Theoretical and Computational Tools for Analyzing the Large-Scale Structure of the Universe

Hand, Nicholas

Theoretical and Computational Tools for Analyzing the Large-Scale Structure of the Universe

By

Nicholas A Hand

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Astrophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Uroš Seljak, Chair
Professor Martin White
Professor Adrian Lee
Dr. Pat McDonald

Fall 2017

Theoretical and Computational Tools for Analyzing the Large-Scale Structure of the Universe

Abstract

Theoretical and Computational Tools for Analyzing the Large-Scale Structure of the Universe

by

Nicholas A Hand

Doctor of Philosophy in Astrophysics

University of California, Berkeley

Professor Uroš Seljak, Chair

The analysis of the large-scale structure (LSS) of the Universe can yield insights into some of the most important questions in contemporary cosmology, and in recent years, has become a data-driven endeavor. With ever-growing data sets, optimal analysis techniques have become essential, not only to extract statistics from data, but also to effectively use computing resources to produce accurate theoretical predictions for those statistics. Future LSS experiments will help answer fundamental questions about our Universe, including the physical nature of dark energy, the mass scale of neutrinos, and the physics of inflation. To do so, improvements must be made to theoretical models as well as the computational tools used to perform such analyses.

This thesis examines multiple aspects of LSS data analysis, presenting novel modeling techniques as well as a software toolkit suitable for analyzing data from the next generation of LSS surveys. First, we present `nbodykit`, an open-source, massively parallel Python toolkit for analyzing LSS data. `nbodykit` is both an interactive and scalable piece of scientific software, providing parallel implementations of many commonly used algorithms in LSS. Its modular design allows researchers to integrate `nbodykit` with their own software to build complex applications to solve specific problems in LSS. Next, we derive an optimal means of using fast Fourier transforms to estimate the multipoles of the line-of-sight dependent power spectrum, eliminating redundancy present in previous estimators in the literature. We also discuss potential advantages of our estimator for future data sets. We then present a novel theoretical model for the redshift-space galaxy power spectrum and demonstrate its accuracy in describing the clustering of galaxies down to scales of $k = 0.4\ h\mathrm{Mpc}^{-1}$. Finally, we analyze the large-scale clustering of quasars from the extended Baryon Oscillation Spectroscopic Survey to constrain the deviation from Gaussian random field initial conditions in the early Universe, known as primordial non-Gaussianity.

Contents

List of Figures **iv**

List of Tables **vii**

Acknowledgments **viii**

1 Introduction **1**
1.1 The standard cosmological model 2
1.2 Large-scale structure 4
1.2.1 Cosmological constraints from LSS data 4
1.2.2 A brief history 7
1.2.3 Present-day and future redshift surveys 8
1.3 Theory 10
1.3.1 The smooth background 10
1.3.2 Density perturbations 13
1.3.3 Full-shape clustering models 16

2 nbodykit: an open-source, massively parallel toolkit for large-scale structure **20**
2.1 Introduction 20
2.2 Overview 23
2.2.1 Initializing `nbodykit` 23
2.2.2 The `nbodykit` Ecosystem 23
2.2.3 A Component-Based Approach 24
2.2.4 Parallelism 26
2.3 Capabilities 27
2.3.1 Cosmology 27
2.3.2 Catalogs 28
2.3.3 Meshes 30
2.3.4 Algorithms 33
2.4 Development Workflow 37
2.4.1 Version Control 37

2.4.2 Automated Testing with MPI Support 38
2.4.3 Use on Personal and HPC Machines 38
2.4.4 Documentation 39
2.5 In Action 41
2.6 Performance Benchmarks 41
2.7 Conclusions 45

3 An optimal FFT-based anisotropic power spectrum estimator 47
3.1 Introduction 47
3.2 Estimators 50
3.2.1 Multipoles 50
3.2.2 Wedges 52
3.2.3 Implementation 53
3.3 Isolating transverse $\mu = 0$ systematics 54
3.3.1 A toy model for fiber assignment 54
3.3.2 Verification with simulations 56
3.3.3 A toy model for window function effects 58
3.4 Statistical properties 61
3.4.1 Covariance 61
3.4.2 Fisher information 66
3.5 Conclusions 67

4 Extending the modeling of the anisotropic galaxy power spectrum to $k = 0.4\ h\mathrm{Mpc}^{-1}$ 70
4.1 Introduction 71
4.2 Simulations 73
4.2.1 RunPB 73
4.2.2 N-series 74
4.2.3 Lettered Challenge Boxes 75
4.3 Analysis methods 75
4.4 The power spectrum model 77
4.4.1 Halo model formalism for galaxies 77
4.4.2 Halo clustering in redshift space 81
4.4.3 Modeling observational effects 90
4.4.4 Model parametrization 93
4.5 Performance of the model 96
4.5.1 RunPB results 96
4.5.2 Independent tests on high resolution mocks 101
4.5.3 Tests on realistic DR12 BOSS CMASS mocks 103
4.5.4 Comparison to published models 106
4.6 Discussion 110
4.7 Conclusion 113

5 Constraints on primordial non-Gaussianity from the clustering of eBOSS quasars 118
5.1 Introduction . . . 118
5.2 Data . . . 120
5.2.1 eBOSS DR14Q sample . . . 120
5.2.2 Completeness weights . . . 121
5.2.3 Synthetic DR14Q catalogs . . . 123
5.3 Theory . . . 125
5.3.1 Local primordial non-Gaussianity . . . 125
5.3.2 Optimal estimators in LSS . . . 126
5.3.3 An optimal estimator for f_{NL} . . . 127
5.4 Analysis Methods . . . 129
5.4.1 Fiducial cosmology . . . 129
5.4.2 Power spectrum estimation . . . 130
5.4.3 Modeling . . . 132
5.4.4 Parameter estimation . . . 134
5.5 Results . . . 136
5.5.1 Quasar bias as a function of redshift . . . 136
5.5.2 Unweighted results . . . 136
5.6 Conclusions and future work . . . 138

A Velocity correlators in the Zel'dovich approximation 143
A.1 P_{01} and P_{11} using LPT . . . 143
A.2 A generalized velocity generating function . . . 144
A.3 The Zel'dovich approximation . . . 146

B Improved HZPT modeling 149
B.1 Dark matter correlators P_{00} and P_{01} . . . 149
B.2 Dark matter radial momentum power spectrum, P_{11} . . . 150
B.3 Halo-matter power spectrum, P^{hm} . . . 151

C Relation between model parameters in the halo model 153
C.1 The fraction of centrals with satellites . . . 153
C.2 The linear bias of centrals with satellites . . . 154
C.3 1-halo term amplitudes . . . 155

Bibliography 157

List of Figures

1.1 A summary of present-day BAO and growth of structure constraints from LSS surveys. 5
1.2 Early galaxy clustering measurements. 8
1.3 The 2D galaxy correlation function and power spectrum from the completed BOSS data set. 9

2.1 The components and interfaces of nbodykit. 25
2.2 A comparison of the effects of interlacing when using the CIC, TSC, and PCS windows. 32
2.3 The performance of the Daubechies and Symlet wavelets in comparison to the CIC, TSC, and PCS windows. 34
2.4 An illustrative example of using Mesh objects in nbodykit. 35
2.5 The source code and components of a galaxy clustering emulator application using nbodykit. 40
2.6 Performance benchmarks for four nbodykit algorithms for a “small” data set (10^6 objects) and a “large” data set (10^7 objects). 42
2.7 The wall-clock time as a function of the number of MPI ranks used for each step in a galaxy clustering emulator application. 43

3.1 The leakage of a transverse $\mu = 0$ systematic into $\mu > 0$ power as a function of the maximum multipole used to reconstruct the observed power $P(k, \mu)$. 56
3.2 Limiting the leakage of a $\mu = 0$ with non-uniform μ bins. 57
3.3 Tests on simulations of the removal of a $\mu = 0$ systematic using non-uniform bins 57
3.4 Removal of a $\mu = 0$ systematic from a density field with a toy model selection function imposed. 60
3.5 The correlation matrix between μ wedges from cubic simulations when using non-uniform and uniform binning schemes. 63
3.6 The correlation matrix between μ wedges from realistic cutsky mock catalogs when using non-uniform and uniform binning schemes. 64
3.7 The condition number of the wedge covariance matrix, as computed analytically and from simulations. 65
3.8 The RSD Fisher information for our wedge estimator using non-uniform μ bins. 67

3.9 The fractional change in the uncertainties in $b_1\sigma_8$ and $f\sigma_8$ when using non-uniform μ wedges and excluding the first bin. 68

4.1 The dependence of the second-order nonlinear effective biases, b_2^{00} and b_2^{01}, on the linear bias b_1 that is used in this chapter. 83
4.2 The accuracy of the dark matter HZPT modeling results used in this work. . . . 84
4.3 The accuracy of the HZPT model for the auto power spectrum of the dark matter radial momentum in comparison to the results from the RunPB simulation. . . . 86
4.4 The deviation of the halo stochasticity $\Lambda(k)$, as defined in equation 4.27, from the Poisson shot noise. 88
4.5 The accuracy of the HZPT model used in this work for the halo-matter cross-correlation, in comparison to the results from the RunPB simulation. 89
4.6 The window function multipoles in configuration space and the effects of the window function on linear Kaiser power spectrum multipoles for the DR12 CMASS NGC survey geometry. 92
4.7 he best-fit monopole, quadrupole, and hexadecapole models as compared to the measurements from the mean of 10 RunPB HOD galaxy realizations at $z = 0.55$. 96
4.8 The 2D posterior distributions for the cosmology parameters $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ obtained from fitting the mean of 10 RunPB HOD galaxy realizations at $z = 0.55$. 98
4.9 Parameter correlations as measured from the posterior distributions when fitting the mean of the 10 RunPB galaxy catalogs. 100
4.10 The mean of the best-fit monopole, quadrupole, and hexadecapole models as compared to the individual measurements from the 21 N-series cubic boxes at $z = 0.5$. 101
4.11 The fractional deviation of the best-fitting $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ values from their true values for each of the seven lettered challenge boxes. 105
4.12 The best-fitting $\ell = 0$, 2, 4, and 6 theory and measurements from a single catalog of the N-series cutsky test suite, which accurately simulates the BOSS DR12 CMASS data set. 107
4.13 The best-fitting $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ parameters from fitting our RSD model to the measured monopole, quadrupole, and hexadecapole from the 84 N-series cutsky mocks. 108
4.14 The best-fitting model and measured simulation points for the monopole, quadrupole, and hexadecapole from the mean of 10 RunPB galaxy catalogs at $z = 0.55$ for all galaxies, all centrals, and isolated centrals with no satellites in the same halo. . 114

