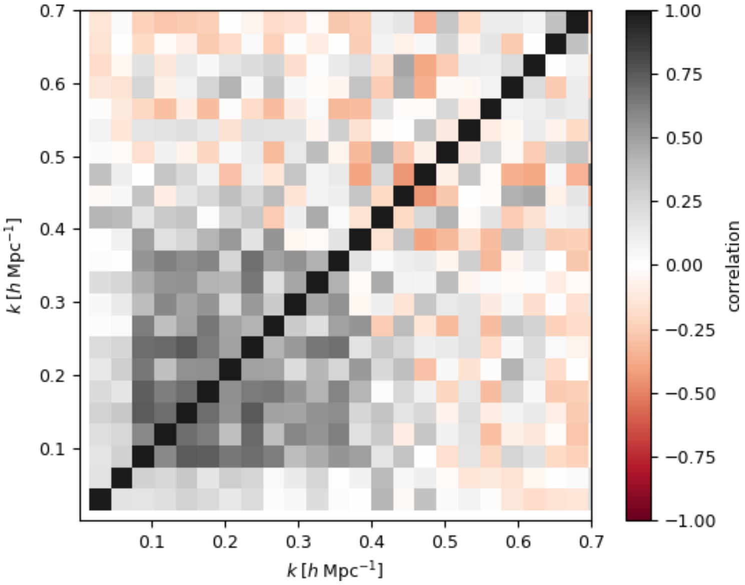


Figure 8.9: Estimated correlation matrix of power spectrum amplitudes with the mean value, normalized using the variance of amplitudes of the power spectrum modes. The low off-diagonal contributions are a clear indication that our method correctly accounted and corrected for otherwise erroneous mode coupling, typically introduced by survey systematic effects and uncertainties. The expected correlations correspond to the lines of sight grid.

Systematic effects, such as survey geometries, often introduce spurious correlations between power-spectrum modes, when not accounted for properly. This is of particular relevance for Ly-α data, which introduces a particular survey geometry through the set of one-dimensional lines of sight. To test for residual correlations between Fourier modes, we estimated the covariance matrix of power-spectrum amplitudes from our ensemble of Markov samples. As shown in Fig. 8.9 this covariance matrix shows a clear diagonal structure with the expected correlations at $k \sim 0.5 - 0.7$ h Mpc^{-1} due to the lines of sight grid. This test shows that our method correctly accounted for survey geometries and other systematic effects that could introduce erroneous mode coupling.

In summary, these tests demonstrate that our method is capable of inferring physically plausible matter density fields with correct power distribution from noisy Ly-α data.

8.7.3 Recovering information between lines of sight

The previous Section describes that inferred density fields are physically plausible realizations of the matter field underlying the Ly-α observations. Here we want to quantify to what accuracy the method recovers the underlying ground truth density field. A particular challenge is to correctly recover the density field in between one-dimensional lines of sight. As illustrated above, our method is capable of recovering the three-dimensional density field from a set of one-dimensional quasar spectra using the Bayesian physical forward modelling approach. In particular, we determine the performance of our algorithm

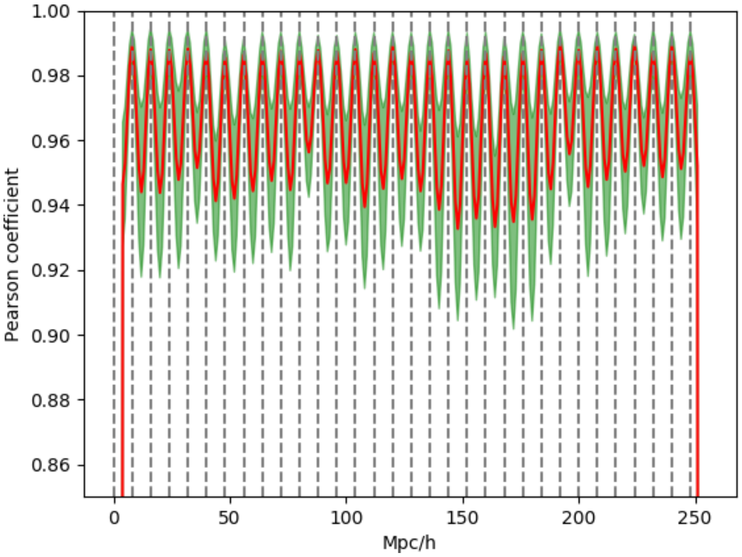
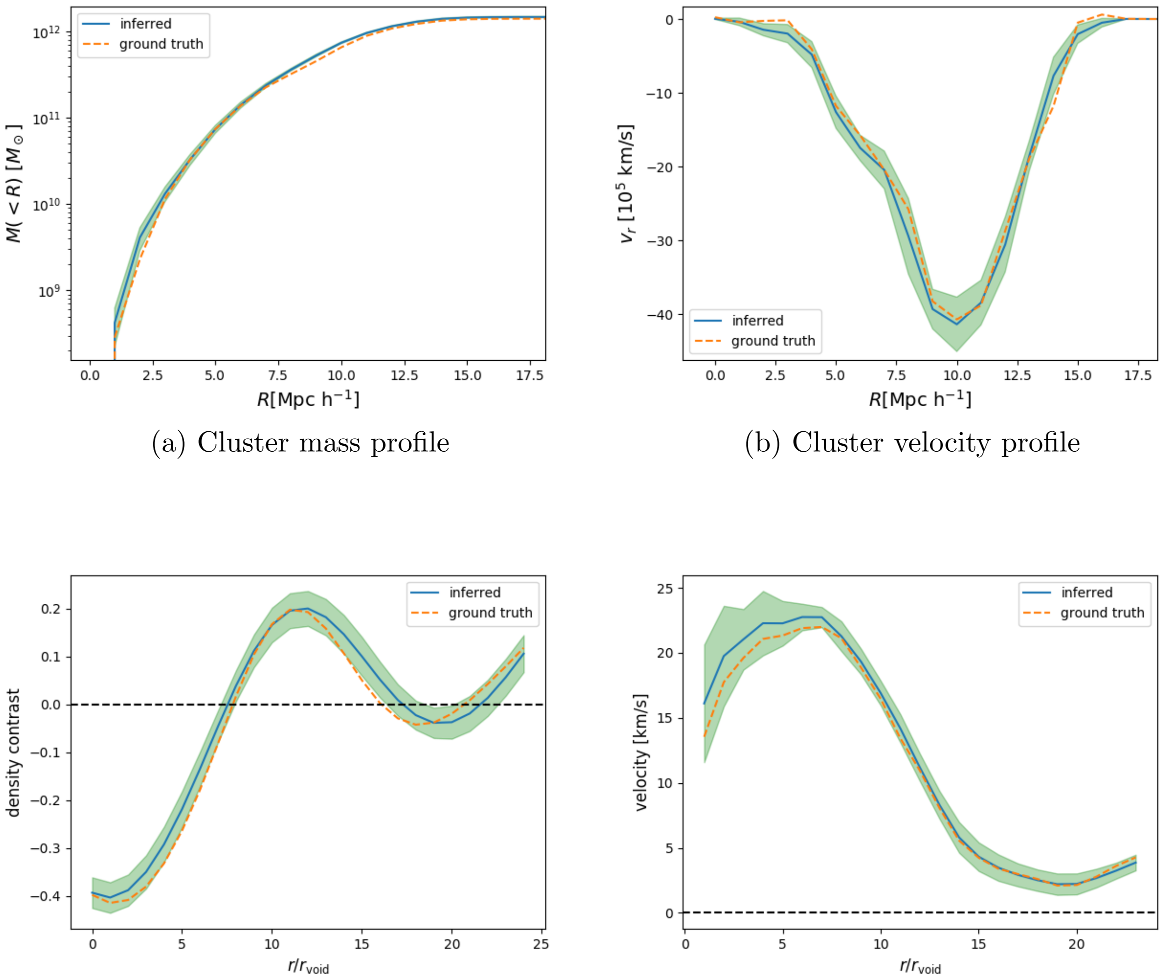


Figure 8.10: Pearson coefficient of the inferred and true density field as a function of the position across lines of sight. The dashed lines indicate the positions of the lines of sight. The Pearson coefficient is > 0.9 at any location in the density field. This, combined with Fig. 8.7 showing that the algorithm recovers the correct amplitudes, indicates that the algorithm can interpolate the information between lines of sight and correctly recover the structures in unobserved regions.

to recover the density field from data by estimating the Pearson correlation coefficient (see Appendix E) between the inferred and ground truth density fields. We estimate the Pearson correlation coefficient as a function of the position on the x-axis, which permits us to track the correlations on and in-between observed lines of sight. Figure 8.10 shows the Pearson coefficient averaged over 300 samples of the chain together with its 1-σ credibility interval. It can be clearly seen that inferred density fields correlate with the ground truth density typically to more than 98% along the lines of sight. But even in-between lines of sight the correlation is on average larger than 90%. These high Pearson correlations demonstrate the capability of our method to recover the cosmic large-scale structure in unobserved regions between observed lines of sight.

8.7.4 Cluster and void profiles

Above we showed that the BORG algorithm recovers physically plausible density fields from Ly-α data that agree with the expected statistical behaviour of a ΛCDM density field. Here we want to test if the algorithm also correctly recovers the properties of individual cosmic structures at their respective locations in the three-dimensional volume. To illustrate this fact, we will study the properties of inferred clusters and voids. In particular, we will show that the algorithm correctly recovers mass density and velocity profiles of cosmic structures.



(a) Cluster mass profile

(b) Cluster velocity profile

(c) Void density profile

(d) Void velocity profile

Figure 8.11: Top: Cluster mass (left) and radial velocity (right) profiles from the inferred ensemble mean density field and the simulation. The algorithm recovers the correct mass profile from the simulation. This demonstrates that the method provides unbiased mass estimates, correcting for the astrophysical bias of the IGM. Bottom: Spherically averaged density (left) and velocity (right) profiles of a void. The method recovers the underlying density and velocity profiles of the simulation.

To test the properties of an inferred cluster, we randomly chose a peak in the final density field. Then we determined the mass and velocity profiles in spherical shells around the identified centre of the cluster. In particular, we estimated the cumulative radial mass profiles as:

$$M(< r) = \frac{1}{N} \sum_{i=0}^{N} m_p \, K^i_{\mathrm{part}}(< r) \tag{8.7}$$

where r is the radial distance from the cluster center, i labels the Markov samples, N is the total number of samples and $K^i_{\mathrm{part}}(< r)$ is the number of simulation particles of Markov sample i inside a sphere of radius r around the cluster center. Finally, $m_p = 287 \times 10^6$ $\mathrm{M}_\odot$ is the particle mass of the simulation, which is obtained as

$$m_p = \frac{\bar{\rho} V_{\mathrm{box}}}{N^{\mathrm{total}}_{\mathrm{part}}} \tag{8.8}$$

where $\bar{\rho}$ is the mean cosmic density, V_{box} is the volume of the inferred density field and $N^{\mathrm{total}}_{\mathrm{part}}$ is the total number of particles used in the simulation. Figure 8.11 shows the cluster mass profile measured from the inferred and true density fields. The inferred mass profile shown here is the average over 60 samples. Figure 8.11 shows that our method provides unbiased mass estimates of cluster mass profiles, becoming an alternative to measuring cluster masses, complementary to weak lensing or X-ray observations.

The method also provides inferred density and velocity profiles of voids. Figure 8.11 shows the density and velocity profiles measured from the inferred and true density fields. The void is defined spherically and the velocity and density profiles are spherically averaged:

$$\rho(r) = \frac{1}{N} \sum_{i=0}^{N} \frac{3 m_p K^i_{\mathrm{part}}(< r)}{4\,\pi r^3} \tag{8.9}$$

$$v(r) = \frac{1}{N} \sum_{i=0}^{N} \frac{1}{K^i_{\mathrm{part}}(< r)} \sum_{p}^{K^i_{\mathrm{part}}(<r)} v_p \tag{8.10}$$

where i runs over the samples, N is the total number of samples, r is the radius of the sphere centered in the volume element with the lowest density, m_p is the particle mass described in eq. (8.8), $K^i_{\mathrm{part}}(< r)$ is the number of particles inside the sphere and v_p are the velocities of the particles provided by the dynamical model. The profiles shown in Fig. 8.11 are obtained from averaging over 60 samples. Voids constitute the dominant volume fraction of the Universe. Since the effect of matter in voids is mitigated, they are ideal to study the diffuse components of the Universe such as dark energy (Granett et al., 2008; Biswas et al., 2010; Lavaux and Wandelt, 2012; Bos et al., 2012) and gravity (Li, 2011; Clampitt et al., 2013; Spolyar et al., 2013; Hamaus et al., 2016). Voids are also interesting tools to constrain the neutrino mass (Massara et al., 2015; Kreisch et al., 2018; Sahlén, 2019; Schuster et al., 2019): the neutrino free-streaming length falls within the range of typical void size. Therefore, the void profiles obtained with this framework provide an alternative tool to constrain the neutrino mass.

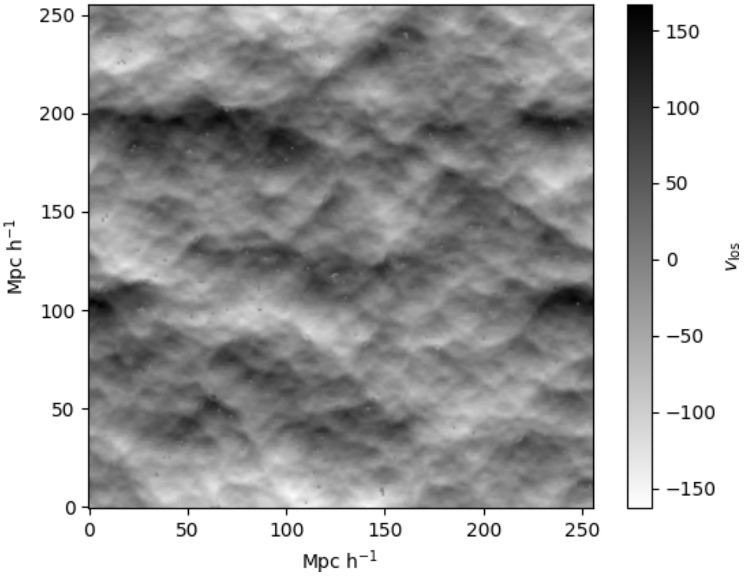


Figure 8.12: Projection of the velocity along lines of sight, which are parallel to the x-axis of the plot.

8.7.5 Velocity field and structure formation

The dynamical model employed in BORG allows to naturally infer the velocity field since it derives from the initial perturbations. The velocity information allows to discriminate between peculiar velocities and the Hubble flow. The combination of velocity and density fields can provide significant information on the formation of structures and galaxies.

Our method provides velocity fields at $z = 2.5$, where this kind of data is usually hard to obtain. Figure 8.12 shows the velocity along the line of sight. The line-of-sight velocity provides a tool to measure the kinetic Sunyaev-Zeldovich effect (Sunyaev and Zeldovich, 1972, 1980; Vishniac, 1987) by cross-correlating the velocity field with the CMB and study the expansion of the Universe.

While the hierarchical structure formation model has been tested in the nearby Universe (Bond et al., 1982; Blumenthal et al., 1984; Davis et al., 1985), our method provides a framework to test it at high redshift. The method provides consistent velocity and density fields, as shown in Fig. 8.13, which shows a zoom-in on the inferred mean density field and the corresponding velocity components (v_x, v_y). Therefore, we can test the hierarchical structure formation model at high-redshift by combining the density and velocity information and testing the predictions of the method with observations. Additionally, these density and velocity fields can provide some insights into galaxy formation and evolution by studying the effect of large-scale structures in galaxy populations (Porqueres et al., 2018). Therefore, this method for the Ly-α forest allows to extend investigations of the nature of galaxies or AGN to the high-redshift Universe.