5.1 The mean density of quasars in the DR14Q sample as a function of redshift for the NGC and SGC regions of the sky. 121
5.2 The measured power spectrum multipoles for the DR14Q sample, as compared to those measured from the mean of the 1,000 EZ mock catalogs. 124

5.3 The window function multipoles in configuration space and the effects of the window function on linear Kaiser power spectrum multipoles for the eBOSS DR14Q NGC survey geometry. 133
5.4 The distribution of the best-fit model parameters for the 1,000 EZ mock catalogs. 140
5.5 The measured power spectrum multipoles for the DR14Q sample and best-fit theoretical model. 141
5.6 The posterior distributions for each of the four model parameters when fitting the DR14Q multipoles. 142

List of Tables

1.1 The *Planck* 2015 cosmological parameters. 3

4.1 The cosmological and simulation parameters for the various N-body simulations used in this chapter. 73
4.2 The halo mass bins used when comparing results from the RunPB simulations to theoretical modeling of halo clustering. 74
4.3 The parameter space of our full-shape RSD fits using the model described in this chapter. 94
4.4 Parameter constraints obtained when fitting the 13-parameter RSD model to the monopole, quadrupole, and hexadecapole measured from the mean of the 10 RunPB galaxy catalogs at $z = 0.55$. 99
4.5 The mean (with expected value subtracted) and standard deviation S of the best-fitting values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ found when fitting the monopole, quadrupole, and hexadecapole from the 21 cubic N-series catalogs. 102
4.6 The best-fitting values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ obtained when fitting our RSD model to the measured monopole, quadrupole, and hexadecapole from the 7 lettered challenge boxes. 104
4.7 The mean and standard deviation of the best-fitting values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ from fits to the 84 N-series cutsky catalogs. 109
4.8 The estimated best-fitting cosmology parameters for the 84 N-cutsky mocks using the model in this chapter as well as the Fourier space models described in Beutler et al. (2017b) and Grieb et al. (2017). 109
4.9 The best-fitting cosmological parameters found when fitting the multipoles of the mean of 10 RunPB galaxy catalogs at $z = 0.55$ when including all galaxies, centrals only, and type A centrals only. 115

5.1 The best-fit linear bias parameters and 1σ errors obtained from fitting the power spectrum model of Section 5.4.3 to the monopole and quadrupole of the NGC and SGC DR14Q samples. 135
5.2 The f_{NL} constraints as measured from the 1,000 EZ mock catalogs and the DR14Q data set for the NGC and SGC sky regions. 137

Acknowledgments

I would like to begin by thanking my research advisor, Uroš Seljak, for his support and direction throughout my PhD studies. He has taught me many things, perhaps most importantly how to approach and think critically about complicated problems. His expertise and intuition has been invaluable to my growth as a scientific researcher. I am also grateful for the freedom he afforded me to tackle and explore my research issues, and crucially, for sharing his culinary masterpieces with me and keeping me as his student despite my utter lack of skiing skills. I would also like to thank the members of my committee for their guidance and hard questions: Martin White, Pat McDonald, and Adrian Lee.

I have been lucky to work with a large number of fantastic collaborators throughout my time at Berkeley. I am indebted to the entire BCCP group for their support and advice. Working as part of such an active research group has been an immense resource, and there are too many people to thank for their support throughout my six years in Berkeley. I have learned so much from Yu Feng, including how to write a proper software package, and he has instilled in me the maxims of a proper software developer. I am grateful to Florian Beutler, Zvonimir Vlah, Yin Li, Emanuele Castorina, Zachary Slepian, and Teppei Okumura for the time, advice, and knowledge they so generously offered me throughout my studies. The BOSS collaboration has been an invaluable resource for me, and in particular, I am grateful to Jan Grieb for his help with my own research. I was also fortunate to work with an incredible group of collaborators before narrowing my PhD focus. Sudeep Das was an incredible mentor early in my studies and his guidance was a large factor in my current career path. Also, thanks to Alexie Leauthaud for sharing her immense expertise with me. I would not have pursued a PhD if not for my amazing undergraduate advisors, David Spergel and Toby Marriage, and the entire ACT collaboration.

There are many people that made Berkeley home for me. I am thankful to my housemates at Del Mar, particularly Kevin Moch, Matt George, Dyas Utomo, Ori Fox, and Emanuele Castorina. I would also like to thank the BADGrads for their support throughout my years. The fellow students in my year are simply the best: Sedona Price, Isaac Shivvers, Dyas Utomo, and Lauren Weiss. The Berkeley Astronomy department has an amazing group of administrative staff who have offered me so much help throughout the years. In particular, Dexter Stewart is absolutely wonderful. I also owe a large debt of gratitude to the many communities supporting the open-source software and tools that were invaluable to my dissertation: `NumPy`, `SciPy`, `pandas`, `IPython`, `Jupyter`, GitHub, Read the Docs, Travis, and Coveralls, to name a few.

Finally, I would not be here today without the constant love and support of my family. Words cannot express how grateful I am. I truly could not have done this without you all, so, simply, thank you. I am particularly grateful for the many care packages and baked goods that brought a bit of Philadelphia to California. I will end with the most important person throughout my time in Berkeley: my wife Caroline. Thank you for always believing in me. I promise to stop talking about redshift-space distortions and Fingers-of-God so much.

Chapter 1

Introduction

The past several decades have brought immense change to the field of cosmology, and in particular, the study of the large-scale structure (LSS) of the Universe. From the earliest realizations that galaxies could be used to trace the matter density field,[1] the size of data sets mapping the three-dimensional Universe has grown from the thousands to the millions, with a factor of ten or more increase expected in the future. Driven by technological advances, the rise of automatic, wide-field galaxy surveys has transformed LSS research into a data-driven science. The aim of the work in this thesis is to improve the tools we use to analyze such data sets and infer properties of the Universe. We focus not only on theoretical techniques, but also numerical and statistical methods, with special consideration of the requirements demanded by future LSS data sets.

To start, we provide a short introduction of the standard cosmological model and how it developed into the consensus paradigm, with a myriad of observations providing evidence. We also provide some context regarding the history of LSS data analysis and outline some of the important theoretical concepts upon which this thesis relies. Individual chapters provide more in-depth formalism when necessary. The primary four chapters of this thesis present related work from several published papers. Chapter 2 presents a massively parallel software toolkit that implements some of the most commonly used algorithms in LSS data analysis. Chapter 3 derives an optimal estimator for the multipoles of the anisotropic power spectrum of a density field and discusses some potential advantages of such an estimator for future data sets. Chapter 4 develops a new theoretical model for the anisotropic galaxy power spectrum and performs stringent tests of the level of theoretical systematics using simulations. Finally, Chapter 5 combines many of the analysis tools described in this thesis to constrain the level of primordial non-Gaussianity in the early Universe using the power spectrum of quasars, as observed by the extended Baryon Oscillation Spectroscopic Survey.

The background material presented in this introduction is collected from a number of sources, including Dodelson (2003), Peacock (1999), Peebles (1980), and Coil (2013). We provide more specific references throughout this chapter when needed.

[1] see the series of papers starting in the 1970s with, e.g., Peebles (1973), Tonry & Davis (1979)

1.1 The standard cosmological model

A variety of observations over the past several decades, including measurements of the cosmic microwave background (CMB), Type Ia supernovae (SNe), and large-scale structure, have provided evidence for a consensus theory for the evolution of the Universe. This theory predicts that approximately 13.8 billion years ago (*Planck* 2015; Planck Collaboration et al. 2016a) the Universe existed in a very hot and dense early state. From this so-called "Big Bang" state, the Universe continued to expand, with the late-time expansion accelerating due to an energy component, referred to as dark energy, whose physical origin is not yet understood. Dark energy comprises roughly 70% of the Universe's present energy budget; the remaining amount includes a small relativistic radiation component (photons and neutrinos) and a larger non-relativistic matter component. The latter is dominated by an effectively collisionless component known as cold dark matter (CDM), comprising roughly 25%. The more familiar baryonic matter, which interacts electromagnetically, makes up the remaining portion (roughly 5%). These various components evolve differently with time (see §1.3.1), and as the Universe has expanded and become less dense, the principal energy component has changed. Thus, the Universe, which began dominated by radiation, has had successive periods of matter and dark energy domination.

The prevailing theory for the initial conditions of the Universe is known as inflation (Guth 1981; Linde 1982; Albrecht & Steinhardt 1982). During this epoch, the Universe expanded exponentially, growing its scale by at least a factor of $\sim 10^{27}$. This rapid expansion stretched quantum fluctuations to a macroscopic level, producing the perturbations that seeded the structure that we observe today. An important consequence of inflation is that it predicts a near homogeneous, isotropic, and spatially flat Universe. These properties have all been observationally confirmed with a high degree of precision, but as of yet, there is no direct, conclusive observational evidence for the inflationary model (see §1.3.2).

Following inflation, the Universe cooled, with the competition between the expansion rate and interaction rates between particles defining the thermal evolution. Eventually, the process of Big Bang nucleosynthesis resulted in the creation of charged atomic nuclei, with roughly 75% of the baryonic mass in the form of hydrogen ions (^{1}H), 25% in the form of helium (^{4}He), and small amounts of deuterium and lithium. Along with these nuclei, electrons, photons, and neutrinos were also present. The charged baryonic matter (protons + electrons) was tightly coupled by their electromagnetic attraction. The free electrons were also tightly coupled to the photons via Thomson scattering. The result was the so-called *photon-baryon fluid*. The pressure of this fluid prevented fluctuations from gravitationally collapsing, instead leading to the production of acoustic waves. These sound waves, known as baryon acoustic oscillations (BAO; see, e.g., Bassett & Hlozek (2010) for a review), propagated until the photons and baryons decoupled. This occurred around the time of "recombination", when the Universe had cooled enough for neutral hydrogen to form, roughly 400,000 years after the Big Bang. Since this time, photons have free streamed from the so-called "surface of last scattering" through the Universe nearly unhindered. It is these photons that we observe today as part of the nearly isotropic 2.7 K cosmic microwave

parameter		value
Present-day baryon density	$\Omega_{\rm b}h^2$	0.02230 ± 0.00014
Present-day cold dark matter density	$\Omega_{\rm c}h^2$	0.1188 ± 0.0010
Angular size of the sound horizon [radians]	$\theta_\star$	104.093 ± 0.030
Optical depth due to reionization	τ	0.066 ± 0.012
Scalar power spectrum index	$n_{\rm s}$	0.9667 ± 0.0040
Amplitude of primordial curvature perturbations	$\ln 10^{10} A_{\rm s}$	3.064 ± 0.023

Table 1.1: The cosmological parameters in the standard, spatially flat ΛCDM model, as measured by Planck Collaboration et al. (2016a) (from Table 4, column 6).

background radiation.

Around 50,000 years after the Big Bang, the Universe's energy density became dominated by matter, and the initial CDM (and after recombination, baryonic) density fluctuations, began to experience substantial growth due to gravitational instability. The fluctuations collapsed under gravity to form the cosmic web: intersecting, nonlinear structures of halos, voids, and filaments that are dominated by CDM. Within dark matter halos, baryonic material cools and collapses to form stars and galaxies. Galaxies can be found in the centers of halos, or in the most massive halos, as satellite galaxies in orbit around the center. The process of structure growth is hierarchical, with smaller objects forming first and merging together to form larger structures. On large scales ($\sim$150 Mpc), the evolution of structure growth is well described by a linear description of the perturbations, but on smaller scales, the nonlinear nature of the fluctuations presents significant theoretical challenges.

The concordance model described here came into focus following one of the most significant physics breakthroughs in recent memory: the discovery of the accelerating expansion rate of the Universe (Riess et al. 1998; Perlmutter et al. 1999). This provided the first direct evidence for the presence of dark energy. Despite its physical origins remaining a mystery, dark energy can be described by a cosmological constant Λ, whose energy density is assumed to be constant in time. Hence, the paradigm described in this section has come to be known as the ΛCDM model. In its standard form, the model assumes the Universe to be spatially flat, with its evolution described by six parameters. The current best constraints on these six parameters are given in Table 1.1, as measured by Planck Collaboration et al. (2016a).

The ΛCDM model is well supported by a number of different observational probes.[2] Most importantly, observations of the anisotropies of the CMB by the *WMAP* (Bennett et al. 2013) and *Planck* (Planck Collaboration et al. 2016a) satellites and the ground-based Atacama Cosmology Telescope (Das et al. 2014) and South Pole Telescope (George et al. 2015) have provided strong constraints on the matter and radiation densities, the angular diameter distance to the surface of last scattering, and the shape and amplitude of the primordial matter power spectrum. Supplementing these constraints, galaxy redshift surveys

[2] For a comprehensive review of the observational probes of dark energy and cosmic acceleration within the context of the ΛCDM model, see Weinberg et al. (2013).

have successively measured the BAO feature on scales of $\sim$150 Mpc, which is imprinted on the clustering of density field tracers by sound waves in the baryon-photon fluid in the early Universe. These measurements have yielded constraints on the expansion rate across a wide redshift range, spanning $z \approx 0.1 - 0.8$ and $z \approx 2.5$ (Sánchez et al. 2006; Beutler et al. 2011; Kazin et al. 2014; Oka et al. 2014; Ross et al. 2015; Alam et al. 2017). Studies of Type Ia SNe not only provided the first evidence for dark energy in Riess et al. (1998); Perlmutter et al. (1999), but have also yielded complementary constraints on the expansion rate and present-day Hubble parameter H_0 (Freedman et al. 2012; Riess et al. 2016; Suzuki et al. 2012; Betoule et al. 2014). The growth of structure can be further constrained by weak gravitational lensing measurements, e.g., Heymans et al. (2013); Troxel et al. (2017), and redshift-space distortion analyses of anisotropic galaxy clustering measurements, e.g., Beutler et al. (2017b); Grieb et al. (2017). When considered together, this diverse set of observations provides compelling evidence for the standard, spatially flat six-parameter ΛCDM model.

1.2 Large-scale structure

1.2.1 Cosmological constraints from LSS data

Measurements of the clustering of luminous objects that trace the underlying matter density field contain a wealth of information about our Universe. Here, we briefly outline some of the analysis methods that have been used successfully to constrain the ΛCDM model in recent years.

Linear redshift-space distortion measurements

Peculiar velocities distort the measured clustering signal in redshift space, boosting the clustering in the direction parallel to the line-of-sight (LOS). The anisotropy is known as redshift space distortions (RSD). On large, linear scales, this distortion can be parametrized by the RSD parameter $\beta = f/b$, where f is the logarithmic growth rate and b is bias of the tracer relative to the matter field (see §1.3.3). Early measurements of β using LOS-dependent clustering include Hamilton (1993); Cole et al. (1995); Tadros et al. (1999); Peacock et al. (2001); Hawkins et al. (2003); These studies provided additional evidence for the ΛCDM model in support of CMB and SNe measurements. Constraints on β can also be used to test general relativity and models of modified gravity, which predict measurable differences in the growth rate f (e.g., Guzzo et al. (2008)).

BAO measurements

Acoustic waves in the baryon-photon fluid in the early Universe imprint a distinctive BAO feature on the clustering of luminous objects in the late-time Universe. Manifesting as a peak in the correlation function and oscillations in the power spectrum, the BAO feature can be used as a “standard ruler” to constrain the expansion rate and properties of dark energy

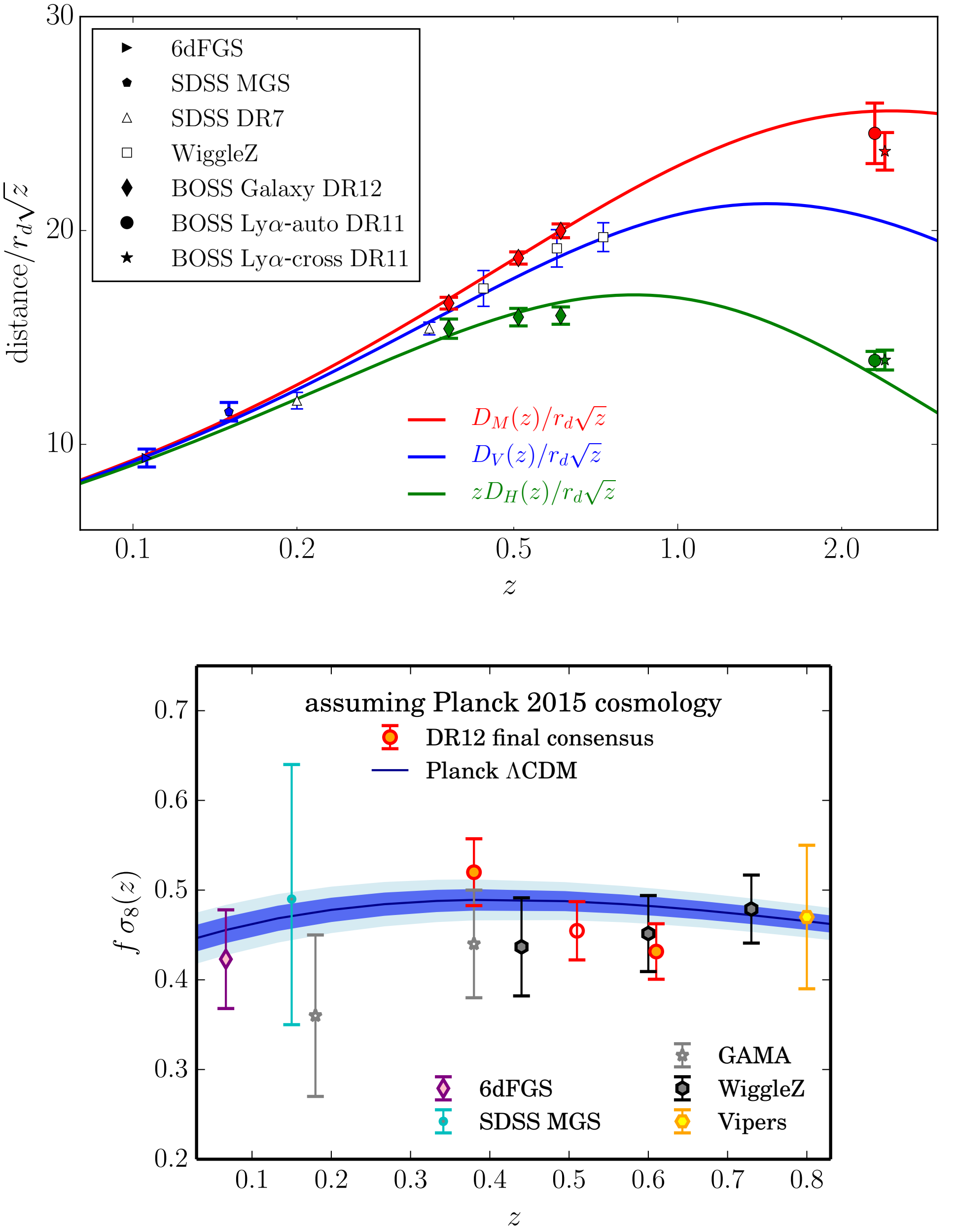


Figure 1.1: A summary of present-day BAO and growth of structure constraints from LSS surveys. *Top*: Distance constraints as function of redshift for a variety of published BAO studies, with *Planck* 2015 predictions shown as the solid lines. *Bottom*: Constraints on the growth rate of structure, parametrized by $f(z)\sigma_8(z)$, from several LSS surveys, with the *Planck* 2015 + general relativity prediction shown as the solid line. *Figure credit*: Alam et al. (2017).

(Eisenstein et al. 1998; Seo & Eisenstein 2003; Blake & Glazebrook 2003). LOS-dependent measurements of the BAO offer the possibility to improve constraints further through the geometric distortion generated when assuming an inaccurate fiducial cosmology, known as the Alcock-Paczynski (AP) effect. The isotropic BAO feature was first detected in Eisenstein et al. (2005) and Cole et al. (2005), and since, both isotropic and anisotropic measurements have been used to constrain the expansion rate over the redshift range $z \approx 0.1 - 0.8$ and $z \approx 2.5$. The results from the completed Baryon Oscillation Spectroscopic Survey (BOSS) (Alam et al. 2017) represent the tightest constraints via BAO to date, with percent-level constraints on the distance scale to $z \sim 0.6$. A summary of the current BAO distance measurements as a function of redshift is presented in the top panel of Figure 1.1, which shows excellent agreement between the prediction of the *Planck* 2015 parameters and measured results.