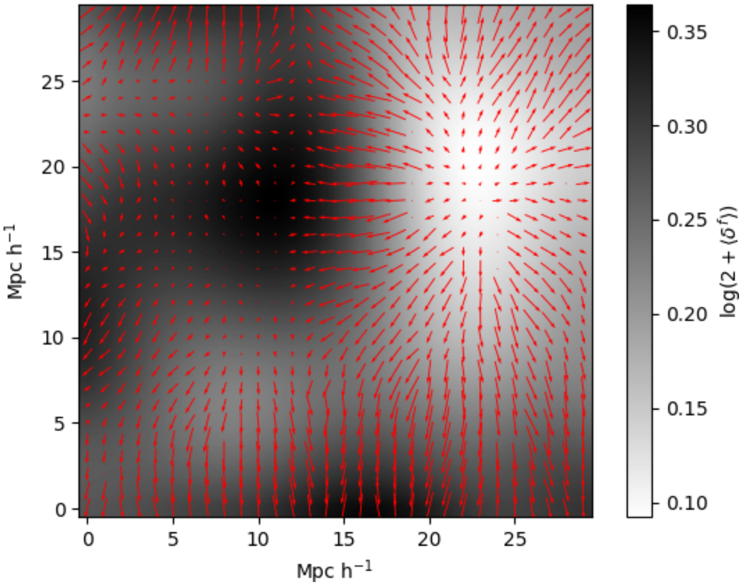


Figure 8.13: Zoom-in on the density field. The vector field shows the velocities on top of the inferred density field, showing matter flowing out of the void and falling into the gravitational potential of the cluster. Therefore, the method provides consistent velocity and density fields.

8.7.6 The astrophysical parameters

Current efforts of the science community aim at constraining the astrophysics of the intergalactic medium (see e.g. Lee, 2012; Rorai et al., 2017; Eilers et al., 2017) since the thermal history of the IGM holds the key to understanding hydrogen reionization at $z > 6$ (Lee, 2012). Particularly, analysis of the Ly-α forest attempt to measure the parameters T_0 and γ. The parameter T_0 is related to the spectral shapes of the ultraviolet background. Therefore, it contains information on the intensity of the ionizing background, which is relevant to understand the nature of the first luminous sources and their impact on the subsequent generation of galaxies (Becker et al., 2015). The parameter γ defines the temperature-density relation indicating whether overdense regions are hotter than the underdense regions ($\gamma > 1$) or vice versa ($\gamma < 1$). While Becker et al. (2007); Bolton et al. (2008); Viel et al. (2009); Rorai et al. (2017) found evidence for $\gamma < 1$, Calura et al. (2012); Garzilli et al. (2012) found $\gamma \approx 1$. Other approaches (Schaye et al., 2000; Theuns and Zaroubi, 2000; McDonald et al., 2001; Lidz et al., 2010; Rudie et al., 2012; Bolton et al., 2014) found $\gamma > 1$, which is more in agreement with the theoretically predicted $\gamma \approx 1.6$ for a post-reionization IGM (Hui and Gnedin, 1997). However, the value of γ is still an open debate since the different values are obtained from different kind of data, affected by systematic uncertainties in a different way (Eilers et al., 2017).

The algorithm presented in this work infers the astrophysical parameters of the FGPA, which are directly related to T_0 and γ. Figure 8.14 shows the posterior distribution and correlations of these parameters and the noise variance σ, obtained from 6800 samples.

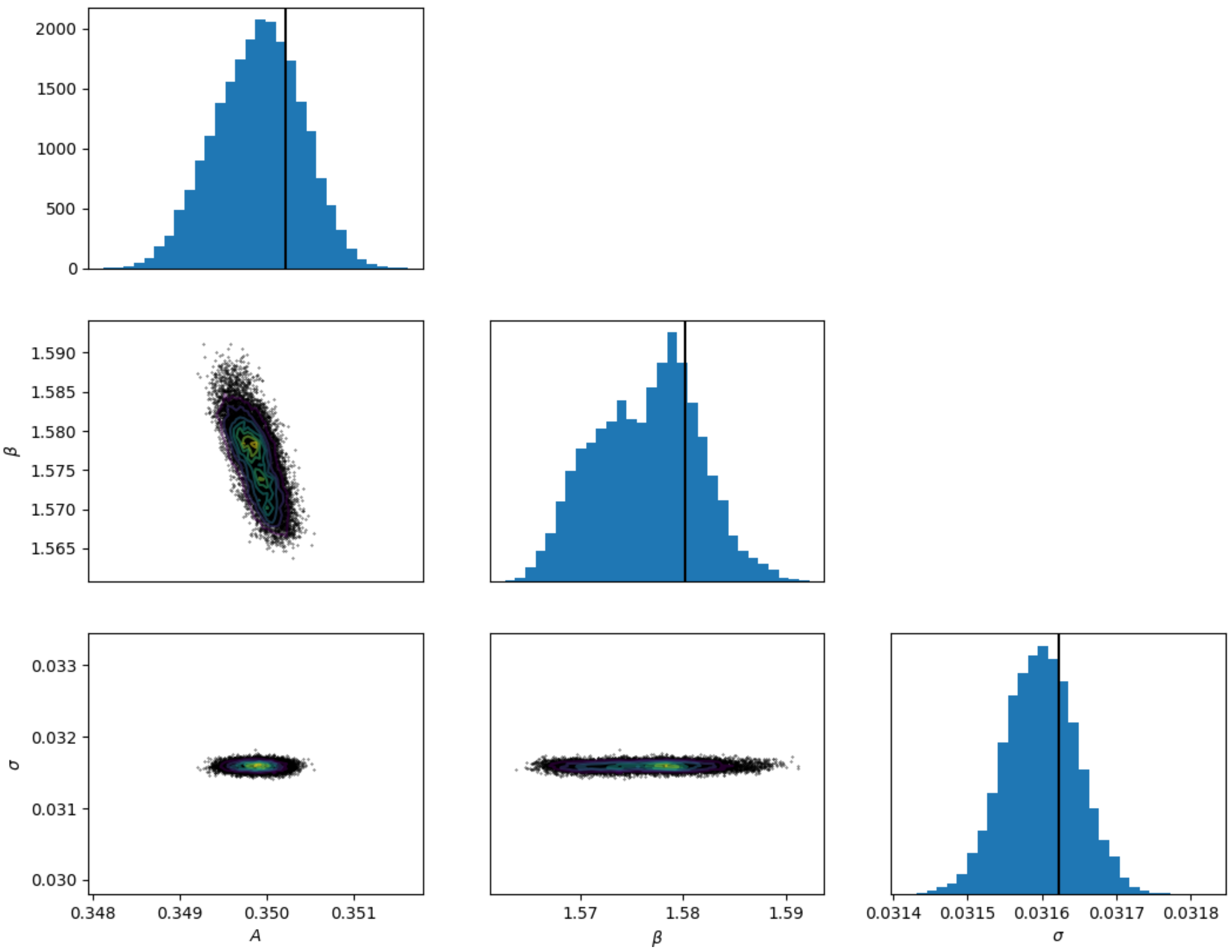


Figure 8.14: Corner plot for the unknown parameters A and β of the FGPA model as well as the standard deviation of the Gaussian noise σ. As indicated in the plot, different panels show different marginal posterior distributions for respective parameters. Black solid lines in uni-variate marginal distributions indicate the true fiducial parameter values. It can be seen that the method correctly infers the underlying parameters and quantifies corresponding uncertainties. It is interesting to remark, that the astrophysical parameters A and β show no a posteriori correlations with the noise standard deviation σ.

This shows that the method recovers the true value of the posterior distribution of the FGPA parameters, corresponding to $T_0 = (206.5 \pm 0.1)10^3\Gamma^{-1/0.7}$K and $\gamma = 1.614 \pm 0.004$, where the uncertainty is derived from the standard deviation of A and β. While the more accurate measurements of γ have an uncertainty above 7% (Rudie et al., 2012; Eilers et al., 2017), this test shows that our method can constrain γ with an uncertainty below 1%. The reason for this tight constraint of γ is probably the fact that higher-order statistics of the density field can break parameter degeneracy (Schmidt et al., 2019). Although constraining γ from real Ly-α forest data will require to model the continuum flux of the quasar, our method holds the promise to shed new light on the temperature-density debate.

Note in Fig. 8.14 that parameters A and β do not show a strong correlation with the noise variance σ, which is jointly sampled. It is also worth noticing that β presents a bimodal distribution. Optimization methods (see e.g. Horowitz et al., 2019) can only explore local extremes of the distribution, becoming sub-optimal to explore multi-modal distributions. However, our MCMC approach can characterize the bimodal distribution and provide the corresponding uncertainties, which are necessary to interpret the data correctly. Therefore, our algorithm provides a framework to constrain the astrophysics of the intergalactic medium jointly with the density field. To the knowledge of the authors, this work presents the first approach that attempts to jointly quantify the astrophysical parameters and the 3d cosmic structure.

8.8 Summary and discussion

Observations of the late time cosmic large-scale structure can provide information on fundamental physics driving the dynamic evolution of our Universe only if we manage to accurately account for non-linear data models and systematic effects of data and connect theory to observations. Detailed physical understanding of our universe, therefore, requires to accompany the current increase of high-quality cosmological data with the development of novel analysis methods capable of making the most of such observations. Traditional analyses of cosmic structures are limited to exploiting only one- and two-point statistics but ignore significant information encoded in higher-order statistics associated with the filamentary features of the late time cosmic matter distribution. To fully account for the information content of the cosmic large-scale structure, we need to follow a field-based approach to extract the entire three-dimensional matter distribution from observations. In this work we are particularly interested in studying the cosmic matter distributions at redshifts $z > 2$ with one-dimensional Lyman-α forest observations.

To extract the full three-dimensional information content from the Ly-α forest, we propose to use a Bayesian physical forward modelling approach to fit a numerical model of gravitational structure formation to data. Building upon the previously presented BORG algorithm, our fully probabilistic framework infers the three-dimensional matter density field and its dynamics from the Lyman-α forest. The method improves considerably over previous approaches (e.g. Kitaura et al., 2012b; Horowitz et al., 2019) by inferring unbiased dark matter fields from the Ly-α forest for the first time. While the previous approaches failed to recover the matter-power spectrum, our method provides the unbiased matter power spectrum at all range of Fourier modes. In particular, the approach followed by Horowitz et al. (2019) produces an underestimation of the power spectrum, corresponding to too low amplitudes of the inferred density field. We have shown that our method recovers the unbiased dark matter distribution and physically plausible density and velocity fields from Lyman-α data.

To make such inferences feasible, we used a likelihood function based on the fluctuating Gunn-Peterson approximation (FGPA). Besides modelling the dynamical evolution and spatial distribution of cosmic matter, our approach also accounts for systematic effects arising from the astrophysical properties of the intergalactic medium, by self-consistently marginalizing out corresponding nuisance parameters of the FGPA model. We have shown that this approach can provide tight constraints on the astrophysical parameters T_0 and γ, which provide significant information about the reionization history of the Universe and the early luminous sources.

Our hierarchical Bayesian framework encodes non-linear dynamical models. This work is based on the Lagrangian Perturbation Theory (LPT). As demonstrated by hydrodynamical simulations, the Ly-α forest arises from regions where the density of matter is within a factor ten of the cosmic mean density and thus is still close to the linear regime of gravitational collapse (Peirani et al., 2014; Sorini et al., 2016).

We tested our approach using realistic mock observations emulating the CLAMATO survey to infer the dark matter density and the velocity field at redshift $z = 2.5$ with a resolution of $1\mathrm{h}^{-1}$ Mpc. These tests demonstrate that our method recovers unbiased reconstructions of the non-linear spatial matter distribution and its power-spectrum from observations of the Ly-α forest. The inferred mean density field presents a high correlation with the true density at any location. This indicates that our method can interpolate the information between lines of sight. While previous approaches tried to address this challenge with polynomial smoothing (Cisewski et al., 2014) or Wiener filtering (Ozbek et al., 2016; Stark et al., 2015b), the dynamical model in our method interpolates the information by accounting for the high-order statistics of the density field.

The dynamical model naturally infers velocity fields jointly with the density field. Therefore, our method provides velocity fields at $z > 2$, where this kind of data is usually hard to obtain. From the velocity fields, we can derive the velocities along the lines of sight or radial velocities. These velocity fields can be cross-correlated with the CMB to detect the kinetic Sunyaev-Zeldovich effect. Additionally, the consistent velocity and density fields provide a framework to test hierarchical structure formation and galaxy formation models at high redshift.

We have shown that the method provides accurate mass and velocity profiles for cosmic structures, such as voids and clusters. Therefore, this method provides an alternative to measure cluster masses, complementary to X-ray and lensing measurements. The method also provides the velocity and density profile of voids. Since the non-linear effects are mitigated in voids, they are sensitive to the diffuse components of the Universe such as dark energy. Therefore, the velocity and density profiles of voids can be used to discriminate between homogeneous dark energy and modifications of gravity.

Since the Ly-α probes the matter distribution down to few megaparsecs, it is sensitive to neutrino masses. This high resolution and the void profiles provided by our algorithm can constrain neutrino masses. Additionally, this high resolution of the Ly-α forest allows inferring the density field at the 1 Mpc scale. By applying new techniques compatible with Ramanah et al. (2019), this high-resolution density field can provide tight constraints of the cosmological parameters by studying structure geometry.

Summarizing, our method clearly demonstrates the feasibility of detailed and physically plausible inferences of three-dimensional large-scale structures at high redshift from the Ly-α forest observations. The proposed approach, therefore, opens a new window to study cosmology and structure formation at high redshifts.

Chapter 9

Summary and conclusions

Summary

The main subject of this thesis is the development of Bayesian methods to infer the three-dimensional (3d) matter distribution at high redshift from cosmological data. These methods aim at providing new paths towards understanding the cosmic dynamics and the formation of structures and galaxies. In particular, I developed a method to infer the 3d matter clustering at $z > 2$ from the Lyman-α forest and derived a robust data model that is insensitive to survey systematics and, therefore, provides unbiased cosmological results even in light of unknown systematics. Further, I investigated the nature of active galactic nuclei (AGN) by studying the relation between AGN and their large-scale structure environment.

The analysis of the cosmic large-scale structure can provide invaluable information on fundamental physics driving the dynamical evolution of our Universe. However, the analysis of cosmological data requires an accurate treatment of systematic effects. The control of systematics is one of the major challenges for the next generation of galaxy surveys, which will map the Universe out to $z \approx 3$. While previous approaches to handle systematic effects relied on generating templates of the expected contaminations (see e.g. Leistedt and Peiris, 2014; Jasche and Lavaux, 2017), the next generation of surveys will be affected by yet unknown foreground and target contaminations. For this reason, I developed a likelihood that can handle data subject to unknown contaminations. The idea behind this likelihood is to marginalise out the unknown large-scale contamination amplitudes. I tested the likelihood with simulated data affected by significant foreground contamination. While the standard Poisson analysis leads to spurious large-scale artefacts in the density field and matter power spectrum, my likelihood can recover the unbiased matter distribution and its corresponding power spectrum.

While the next generation of surveys (LSST and Euclid) will focus on observing galaxies, the Ly-α forest provides complementary information and higher spatial resolution than can be achieved with flux-limited galaxy sampling at $z > 2$. Besides probing the small scales, which are sensitive to neutrino masses and dark matter models, the Ly-α traces underdense regions of the Universe, which have the potential to discriminate between dark energy models and modifications of gravity. For these reasons, I developed a method to infer the 3d density field from the Ly-α forest in quasar spectra. Since the nature of quasars and, more generally, AGN is still not well understood, I investigated the relation between AGN and their large-scale structure environment to achieve a better understanding of AGN formation and evolution. This research found evidence of an evolutionary transition between two spectral types of AGN (Transition objects evolving into LINERs), which confirmed the hypothesis of Constantin and Vogeley (2006).