Full-shape clustering measurements

Modeling the broadband clustering signal adds further constraining power, albeit at the cost of additional theoretical complexity. The previously described methods generally marginalize over the broadband shape to avoid such complexities. Difficulties include the modeling of nonlinear velocity effects, the nonlinear evolution of the dark matter density, and the scale-dependent biasing between the observed tracers and the matter field (Scoccimarro 2004; Okumura & Jing 2011; Jennings 2012; Kwan et al. 2012). However, full-shape analyses offer the chance to measure the growth rate of structure and further constrain the expansion rate through the AP effect (Shoji et al. 2009). Early examples of full-shape analyses include Percival et al. (2001); Tegmark et al. (2004), with more recent results from BOSS (e.g., Beutler et al. 2017b; Grieb et al. 2017), WiggleZ (Blake et al. 2012; Contreras et al. 2013), and VIPERS (de la Torre et al. 2013). Growth of structure measurements are not currently as sensitive as BAO constraints, with the tightest constraint of order ∼5%. We show a summary of the present-day measurements of the growth rate as a function of redshift in the bottom panel of Figure 1.1. The results agree remarkably well with the ΛCDM prediction using the *Planck* 2015 parameters.

Beyond the growth rate and expansion history, broadband clustering also encodes information about other important physics. Neutrinos with non-zero mass induce a scale-dependent suppression of the clustering amplitude, as massive neutrinos affect the expansion rate, but free stream out of matter perturbations while still relativistic. While neutrino oscillation experiments are sensitive to the mass differences between states, full-shape clustering measurements constrain the sum of the neutrino masses, e.g., Lesgourgues & Pastor (2006); Beutler et al. (2014a). In addition, the broadband clustering on large scales contains signatures of primordial non-Gaussianity, the deviation from Gaussian random field initial conditions in the early Universe. Primordial non-Gaussianity of the local type can be detected through the scale-dependent bias it introduces on large scales (Dalal et al. 2008; Slosar et al. 2008; Desjacques & Seljak 2010).

Several chapters in this thesis focus on the broadband clustering signal. Chapter 3 presents an optimal estimator for measuring the anisotropic power spectrum from survey data using fast Fourier transforms. In Chapter 4, we describe a new model for the galaxy power spectrum in redshift space and test its constraining power when modeling the broadband clustering signal using simulations. Finally, in Chapter 5, we constrain local primordial non-Gaussianity by modeling the large-scale power spectrum of quasars from the extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016).

1.2.2 A brief history

The roots of the study of large-scale structure today lie in the realization that observations of "spiral nebulae" in the early 20th century were actually observations of extragalactic galaxies which were not distributed uniformly in space (Hubble 1926, 1934). Following this groundbreaking work by Hubble, technological progress would hinder further advances until the publications of the larger galaxy catalogs of Shane & Wirtanen (1967) and Zwicky et al. (1961). Subsequent analyses focused on measurements of the angular correlation function (due to a lack of accurate distances), including a series of important papers beginning with Peebles (1973). Advances in theoretical modeling coupled with the realization that galaxies could be used as tracers of the matter density field led to the first robust cosmological inferences from LSS data (Zel'dovich 1970; Davis et al. 1977; Davis & Peebles 1977; Peebles 1980; Davis & Peebles 1983; Maddox et al. 1990; Baumgart & Fry 1991; Park et al. 1992). We show examples of two such early analyses in Figure 1.2. The left panel shows the early angular correlation function measurements from Davis et al. (1977) for the Shane & Wirtanen (1967) and Zwicky et al. (1961) catalogs, while the right panel plots the correlation function perpendicular and parallel to the line-of-sight using 2,400 galaxies from the first CfA redshift survey (Huchra et al. 1983). This latter measurement represents one of the earliest measurements of three-dimensional clustering using galaxy redshift information and provided evidence for the anisotropy introduced by peculiar velocities.

As the sizes of available galaxy data sets increased, the interpretation of the clustering results spurred further theoretical advances. In particular, a number of works studied the relationship between the galaxy and matter density fields, known as "galaxy bias", to help resolve discrepancies between observations and theoretical predictions (Kaiser 1984; Davis et al. 1985; Rees 1985; Cole & Kaiser 1989). The impact of peculiar velocities on observed clustering motivated several milestone theoretical studies on redshift-space distortions (Kaiser 1987; Davis & Peebles 1983; Hamilton 1992), with numerous applications to early redshift surveys (Cole et al. 1995; Loveday et al. 1996; Tadros et al. 1999). Furthermore, early LSS clustering analyses demonstrated that the prevailing theory of the time (a flat, matter-dominated Universe) could not explain the observed data, providing some of the earliest evidence for the ΛCDM model (Efstathiou et al. 1990; Vogeley et al. 1992; Krauss & Turner 1995; Ostriker & Steinhardt 1995).

The rise of multi-fiber spectrograph technology enabled galaxy surveys to measure redshifts for hundreds of thousands of objects, pioneered by the Two-degree Field Galaxy Redshift Survey (2dFGRS; Colless et al. 2001) and the Sloan Digital Sky Survey (SDSS; York

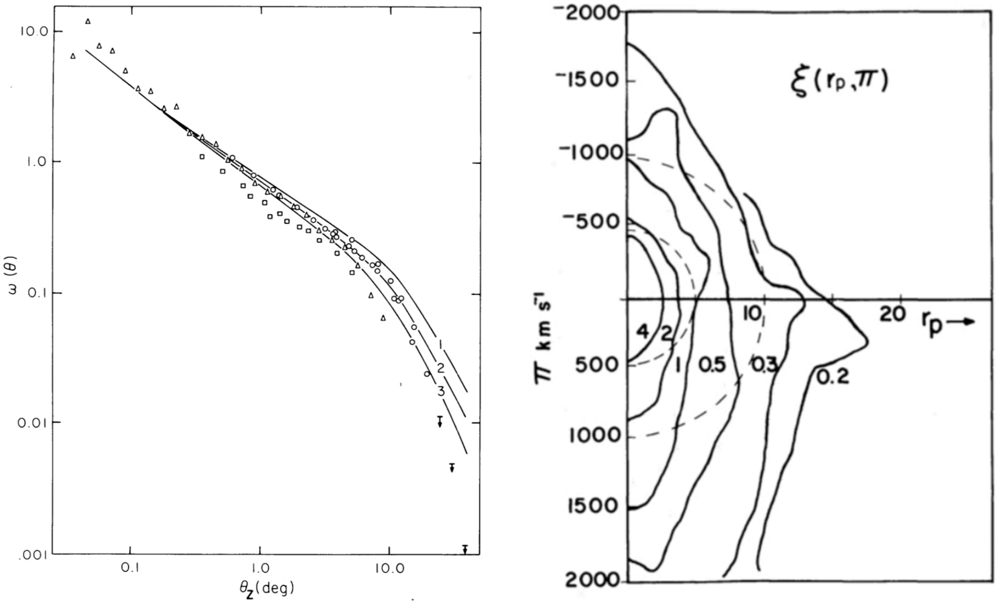


Figure 1.2: *Left:* an early measurement of the angular correlation function $w(\theta)$, using the Shane-Wirtanen and Zwicky galaxy catalogs (as presented in Davis et al. 1977). *Right:* a 2D correlation function measurement $\xi(r_\mathrm{p}, \pi)$ using data (contours) from the first CfA redshift survey of 2,400 galaxies (as presented in Davis & Peebles 1983). Evidence for anisotropy due to the effects of galaxy peculiar velocities can be seen.

et al. 2000). These surveys further cemented the ΛCDM paradigm with the measurement of the 2dFGRS galaxy power spectrum (Percival et al. 2001) and the first detections of the BAO from two-point clustering using the SDSS (Eisenstein et al. 2005) and 2dFGRS (Cole et al. 2005) data sets.

1.2.3 Present-day and future redshift surveys

Several redshift surveys currently underway (or recently completed) were designed to optimize constraints of the expansion rate and properties of dark energy using measurements of the BAO in galaxy clustering. In particular, the recently completed Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) represents the current state of the art data set in LSS and has measured redshifts for more than 1.5 million galaxies across the redshift range $0.2 < z < 0.75$. In recent results, Alam et al. (2017) uses measurements of the BAO feature to constrain the expansion rate at the percent level at redshifts of $z = 0.38$, 0.51, and 0.61. Figure 1.3 shows the exquisite precision of the BOSS measurements at $z = 0.51$ for the correlation function and power spectrum, where redshift-space distortions and the BAO can both be clearly seen. BOSS has also measured the BAO at $z \approx 2.5$ using the Lyman-α forest to trace the matter density field (Font-Ribera et al. 2014; Delubac et al. 2015). The BAO constraints from the BOSS galaxy sample are complemented by similar constraints at

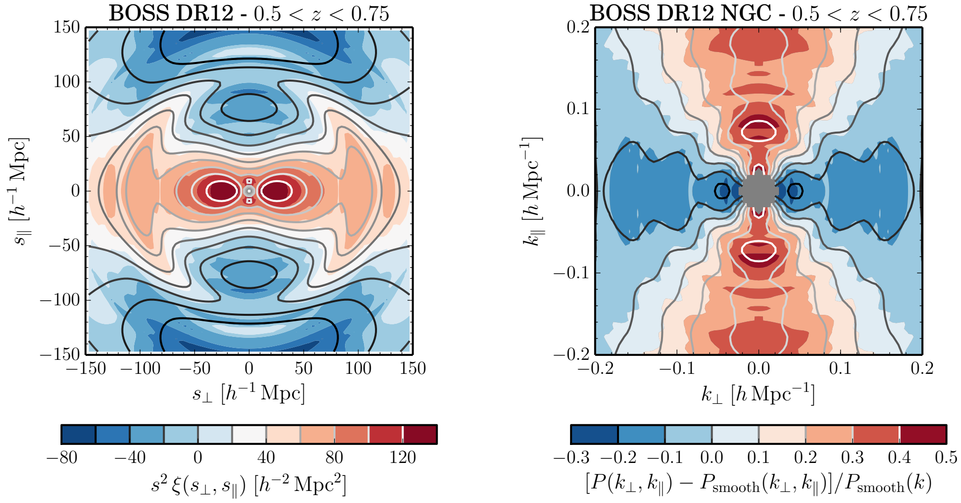


Figure 1.3: The 2D galaxy correlation function (*left*) and power spectrum (*right*) in the directions perpendicular and parallel to the line-of-sight for the completed BOSS data set (as presented in Alam et al. 2017). The color scale shows the data and the contours show the prediction of the best-fit model. The anisotropy of the contours is due to a combination of redshift-space distortions and the Alcock-Paczynski effect. The BAO ring is evident in the correlation function on scales of ~100 h^{-1}Mpc in the left panel. To better distinguish oscillations in the power spectrum due to BAO, the right panel plots the 2D power spectrum divided by the best-fit smooth component (without oscillations).

lower redshift by the 6dF Redshift Survey (Beutler et al. 2011) and at higher redshift by the WiggleZ survey (Blake et al. 2011a,b).

There are a number of ongoing and upcoming experiments that aim to shed light on some of the remaining open questions in cosmology and large-scale structure. The primary goal driving the design of these surveys is the the origin of cosmic acceleration. Is the cosmological constant model for dark energy, as is assumed in the ΛCDM model, correct, or does its energy density vary in time and space? Or, does the accelerated expansion arise from a breakdown of general relativity on cosmological scales? Other goals include the detection of neutrino mass and the neutrino hierarchy of states and the physics of inflation and the early Universe, including the detection of primordial non-Gaussianity. Ongoing (so-called Stage III) surveys attempting to answer these questions include eBOSS (Dawson et al. 2016), the Dark Energy Survey (DES; Diehl et al. 2014), and the Hobby Eberly Telescope Dark Energy Experiment (HETDEX; Hill et al. 2008). Future surveys (Stage IV) include the Dark Energy Spectroscopic Instrument (DESI; Levi et al. 2013), the Subaru Prime Focus Spectrograph (PFS; Takada et al. 2014), 4MOST (de Jong et al. 2014), and the space-based *Euclid* satellite (Laureijs et al. 2011) and Wide Field Infrared Survey Telescope (WFIRST; Spergel et al. 2015).

1.3 Theory

In this section, we present the theoretical framework for the analysis of LSS data. We discuss the smooth background Universe in §1.3.1 and the creation, evolution, and measurement of density fluctuations in §1.3.2. In §1.3.3, we describe some of the key elements of modeling full-shape broadband clustering, including galaxy bias, redshift-space distortions, and nonlinear gravitational evolution.

1.3.1 The smooth background

In general relativity (GR), a homogeneous and isotropic expanding Universe can be described via the Friedmann-Lemaítre-Robertson-Walker (FLRW) metric, where the spacetime line element is given by[3]

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi)\right], \tag{1.1}$$

where t is the time coordinate, r, θ, and ϕ are the spatial coordinates in a spherical coordinate system, and $a(t)$ is the dimensionless scale factor which describes the expansion of the Universe. The curvature constant K defines the spatial geometry, with $K = 0$, $K > 0$, and $K < 0$ corresponding to a flat, closed, and open Universe, respectively.

The scale factor $a(t)$ relates the comoving coordinate system, which is fixed with respect to the background expansion, to the corresponding physical coordinates. We can relate comoving separations $\boldsymbol{x}$ to physical distances $\boldsymbol{r}$ as

$$\boldsymbol{r} = a(t)\,\boldsymbol{x}, \text{ and } \boldsymbol{v} = \dot{\boldsymbol{r}} = \dot{a}(t)\,\boldsymbol{x} + \boldsymbol{v}_{\mathrm{pec}}, \tag{1.2}$$

where $\boldsymbol{v}$ is the physical velocity field, which is composed of a contribution from the background expansion and a residual peculiar velocity $\boldsymbol{v}_{\mathrm{pec}}$. The normalization of the scale factor is chosen to be unity today, $a(t_0) \equiv 1$, where t_0 refers to the present day. Both the time coordinate t and the scale factor are monotonically increasing quantities and can be used to refer to specific points in the Universe's past or future. Another common related quantity is the cosmological redshift z, which is defined as the amount that the wavelength of an emitted photon stretches due to the background expansion. It is given by

$$1 + z = \frac{\lambda_{\mathrm{obs}}}{\lambda_{\mathrm{emit}}} = \frac{1}{a}, \tag{1.3}$$

where λ_{emit} is the wavelength that the light was emitted at, and λ_{obs} is the wavelength it was observed at (at $a_{\mathrm{obs}} = 1$). For cosmological observations, redshift is used as the radial coordinate. Light that is observed with a redshift z was emitted in the past when the Universe's scale factor was $a = (1 + z)^{-1}$.

[3] Note that we assume a unit system using $c = 1$.

It is also convenient to define a radial coordinate χ such that

$$\mathrm{d}\chi = \frac{\mathrm{d}r}{\sqrt{1-Kr^2}}. \tag{1.4}$$

This allows the metric of equation 1.1 to be written as

$$\mathrm{d}s^2 = \mathrm{d}t^2 + a^2(t)\left[\mathrm{d}\chi^2 + S_K^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi)\right], \tag{1.5}$$

where S_K is defined to be

$$S_K(\chi) = \begin{cases} K^{-1/2}\sin(\sqrt{K}\chi) & \text{if } K > 0, \\ \chi & \text{if } K = 0, \\ |K|^{-1/2}\sinh(\sqrt{|K|}\chi) & \text{if } K < 0. \end{cases} \tag{1.6}$$

Under the assumptions of isotropy and homogeneity, the smooth background Universe can be described as a perfect fluid with density ρ and pressure p. Using the Einstein equations, the evolution of the scale factor, the so-called Friedmann equations, are given by

$$\left(\frac{\dot{a}}{a}\right)^2 \equiv H^2(a) = \frac{8\pi G}{3}\rho - \frac{K}{a^2}, \tag{1.7}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p), \tag{1.8}$$

where the Hubble parameter, $H(a) \equiv \dot{a}/a$, is a measure of the expansion rate of the Universe. The present-day Hubble parameter is denoted as H_0 and is often expressed in terms of the dimensionless Hubble parameter h as $H_0 = 100h$ km s^{-1} Mpc^{-1}.

We can model each component of the density as a fluid with equation of state, $p = w\rho$, where w is constant. Using energy-momentum conservation, the density evolution with scale factor is

$$\rho \propto a^{-3(1+w)}, \tag{1.9}$$

where the equation-of-state parameter w depends on the particle species. The relevant values are $w = 0$ for non-relativistic, collisionless matter, $w = 1/3$ for radiation, $w = -1$ for a cosmological constant, and $w < -1/3$ for more general forms of dark energy. Equation 1.8 shows that accelerated expansion occurs when $w < -1/3$. Each component of the total density evolves differently with time, leading to periods of time where different components dominate the total density.

It is useful to consider the value of the total density in equation 1.7 for a flat geometry ($K = 0$), which is known as the critical density,

$$\rho_{\text{crit}}(z) \equiv \frac{3H(z)}{8\pi G}. \tag{1.10}$$

We can normalize the density of each component of the density by this critical density to yield the density parameters,

$$\Omega_i(z) \equiv \frac{\rho_i(z)}{\rho_{\rm crit}(z)}, \tag{1.11}$$

and define the total density parameter as $\Omega_{\rm tot} = \Omega_{\rm m} + \Omega_{\rm r} + \Omega_{\rm de}$. We can define a similar quantity for the curvature,

$$\Omega_{\rm k}(z) \equiv -\frac{K}{H^2(z)}(1+z)^2, \tag{1.12}$$

such that equation 1.7 now reads $\sum_i \Omega_i + \Omega_{\rm k} = \Omega_{\rm tot} + \Omega_{\rm k} = 1$. It is useful to express this equation in terms of present-day quantities,

$$H^2(z) = H_0^2\left[\Omega_{\rm m,0}(1+z)^3 + \Omega_{\rm r,0}(1+z)^4 + \Omega_{\rm de,0}(1+z)^{3(1+w)} + \Omega_{\rm k,0}(1+z)^2\right]. \tag{1.13}$$

In a flat ΛCDM model with $w_{\rm de} = w_\Lambda = -1$ and $K = 0$, this equation simplifies to

$$H^2(z) = H_0^2\left[\Omega_{\rm m,0}(1+z)^3 + \Omega_{\rm r,0}(1+z)^4 + \Omega_\Lambda\right]. \tag{1.14}$$

More general forms of dark energy allow its equation-of-state parameter to vary with time. A common, phenomenological parametrization for $w_{\rm de}(z)$ is (Chevallier & Polarski 2001; Linder 2003)

$$w_{\rm de}(z) = w_0 + w_a(1-a) = w_0 + w_a\frac{z}{1+z}. \tag{1.15}$$

Equation 1.14 plays an important role in determining distances in cosmology. The line-of-sight comoving distance to an object can be calculated by noting that photons travel on null geodesics ($\mathrm{d}s^2 = 0$), such that the comoving distance is given by

$$\chi(z) = \int_0^t \frac{dt'}{a(t')} = \int_0^z \frac{\mathrm{d}z'}{H(z')}, \tag{1.16}$$

where we have used the fact that $H = (\mathrm{d}a/dt)/a$ and $\mathrm{d}a = -a^2\mathrm{d}z$.

Another important distance quantity in LSS, and in particular, the analysis of the BAO as a standard ruler, is the angular diameter distance. It gives the ratio of an object's physical size to its observed angular size. From equation 1.5, we can see that an object with an angular size of $\delta\phi$ has a corresponding physical size given by

$$\delta l = a(z)S_K(\chi(z))\delta\phi. \tag{1.17}$$

By comparison with the analogous relation in flat Euclidean space, we see that the angular diameter distance must correspond to the pre-factor in this equation, given by

$$D_A(z) \equiv a(z)S_K(\chi(z)) = (1+z)^{-1}S_K(\chi(z)). \tag{1.18}$$

In a flat geometry, the angular diameter distance is simply related to the comoving distance as $D_A(z) = \chi/(1+z)$.

1.3.2 Density perturbations

This section introduces the methods by which we statistically describe density fluctuations in the Universe (§1.3.2) and provides additional background about the creation and evolution of these fluctuations throughout the Universe's history (§1.3.2).