Capturing the full information content of the Ly-α forest requires a three-dimensional analysis of the matter distribution. Although most of the previous analyses of the Ly-α forest focused on analysing the power spectrum, a significant amount of information is associated with the filamentary structure of the cosmic web due to the non-linear gravitational clustering. To capture the high-order statistics corresponding to the filamentary structure, I developed a Bayesian framework to infer the 3d dark matter density field at high redshift by extending the BORG algorithm, equipping it with a data model for the Ly-α forest. Improving on previous attempts to infer the 3d cosmic web from the Ly-α forest (Kitaura et al., 2012b; Horowitz et al., 2019), my method recovers the underlying spatial matter distribution and its power spectrum. To make this inference feasible, I developed a likelihood function based on the fluctuating Gunn-Peterson approximation and accounted for the bias introduced by the unknown astrophysics of the intergalactic medium (IGM). In particular, the method constrains the temperature-density relation of the neutral hydrogen and the spectral shape of the first luminous sources. The capability of the method to constrain those parameters has the potential to contribute to the current debate on the thermal history of the IGM, which encodes valuable information to understand the reionization.

Summarising, this thesis provides new methods to analyse the 3d cosmic large-scale structure at high-redshifts from cosmological data. By contributing to optimising the scientific return of current and future galaxy surveys and providing a framework to infer the matter distribution from the Ly-α forest, this work opens a new angle to put the standard cosmological model to an ultimate test with next-generation data.

Outlook

The statistical framework for the analysis of 3d matter clustering at high redshift developed in this thesis provides the dark-matter density and its dynamics with high accuracy. My approach, therefore, provides a completely new access to matter clustering at high redshifts and permits to draw conclusions about cosmological models and fundamental physics from data. An incomplete list of possible future applications to test fundamental physics is given in the following:

Constraining neutrino masses

During the radiation dominated epoch, neutrinos free-stream as relativistic particles. The effect of the free-streaming is that neutrino masses modify the gravitational potential at small scales, affecting the growth rate of structures. By probing the cosmic web down to 1 Mpc, the Ly-α forest is sensitive to neutrino masses. Therefore, my inferred density fields can be used to constrain neutrino masses.

Specifically, the size of voids is sensitive to neutrino masses (Kreisch et al., 2018) and the three-dimensional density field allows detecting ten times more voids than directly from the noisy data (Leclercq, 2014). Therefore, neutrino masses can be constrained from measuring the abundance of voids of different sizes and the void-void power spectrum, combining them with the measurement of the cluster mass function in the high-resolution density field.

Alcock-Paczynski test to infer cosmological parameters

The Alcock-Paczynski (AP) consists of a geometrical test derived from the cosmological principle: in an isotropic Universe, on average cosmic structures are isotropic. This can be used to constrain the comoving-redshift coordinate transformation and infer cosmological parameters (Ramanah et al., 2019).

While the BAO analysis has a sampling rate of every 150 Mpc, the AP test can extract information also from smaller scales (Ramanah et al., 2019). By implementing this approach into the Ly-α framework, it can provide tighter constraints of the matter density Ω_m and the dark energy equation of state w and w_a, exploiting the high-resolution of the Ly-α forest.

Structure formation and galaxy evolution at high redshift

The hierarchical structure formation model predicts that structures are formed by accretion and merging of smaller objects. This model has been tested in the nearby Universe but can it explain matter clustering at high z?

Since the BORG framework infers density fields and velocities, it can provide physically plausible formation histories at high redshift. This allows testing structure formation models by comparing the merger trees of clusters to the predictions of structure formation models.

Current efforts of the science community address the question of how galaxies acquire their present-day properties (LSST Science Collaboration et al., 2009). My study on AGN can be extended with the Ly-α framework to higher redshifts to analyse earlier stages in the formation of AGN. The inferred density can be used to test galaxy formation models by comparing observations of LSST and Euclid to the predictions of simulations (star-forming rate, cold accretion, starburst, etc.).

Constrained simulations at high redshift

Large-scale structure simulations have been used to predict structure formation models in the non-linear regime. However, the comparison with data requires that the simulation covers a large representative volume, which comes at the expense of mass resolution. To overcome this problem, one can use simulations from the inferred initial conditions provided by the BORG framework, which are representative of the actual Universe and their predictions can be directly compared to data.

The aim of these simulations is to study the physics of the IGM by comparing the data to the predictions of cold accretion, the impact of AGN feedback and the properties of the gas in filamentary structures.

Further development of the algorithm

The analysis of the Ly-α forest required the implementation of an infrastructure to analyse line-of-sight information in the BORG framework. This includes building a projector to predict the density along the line of sight and proving that the BORG framework can interpolate the information between lines of sight to recover the filamentary structure of the cosmic web. This infrastructure can be used for other large-scale structure probes that require line of sight observations. Some examples of these probes are 21 cm and cosmic shear. Here we discuss the extension for the analysis of the cosmic shear since the next generation of surveys (LSST and Euclid) will provide precise measurements of weak lensing of galaxies.

When the light from distant galaxies travels through the Universe, it is deflected by the gravitational potential of the large-scale structures. This results in a coherent distortion of galaxy images, which traces the matter distribution along the line of sight.

Alsing et al. (2016, 2017) presented a tomographic analysis of the cosmic shear, focusing on the two-point statistics. However, a significant part of the information is transported to high-order correlations. Heavens (2003) introduced a 3d analysis of the cosmic shear, which avoids the loss of information due to tomographic redshift binning. The implementation of the 3d analysis of the cosmic shear to the BORG framework will provide a high-order statistics analysis to infer the large-scale density field and its dynamics from the weak lensing measurements.

Appendix A

Leapfrog integrator

The BORG framework employs a dynamical model for structure formation, which is applied to dark matter particles populating the density field. Evolving the density field requires to integrate the equations of motion of the particles with a discretised time. The numerical integrator needs to be accurate to guarantee high acceptance rates by conserving the Hamiltonian. In addition, the integrator needs to be reversible and symplectic to ensure detailed balance. For these reasons, the integrator implemented in BORG is the leapfrog integrator (Duane et al., 1987).

The leapfrog scheme can be written as

$$\begin{aligned} p_i\left(t+\frac{\epsilon}{2}\right) &= p_i\left(t\right) - \frac{\epsilon}{2}\frac{\partial\Psi\left(\delta^{\mathrm{ic}}\right)}{\partial\delta_i^{\mathrm{ic}}}\left(\delta_i^{\mathrm{ic}}\left(t\right)\right), && \text{(A.1)} \\ \delta_i^{\mathrm{ic}}\left(t+\epsilon\right) &= \delta_i^{\mathrm{ic}}\left(t\right) + \epsilon\frac{p_i\left(t+\frac{\epsilon}{2}\right)}{m_i}, && \text{(A.2)} \\ p_i\left(t+\epsilon\right) &= p_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial\Psi\left(\delta^{\mathrm{ic}}\right)}{\partial\delta_i^{\mathrm{ic}}}\left(\delta_i^{\mathrm{ic}}\left(t+\epsilon\right)\right), && \text{(A.3)} \end{aligned}$$

where m_i is the element of the mass matrix at position x.

The equations of motion are integrated by making n steps with a finite step size ϵ. To prevent resonant trajectories, time steps ϵ are randomised, drawing ϵ from a uniform distribution.

Appendix B

Cloud-in-cell scheme

The cloud-in-cell algorithm (Hockney and Eastwood, 1988) is used in the BORG framework to compute the density fields from particle distributions. This method defines a procedure to assign quantities carried by particles to a grid. The idea behind this is to assume that particles have a 'shape' S that intersects the grid. In a one-dimensional case, a fraction of a particle at x_p is assigned to a cell x_c of the grid as

$$W(x_p - x_c) \equiv \int_{x_c-\Delta x/2}^{x_c+\Delta x/2} S(x' - x_p)\,\mathrm{d}x'\,. \tag{B.1}$$

with Δx being the size of the cell. In three dimensions, $W(\mathbf{x}_p - \mathbf{x}_c) = W(x_p - x_c)W(y_p - y_c)W(z_p - z_c)$. Therefore, a quantity A carried by particles as A_p can be computed on the grid at a position $\mathbf{x_c}$ as

$$A(\mathbf{x_c}) = \sum_{\{\mathbf{x}_p\}} A_p W(\mathbf{x}_p - \mathbf{x}_c). \tag{B.2}$$

In the case of the dynamical model, the quantity A is the dark matter mass. In the cloud-in-cell scheme, the shape of the particles is defined as a rectangular parallelepiped, involving the eight nearest cells (i.e. volume elements) for each particle.

Appendix C

Hamiltonian equations for the large-scale structure

The BORG framework employs a Hamiltonian Monte Carlo (HMC) algorithm to generate realisations of the density field. The HMC algorithm explores the parameter space efficiently by using the gradient information: the gradients indicate the direction towards the states with higher probability. Therefore, the HMC method implemented in BORG requires to compute the gradients of the posterior distribution with respect to the initial density field. In this appendix, I will derive the gradients of the prior distribution (Section 5.2) and likelihood distribution for galaxy surveys and Ly-α forest (Section 5.4).

Prior

The prior potential is given by the Gaussian initial conditions

$$\psi_{\text{prior}}\left(\delta^{\text{ic}}\right) = \frac{1}{2}\sum_{xy} \delta^{\text{ic}}_x S^{-1}_{xy} \delta^{\text{ic}}_y . \tag{C.1}$$

where the indices x and y label the volume elements and S_{xy} is the covariance matrix. The gradient with respect to the initial density is given by

$$\frac{\partial \psi_{\text{prior}}\left(\delta^{\text{ic}}\right)}{\partial \delta^{\text{ic}}_x} = \sum_y S^{-1}_{xy} \delta^{\text{ic}}_y \tag{C.2}$$

.

Galaxy redshift surveys

The likelihood potential for galaxy redshift surveys is derived from eq. (5.14):

$$\Psi_{\text{likelihood}}\left(\delta^{\text{ic}}\right) = \sum_{x,l} R^l_x \bar{N}^l \left(1 + B^l\left[M\left(a, \delta^{\text{ic}}\right)\right]_x\right) - \sum_{x,l} N^l_x \ln\left(R^l_x \bar{N}^l \left(1 + B^l\left[M\left(a, \delta^{\text{ic}}\right)\right]_x\right)\right), \tag{C.3}$$

where x labels the volume elements, l indicates the magnitude bin, R is the linear survey response, $\bar{N}$ is the expected number of galaxies, B^l is the bias model, and $M(a,\delta)$ is the dynamical model.

The forces corresponding to this likelihood term are

$$\begin{aligned} \frac{\partial\Psi_{\text{likelihood}}\left(\delta^{\text{ic}}\right)}{\partial\delta_y^{\text{ic}}} &= \sum_l \left(1 - \frac{1}{R_y^l\bar{N}^l\left(1 + B^l M\left(a,\delta^{\text{ic}}\right)_y\right)}\right) \\ &\times\ R_y^l\bar{N}^l B^l \frac{\partial M\left(a,\delta^{\text{ic}}\right)}{\partial\delta_y^{\text{ic}}} \end{aligned} \tag{C.4}$$

where the derivative of the dynamical model depends on the cloud-in-cell kernel.

Lyman-α forest

The potential corresponding to the likelihood for the Lyman-α forest is derived from eq. (5.16):

$$\begin{aligned} \psi_{\text{likelihood}}(\delta^{\text{ic}}) &= \sum_n \sum_x^N \frac{\Big((F_n)_x - \exp[A(1 + M(a,\delta^{\text{ic}})_x)^\beta]\Big)^2}{2\sigma^2} \\ &+ \frac{1}{2^N}\sum_n \ln(2\pi\sigma^2)^N \end{aligned} \tag{C.5}$$

where n labels the line of sight, x runs over the voxels intersected by the n-th line of sight, F is the transmitted flux fraction, N is the number of volume elements in that line of sight, σ is the variance of the noise and A, β are the astrophysical parameters.

The gradient is given by

$$\begin{aligned} \frac{\partial\psi_{\text{likelihood}}(\delta^{\text{ic}}))}{\partial\delta_x^{\text{ic}}} &= \sum_n \frac{(F_n)_x - \exp[-A(1 + M(a,\delta^{\text{ic}})_x)^\beta]}{\sigma^2} \\ &\times\ A\beta\big(1 + M(a,\delta^{\text{ic}})_x\big)^{\beta-1} \\ &\times\ \exp\Big[-A\big(1 + M(a,\delta^{\text{ic}})_x\big)^\beta\Big] \\ &\times\ \frac{\partial M(a,\delta^{\text{ic}})}{\partial\delta_x^{\text{ic}}}. \end{aligned} \tag{C.6}$$

where the derivative of the dynamical model depends on the cloud-in-cell kernel.

Hamiltonian equations

Finally, the equations of motion of the Hamiltonian system can be written as

$$\frac{\mathrm{d}\delta_i^{\mathrm{ic}}}{\mathrm{d}t} = \sum_j M_{ij}^{-1} p_j, \tag{C.7}$$

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = -\sum_j S_{ij}^{-1} \delta_j^{\mathrm{ic}} + \frac{\partial \psi_{\mathrm{likelihood}}\left(\delta^{\mathrm{ic}}\right)}{\partial \delta_j^{\mathrm{ic}}} \tag{C.8}$$

where M_{ij} is the Hamiltonian mass matrix.

Appendix D

Aligning observed AGN with inferred density fields

In order to study the large-scale environment of observed AGN, we needed to relate them to our reconstructions of the three-dimensional density field. This means that it was necessary to map their coordinates into our analysis domain.

Since the three-dimensional density fields have been inferred with respect to the CMB restframe, we first had to adjust redshifts of observed objects correspondingly in the MPA/JHU catalog. The transformation is achieved by accounting for the observers velocity with respect to the CMB $v_{\rm los}$ along the line of sight $v_{\rm los}$:

$$z = z_{\rm obs} + v_{\rm los}^{\rm CMB}/c, \tag{D.1}$$

where the line of sight velocity is given by Tully (2007):

$$v_{\rm los}^{\rm CMB} = -25\cos l\cos b - 246\sin l\cos b + 277\sin b, \tag{D.2}$$

with (l, b) being the galactic coordinates and $v_{\rm los}$ is obtained from the projection of $v_{\rm CMB} = (-25, -246, 277)$ km s^{-1} which is the relative velocity of the Sun with respect to the CMB in Galactic coordinates. We then translated corrected redshifts into comoving distances d by solving the equation:

$$d = d_H \int_0^z \frac{dz'}{E(z')}, \tag{D.3}$$

where we have used the cosmological parameters given in Section 6.2. Finally we transformed spherical to Cartesian coordinates via:

$$\mathbf{r} = d(\cos l\cos b, \cos b\sin l, \sin b). \tag{D.4}$$

Corresponding grid positions of the analysis domain are then obtained by rescaling the coordinate vector $\mathbf{r}$ to the voxel index as $\mathbf{r}_{\rm pix} = (N/L)\mathbf{r}$ where N is the number of voxels in each dimension and L is the size of the reconstruction domain. Since the observer is at the center of the box, $\mathbf{r}$ is shifted by $L/2$.

Appendix E

Pearson's correlation coefficient

Pearson's correlation coefficient is a measurement of the linear correlation between two variables. It has a value between 1 and -1, with 1 indicating linear positive correlation and -1 anticorrelation. Pearson's coefficient is defined as

$$r_{xy} = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\left[\sum_i (x_i - \langle x \rangle)^2 \right]^{1/2} \left[\sum_i (y_i - \langle y \rangle)^2 \right]^{1/2}} \tag{E.1}$$

where i labels the samples and $\langle x \rangle$ indicates the sample mean of x.

Bibliography

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., Aleksić, J., Allam, S., Allen, S., Amara, A., Annis, J., Asorey, J., and Avila, S. (2018). Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *Phys. Rev. D*, 98(4):043526.

Adriani, O., Aguilar-Bentez, M., P Ahlen, S., Akbari, H., Alcaraz, J., Aloisio, A., Alverson, G., Alviggi, M., Ambrosi, G., An, Q., Anderhub, H., L Anderson, A., P Andreev, V., Angelov, T., Antonov, L., Antreasyan, D., Arce, P., Arefev, A., G Atamanchuk, A., and C C Van der Zwaan, B. (1992). Determination of the number of light neutrino species.