Two-point clustering

We start by defining the overdensity field in configuration space, $\delta(\boldsymbol{x})$, as

$$\delta(\boldsymbol{x}) \equiv \frac{n(\boldsymbol{x}) - \bar{n}}{\bar{n}}, \tag{1.19}$$

where $n(\boldsymbol{x})$ is the spatial number density of the objects of interest and $\bar{n}$ is the mean density. We can Fourier transform this field to express it as a function of wavenumber $\boldsymbol{k}$ using the following Fourier convention

$$\delta(\boldsymbol{k}) = \int d^3x \; \delta(\boldsymbol{x}) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{1.20}$$

$$\delta(\boldsymbol{x}) = \int \frac{d^3k}{(2\pi)^3} \; \delta(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{1.21}$$

We can measure the statistical properties of $\delta(\boldsymbol{x})$ via N-point correlators, $\langle \delta_1 \delta_2 \ldots \delta_N \rangle$. Here, $\langle \ldots \rangle$ denotes an ensemble average but as we only have a single realization of the Universe, this is replaced observationally with a spatial average. The statistical properties of a Gaussian field are fully specified by its two-point statistics, either the correlation function in configuration space or the power spectrum in Fourier space. The latter is defined as the covariance of the overdensity field in Fourier space,

$$\langle \delta(\boldsymbol{k}) \delta^\star(\boldsymbol{k}') \rangle = (2\pi)^3 \delta_D(\boldsymbol{k} - \boldsymbol{k}') P(\boldsymbol{k}), \tag{1.22}$$

where δ_D is the 3D Dirac delta function and $\delta^\star(\boldsymbol{k})$ denotes the complex conjugate of the Fourier modes of the overdensity field. The correlation function is the Fourier transform of the power spectrum,

$$\xi(\boldsymbol{r}) \equiv \langle \delta(\boldsymbol{x}) \delta(\boldsymbol{x} + \boldsymbol{r}) \rangle = \int \frac{d^3k}{(2\pi^3)} \; P(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}. \tag{1.23}$$

Under assumptions of isotropy and homogeneity, these two-point statistics become a function of amplitude only, $P(\boldsymbol{k}) = P(k)$ and $\xi(\boldsymbol{r}) = \xi(r)$. In this case, the correlation function can be calculated as

$$\xi(r) = \frac{1}{2\pi^2} \int \mathrm{d}k \; k^2 P(k) j_0(kr), \tag{1.24}$$

where j_0 is the spherical Bessel function of order zero. As the power spectrum has units of volume, it is convenient to define a dimensionless quantity,

$$\Delta^2(k) = \frac{k^3}{2\pi^2} P(k). \tag{1.25}$$

Intuitively, the power spectrum can be viewed as measuring the variance of the density fluctuations on a given scale, while the correlation function can be understood as the number of pairs of objects at a given separation in excess of a random distribution.

A Gaussian density field is fully described by its two-point statistics, and the initial density fluctuations in the early Universe are expected be very close to Gaussian. Initially, the density fluctuations were small ($|\delta| \ll 1$) and can be described via linear approximations to the gravitational evolution equations. The density fluctuations in Fourier space also remain independent, as evidenced by the Dirac delta function in equation 1.22. However, this approximation only holds on large scales when considering the evolved density field of the late-time Universe. In this case, different scales become correlated via nonlinear gravitational evolution. This mode coupling generates non-zero higher order correlators, most importantly the three-point (bispectrum) and four-point functions (trispectrum).

Initial conditions and the transfer function

Inflation (Guth 1981; Linde 1982; Albrecht & Steinhardt 1982) is widely accepted as the most likely scenario for the generation of density fluctuations in the early Universe. While we are yet to find conclusive observational evidence, inflation does make several predictions that can been observationally tested. These predictions can be summarized as:

- The initial density fluctuations were Gaussian, possibly with a small level of primordial non-Gaussianity (PNG), and can be described by the primordial power spectrum. The *Planck* 2015 results placed strong constraints on PNG; they constrained local type PNG to be $f_{\mathrm{NL}} = 0.8 \pm 5.0$. We explore PNG in more detail in Chapter 5 using the quasar power spectrum to constrain local type PNG.

- The initial perturbations are expected to be adiabatic, with only a single degree of freedom, and can be described entirely by the density power spectrum. The *Planck* 2015 results constrain the alternative, isocurvature modes, to be extremely small.

- The (scalar) primordial power spectrum is nearly scale-invariant and described as a power law, as given by

$$\Delta_{\mathrm{s}}^2(k) = A_{\mathrm{s}} \left(\frac{k}{k_0} \right)^{n_{\mathrm{s}}-1}, \tag{1.26}$$

where $\Delta_{\rm s}^2$ is the dimensionless primordial power spectrum (defined in equation 1.25), $A_{\rm s}$ is the freely varying amplitude, k_0 is an arbitrary pivot scale, and $n_{\rm s}$ is the spectral index. The spectral index is defined as

$$n_{\rm s}(k) - 1 = \frac{{\rm d}\ln\Delta_{\rm s}^2(k)}{{\rm d}\ln k}. \tag{1.27}$$

Here, "scale-invariant" refers to the primordial fluctuations in potential and translates to $n_{\rm s} \approx 1$ for the density perturbations. The *Planck* 2015 results constrain the spectral index with unprecedented precision, finding $n_{\rm s} = 0.968 \pm 0.006$.

- Inflation produces tensor perturbations in addition to scalar fluctuations. The tensor fluctuations correspond to gravitational waves, which can be detected through a measurement of the B-mode polarization of the CMB. No such detection has yet been measured, but if found, would provide strong evidence for the accuracy of the inflationary paradigm.

The evolution of the initial density fluctuations is a complicated process that requires using the linearized Einstein equations and a perturbed FLRW spacetime metric.[4] Here, we qualitatively describe the main aspects of linear evolution. We ignore the added complexities of nonlinear evolution, which must require the use of perturbative approaches, numerical simulations, or fitting formulae to model accurately (see §1.3.3 for a discussion of nonlinearities).

Following inflation, the comoving horizon increased with time as the Universe expanded, and fluctuations of increasing scale entered the horizon. Modes within the horizon were causally connected and able to evolve with time. However, this evolution depended on the dominant component of the total energy density. In the early Universe when radiation dominated, the pressure of the radiation impeded growth of fluctuations, but once non-relativistic matter became the major component, fluctuations were able to grow more substantially, with growth proportional to the scale factor. Once dark energy began to dominates at late times, the growth slowed once again. The critical scale for understanding the evolution of perturbations is the horizon scale at the transition from radiation to matter domination, $k_{\rm eq}$. Fluctuations smaller than this scale entered the horizon during radiation domination and had their growth impeded, while larger modes entered during matter domination and experienced immediate growth. This leads to a turnover in the power spectrum where $k = k_{\rm eq}$.

The transfer function encodes this complicated evolution, relating the power spectrum of the evolved perturbations to the primordial power spectrum,

$$P(k, a) \propto \sigma_8^2 D^2(a) T^2(k) k^{n_{\rm s}}, \tag{1.28}$$

where $k^{n_{\rm s}}$ is the primordial power spectrum, $T(k)$ is the transfer function (independent of redshift), and the redshift evolution is encoded in the growth function $D(a)$. The power

[4] We refer to the reader to the review of Bardeen (1980) and the textbook treatment of Dodelson (2003) for more details.

spectrum normalization is parametrized by the value of σ_8^2, which expresses the variance of the density fluctuations at $z = 0$ smoothed on a comoving scale of $R = 8\ h^{-1}$Mpc.

While analytic approximations for the transfer function do exist, (Bardeen et al. 1986; Eisenstein et al. 1998), it is now typical to use software to numerically solve the full set of Einstein and Boltzmann equations, e.g., CAMB (Lewis et al. 2000) or CLASS (Lesgourgues 2011), in order to achieve a high degree of accuracy. Nevertheless, it is informative to consider the asymptotic behavior of $T(k)$ for scales much larger and smaller than $k_{\rm eq}$,

$$T(k) \propto \begin{cases} 1 & \text{if } k \ll k_{\rm eq}, \\ k^{-2}\ln k & \text{if } k \gg k_{\rm eq}, \end{cases} \tag{1.29}$$

From these limits we see that the scale dependence of evolved fluctuations on large scales mirrors that of the primordial power spectrum.

The growth function $D(a)$ specifies the redshift evolution of the amplitude of the power spectrum. In the matter-dominated epoch, the growth is simply given by $D(a) \propto a$. However, the growth is more complicated when considering the late-time evolution when dark energy is the dominant component. In general, the growth function is given by (Dodelson 2003)

$$D(a) = \frac{5\Omega_{\rm m,0}}{2}\frac{H(a)}{H_0}\int_0^a \frac{{\rm d}a'}{(a')^3}\frac{H_0^3}{H^3(a')}, \tag{1.30}$$

where we have assumed that radiation is negligible, and the normalization is such that $D(a) = a$ in the matter-dominated era ($z \approx 10$). For a ΛCDM model, growth at late times is hampered by dark energy, and $D(a)$ increases slower than in the case of $\Omega_{\rm m,0} = 1$.

1.3.3 Full-shape clustering models

In this section, we describe the key elements of a full-shape clustering analysis from a modeling perspective, including the treatment of biasing (§1.3.3), redshift-space distortions (§1.3.3), and nonlinearities (§1.3.3).

Bias

The luminous objects that we observe in the late-time Universe, e.g., galaxies or quasars, are "biased" tracers of the underlying matter density field. In this context, bias refers to the relationship between the spatial distribution of the tracer and that of the underlying matter field. In principle, this relationship can be extremely complicated, with stochastic and non-local contributions. For a comprehensive review we refer the reader to Fry & Gaztanaga (1993) and Desjacques et al. (2016).

In its most general form, bias can be viewed as a functional of the underlying matter field, such that

$$\delta_{\rm g}(\boldsymbol{x}) = F[\delta(\boldsymbol{x}')], \tag{1.31}$$

where $\delta_{\rm g}$ is the overdensity field of the tracer, and δ is the matter overdensity. The relationship can be non-local, where δ_g at position $\boldsymbol{x}$ depends on the value of δ at $\boldsymbol{x}'$. However, a usual approach is to perturbatively Taylor expand the tracer overdensity as a function of the local matter density as (Fry & Gaztanaga 1993)

$$\delta_{\rm g}(\boldsymbol{x}) \approx \sum_{i=0}^{\infty} \frac{b_i}{i!}(\delta^i(\boldsymbol{x}) - \langle \delta^i \rangle), \tag{1.32}$$

where b_i represent the bias parameters. Another common simplifying assumption is that the bias relationship is linear on sufficiently large scales: $\delta_{\rm g} = b_1 \delta$. In this case, the two-point clustering statistics are boosted by a constant offset,

$$P_{\rm g}(k) = b_1^2 P(k), \text{ and } \xi_{\rm g}(r) = b_1^2 \xi(r). \tag{1.33}$$

In Chapter 4, we extend the modeling of the broadband power spectrum of galaxies to small scales, where the simple linear bias relationship is no longer accurate. We rely on the more recent theoretical work of McDonald & Roy (2009); Baldauf et al. (2012); Saito et al. (2014) to include both nonlocal and nonlinear contributions to the bias, which have been demonstrated to improve the accuracy of theoretical models. Furthermore, discrete tracers are also stochastic tracers of the matter field. This stochasticity is typically modeled as the Poisson shot noise of the sample ($\bar{n}^{-1}$, where $\bar{n}$ is the number density), but there can be significant deviations that lead to additional complicated scale-dependence in the relationship between the tracer and matter density fields (e.g., Baldauf et al. 2013). Modeling these deviations plays an important role in our work presented in Chapter 4.