Alam, S., Albareti, F. D., Allende Prieto, C., Anders, F., Anderson, S. F., Anderton, T., Andrews, B. H., Armengaud, E., Aubourg, É., Bailey, S., and et al. (2015). The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *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., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B. A., Rodríguez-Torres, S. A., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martín, J. A., Saito, S., Salazar-Albornoz, S., Samushia, L., Sánchez, A. G., Satpathy, S., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Slosar, A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Vazquez, J. A., Verde, L., Wake, D. A., Wang, Y., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yèche, C., Zehavi, I., Zhai, Z., and Zhao, G.-B. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *MNRAS*, 470:2617–2652.

Alam, S., Ho, S., and Silvestri, A. (2016). Testing deviations from ΛCDM with growth rate measurements from six large-scale structure surveys at z = 0.06-1. *MNRAS*, 456(4):3743–3756.

Albrecht, A. and Steinhardt, P. J. (1982). Cosmology for grand unified theories with radiatively induced symmetry breaking. *Physical Review Letters*, 48:1220–1223.

Alpher, R. A., Bethe, H., and Gamow, G. (1948). The Origin of Chemical Elements. *Physical Review*, 73:803–804.

Alsing, J., Heavens, A., and Jaffe, A. H. (2017). Cosmological parameters, shear maps and power spectra from CFHTLenS using Bayesian hierarchical inference. *MNRAS*, 466:3272–3292.

Alsing, J., Heavens, A., Jaffe, A. H., Kiessling, A., Wandelt, B., and Hoffmann, T. (2016). Hierarchical cosmic shear power spectrum inference. *MNRAS*, 455:4452–4466.

Amendola, L., Appleby, S., Avgoustidis, A., Bacon, D., Baker, T., Baldi, M., Bartolo, N., Blanchard, A., Bonvin, C., Borgani, S., Branchini, E., Burrage, C., Camera, S., Carbone, C., Casarini, L., Cropper, M., de Rham, C., Dietrich, J. P., Di Porto, C., Durrer, R., Ealet, A., Ferreira, P. G., Finelli, F., García-Bellido, J., Giannantonio, T., Guzzo, L., Heavens, A., Heisenberg, L., Heymans, C., Hoekstra, H., Hollenstein, L., Holmes, R., Hwang, Z., Jahnke, K., Kitching, T. D., Koivisto, T., Kunz, M., La Vacca, G., Linder, E., March, M., Marra, V., Martins, C., Majerotto, E., Markovic, D., Marsh, D., Marulli, F., Massey, R., Mellier, Y., Montanari, F., Mota, D. F., Nunes, N. J., Percival, W., Pettorino, V., Porciani, C., Quercellini, C., Read, J., Rinaldi, M., Sapone, D., Sawicki, I., Scaramella, R., Skordis, C., Simpson, F., Taylor, A., Thomas, S., Trotta, R., Verde, L., Vernizzi, F., Vollmer, A., Wang, Y., Weller, J., and Zlosnik, T. (2018). Cosmology and fundamental physics with the Euclid satellite. *Living Reviews in Relativity*, 21:2.

Anderson, L., Aubourg, E., Bailey, S., Bizyaev, D., Blanton, M., Bolton, A. S., Brinkmann, J., Brownstein, J. R., Burden, A., Cuesta, A. J., da Costa, L. A. N., Dawson, K. S., de Putter, R., Eisenstein, D. J., Gunn, J. E., Guo, H., Hamilton, J.-C., Harding, P., Ho, S., Honscheid, K., Kazin, E., Kirkby, D., Kneib, J.-P., Labatie, A., Loomis, C., Lupton, R. H., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., McBride, C. K., Mehta, K. T., Mena, O., Montesano, F., Muna, D., Nichol, R. C., Nuza, S. E., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J., Pâris, I., Percival, W. J., Petitjean, P., Prada, F., Reid, B., Roe, N. A., Ross, A. J., Ross, N. P., Samushia, L., Sánchez, A. G., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Skibba, R. A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Verde, L., Wagner, C., Wake, D. A., Weaver, B. A., Weinberg, D. H., White, M., Xu, X., Yèche, C., Zehavi, I., and Zhao, G.-B. (2012). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. *MNRAS*, 427:3435–3467.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43.

Antonucci, R. (1993). Unified models for active galactic nuclei and quasars. *ARA&A*, 31:473–521.

Bahcall, J. N., Bergeron, J., Boksenberg, A., Hartig, G. F., Jannuzi, B. T., Kirhakos, S., Sargent, W. L. W., Savage, B. D., Schneider, D. P., and Turnshek, D. A. (1993). The Hubble Space Telescope Quasar Absorption Line Key Project. I. First Observational Results, Including Lyman-Alpha and Lyman-Limit Systems. *ApJS*, 87:1.

Bahcall, J. N., Bergeron, J., Boksenberg, A., Hartig, G. F., Jannuzi, B. T., Kirhakos, S., Sargent, W. L. W., Savage, B. D., Schneider, D. P., and Turnshek, D. A. (1996). The Hubble Space Telescope Quasar Absorption Line Key Project. VII. Absorption Systems at Z abs &lt;= 1.3. *ApJ*, 457:19.

Baldwin, J. A., Phillips, M. M., and Terlevich, R. (1981). Classification parameters for the emission-line spectra of extragalactic objects. *PASP*, 93:5–19.

Balogh, M., Eke, V., Miller, C., Lewis, I., Bower, R., Couch, W., Nichol, R., Bland-Hawthorn, J., Baldry, I. K., Baugh, C., Bridges, T., Cannon, R., Cole, S., Colless, M., Collins, C., Cross, N., Dalton, G., de Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Gomez, P., Gray, A., Hawkins, E., Jackson, C., Lahav, O., Lumsden, S., Maddox, S., Madgwick, D., Norberg, P., Peacock, J. A., Percival, W., Peterson, B. A., Sutherland, W., and Taylor, K. (2004). Galaxy ecology: groups and low-density environments in the SDSS and 2dFGRS. *MNRAS*, 348:1355–1372.

Bardeen, J. M., Bond, J. R., Kaiser, N., and Szalay, A. S. (1986). The statistics of peaks of Gaussian random fields. *ApJ*, 304:15–61.

Bardeen, J. M., Steinhardt, P. J., and Turner, M. S. (1983). Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *Phys. Rev. D*, 28:679–693.

Basilakos, S. and Nesseris, S. (2016). Testing Einstein's gravity and dark energy with growth of matter perturbations: Indications for new physics? *Phys. Rev. D*, 94(12):123525.

Baugh, C. M., Gaztanaga, E., and Efstathiou, G. (1995). A comparison of the evolution of density fields in perturbation theory and numerical simulations - II. Counts-in-cells analysis. *MNRAS*, 274(4):1049–1070.

Baumann, D. (2009). TASI Lectures on Inflation. *arXiv e-prints*, page arXiv:0907.5424.

Bautista, J. E., Busca, N. G., Guy, J., Rich, J., Blomqvist, M., du Mas des Bourboux, H., Pieri, M. M., Font-Ribera, A., Bailey, S., and Delubac, T. (2017). Measurement of baryon acoustic oscillation correlations at z = 2.3 with SDSS DR12 Lyα-Forests. *A&A*, 603:A12.

Becker, G. D., Bolton, J. S., and Lidz, A. (2015). Reionisation and High-Redshift Galaxies: The View from Quasar Absorption Lines. *PASA*, 32:e045.

Becker, G. D., Rauch, M., and Sargent, W. L. W. (2007). The Evolution of Optical Depth in the Lyα Forest: Evidence Against Reionization at z~6. *ApJ*, 662(1):72–93.

Bernardeau, F., Colombi, S., Gaztañaga, E., and Scoccimarro, R. (2002). Large-scale structure of the Universe and cosmological perturbation theory. *Phys. Rev. D*, 367(1-3):1–248.

Bertschinger, E., Dekel, A., Faber, S. M., Dressler, A., and Burstein, D. (1990). Potential, Velocity, and Density Fields from Redshift-Distance Samples: Application: Cosmography within 6000 Kilometers per Second. *ApJ*, 364:370.

Bird, S., Peiris, H. V., Viel, M., and Verde, L. (2011). Minimally parametric power spectrum reconstruction from the Lyman α forest. *MNRAS*, 413:1717–1728.

Biswas, R., Alizadeh, E., and Wandelt, B. D. (2010). Voids as a precision probe of dark energy. *Phys. Rev. D*, 82(2):023002.

Blomqvist, M., du Mas des Bourboux, H., Busca, N. G., de Sainte Agathe, V., Rich, J., Balland, C., Bautista, J. E., Dawson, K., Font-Ribera, A., and Guy, J. (2019). Baryon acoustic oscillations from the cross-correlation of Lyα absorption and quasars in eBOSS DR14. *arXiv e-prints*, page arXiv:1904.03430.

Blumenthal, G. R., Faber, S. M., Primack, J. R., and Rees, M. J. (1984). Formation of galaxies and large-scale structure with cold dark matter. *Nature*, 311:517–525.

Boera, E., Becker, G. D., Bolton, J. S., and Nasir, F. (2019). Revealing Reionization with the Thermal History of the Intergalactic Medium: New Constraints from the Lyα Flux Power Spectrum. *ApJ*, 872(1):101.

Bolton, J. S., Becker, G. D., Haehnelt, M. G., and Viel, M. (2014). A consistent determination of the temperature of the intergalactic medium at redshift z = 2.4. *MNRAS*, 438(3):2499–2507.

Bolton, J. S., Viel, M., Kim, T. S., Haehnelt, M. G., and Carswell, R. F. (2008). Possible evidence for an inverted temperature-density relation in the intergalactic medium from the flux distribution of the Lyα forest. *MNRAS*, 386(2):1131–1144.

Bond, J. R., Szalay, A. S., and Turner, M. S. (1982). Formation of galaxies in a gravitino-dominated universe. *Physical Review Letters*, 48:1636–1639.

Bos, E. G. P., van de Weygaert, R., Dolag, K., and Pettorino, V. (2012). The darkness that shaped the void: dark energy and cosmic voids. *MNRAS*, 426(1):440–461.

Bouchet, F. R., Colombi, S., Hivon, E., and Juszkiewicz, R. (1995). Perturbative Lagrangian approach to gravitational instability. *A&A*, 296:575.

Bovy, J., Myers, A. D., Hennawi, J. F., Hogg, D. W., McMahon, R. G., Schiminovich, D., Sheldon, E. S., Brinkmann, J., Schneider, D. P., and Weaver, B. A. (2012). Photometric Redshifts and Quasar Probabilities from a Single, Data-driven Generative Model. *ApJ*, 749:41.

Boyle, A. and Komatsu, E. (2018). Deconstructing the neutrino mass constraint from galaxy redshift surveys. *Journal of Cosmology and Astro-Particle Physics*, 2018(3):035.

Brenier, Y., Frisch, U., Hénon, M., Loeper, G., Matarrese, S., Mohayaee, R., and Sobolevskiĭ, A. (2003). Reconstruction of the early Universe as a convex optimization problem. *MNRAS*, 346(2):501–524.

Buchert, T., Melott, A. L., and Weiss, A. G. (1994). Testing higher-order Lagrangian perturbation theory against numerical simulations I. Pancake models. *A&A*, 288:349–364.

Busca, N. G., Delubac, T., Rich, J., Bailey, S., Font-Ribera, A., Kirkby, D., Le Goff, J. M., Pieri, M. M., Slosar, A., and Aubourg, É. (2013). Baryon acoustic oscillations in the Lyα forest of BOSS quasars. *A&A*, 552:A96.

Calura, F., Tescari, E., D'Odorico, V., Viel, M., Cristiani, S., Kim, T. S., and Bolton, J. S. (2012). The Lyman α forest flux probability distribution at z&gt;3. *MNRAS*, 422(4):3019–3036.

Carroll, S. M. (2004). *Spacetime and geometry. An introduction to general relativity.*

Cen, R., Miralda-Escudé, J., Ostriker, J. P., and Rauch, M. (1994). Gravitational Collapse of Small-Scale Structure as the Origin of the Lyman-Alpha Forest. *ApJ*, 437:L9.

Charlot, S. and Longhetti, M. (2001). Nebular emission from star-forming galaxies. *MNRAS*, 323:887–903.

Chen, C.-T. J., Hickox, R. C., Alberts, S., Brodwin, M., Jones, C., Murray, S. S., Alexander, D. M., Assef, R. J., Brown, M. J. I., Dey, A., Forman, W. R., Gorjian, V., Goulding, A. D., Le Floc'h, E., Jannuzi, B. T., Mullaney, J. R., and Pope, A. (2013). A Correlation between Star Formation Rate and Average Black Hole Accretion in Star-forming Galaxies. *ApJ*, 773:3.

Christlein, D. and Zabludoff, A. I. (2004). Can Early-Type Galaxies Evolve from the Fading of the Disks of Late-Type Galaxies? *ApJ*, 616:192–198.

Cisewski, J., Croft, R. A. C., Freeman, P. E., Genovese, C. R., Khandai, N., Ozbek, M., and Wasserman, L. (2014). Non-parametric 3D map of the intergalactic medium using the Lyman-alpha forest. *MNRAS*, 440:2599–2609.

Clampitt, J., Cai, Y.-C., and Li, B. (2013). Voids in modified gravity: excursion set predictions. *MNRAS*, 431(1):749–766.

Clowe, D., Bradač, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C., and Zaritsky, D. (2006). A Direct Empirical Proof of the Existence of Dark Matter. *ApJ*, 648(2):L109–L113.

Clowe, D., Gonzalez, A., and Markevitch, M. (2004). Weak-Lensing Mass Reconstruction of the Interacting Cluster 1E 0657-558: Direct Evidence for the Existence of Dark Matter. *ApJ*, 604:596–603.

Cole, S. and Kaiser, N. (1989). Biased clustering in the cold dark matter cosmogony. *MNRAS*, 237:1127–1146.

Colless, M., Peterson, B. A., Jackson, C., Peacock, J. A., Cole, S., Norberg, P., Baldry, I. K., Baugh, C. M., Bland-Hawthorn, J., Bridges, T., Cannon, R., Collins, C., Couch, W., Cross, N., Dalton, G., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Sutherland, W., and Taylor, K. (2003). The 2dF Galaxy Redshift Survey: Final Data Release. *arXiv e-prints*, pages astro–ph/0306581.

Colombi, S., Dodelson, S., and Widrow, L. M. (1996). Large-Scale Structure Tests of Warm Dark Matter. *ApJ*, 458:1.

Constantin, A., Hoyle, F., and Vogeley, M. S. (2008). Active Galactic Nuclei in Void Regions. *ApJ*, 673:715–729.

Constantin, A. and Vogeley, M. S. (2006). The Clustering of Low-Luminosity Active Galactic Nuclei. *ApJ*, 650:727–748.

Coziol, R., Andernach, H., Torres-Papaqui, J. P., Ortega-Minakata, R. A., and Moreno del Rio, F. (2017). What sparks the radio-loud phase of nearby quasars? *MNRAS*, 466:921–944.

Croft, R. A. C., Weinberg, D. H., Katz, N., and Hernquist, L. (1998). Cosmology from the structure of the lya forest. In Mueller, V., Gottloeber, S., Muecket, J. P., and Wambsganss, J., editors, *Large Scale Structure: Tracks and Traces*, pages 69–75.

Cruzen, S., Wehr, T., Weistrop, D., Angione, R. J., and Hoopes, C. (2002). Spectroscopy of Galaxies in the Bootes Void. *AJ*, 123:142–158.

Davé, R. (2001). The Evolution of the Lyman Alpha Forest from z ˜3 –&gt; 0. In von Hippel, T., Simpson, C., and Manset, N., editors, *Astrophysical Ages and Times Scales*, volume 245 of *Astronomical Society of the Pacific Conference Series*, page 561.

Davé, R., Hernquist, L., Katz, N., and Weinberg, D. H. (1999). The Low-Redshift Lyα Forest in Cold Dark Matter Cosmologies. *ApJ*, 511(2):521–545.

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

Dawson, K. S., Kneib, J.-P., Percival, W. J., Alam, S., Albareti, F. D., Anderson, S. F., Armengaud, E., Aubourg, É., Bailey, S., Bautista, J. E., Berlind, A. A., Bershady, M. A., Beutler, F., Bizyaev, D., Blanton, M. R., Blomqvist, M., Bolton, A. S., Bovy, J., Brandt, W. N., Brinkmann, J., Brownstein, J. R., Burtin, E., Busca, N. G., Cai, Z., Chuang, C.-H., Clerc, N., Comparat, J., Cope, F., Croft, R. A. C., Cruz-Gonzalez, I., da Costa, L. N., Cousinou, M.-C., Darling, J., de la Macorra, A., de la Torre, S., Delubac, T., du Mas des Bourboux, H., Dwelly, T., Ealet, A., Eisenstein, D. J., Eracleous, M., Escoffier, S., Fan, X., Finoguenov, A., Font-Ribera, A., Frinchaboy, P., Gaulme, P., Georgakakis, A., Green, P., Guo, H., Guy, J., Ho, S., Holder, D., Huehnerhoff, J., Hutchinson, T., Jing, Y., Jullo, E., Kamble, V., Kinemuchi, K., Kirkby, D., Kitaura, F.-S., Klaene, M. A., Laher, R. R., Lang, D., Laurent, P., Le Goff, J.-M., Li, C., Liang, Y., Lima, M., Lin, Q., Lin, W., Lin, Y.-T., Long, D. C., Lundgren, B., MacDonald, N., Geimba Maia, M. A., Malanushenko, E., Malanushenko, V., Mariappan, V., McBride, C. K., McGreer, I. D., Ménard, B., Merloni, A., Meza, A., Montero-Dorta, A. D., Muna, D., Myers, A. D., Nandra, K., Naugle, T., Newman, J. A., Noterdaeme, P., Nugent, P., Ogando, R., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Peacock, J. A., Petitjean, P., Pieri, M. M., Pisani, A., Prada, F., Prakash, A., Raichoor, A., Reid, B., Rich, J., Ridl, J., Rodriguez-Torres, S., Carnero Rosell, A., Ross, A. J., Rossi, G., Ruan, J., Salvato, M., Sayres, C., Schneider, D. P., Schlegel, D. J., Seljak, U., Seo, H.-J., Sesar, B., Shandera, S., Shu, Y., Slosar, A., Sobreira, F., Streblyanska, A., Suzuki, N., Taylor, D., Tao, C., Tinker, J. L., Tojeiro, R., Vargas-Magaña, M., Wang, Y., Weaver, B. A., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yeche, C., Zhai, Z., Zhao, C., Zhao, G.-b., Zheng, Z., Ben Zhu, G., and Zou, H. (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. *Astrophysical Journal*, 151:44.

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., Blomqvist, M., Bolton, A. S., Borde, A., Bovy, J., Brandt, W. N., Brewington, H., Brinkmann, J., Brown, P. J., Brownstein, J. R., Bundy, K., Busca, N. G., Carithers, W., Carnero, A. R., Carr, M. A., Chen, Y., Comparat, J., Connolly, N., Cope, F., Croft, R. A. C., Cuesta, A. J., da Costa, L. N., Davenport, J. R. A., Delubac, T., de Putter, R., Dhital, S., Ealet, A., Ebelke, G. L., Eisenstein, D. J., Escoffier, S., Fan, X., Filiz Ak, N., Finley, H., Font-Ribera, A., Génova-Santos, R., Gunn, J. E., Guo, H., Haggard, D., Hall, P. B., Hamilton, J.-C., Harris, B., Harris, D. W., Ho, S., Hogg, D. W., Holder, D., Honscheid, K., Huehnerhoff, J., Jordan, B., Jordan, W. P., Kauffmann, G., Kazin, E. A., Kirkby, D., Klaene, M. A., Kneib, J.-P., Le Goff, J.-M., Lee, K.-G., Long, D. C., Loomis, C. P., Lundgren, B., Lupton, R. H., Maia, M. A. G., Makler, M., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., Margala, D., Masters, K. L., McBride, C. K., McDonald, P., McGreer, I. D., McMahon, R. G., Mena, O., Miralda-Escudé, J., Montero-Dorta, A. D., Montesano, F., Muna, D., Myers, A. D., Naugle, T., Nichol, R. C., Noterdaeme, P., Nuza, S. E., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Owen, R., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Percival, W. J., Pérez-Fournon, I., Pérez-Ràfols, I., Petitjean, P., Pfaffenberger, R., Pforr, J., Pieri, M. M., Prada, F., Price-Whelan, A. M., Raddick, M. J., Rebolo, R., Rich, J., Richards, G. T., Rockosi, C. M., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martin, J. A., Samushia, L., Sánchez, A. G., Sayres, C., Schmidt, S. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Shelden, A., Sheldon, E., Shen, Y., Shu, Y., Slosar, A., Smee, S. A., Snedden, S. A., Stauffer, F., Steele, O., Strauss, M. A., Streblyanska, A., Suzuki, N., Swanson, M. E. C., Tal, T., Tanaka, M., Thomas, D., Tinker, J. L., Tojeiro, R., Tremonti, C. A., Vargas Magaña, M., Verde, L., Viel, M., Wake, D. A., Watson, M., Weaver, B. A., Weinberg, D. H., Weiner, B. J., West, A. A., White, M., Wood-Vasey, W. M., Yeche, C., Zehavi, I., Zhao, G.-B., and Zheng, Z. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. *AJ*, 145:10.

Décamp, D., Deschizeaux, B., P Lees, J., N Minard, M., M Crespo, J., Delfino, M., Fernandez, E., Martinez, M., Miquel, R., M Mir, L., Orteu, S., Pacheco Pages, A., Perlas, J., Tubau, E., G Catanesi, M., De Palma, M., Farilla, A., Iaselli, G., Maggi, G., and Zobernig, G. (1990). A precise determination of the number of families with light neutrinos and of the z boson partial widths.

Dekel, A., Eldar, A., Kolatt, T., Yahil, A., Willick, J. A., Faber, S. M., Courteau, S., and Burstein, D. (1999). POTENT Reconstruction from Mark III Velocities. *ApJ*, 522(1):1–38.

Delubac, T., Bautista, J. E., Busca, N. G., Rich, J., Kirkby, D., Bailey, S., Font-Ribera, A., Slosar, A., Lee, K.-G., and Pieri, M. M. (2015). Baryon acoustic oscillations in the Lyα forest of BOSS DR11 quasars. *A&A*, 574:A59.

DESI Collaboration, Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Allende Prieto, C., Annis, J., Bailey, S., and Balland, C. (2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*, page arXiv:1611.00036.

Desjacques, V., Jeong, D., and Schmidt, F. (2018). Large-scale galaxy bias. *Phys. Rep.*, 733:1–193.

Dodelson, S. (2003). *Modern cosmology.*

du Mas des Bourboux, H., Le Goff, J.-M., Blomqvist, M., Busca, N. G., Guy, J., Rich, J., Yèche, C., Bautista, J. E., Burtin, É., and Dawson, K. S. (2017). Baryon acoustic oscillations from the complete SDSS-III Lyα-quasar cross-correlation function at z = 2.4. *A&A*, 608:A130.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195:216–222.

Efstathiou, G. (1999). Constraining the equation of state of the Universe from distant Type Ia supernovae and cosmic microwave background anisotropies. *MNRAS*, 310(3):842–850.

Eilers, A.-C., Hennawi, J. F., and Lee, K.-G. (2017). Joint Bayesian Estimation of Quasar Continua and the Lyα Forest Flux Probability Distribution Function. *ApJ*, 844:136.

Einasto, J. (2009). Dark Matter. *arXiv e-prints*, page arXiv:0901.0632.

Einasto, M., Saar, E., Martínez, V. J., Einasto, J., Liivamägi, L. J., Tago, E., Starck, J.-L., Müller, V., Heinämäki, P., Nurmi, P., Paredes, S., Gramann, M., and Hütsi, G. (2008). Toward Understanding Rich Superclusters. *ApJ*, 685:83–104.

Einasto, M., Suhhonenko, I., Heinämäki, P., Einasto, J., and Saar, E. (2005). Environmental enhancement of DM haloes. *A&A*, 436:17–24.

Eisenstein, D. J. and Hu, W. (1998). Baryonic Features in the Matter Transfer Function. *ApJ*, 496:605–+.

Eisenstein, D. J. and Hu, W. (1999). Power Spectra for Cold Dark Matter and Its Variants. *ApJ*, 511:5–15.

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., and et al. (2011). SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. *Astrophysical Journal*, 142:72.

Elsner, F., Leistedt, B., and Peiris, H. V. (2017). Unbiased pseudo-C power spectrum estimation with mode projection. *MNRAS*, 465:1847–1855.

Eriksen, H. K., Dickinson, C., Jewell, J. B., Banday, A. J., Górski, K. M., and Lawrence, C. R. (2008). The Joint Large-Scale Foreground-CMB Posteriors of the 3 Year WMAP Data. *ApJ*, 672:L87.

Ferrarese, L. and Merritt, D. (2000). A Fundamental Relation between Supermassive Black Holes and Their Host Galaxies. *ApJ*, 539:L9–L12.

Filippenko, A. V. and Terlevich, R. (1992). O-star photoionization models of liners with weak forbidden O I 6300 A emission. *ApJ*, 397:L79–L82.

Flores, R. A. and Primack, J. R. (1994). Observational and Theoretical Constraints on Singular Dark Matter Halos. *ApJ*, 427:L1.

Forero-Romero, J. E., Hoffman, Y., Gottlöber, S., Klypin, A., and Yepes, G. (2009). A dynamical classification of the cosmic web. *MNRAS*, 396:1815–1824.

Frenk, C. S. and White, S. D. M. (2012). Dark matter and cosmic structure. *Annalen der Physik*, 524(9-10):507–534.

Friedmann, A. (1922). Über die Krümmung des Raumes. *Zeitschrift fur Physik*, 10:377–386.

Friedmann, A. (1924). Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes. *Zeitschrift fur Physik*, 21:326–332.

Frisch, U., Matarrese, S., Mohayaee, R., and Sobolevski, A. (2002). A reconstruction of the initial conditions of the Universe by optimal mass transportation. *Nature*, 417(6886):260–262.

Gallerani, S., Kitaura, F. S., and Ferrara, A. (2011). Cosmic density field reconstruction from Lyα forest data. *MNRAS*, 413:L6–L10.

Garzilli, A., Bolton, J. S., Kim, T. S., Leach, S., and Viel, M. (2012). The intergalactic medium thermal history at redshift z = 1.7-3.2 from the Lyα forest: a comparison of measurements using wavelets and the flux distribution. *MNRAS*, 424(3):1723–1736.

Geller, M. J. (1990). Mapping the universe - Slices and bubbles. *Mercury*, 19:66–76.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. (2005). HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *ApJ*, 622:759–771.

Granett, B. R., Neyrinck, M. C., and Szapudi, I. (2008). An Imprint of Superstructures on the Microwave Background due to the Integrated Sachs-Wolfe Effect. *ApJ*, 683(2):L99.

Gregg, M. D. (1995). The Infrared Diameter-Velocity Dispersion Relation for Elliptical Galaxies. *AJ*, 110:1052.

Gregory, P. (2010). *Bayesian Logical Data Analysis for the Physical Sciences*.

Gregory, S. A. and Thompson, L. A. (1978). The Coma/A1367 supercluster and its environs. *ApJ*, 222:784–799.

Gregory, S. A., Thompson, L. A., and Tifft, W. G. (1981). The Perseus supercluster. *ApJ*, 243:411–426.

Gunn, J. E. and Peterson, B. A. (1965). On the Density of Neutral Hydrogen in Intergalactic Space. *ApJ*, 142:1633–1641.

Guth, A. H. (1981). Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D*, 23:347–356.

Guth, A. H. and Pi, S.-Y. (1982). Fluctuations in the new inflationary universe. *Physical Review Letters*, 49:1110–1113.

Hahn, O., Porciani, C., Carollo, C. M., and Dekel, A. (2007). Properties of dark matter haloes in clusters, filaments, sheets and voids. *MNRAS*, 375:489–499.

Hamaus, N., Pisani, A., Sutter, P. M., Lavaux, G., Escoffier, S., Wand elt, B. D., and Weller, J. (2016). Constraints on Cosmology and Gravity from the Dynamics of Voids. *Phys. Rev. Lett.*, 117(9):091302.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hata, N., Scherrer, R. J., Steigman, G., Thomas, D., and Walker, T. P. (1996). Predicting Big Bang Deuterium. *ApJ*, 458:637.

He, S., Alam, S., Ferraro, S., Chen, Y.-C., and Ho, S. (2018). The detection of the imprint of filaments on cosmic microwave background lensing. *Nature Astronomy*, 2(5):401–406.

Heavens, A. (2003). 3D weak lensing. *MNRAS*, 343:1327–1334.

Heckman, T. M. and Best, P. N. (2014). The Coevolution of Galaxies and Supermassive Black Holes: Insights from Surveys of the Contemporary Universe. *ARA&A*, 52:589–660.

Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., Tremonti, C., and White, S. D. M. (2004). Present-Day Growth of Black Holes and Bulges: The Sloan Digital Sky Survey Perspective. *ApJ*, 613:109–118.

Hernquist, L., Katz, N., Weinberg, D. H., and Miralda-Escudé, J. (1996). The Lyman-Alpha Forest in the Cold Dark Matter Model. *ApJ*, 457:L51.

Hinshaw, G., Nolta, M. R., Bennett, C. L., Bean, R., Doré, O., Greason, M. R., Halpern, M., Hill, R. S., Jarosik, N., Kogut, A., Komatsu, E., Limon, M., Odegard, N., Meyer, S. S., Page, L., Peiris, H. V., Spergel, D. N., Tucker, G. S., Verde, L., Weiland, J. L., Wollack, E., and Wright, E. L. (2007). Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Temperature Analysis. *ApJS*, 170:288–334.

Ho, L. C., Filippenko, A. V., and Sargent, W. L. W. (2003). A Search for “Dwarf” Seyfert Nuclei. VI. Properties of Emission-Line Nuclei in Nearby Galaxies. *ApJ*, 583:159–177.

Ho, S., Agarwal, N., Myers, A. D., Lyons, R., Disbrow, A., Seo, H.-J., Ross, A., Hirata, C., Padmanabhan, N., O’Connell, R., Huff, E., Schlegel, D., Slosar, A., Weinberg, D., Strauss, M., Ross, N. P., Schneider, D. P., Bahcall, N., Brinkmann, J., Palanque-Delabrouille, N., and Yèche, C. (2015). Sloan Digital Sky Survey III photometric quasar clustering: probing the initial conditions of the Universe. *J. Cosmology Astropart. Phys.*, 5:040.

Ho, S., Cuesta, A., Seo, H.-J., de Putter, R., Ross, A. J., White, M., Padmanabhan, N., Saito, S., Schlegel, D. J., Schlafly, E., Seljak, U., Hernández- Monteagudo, C., Sánchez, A. G., Percival, W. J., Blanton, M., Skibba, R., Schneider, D., Reid, B., Mena, O., Viel, M., Eisenstein, D. J., Prada, F., Weaver, B. A., Bahcall, N., Bizyaev, D., Brewinton, H., Brinkman, J., Nicolaci da Costa, L., Gott, J. R., Malanushenko, E., Malanushenko, V., Nichol, B., Oravetz, D., Pan, K., Palanque-Delabrouille, N., Ross, N. P., Simmons, A., de Simoni, F., Snedden, S., and Yeche, C. (2012). Clustering of Sloan Digital Sky Survey III Photometric Luminous Galaxies: The Measurement, Systematics, and Cosmological Implications. *ApJ*, 761:14.

Hockney, R. W. and Eastwood, J. W. (1988). *Computer simulation using particles.*

Hoekstra, H. (2005). Mapping the Dark Matter Using Weak Lensing. In Colless, M., Staveley-Smith, L., and Stathakis, R. A., editors, *Maps of the Cosmos*, volume 216 of *IAU Symposium*, page 140.