Redshift-space distortions

The three-dimensional distribution of luminous objects, as measured by redshift surveys, are distorted in the radial direction. These so-called redshift-space distortions result from the fact that the measured redshift of objects is also sensitive to its peculiar velocity through the Doppler effect, and in turn, the measured redshift is used to infer the line-of-sight distance to the object. Because the peculiar velocity field is sourced by the large-scale gravitational potential, the clustering of objects in redshift space includes an anisotropic signal containing information about the rate of structure growth in the Universe.

On large scales, objects have moderate infall velocities, leading to a compression of the clustering signal along the line-of-sight. In a seminal paper, Kaiser (1987) derived the linear-order prediction for this effect. The isotropic clustering in real space becomes anisotropic with a boost in clustering along the line-of-sight,

$$\delta^S(k, \mu) = (1 + f\mu^2)\delta(k), \tag{1.34}$$

where δ^S and δ are the overdensity fields in redshift and real space, respectively, $f = {\rm d}\ln D/{\rm d}\ln a$ is the logarithmic growth rate, and $\mu = \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}$ is the cosine of the angle between the Fourier wavevector and the line-of-sight. For a biased tracer, the linear redshift-space power spectrum becomes

$$P^S_{\rm g}(k,\mu) = \left[1+\beta\mu^2\right]^2\ b_1^2 P(k), \tag{1.35}$$

where $P^S_{\rm g}$ is the redshift-space power spectrum of the tracer, $P(k)$ is the matter power spectrum in real space, and we have defined the linear RSD parameter $\beta \equiv f/b_1$, where b_1 is the linear bias of the tracer. The anisotropic form of equation 1.35 suggests expanding $P^S(k,\mu)$ into a basis of Legendre polynomials $\mathcal{L}_\ell(\mu)$ as

$$P^S(k,\mu) = \sum_{\ell=0,2,4} P_\ell(k)\mathcal{L}_\ell(\mu) = P_0(k)\mathcal{L}_0(\mu) + P_2(k)\mathcal{L}_2(\mu) + P_4\mathcal{L}_4(\mu), \tag{1.36}$$

where we have introduced the power spectrum multipoles $P_\ell(k)$, and P_0, P_2, and P_4 are known as the monopole, quadrupole, and hexadecapole, respectively. From equation 1.36 we can see that linear RSD can be fully described by the $\ell = 0$, 2, and 4 multipoles. Measurements of β and RSD modeling in general constrain the growth of structure through their sensitivity to the growth rate f. These measurements can test for deviations from the ΛCDM model and general relativity by comparing to the general relativity prediction of $f \simeq [\Omega_{\rm m}(z)]^{0.55}$.

On small scales, additional distortions are generated by the large, nonlinear virial motions of objects within their dark matter halos, which is known as the Finger-of-God (FoG) effect (Jackson 1972). The FoG effect elongates structures along the line-of-sight, damping the clustering on small scales. It is typically modeled by multiplying the power spectrum of equation 1.35 with a Gaussian or Lorentzian damping factor, motivated by the shape of the pairwise velocity distribution function (Peacock & Dodds 1994; Park et al. 1994; Ballinger et al. 1996).

For a precise full-shape analysis, the RSD description must go beyond the linear-order predictions of Kaiser (1987), which is known to break down on relatively large scales, e.g., Okumura & Jing (2011); Jennings (2012); Kwan et al. (2012). In Chapter 4, our description of the redshift-space power spectrum uses the distribution function approach of Seljak & McDonald (2011) to model the nonlinear mapping from real to redshift space. Our approach is complementary to other models in the literature that account for the nonlinear coupling between the density and velocity fields in redshift space, e.g., Scoccimarro (2004); Taruya et al. (2010). We focus on large scales in Chapter 5 when modeling the quasar power spectrum, and thus, we are able to use the simple Kaiser power spectrum model of equation 1.35.

Nonlinear evolution

Linear clustering models break down on relatively large scales, with a number of nonlinear processes becoming important on small scales where density perturbations become large ($\delta \gg 1$). For the power spectrum, this break down occurs on scales even as large as $k \sim 0.1\ h\text{Mpc}^{-1}$. Using a wider range of scales when modeling a clustering measurement is generally beneficial, as the statistical precision of the measurements is typically higher on smaller scales. Thus, there is strong motivation to model small scales, despite the added challenges due to nonlinearities. This motivation is the driving force behind the modeling presented in Chapter 4, which presents a power spectrum model accurate well into the nonlinear regime, $k \sim 0.4\ h\mathrm{Mpc}^{-1}$.

In general, there are three common approaches to modeling the effects of nonlinearities on clustering measurements, which can be summarized as:

- *Simulation-based methods:* N-body simulations directly solve the nonlinear evolution of the density field via numerical methods (see, e.g., Springel 2005). They represent the most accurate method for theoretical predictions but are computationally expensive. This computational cost prevents their use in parameter inference from clustering measurements. In recent years, a class of quasi-N-body schemes, known as particle mesh solvers, has been developed to reduce the computational cost of simulations (Merz et al. 2005; Tassev et al. 2013; White et al. 2014; Feng et al. 2016). These methods allow the fast generation of catalogs of biased tracers from approximate dark matter density fields.

- *Perturbation theory:* Nonlinear gravitational dynamics can be modeled with perturbative approaches that expand the nonlinear density field in terms of powers of linear quantities. Known as cosmological perturbation theory, the topic is well-studied with a wide variety of complementary approaches (see Bernardeau et al. (2002) for a review). Some of the more popular approaches in recent work include renormalized perturbation theory (Crocce & Scoccimarro 2006a,b; Taruya et al. 2012) and Convolution Lagrangian Perturbation Theory (Carlson et al. 2013).

- *Fitting functions:* The use of fitting formulae, which are often calibrated on N-body simulations or use a physically motivated ansatz, is a quick and practical method to include the effects of nonlinear evolution on clustering statistics. Prominent examples include the Halofit (Smith et al. 2003) and Halo-Zel'dovich Perturbation Theory (Seljak & Vlah 2015) prescriptions for the nonlinear matter power spectrum.

Each of these methods has its own advantages and disadvantages, and we rely on a combination of all three approaches in Chapter 4 to accurately model the redshift-space galaxy power spectrum down to scales of $k \sim 0.4\ h\mathrm{Mpc}^{-1}$.

Chapter 2

nbodykit: an open-source, massively parallel toolkit for large-scale structure

In this chapter, we present `nbodykit`, an open-source, massively parallel Python toolkit for analyzing large-scale structure (LSS) data. Using Python bindings of the Message Passing Interface (MPI), we provide parallel implementations of many commonly used algorithms in LSS. `nbodykit` is both an interactive and scalable piece of scientific software, performing well in a supercomputing environment while still taking advantage of the interactive tools provided by the Python ecosystem. Existing functionality includes estimators of the power spectrum, 2 and 3-point correlation functions, a Friends-of-Friends grouping algorithm, mock catalog creation via the halo occupation distribution technique, and approximate N-body simulations via the FastPM scheme. The package also provides a set of distributed data containers, insulated from the algorithms themselves, that enable `nbodykit` to provide a unified treatment of both simulation and observational data sets. `nbodykit` can be easily deployed in a high performance computing environment, overcoming some of the traditional difficulties of using Python on supercomputers. We provide performance benchmarks illustrating the scalability of the software. The modular, component-based approach of `nbodykit` allows researchers to easily build complex applications using its tools. The package is extensively documented at `http://nbodykit.readthedocs.io`, which also includes an interactive set of example recipes for new users to explore. As open-source software, we hope `nbodykit` provides a common framework for the community to use and develop in confronting the analysis challenges of future LSS surveys.

2.1 Introduction

The analysis of LSS data sets has played a pivotal role in establishing the current concordance paradigm in modern cosmology, the ΛCDM model. From the earliest galaxy surveys (Davis et al. 1985; Maddox et al. 1990), comparisons between the theoretical predictions for the distribution of matter in the Universe and observations have proven to be a valuable tool. Indeed, LSS observations, in combination with cosmic microwave background (CMB) measurements, provided some of the earliest evidence for the ΛCDM model, e.g., Efstathiou et al. (1990); Krauss & Turner (1995); Ostriker & Steinhardt (1995). Interest in LSS surveys increased immensely following the first direct evidence for cosmic acceleration (Riess et al. 1998; Perlmutter et al. 1999), as it was realized that the baryon acoustic oscillation (BAO) feature imprinted on large-scale clustering provided a "standard ruler" to map the expansion history (Eisenstein et al. 1998; Blake & Glazebrook 2003; Seo & Eisenstein 2003). From its first measurements (Cole et al. 2005; Eisenstein et al. 2005) to more recent studies (Font-Ribera et al. 2014; Delubac et al. 2015; Alam et al. 2017; Slepian et al. 2017), the BAO has proved to be a valuable probe of cosmic acceleration, enabling the most precise measurements of the expansion history of the Universe over a wide redshift range. Analyses of these data sets have also pushed us closer to answering other important questions in contemporary cosmology, including deviations from General Relativity (Mueller et al. 2016), the neutrino mass scale (Lesgourgues & Pastor 2006; Beutler et al. 2014a), and the existence of primordial non-Gaussianity (Slosar et al. 2008; Desjacques & Seljak 2010).

The foundations of the numerical methods used in LSS data analysis today go back several decades. Hockney & Eastwood (1981) discussed several important computer simulation methods, including but not limited to mass assignment interpolation windows and the interlacing technique for reducing aliasing. The Friends-of-Friends (FOF) algorithm for identifying halos from a numerical simulation was first utilized in Davis et al. (1985). The most commonly used clustering estimators for the two-point correlation function (2PCF) and power spectrum were first developed in Landy & Szalay (1993) and Feldman et al. (1994), respectively, and techniques to measure anisotropic clustering via a multipole basis were first used around the same time, e.g., Cole et al. (1995). Other modern, well-established numerical techniques include N-body simulation methods, e.g., Springel et al. (2001); Springel (2005), and the use of KD-trees in correlation function estimators (Moore et al. 2001).