Horowitz, B., Lee, K.-G., White, M., Krolewski, A., and Ata, M. (2019). TARDIS Paper I: A Constrained Reconstruction Approach to Modeling the z˜2.5 Cosmic Web Probed by Lyman-alpha Forest Tomography. *arXiv e-prints*, page arXiv:1903.09049.

Huchra, J. P., Geller, M. J., de Lapparent, V., and Burg, R. (1988). The CFA Redshift Survey. In Audouze, J., Pelletan, M. C., Szalay, A., Zel'dovich, Y. B., and Peebles, P. J. E., editors, *Large Scale Structures of the Universe*, volume 130 of *IAU Symposium*, page 105.

Hui, L. and Gnedin, N. Y. (1997). Equation of state of the photoionized intergalactic medium. *MNRAS*, 292(1):27–42.

Huterer, D., Cunha, C. E., and Fang, W. (2013). Calibration errors unleashed: effects on cosmological parameters and requirements for large-scale structure surveys. *MNRAS*, 432:2945–2961.

Huterer, D., Kirkby, D., Bean, R., Connolly, A., Dawson, K., Dodelson, S., Evrard, A., Jain, B., Jarvis, M., and Linder, E. (2015). Growth of cosmic structure: Probing dark energy beyond expansion. *Astroparticle Physics*, 63:23–41.

Huterer, D. and Turner, M. S. (1999). Prospects for probing the dark energy via supernova distance measurements. *Phys. Rev. D*, 60(8):081301.

Hwang, H. S., Park, C., Elbaz, D., and Choi, Y.-Y. (2012). Activity in galactic nuclei of cluster and field galaxies in the local universe. *A&A*, 538:A15.

Ivezic, Z., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, R., Angeli, G., Ansari, R., Antilogus, P., Arndt, K. T., Astier, P., Aubourg, E., Axelrod, T., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauman, B. J., Beaumont, S., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Blanc, G., Blandford, R. D., Bloom, J. S., Bogart, J., Borne, K., Bosch, J. F., Boutigny, D., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Chandrasekharan, S., Chesley, S., Cheu, E. C., Chiang, J., Claver, C. F., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daubard, G., Daues, G., Delgado, F., Digel, S., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Eracleous, M., Ferguson, H., Frank, J., Freemon, M., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gibson, R. R., Gilmore, D. K., Glanzman, T., Goodenow, I., Gressler, W. J., Gris, P., Guyonnet, A., Hascall, P. A., Haupt, J., Hernandez, F., Hogan, C., Huang, D., Huffer, M. E., Innes, W. R., Jacoby, S. H., Jain, B., Jee, J., Jernigan, J. G., Jevremovic, D., Johns, K., Jones, R. L., Juramy-Gilles, C., Juric, M., Kahn, S. M., Kalirai, J. S., Kallivayalil, N., Kalmbach, B., Kantor, J. P., Kasliwal, M. M., Kessler, R., Kirkby, D., Knox, L., Kotov, I., Krabbendam, V. L., Krughoff, S., Kubanek, P., Kuczewski, J., Kulkarni, S., Lambert, R., Le Guillou, L., Levine, D., Liang, M., Lim, K., Lintott, C., Lupton, R. H., Mahabal, A., Marshall, P., Marshall, S., May, M., McKercher, R., Migliore, M., Miller, M., Mills, D. J., Monet, D. G., Moniez, M., Neill, D. R., Nief, J., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., Ortiz, S., Owen, R. E., Pain, R., Peterson, J. R., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A., Regnault, N., Ridgway, S. T., Ritz, S., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schindler, R. H., Schneider, D. P., Schumacher, G., Sebag, J., Sembroski, G. H., Seppala, L. G., Shipsey, I., Silvestri, N., Smith, J. A., Smith, R. C., Strauss, M. A., Stubbs, C. W., Sweeney, D., Szalay, A., Takacs, P., Thaler, J. J., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Xin, B., Walkowicz, L., Walter, C. W., Wang, D. L., Warner, M., Willman, B., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Yoachim, P., Zhan, H., and for the LSST Collaboration (2008). LSST: from Science Drivers to Reference Design and Anticipated Data Products. *ArXiv e-prints*.

Jasche, J. and Kitaura, F. S. (2010). Fast Hamiltonian sampling for large-scale structure inference. *MNRAS*, 407:29–42.

Jasche, J. and Lavaux, G. (2017). Bayesian power spectrum inference with foreground and target contamination treatment. *A&A*, 606:A37.

Jasche, J. and Lavaux, G. (2018). Physical Bayesian modelling of the non-linear matter distribution: new insights into the Nearby Universe. *ArXiv e-prints*.

Jasche, J., Leclercq, F., and Wandelt, B. D. (2015). Past and present cosmic structure in the SDSS DR7 main sample. *J. Cosmology Astropart. Phys.*, 1:036.

Jasche, J. and Wandelt, B. D. (2012). Bayesian inference from photometric redshift surveys. *MNRAS*, 425:1042–1056.

Jasche, J. and Wandelt, B. D. (2013). Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *MNRAS*, 432:894–913.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, 186(1007):453–461.

Jiang, N., Wang, H., Mo, H., Dong, X.-B., Wang, T., and Zhou, H. (2016). Differences in Halo-scale Environments between Type 1 and Type 2 AGNs at Low Redshift. *ApJ*, 832:111.

Karhunen, K., Kotilainen, J. K., Falomo, R., and Bettoni, D. (2014). Low-redshift quasars in the SDSS Stripe 82. The local environments. *MNRAS*, 441:1802–1816.

Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D. M., Ridgway, S. E., Brinkmann, J., Fukugita, M., Hall, P. B., Ivezić, Ž., Richards, G. T., and Schneider, D. P. (2003). The host galaxies of active galactic nuclei. *MNRAS*, 346:1055–1077.

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., and Trevena, J. (2001). Theoretical Modeling of Starburst Galaxies. *ApJ*, 556:121–140.

Kirshner, R. P., Oemler, Jr., A., Schechter, P. L., and Shectman, S. A. (1981). A million cubic megaparsec void in Bootes. *ApJ*, 248:L57–L60.

Kitaura, F.-S., Angulo, R. E., Hoffman, Y., and Gottlöber, S. (2012a). Estimating cosmic velocity fields from density fields and tidal tensors. *MNRAS*, 425(4):2422–2435.

Kitaura, F. S. and Enßlin, T. A. (2008). Bayesian reconstruction of the cosmological large-scale structure: methodology, inverse algorithms and numerical optimization. *MNRAS*, 389:497–544.

Kitaura, F.-S., Gallerani, S., and Ferrara, A. (2012b). Multiscale inference of matter fields and baryon acoustic oscillations from the Lyα forest. *MNRAS*, 420:61–74.

Kitaura, F. S., Jasche, J., Li, C., Enßlin, T. A., Metcalf, R. B., Wandelt, B. D., Lemson, G., and White, S. D. M. (2009). Cosmic cartography of the large-scale structure with Sloan Digital Sky Survey data release 6. *MNRAS*, 400:183–203.

Köhlinger, F., Viola, M., Joachimi, B., Hoekstra, H., van Uitert, E., Hildebrandt, H., Choi, A., Erben, T., Heymans, C., and Joudaki, S. (2017). KiDS-450: the tomographic weak lensing power spectrum and constraints on cosmological parameters. *MNRAS*, 471(4):4412–4435.

Krauss, L. M. and Romanelli, P. (1990). Big Bang Nucleosynthesis: Predictions and Uncertainties. *ApJ*, 358:47.

Kreisch, C. D., Pisani, A., Carbone, C., Liu, J., Hawken, A. J., Massara, E., Spergel, D. N., and Wandelt, B. D. (2018). Massive Neutrinos Leave Fingerprints on Cosmic Voids. *ArXiv e-prints*.

Lahav, O., Fisher, K. B., Hoffman, Y., Scharf, C. A., and Zaroubi, S. (1994). Wiener Reconstruction of All-Sky Galaxy Surveys in Spherical Harmonics. *ApJ*, 423:L93.

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., Garilli, B., Gondoin, P., Guzzo, L., Hoar, J., Hoekstra, H., Holmes, R., Kitching, T., Maciaszek, T., Mellier, Y., Pasian, F., Percival, W., Rhodes, J., Saavedra Criado, G., Sauvage, M., Scaramella, R., Valenziano, L., Warren, S., Bender, R., Castander, F., Cimatti, A., Le Fèvre, O., Kurki-Suonio, H., Levi, M., Lilje, P., Meylan, G., Nichol, R., Pedersen, K., Popa, V., Rebolo Lopez, R., Rix, H. W., Rottgering, H., Zeilinger, W., Grupp, F., Hudelot, P., Massey, R., Meneghetti, M., Miller, L., Paltani, S., Paulin-Henriksson, S., Pires, S., Saxton, C., Schrabback, T., Seidel, G., Walsh, J., Aghanim, N., Amendola, L., Bartlett, J., Baccigalupi, C., Beaulieu, J. P., Benabed, K., Cuby, J. G., Elbaz, D., Fosalba, P., Gavazzi, G., Helmi, A., Hook, I., Irwin, M., Kneib, J. P., Kunz, M., Mannucci, F., Moscardini, L., Tao, C., Teyssier, R., Weller, J., Zamorani, G., Zapatero Osorio, M. R., Boulade, O., Foumond, J. J., Di Giorgio, A., Guttridge, P., James, A., Kemp, M., Martignac, J., Spencer, A., Walton, D., Blümchen, T., Bonoli, C., Bortoletto, F., Cerna, C., Corcione, L., Fabron, C., Jahnke, K., Ligori, S., Madrid, F., Martin, L., Morgante, G., Pamplona, T., Prieto, E., Riva, M., Toledo, R., Trifoglio, M., Zerbi, F., Abdalla, F., Douspis, M., Grenet, C., Borgani, S., Bouwens, R., Courbin, F., Delouis, J. M., Dubath, P., Fontana, A., Frailis, M., Grazian, A., Koppenhöfer, J., Mansutti, O., Melchior, M., Mignoli, M., Mohr, J., Neissner, C., Noddle, K., Poncet, M., Scodeggio, M., Serrano, S., Shane, N., Starck, J. L., Surace, C., Taylor, A., Verdoes-Kleijn, G., Vuerli, C., Williams, O. R., Zacchei, A., Altieri, B., Escudero Sanz, I., Kohley, R., Oosterbroek, T., Astier, P., Bacon, D., Bardelli, S., Baugh, C., Bellagamba, F., Benoist, C., Bianchi, D., Biviano, A., Branchini, E., Carbone, C., Cardone, V., Clements, D., Colombi, S., Conselice, C., Cresci, G., Deacon, N., Dunlop, J., Fedeli, C., Fontanot, F., Franzetti, P., Giocoli, C., Garcia-Bellido, J., Gow, J., Heavens, A., Hewett, P., Heymans, C., Holland, A., Huang, Z., Ilbert, O., Joachimi, B., Jennins, E., Kerins, E., Kiessling, A., Kirk, D., Kotak, R., Krause, O., Lahav, O., van Leeuwen, F., Lesgourgues, J., Lombardi, M., Magliocchetti, M., Maguire, K., Majerotto, E., Maoli, R., Marulli, F., Maurogordato, S., McCracken, H., McLure, R., Melchiorri, A., Merson, A., Moresco, M., Nonino, M., Norberg, P., Peacock, J., Pello, R., Penny, M., Pettorino, V., Di Porto, C., Pozzetti, L., Quercellini, C., Radovich, M., Rassat, A., Roche, N., Ronayette, S., Rossetti, E., Sartoris, B., Schneider, P., Semboloni, E., Serjeant, S., Simpson, F., Skordis, C., Smadja, G., Smartt, S., Spano, P., Spiro, S., Sullivan, M., Tilquin, A., Trotta, R., Verde, L., Wang, Y., Williger, G., Zhao, G., Zoubian, J., and Zucca, E. (2011). Euclid Definition Study Report. *arXiv e-prints*, page arXiv:1110.3193.

Lavaux, G. (2008). Lagrangian reconstruction of cosmic velocity fields. *Physica D Nonlinear Phenomena*, 237(14-17):2139–2144.

Lavaux, G. and Hudson, M. J. (2011). The 2M++ galaxy redshift catalogue. *MNRAS*, 416:2840–2856.

Lavaux, G. and Jasche, J. (2016). Unmasking the masked Universe: the 2M++ catalogue through Bayesian eyes. *MNRAS*, 455:3169–3179.

Lavaux, G. and Wandelt, B. D. (2012). Precision Cosmography with Stacked Voids. *ApJ*, 754(2):109.

Layzer, D. (1956). A new model for the distribution of galaxies in space. *AJ*, 61:383.

Leclercq, F. (2014). Bayesian inference of dark matter voids in galaxy surveys. *ArXiv e-prints*.

Leclercq, F., Jasche, J., Gil-Marín, H., and Wandelt, B. (2013). One-point remapping of Lagrangian perturbation theory in the mildly non-linear regime of cosmic structure formation. *J. Cosmology Astropart. Phys.*, 11:048.

Leclercq, F., Jasche, J., Sutter, P. M., Hamaus, N., and Wandelt, B. (2015). Dark matter voids in the SDSS galaxy survey. *J. Cosmology Astropart. Phys.*, 3:047.

Lee, K.-G. (2012). Systematic Continuum Errors in the Lyα Forest and the Measured Temperature-Density Relation. *ApJ*, 753:136.

Lee, K.-G. (2016). Lyα Forest Tomography of the z &gt; 2 Cosmic Web. In van de Weygaert, R., Shandarin, S., Saar, E., and Einasto, J., editors, *The Zeldovich Universe: Genesis and Growth of the Cosmic Web*, volume 308 of *IAU Symposium*, pages 360–363.

Lee, K.-G., Bailey, S., Bartsch, L. E., Carithers, W., Dawson, K. S., Kirkby, D., Lundgren, B., Margala, D., Palanque-Delabrouille, N., Pieri, M. M., Schlegel, D. J., Weinberg, D. H., Yèche, C., Aubourg, É., Bautista, J., Bizyaev, D., Blomqvist, M., Bolton, A. S., Borde, A., Brewington, H., Busca, N. G., Croft, R. A. C., Delubac, T., Ebelke, G., Eisenstein, D. J., Font-Ribera, A., Ge, J., Hamilton, J.-C., Hennawi, J. F., Ho, S., Honscheid, K., Le Goff, J.-M., Malanushenko, E., Malanushenko, V., Miralda-Escudé, J., Myers, A. D., Noterdaeme, P., Oravetz, D., Pan, K., Pâris, I., Petitjean, P., Rich, J., Rollinde, E., Ross, N. P., Rossi, G., Schneider, D. P., Simmons, A., Snedden, S., Slosar, A., Spergel, D. N., Suzuki, N., Viel, M., and Weaver, B. A. (2013). The BOSS Lyα Forest Sample from SDSS Data Release 9. *AJ*, 145:69.

Lee, K.-G., Krolewski, A., White, M., Schlegel, D., Nugent, P. E., Hennawi, J. F., Müller, T., Pan, R., Prochaska, J. X., Font-Ribera, A., Suzuki, N., Glazebrook, K., Kacprzak, G. G., Kartaltepe, J. S., Koekemoer, A. M., Le Fèvre, O., Lemaux, B. C., Maier, C., Nanayakkara, T., Rich, R. M., Sanders, D. B., Salvato, M., Tasca, L., and Tran, K.-V. H. (2018). First Data Release of the COSMOS Lyα Mapping and Tomography Observations: 3D Lyα Forest Tomography at 2.05 &lt; z &lt; 2.55. *The Astrophysical Journal Supplement Series*, 237:31.

Leistedt, B. and Peiris, H. V. (2014). Exploiting the full potential of photometric quasar surveys: optimal power spectra through blind mitigation of systematics. *MNRAS*, 444:2–14.

Lesgourgues, J. and Pastor, S. (2006). Massive neutrinos and cosmology. *Phys. Rep.*, 429(6):307–379.

Levi, M., Bebek, C., Beers, T., Blum, R., Cahn, R., Eisenstein, D., Flaugher, B., Honscheid, K., Kron, R., and Lahav, O. (2013). The DESI Experiment, a whitepaper for Snowmass 2013. *arXiv e-prints*, page arXiv:1308.0847.

Li, B. (2011). Voids in coupled scalar field cosmology. *MNRAS*, 411(4):2615–2627.

Li, C., Kauffmann, G., Wang, L., White, S. D. M., Heckman, T. M., and Jing, Y. P. (2006). The clustering of narrow-line AGN in the local Universe. *MNRAS*, 373:457–468.

Lidz, A., Faucher-Giguère, C.-A., Dall'Aglio, A., McQuinn, M., Fechner, C., Zaldarriaga, M., Hernquist, L., and Dutta, S. (2010). A Measurement of Small-scale Structure in the 2.2 &lt;= z &lt;= 4.2 Lyα Forest. *ApJ*, 718(1):199–230.

Lietzen, H., Heinämäki, P., Nurmi, P., Liivamägi, L. J., Saar, E., Tago, E., Takalo, L. O., and Einasto, M. (2011). Large-scale environments of z $<$ 0.4 active galaxies. *A&A*, 535:A21.

Lietzen, H., Heinämäki, P., Nurmi, P., Tago, E., Saar, E., Liivamägi, J., Tempel, E., Einasto, M., Einasto, J., Gramann, M., and Takalo, L. O. (2009). Environments of nearby quasars in Sloan Digital Sky Survey. *A&A*, 501:145–155.

Linde, A. D. (1982). A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Physics Letters B*, 108:389–393.

Liu, J., Chen, X., and Ji, X. (2017). Current status of direct dark matter detection experiments. *Nature Physics*, 13(3):212–216.

Loeb, A. and Furlanetto, S. R. (2013). *The First Galaxies in the Universe.*

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D., Asztalos, S. J., and Axelrod, T. S. (2009). LSST Science Book, Version 2.0. *arXiv e-prints*, page arXiv:0912.0201.

Lynds, R. (1971). The Absorption-Line Spectrum of 4c 05.34. *ApJ*, 164:L73.

Mackay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms.*

Markevitch, M., Gonzalez, A. H., Clowe, D., Vikhlinin, A., Forman, W., Jones, C., Murray, S., and Tucker, W. (2004). Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657-56. *ApJ*, 606(2):819–824.

Martínez, V. J. and Saar, E. (2003). *Statistics of galaxy clustering*, pages 143–160.

Massara, E., Villaescusa-Navarro, F., Viel, M., and Sutter, P. M. (2015). Voids in massive neutrino cosmologies. *Journal of Cosmology and Astro-Particle Physics*, 2015(11):018.

Mather, J. C., Cheng, E. S., Cottingham, D. A., Eplee, Jr., R. E., Fixsen, D. J., Hewagama, T., Isaacman, R. B., Jensen, K. A., Meyer, S. S., Noerdlinger, P. D., Read, S. M., Rosen, L. P., Shafer, R. A., Wright, E. L., Bennett, C. L., Boggess, N. W., Hauser, M. G., Kelsall, T., Moseley, Jr., S. H., Silverberg, R. F., Smoot, G. F., Weiss, R., and Wilkinson, D. T. (1994). Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. *ApJ*, 420:439–444.

McConnachie, A., Babusiaux, C., Balogh, M., Driver, S., Côté, P., Courtois, H., Davies, L., Ferrarese, L., Gallagher, S., and Ibata, R. (2016). The Detailed Science Case for the Maunakea Spectroscopic Explorer: the Composition and Dynamics of the Faint Universe. *arXiv e-prints*, page arXiv:1606.00043.

McDonald, P., Miralda-Escudé, J., Rauch, M., Sargent, W. L. W., Barlow, T. A., and Cen, R. (2001). A Measurement of the Temperature-Density Relation in the Intergalactic Medium Using a New Lyα Absorption-Line Fitting Method. *ApJ*, 562(1):52–75.

Merloni, A. and Heinz, S. (2013). *Evolution of Active Galactic Nuclei*, page 503.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Cosmology and Astroparticle Physics*, 21:1087–1092.

Miralda-Escudé, J., Cen, R., Ostriker, J. P., and Rauch, M. (1995). Hydrodynamic Simulations of the Lyman-Alpha Forest in a Theory of Gravitational Collapse. In Meylan, G., editor, *QSO Absorption Lines*, page 427.

Mishra-Sharma, S., Alonso, D., and Dunkley, J. (2018). Neutrino masses and beyond-Λ CDM cosmology with LSST and future CMB experiments. *Phys. Rev. D*, 97(12):123544.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation*.

Mo, H., van den Bosch, F. C., and White, S. (2010). *Galaxy Formation and Evolution*.

Mohayaee, R. and Sobolevskiĭ, A. (2008). The Monge Ampère Kantorovich approach to reconstruction in cosmology. *Physica D Nonlinear Phenomena*, 237(14-17):2145–2150.

Monaco, P., Di Dio, E., and Sefusatti, E. (2018). A blind method to recover the mask of a deep galaxy survey. *arXiv e-prints*, page arXiv:1812.02104.

Moore, B. (1994). Evidence against dissipation-less dark matter from observations of galaxy haloes. *Nature*, 370(6491):629–631.

Moore, B., Quinn, T., Governato, F., Stadel, J., and Lake, G. (1999). Cold collapse and the core catastrophe. *MNRAS*, 310(4):1147–1152.

Moutarde, F., Alimi, J.-M., Bouchet, F. R., Pellat, R., and Ramani, A. (1991). Precollapse scale invariance in gravitational instability. *ApJ*, 382:377–381.

Mukhanov, V. (2005). *Physical Foundations of Cosmology.*

Myers, A. D., Palanque-Delabrouille, N., Prakash, A., Pâris, I., Yeche, C., Dawson, K. S., Bovy, J., Lang, D., Schlegel, D. J., and Newman, J. A. (2015). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Quasar Target Selection. *ApJS*, 221(2):27.

Nakamura, T. and Chiba, T. (2001). Determining the Equation of State of the Expanding Universe Using a New Independent Variable. *ApJ*, 550(1):1–6.

Nasir, F., Bolton, J. S., and Becker, G. D. (2016). Inferring the IGM thermal history during reionization with the Lyman α forest power spectrum at redshift z 5. *MNRAS*, 463(3):2335–2347.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1997). A Universal Density Profile from Hierarchical Clustering. *ApJ*, 490(2):493–508.

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. *Technical Report CRG-TR-93-1.*

Neal, R. M. (2000). Slice Sampling. *arXiv e-prints*, page physics/0009028.

Neyrinck, M. C., Aragón-Calvo, M. A., Jeong, D., and Wang, X. (2014). A halo bias function measured deeply into voids without stochasticity. *MNRAS*, 441:646–655.

Nusser, A. and Dekel, A. (1992). Tracing Large-Scale Fluctuations Back in Time. *ApJ*, 391:443.

Ozbek, M., Croft, R. A. C., and Khandai, N. (2016). Large-scale 3D mapping of the intergalactic medium using the Lyman α forest. *MNRAS*, 456:3610–3623.

Palanque-Delabrouille, N., Yèche, C., Baur, J., Magneville, C., Rossi, G., Lesgourgues, J., Borde, A., Burtin, E., LeGoff, J.-M., and Rich, J. (2015). Neutrino masses and cosmology with Lyman-alpha forest power spectrum. *Journal of Cosmology and Astro-Particle Physics*, 2015(11):011.

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

Peacock, J. A. and Dodds, S. J. (1996). Non-linear evolution of cosmological power spectra. *MNRAS*, 280:L19–L26.

Peacock, J. A. and Smith, R. E. (2000). Halo occupation numbers and galaxy bias. *MNRAS*, 318:1144–1156.

Peebles, P. J. E. (1980). *The large-scale structure of the universe.*

Peeples, M. S., Weinberg, D. H., Davé, R., Fardal, M. A., and Katz, N. (2010). Pressure support versus thermal broadening in the Lyman α forest - I. Effects of the equation of state on longitudinal structure. *MNRAS*, 404:1281–1294.

Peirani, S., Weinberg, D. H., Colombi, S., Blaizot, J., Dubois, Y., and Pichon, C. (2014). LyMAS: Predicting Large-scale Lyα Forest Statistics from the Dark Matter Density Field. *ApJ*, 784:11.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., and Groom, D. E. (1999). Measurements of Ω and Λ from 42 High-Redshift Supernovae. *ApJ*, 517(2):565–586.

Pimbblet, K. A., Shabala, S. S., Haines, C. P., Fraser-McKelvie, A., and Floyd, D. J. E. (2013). The drivers of AGN activity in galaxy clusters: AGN fraction as a function of mass and environment. *MNRAS*, 429:1827–1839.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., Baccigalupi, C., Banday, A. J., and et al. (2014). Planck 2013 results. XVI. Cosmological parameters. *A&A*, 571:A16.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., and Bartlett, J. G. (2016a). Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts. *A&A*, 594:A24.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. (2016b). Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hand ley, W., Herranz, D., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wand elt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. (2018). Planck 2018 results. VI. Cosmological parameters. *arXiv e-prints*, page arXiv:1807.06209.

Planck Collaboration, Akrami, Y., Arroja, F., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bond, J. R., Borrill, J., Bouchet, F. R., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Casaponsa, B., Challinor, A., Chiang, H. C., Colombo, L. P. L., Combet, C., Crill, B. P., Cuttaia, F., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Fergusson, J., Fernand ez-Cobos, R., Finelli, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hand ley, W., Hansen, F. K., Herranz, D., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Jung, G., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lindholm, V., López-Caniego, M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mand olesi, N., Marcos-Caballero, A., Maris, M., Martin, P. G., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meerburg, P. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Miville-Deschênes, M. A., Molinari, D., Moneti, A., Montier, L., Morgante, G., Moss, A., Münchmeyer, M., Natoli, P., Oppizzi, F., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Perrotta, F., Pettorino, V., Piacentini, F., Polenta, G., Puget, J. L., Rachen, J. P., Racine, B., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Savelainen, M., Scott, D., Shellard, E. P. S., Shiraishi, M., Sirignano, C., Sirri, G., Smith, K., Spencer, L. D., Stanco, L., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valiviita, J., Van Tent, B., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., Zacchei, A., and Zonca, A. (2019). Planck 2018 results. IX. Constraints on primordial non-Gaussianity. *arXiv e-prints*, page arXiv:1905.05697.

Pope, A. C., Matsubara, T., Szalay, A. S., Blanton, M. R., Eisenstein, D. J., Gray, J., Jain, B., Bahcall, N. A., Brinkmann, J., and Budavari, T. (2004). Cosmological Parameters from Eigenmode Analysis of Sloan Digital Sky Survey Galaxy Redshifts. *ApJ*, 607(2):655–660.

Porqueres, N., Jasche, J., Enßlin, T. A., and Lavaux, G. (2018). Imprints of the large-scale structure on AGN formation and evolution. *A&A*, 612:A31.

Porqueres, N., Jasche, J., Lavaux, G., and Enßlin, T. (2019a). Inferring high redshift large-scale structure dynamics from the Lyman-alpha forest. *arXiv e-prints*, page arXiv:1907.02973.

Porqueres, N., Kodi Ramanah, D., Jasche, J., and Lavaux, G. (2019b). Explicit Bayesian treatment of unknown foreground contaminations in galaxy surveys. *A&A*, 624:A115.

Racca, G. D., Laureijs, R., Stagnaro, L., Salvignol, J.-C., Lorenzo Alvarez, J., Saavedra Criado, G., Gaspar Venancio, L., Short, A., Strada, P., Bönke, T., Colombo, C., Calvi, A., Maiorano, E., Piersanti, O., Prezelus, S., Rosato, P., Pinel, J., Rozemeijer, H., Lesna, V., Musi, P., Sias, M., Anselmi, A., Cazaubiel, V., Vaillon, L., Mellier, Y., Amiaux, J., Berthé, M., Sauvage, M., Azzollini, R., Cropper, M., Pottinger, S., Jahnke, K., Ealet, A., Maciaszek, T., Pasian, F., Zacchei, A., Scaramella, R., Hoar, J., Kohley, R., Vavrek, R., Rudolph, A., and Schmidt, M. (2016). The Euclid mission design. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 9904 of *Proc. SPIE*, page 99040O.

Ramanah, D. K., Lavaux, G., Jasche, J., and Wandelt, B. D. (2019). Cosmological inference from Bayesian forward modelling of deep galaxy redshift surveys. *A&A*, 621:A69.

Refregier, A., Amara, A., Kitching, T. D., Rassat, A., Scaramella, R., Weller, J., and Euclid Imaging Consortium, f. t. (2010). Euclid Imaging Consortium Science Book. *arXiv e-prints*, page arXiv:1001.0061.

Riess, A. G., Casertano, S., Yuan, W., Macri, L., Bucciarelli, B., Lattanzi, M. G., MacKenty, J. W., Bowers, J. B., Zheng, W., Filippenko, A. V., Huang, C., and Anderson, R. I. (2018). Milky Way Cepheid Standards for Measuring Cosmic Distances and Application to Gaia DR2: Implications for the Hubble Constant. *ApJ*, 861:126.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., and Kirshner, R. P. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *AJ*, 116(3):1009–1038.

Riess, A. G., Macri, L. M., Hoffmann, S. L., Scolnic, D., Casertano, S., Filippenko, A. V., Tucker, B. E., Reid, M. J., Jones, D. O., and Silverman, J. M. (2016). A 2.4% Determination of the Local Value of the Hubble Constant. *ApJ*, 826(1):56.

Rorai, A., Becker, G. D., Haehnelt, M. G., Carswell, R. F., Bolton, J. S., Cristiani, S., D'Odorico, V., Cupani, G., Barai, P., and Calura, F. (2017). Exploring the thermal state of the low-density intergalactic medium at z = 3 with an ultrahigh signal-to-noise QSO spectrum. *MNRAS*, 466(3):2690–2709.

Ross, A. J., Beutler, F., Chuang, C.-H., Pellejero-Ibanez, M., Seo, H.-J., Vargas-Magaña, M., Cuesta, A. J., Percival, W. J., Burden, A., Sánchez, A. G., Grieb, J. N., Reid, B., Brownstein, J. R., Dawson, K. S., Eisenstein, D. J., Ho, S., Kitaura, F.-S., Nichol, R. C., Olmstead, M. D., Prada, F., Rodríguez-Torres, S. A., Saito, S., Salazar-Albornoz, S., Schneider, D. P., Thomas, D., Tinker, J., Tojeiro, R., Wang, Y., White, M., and Zhao, G.-b. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: observational systematics and baryon acoustic oscillations in the correlation function. *MNRAS*, 464:1168–1191.

Ross, A. J., Ho, S., Cuesta, A. J., Tojeiro, R., Percival, W. J., Wake, D., Masters, K. L., Nichol, R. C., Myers, A. D., de Simoni, F., Seo, H. J., Herná ndez-Monteagudo, C., Crittenden, R., Blanton, M., Brinkmann, J., da Costa, L. A. N., Guo, H., Kazin, E., Maia, M. A. G., Maraston, C., Padmanabhan, N., Prada, F., Ramos, B., Sanchez, A., Schlafly, E. F., Schlegel, D. J., Schneider, D. P., Skibba, R., Thomas, D., Weaver, B. A., White, M., and Zehavi, I. (2011). Ameliorating systematic uncertainties in the angular clustering of galaxies: a study using the SDSS-III. *MNRAS*, 417:1350–1373.

Rossi, G. (2014). Neutrino mass from the Lyman-Alpha forest. *arXiv e-prints*, page arXiv:1406.5411.

Rossi, G., Yèche, C., Palanque-Delabrouille, N., and Lesgourgues, J. (2015). Constraints on dark radiation from cosmological probes. *Phys. Rev. D*, 92(6):063505.

Rubin, V. C. and Ford, Jr., W. K. (1970). Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *Astrophysical Journal*, 159:379.

Rudie, G. C., Steidel, C. C., and Pettini, M. (2012). The Temperature-Density Relation in the Intergalactic Medium at Redshift langzrang = 2.4. *ApJ*, 757(2):L30.

Rusu, C. E., Wong, K. C., Bonvin, V., Sluse, D., Suyu, S. H., Fassnacht, C. D., Chan, J. H. H., Hilbert, S., Auger, M. W., and Sonnenfeld, A. (2019). H0LiCOW XII. Lens mass model of WFI2033-4723 and blind measurement of its time-delay distance and H_0. *arXiv e-prints*, page arXiv:1905.09338.

Ryden, B. (2003). *Introduction to cosmology.*

Sahlén, M. (2019). Cluster-void degeneracy breaking: Neutrino properties and dark energy. *Phys. Rev. D*, 99(6):063525.

Sahni, V. and Starobinsky, A. (2006). Reconstructing Dark Energy. *International Journal of Modern Physics D*, 15(12):2105–2132.

Saini, T. D., Raychaudhury, S., Sahni, V., and Starobinsky, A. A. (2000). Reconstructing the Cosmic Equation of State from Supernova Distances. *Phys. Rev. Lett.*, 85(6):1162–1165.

Schaye, J., Theuns, T., Rauch, M., Efstathiou, G., and Sargent, W. L. W. (2000). The thermal history of the intergalactic medium*. *MNRAS*, 318(3):817–826.

Scheuer, P. A. G. (1965). A Sensitive Test for the Presence of Atomic Hydrogen in Intergalactic Space. *Nature*, 207(5000):963.

Schlegel, D. J., Finkbeiner, D. P., and Davis, M. (1998). Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *ApJ*, 500:525–553.

Schmidt, F., Elsner, F., Jasche, J., Nguyen, N. M., and Lavaux, G. (2019). A rigorous EFT-based forward model for large-scale structure. *J. Cosmology Astropart. Phys.*, 2019(1):042.

Schuecker, P. and Ott, H.-A. (1991). Scales of structures and homogeneity in the universe. *ApJ*, 378:L1–L4.

Schuster, N., Hamaus, N., Pisani, A., Carbone, C., Kreisch, C. D., Pollina, G., and Weller, J. (2019). The bias of cosmic voids in the presence of massive neutrinos. *arXiv e-prints*, page arXiv:1905.00436.

Scoccimarro, R. (2000). The Bispectrum: From Theory to Observations. *ApJ*, 544:597–615.

Scranton, R., Johnston, D., Dodelson, S., Frieman, J. A., Connolly, A., Eisenstein, D. J., Gunn, J. E., Hui, L., Jain, B., Kent, S., Loveday, J., Narayanan, V., Nichol, R. C., O'Connell, L., Scoccimarro, R., Sheth, R. K., Stebbins, A., Strauss, M. A., Szalay, A. S., Szapudi, I., Tegmark, M., Vogeley, M., Zehavi, I., Annis, J., Bahcall, N. A., Brinkman, J., Csabai, I., Hindsley, R., Ivezic, Z., Kim, R. S. J., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T., Munn, J., Peoples, J., Pier, J., Richards, G. T., Rockosi, C., Schlegel, D., Schneider, D. P., Stoughton, C., Tucker, D. L., Yanny, B., and York, D. G. (2002). Analysis of Systematic Effects and Statistical Uncertainties in Angular Clustering of Galaxies from Early Sloan Digital Sky Survey Data. *ApJ*, 579:48–75.

Seljak, U. (2000). Analytic model for galaxy and dark matter clustering. *MNRAS*, 318:203–213.

Seljak, U., Slosar, A., and McDonald, P. (2006). Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints. *Journal of Cosmology and Astro-Particle Physics*, 2006(10):014.

Shafieloo, A., Alam, U., Sahni, V., and Starobinsky, A. A. (2006). Smoothing supernova data to reconstruct the expansion history of the Universe and its age. *MNRAS*, 366(3):1081–1095.

Shectman, S. A., Landy, S. D., Oemler, A., Tucker, D. L., Lin, H., Kirshner, R. P., and Schechter, P. L. (1996). The Las Campanas Redshift Survey. *ApJ*, 470:172.

Silverman, J., Hasinger, G., Brusa, M., Mainieri, V., Cappelluti, N., Finoguenov, A., Brunner, H., Comastri, A., Gilli, R., Vignali, C., Lilly, S., Kovac, K., Mainieri, V., Zamorani, G., Le Fevre, O., Contini, T., Bolzonella, M., and Scodeggio, M. (2008). A 3D map of the AGN distribution and the relation to the zCOSMOS density field. In *The X-ray Universe 2008*, page 154.

Simon, P., Taylor, A. N., and Hartlap, J. (2009). Unfolding the matter distribution using three-dimensional weak gravitational lensing. *MNRAS*, 399(1):48–68.

Skrutskie, M. F., Cutri, R. M., Stiening, R., Weinberg, M. D., Schneider, S., Carpenter, J. M., Beichman, C., Capps, R., Chester, T., Elias, J., Huchra, J., Liebert, J., Lonsdale, C., Monet, D. G., Price, S., Seitzer, P., Jarrett, T., Kirkpatrick, J. D., Gizis, J. E., Howard, E., Evans, T., Fowler, J., Fullmer, L., Hurt, R., Light, R., Kopan, E. L., Marsh, K. A., McCallon, H. L., Tam, R., Van Dyk, S., and Wheelock, S. (2006). The Two Micron All Sky Survey (2MASS). *Astrophysical Journal*, 131:1163–1183.

Slosar, A., Font-Ribera, A., Pieri, M. M., Rich, J., Le Goff, J.-M., Aubourg, É., Brinkmann, J., Busca, N., Carithers, B., and Charlassier, R. (2011). The Lyman-α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data. *Journal of Cosmology and Astro-Particle Physics*, 2011(9):001.

Smith, M. S., Kawano, L. H., and Malaney, R. A. (1993). Experimental, Computational, and Observational Analysis of Primordial Nucleosynthesis. *ApJS*, 85:219.

Smith, R. E., Peacock, J. A., Jenkins, A., White, S. D. M., Frenk, C. S., Pearce, F. R., Thomas, P. A., Efstathiou, G., and Couchman, H. M. P. (2003). Stable clustering, the halo model and non-linear cosmological power spectra. *MNRAS*, 341:1311–1332.

Sorini, D., Oñorbe, J., Lukić, Z., and Hennawi, J. F. (2016). Modeling the Lyα Forest in Collisionless Simulations. *ApJ*, 827:97.

Spolyar, D., Sahlén, M., and Silk, J. (2013). Topology and Dark Energy: Testing Gravity in Voids. *Phys. Rev. Lett.*, 111(24):241103.

Stark, C. W., Font-Ribera, A., White, M., and Lee, K.-G. (2015a). Finding high-redshift voids using Lyman α forest tomography. *MNRAS*, 453:4311–4323.

Stark, C. W., White, M., Lee, K.-G., and Hennawi, J. F. (2015b). Protocluster discovery in tomographic Ly α forest flux maps. *MNRAS*, 453:311–327.

Starobinsky, A. A. (1982). Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations. *Physics Letters B*, 117:175–178.

Strateva, I., Ivezić, Ž., Knapp, G. R., Narayanan, V. K., Strauss, M. A., Gunn, J. E., Lupton, R. H., Schlegel, D., Bahcall, N. A., Brinkmann, J., Brunner, R. J., Budavári, T., Csabai, I., Castander, F. J., Doi, M., Fukugita, M., Győry, Z., Hamabe, M., Hennessy, G., Ichikawa, T., Kunszt, P. Z., Lamb, D. Q., McKay, T. A., Okamura, S., Racusin, J., Sekiguchi, M., Schneider, D. P., Shimasaku, K., and York, D. (2001). Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. *AJ*, 122:1861–1874.

Strauss, M. A., Weinberg, D. H., Lupton, R. H., Narayanan, V. K., Annis, J., Bernardi, M., Blanton, M., Burles, S., Connolly, A. J., Dalcanton, J., Doi, M., Eisenstein, D., Frieman, J. A., Fukugita, M., Gunn, J. E., Ivezić, Ž., Kent, S., Kim, R. S. J., Knapp, G. R., Kron, R. G., Munn, J. A., Newberg, H. J., Nichol, R. C., Okamura, S., Quinn, T. R., Richmond, M. W., Schlegel, D. J., Shimasaku, K., SubbaRao, M., Szalay, A. S., Vanden Berk, D., Vogeley, M. S., Yanny, B., Yasuda, N., York, D. G., and Zehavi, I. (2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. *AJ*, 124:1810–1824.

Sudevan, V., Aluri, P. K., Yadav, S. K., Saha, R., and Souradeep, T. (2017). Improved Diffuse Foreground Subtraction with the ILC Method: CMB Map and Angular Power Spectrum Using Planck and WMAP Observations. *ApJ*, 842:62.

Sunyaev, R. A. and Zeldovich, I. B. (1980). The velocity of clusters of galaxies relative to the microwave background - The possibility of its measurement. *MNRAS*, 190:413–420.

Sunyaev, R. A. and Zeldovich, Y. B. (1972). The Observations of Relic Radiation as a Test of the Nature of X-Ray Radiation from the Clusters of Galaxies. *Comments on Astrophysics and Space Physics*, 4:173.

Szalay, A. S., Jain, B., Matsubara, T., Scranton, R., Vogeley, M. S., Connolly, A., Dodelson, S., Eisenstein, D., Frieman, J. A., and Gunn, J. E. (2003). Karhunen-Loève Estimation of the Power Spectrum Parameters from the Angular Distribution of Galaxies in Early Sloan Digital Sky Survey Data. *ApJ*, 591(1):1–11.

Tegmark, M., Blanton, M. R., Strauss, M. A., Hoyle, F., Schlegel, D., Scoccimarro, R., Vogeley, M. S., Weinberg, D. H., Zehavi, I., and Berlind, A. (2004). The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey. *ApJ*, 606(2):702–740.

Tegmark, M., Dodelson, S., Eisenstein, D. J., Narayanan, V., Scoccimarro, R., Scranton, R., Strauss, M. A., Connolly, A., Frieman, J. A., and Gunn, J. E. (2002). The Angular Power Spectrum of Galaxies from Early Sloan Digital Sky Survey Data. *ApJ*, 571(1):191–205.

Tegmark, M. and Efstathiou, G. (1996). A method for subtracting foregrounds from multifrequency CMB sky maps**. *MNRAS*, 281:1297–1314.

Tegmark, M., Hamilton, A. J. S., Strauss, M. A., Vogeley, M. S., and Szalay, A. S. (1998). Measuring the Galaxy Power Spectrum with Future Redshift Surveys. *ApJ*, 499:555–576.

Tempel, E., Einasto, J., Einasto, M., Saar, E., and Tago, E. (2009). Anatomy of luminosity functions: the 2dFGRS example. *A&A*, 495:37–51.

Theuns, T. and Zaroubi, S. (2000). A wavelet analysis of the spectra of quasi-stellar objects. *MNRAS*, 317(4):989–995.

Tully, R. B. (2007). Our CMB Motion: The Role of the Local Void. In Metcalfe, N. and Shanks, T., editors, *Cosmic Frontiers*, volume 379 of *Astronomical Society of the Pacific Conference Series*, page 24.

Tully, R. B. and Fisher, J. R. (1977). A new method of determining distances to galaxies. *A&A*, 54:661–673.

Vansyngel, F., Wandelt, B. D., Cardoso, J.-F., and Benabed, K. (2016). Semi-blind Bayesian inference of CMB map and power spectrum. *A&A*, 588:A113.

Vettolani, G., Zucca, E., Zamorani, G., Cappi, A., Merighi, R., Mignoli, M., Stirpe, G. M., MacGillivray, H., Collins, C., and Balkowski, C. (1997). The ESO Slice Project (ESP) galaxy redshift survey. I. Description and first results. *A&A*, 325:954–960.

Viel, M., Becker, G. D., Bolton, J. S., and Haehnelt, M. G. (2013). Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data. *Phys. Rev. D*, 88(4):043502.

Viel, M., Bolton, J. S., and Haehnelt, M. G. (2009). Cosmological and astrophysical constraints from the Lyman α forest flux probability distribution function. *MNRAS*, 399(1):L39–L43.

Viel, M., Haehnelt, M. G., and Lewis, A. (2006). The Lyman α forest and WMAP year three. *MNRAS*, 370(1):L51–L55.

Vishniac, E. T. (1987). Reionization and Small-Scale Fluctuations in the Microwave Background. *ApJ*, 322:597.

Weller, J. and Albrecht, A. (2001). Opportunities for Future Supernova Studies of Cosmic Acceleration. *Phys. Rev. Lett.*, 86(10):1939–1942.

Weller, J. and Albrecht, A. (2002). Future supernovae observations as a probe of dark energy. *Phys. Rev. D*, 65(10):103512.

Weymann, R. J., Jannuzi, B. T., Lu, L., Bahcall, J. N., Bergeron, J., Boksenberg, A., Hartig, G. F., Kirhakos, S., Sargent, W. L. W., and Savage, B. D. (1998). The Hubble Space Telescope Quasar Absorption Line Key Project. XIV. The Evolution of Lyα Absorption Lines in the Redshift Interval z = 0-1.5. *ApJ*, 506(1):1–18.

Williger, G. M., Carswell, R. F., Weymann, R. J., Jenkins, E. B., Sembach, K. R., Tripp, T. M., Davé, R., Haberzettl, L., and Heap, S. R. (2010). The low-redshift Lyα forest towards 3C 273. *MNRAS*, 405(3):1736–1758.

Wright, E. L., Mather, J. C., Fixsen, D. J., Kogut, A., Shafer, R. A., Bennett, C. L., Boggess, N. W., Cheng, E. S., Silverberg, R. F., Smoot, G. F., and Weiss, R. (1994). Interpretation of the COBE FIRAS CMBR spectrum. *ApJ*, 420:450–456.

Yèche, C., Palanque-Delabrouille, N., Baur, J., and du Mas des Bourboux, H. (2017). Constraints on neutrino masses from Lyman-alpha forest power spectrum with BOSS and XQ-100. *Journal of Cosmology and Astro-Particle Physics*, 2017(6):047.

Zaroubi, S., Hoffman, Y., and Dekel, A. (1999). Wiener Reconstruction of Large-Scale Structure from Peculiar Velocities. *ApJ*, 520:413–425.

Zeldovich, I. B., Einasto, J., and Shandarin, S. F. (1982). Giant voids in the universe. *Nature*, 300:407–413.

Zel'dovich, Y. B. (1970). Gravitational instability: An approximate theory for large density perturbations. *A&A*, 5:84–89.

Zwicky, F. (1933). Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127.

Acknowledgement

First, I would like to thank Torsten Enßlin for giving me the opportunity to study cosmology at the Max-Planck Institute for Astrophysics (MPA). He gave me the freedom and opportunity to develop my projects.

I would like to express my sincere gratitude to Jens Jasche for his invaluable support, guidance and encouragement. Jens has been a very friendly and supportive advisor, always ready to discuss the project or provide career advice. I deeply appreciate the energy and dedication that he invested in this work despite the distance. His engagement and motivation made working with him both delightful and inspiring. I enjoyed every single of our scientific discussions and valued his advice and feedback on my talks and scientific writing. He was always ready to listen to me and showed great interest in my development as a scientist. I would like to thank him for sharing his knowledge and optimism, always pushing me forward and encouraging me.

I am also very grateful to Guilhem Lavaux for his support in the implementation of the algorithms and for sharing his knowledge of programming. Guilhem always found the time to look at my code and to give advice on how to improve it. I am especially thankful for his help to improve the numerical scalability of the Ly-α forest algorithm.

I want to thank all the members of the Aquila Consortium for creating a very productive environment during our group meetings and for being always open to discuss each other's projects.

I would like to thank Doogesh Kodi Ramanah for the fruitful team dynamics during our collaboration for the robust likelihood.

Also, I am thankful to the members of the Information Field Theory group for creating a pleasant group atmosphere at the MPA and for the interesting and entertaining conversations at lunch.

Finally, I would like to thank my partner Andreas, my close friend Joel, and my parents Ismael and Olga for their patience, encouragement and support during this thesis.