Recent years have brought important updates to these analysis techniques. Advances in LSS observations, with increased sample sizes and statistical precision, have driven the development of new statistical estimators, while also increasing modeling complexities and creating a need to reduce wall-clock times. Recently, we have seen faster power spectrum and 2PCF multipole estimators (Yamamoto et al. 2006; Scoccimarro 2015; Bianchi et al. 2015; Slepian & Eisenstein 2015b, 2016; Hand et al. 2017b) and improved FOF algorithms (Springel 2005; Behroozi et al. 2013; Feng & Modi 2017). Highly optimized software, e.g., `Corrfunc` (Sinha & Garrison 2017), is also becoming increasingly common. New statistical estimators, e.g., Slepian & Eisenstein (2015a, 2017); Castorina & White (2017), are being developed to extract as much information as possible from LSS surveys. The rise of particle mesh simulation methods (Merz et al. 2005; Tassev et al. 2013; White et al. 2014; Feng et al. 2016) has offered a computationally cheaper alternative to running full N-body simulations. Finally, tools have emerged to help deal with the growing complexities of modeling the connection between halos and galaxies (Hearin et al. 2017). These examples represent just a sampling of the recent updates to LSS data analysis and modeling techniques.

The well-established foundation of LSS numerical methods suggests the community could benefit from a standard software package providing implementations of these methods. Such a package would also serve as a common framework for users as they incorporate future extensions and advancements. Given the already rising wall-clock times of current analyses and the expected volume of data from next-generation LSS surveys, scaling performance should also be a key priority.

Several computing trends in the past few years have emerged to help make such a software package possible. First, the Python programming language[1] has emerged as the most popular language in the field of astronomy (Momcheva & Tollerud 2015; NSF 2017), and the astropy[2] package (Astropy Collaboration et al. 2013) has led the development of an astronomy-focused Python ecosystem. Python's elegant syntax and dynamic nature make the language easy to learn and work with. Combined with its object-oriented focus and the larger ecosystem containing SciPy[3] (Jones et al. 2001–2017), NumPy[4] (van der Walt et al. 2011), IPython[5] (Perez & Granger 2007), and Jupyter[6] (Thomas et al. 2016), Python is well-suited for both rapid application development and use in scientific research. Second, the availability and performance of large-scale computing resources continues to grow, and initiatives, e.g., The Exascale Computing Project,[7] have been established to ensure the sustainability of this trend. At the same time, solutions to the traditional barriers to using Python on massively parallel, high-performance computing (HPC) machines have been developed. The mpi4py package (Dalcín et al. 2008; Dalcin et al. 2011) has facilitated the development of parallel Python applications by providing bindings of the Message Passing Interface (MPI) standard. Furthermore, tools have been developed, e.g., Feng & Hand (2016), to alleviate the start-up bottleneck encountered when launching Python applications on HPC systems.

Motivated by these recent developments, we present the first public release of nbodykit (v0.3.0), an open-source, parallel toolkit written in Python for use in the analysis of LSS data. Designed for use on HPC machines, nbodykit includes fully parallel implementations of a canonical set of LSS algorithms. It also includes a set of distributed and extensible data containers, which can support a wide variety of data formats and large volumes of data. These data containers are insulated from the algorithms themselves, allowing nbodykit to be used for either simulation or observational data sets. We have balanced the scalable nature of nbodykit with an object-oriented, component-based design that also facilitates interactive use. This allows researchers to take advantage of interactive Python tools, e.g., the Jupyter notebook, as well as integrate nbodykit components with their own software to build larger applications that solve specific problems in LSS.

nbodykit has been developed, tested, and deployed on the Edison and Cori Cray supercomputers at the National Energy Research Scientific Computing Center (NERSC) and

[1] http://python.org

[2] http://www.astropy.org

[3] https://www.scipy.org

[4] http://www.numpy.org

[5] https://ipython.org

[6] http://jupyter.org

[7] https://www.exascaleproject.org

has been utilized in several published research studies (Hand et al. 2017b,c; Ding et al. 2017; Pinol et al. 2017; Schmittfull et al. 2017; Modi et al. 2017; Feng et al. 2016; Waters et al. 2016). Since its start, it has been developed on GitHub as open-source software at https://github.com/bccp/nbodykit.

The objective of this chapter is to provide an overview of the nbodykit software and familiarize the community with some of its capabilities. We hope that researchers find nbodykit to be a useful tool in their scientific work and in the spirit of open science, that it continues to grow via community contributions. Extensive documentation and tutorials are available at http://nbodykit.readthedocs.io, and we do not aim to provide such detailed documentation in this work. The documentation also includes instructions for launching an interactive environment containing a set of example recipes. This allows new users to explore nbodykit without setting up their own nbodykit installation.

This chapter is organized as follows. We provide a broad overview of nbodykit in Section 2.2 and discuss a more detailed list of its capabilities in Section 2.3. We describe our development process and deployment strategy for nbodykit in Section 2.4. Section 2.5 presents an illustrative example use case, and Section 2.6 outlines performance benchmarks for various algorithms. Finally, we conclude and summarize in Section 2.7.

2.2 Overview

2.2.1 Initializing nbodykit

A core design goal of nbodykit is maintaining an interactive user experience, allowing the user to quickly experiment and to prototype new analysis pipelines while still leveraging the power of parallel processing when necessary. We adopt a "lab" framework for nbodykit, where all of the necessary data containers and algorithms can be imported from the nbodykit.lab module. Furthermore, we utilize Python's logging module to print messages at runtime, which allows users to track the progress of the application in real time. Typically, applications using nbodykit begin with the following statements:

```
from nbodykit.lab import *
from nbodykit import setup_logging

setup_logging()
```

2.2.2 The nbodykit Ecosystem

nbodykit is explicitly maintained as a pure Python package. However, it depends on several compiled extension packages that each focus on more specialized tasks. This approach enables nbodykit to describe higher-level abstractions in Python and retains the readability, syntax, and user interface benefits of the Python language. For computationally expensive sections of the code base, we use the compiled extension packages for speed. With the emergence of Python package managers such as Anaconda,[8] the availability of binary versions of these compiled packages for different operating systems has sufficiently eased most installation issues in our experience (see Section 2.4.3).

Below, we describe some of the more important dependencies of `nbodykit`, each of which is focused on solving a particular problem:

- `pfft-python`: a Python binding of the `PFFT` software (Pippig 2013), which computes parallel fast Fourier transforms (FFTs) (Feng 2017d).
- `pmesh`: particle mesh calculations, including density field interpolation and discrete parallel FFTs via `pfft-python` (Feng 2017e).
- `bigfile`: a reproducible, massively parallel input/output (IO) library for large, hierarchical data sets (Feng 2017a).
- `kdcount`: spatial indexing operations via KD-trees (Feng 2017c).
- `classylss`: a Python binding of the `CLASS` Boltzmann solver (Hand & Feng 2017).
- `fastpm-python`: a Python implementation of the FastPM scheme for quasi N-body simulations (Feng 2017b; Feng et al. 2016).
- `Corrfunc`: a set of high-performance routines for computing pair counting statistics (Sinha & Garrison 2017).
- `Halotools`: a package to build and test models of the galaxy-halo connection (Hearin et al. 2017).

2.2.3 A Component-Based Approach

The design of `nbodykit` focuses on a modular, component-based approach. The *components* are exposed to the user as a set of Python classes and functions, and users can combine these components to build their specific applications. This design differs from the more commonly used alternative in cosmology software, which is a monolithic application controlled by a single configuration file, e.g., as in `CAMB` (Lewis et al. 2000), `CLASS` (Blas et al. 2011), and `Gadget` (Springel et al. 2001). We note that modular, object-oriented designs are becoming more popular recently, e.g., `astropy`, the `yt project` (Turk et al. 2011), and `Halotools` (Hearin et al. 2017). During the development process, we have found that a component-based approach offers greater freedom and flexibility to build complex applications with `nbodykit`.

[8] https://anaconda.com

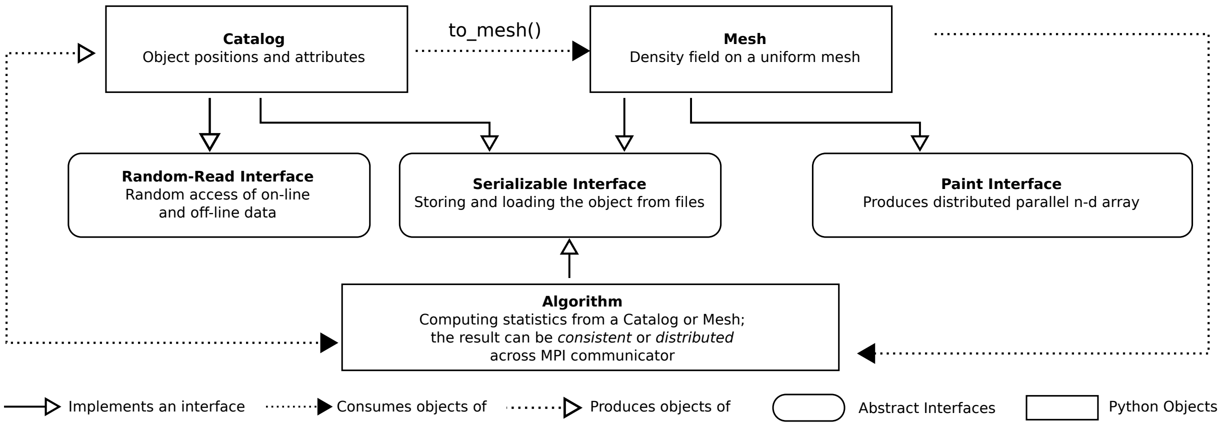

Catalog
Object positions and attributes
to_mesh()
Mesh
Density field on a uniform mesh
Random-Read Interface
Random access of on-line and off-line data
Serializable Interface
Storing and loading the object from files
Paint Interface
Produces distributed parallel n-d array
Algorithm
Computing statistics from a Catalog or Mesh; the result can be *consistent* or *distributed* across MPI communicator
Implements an interface
Consumes objects of
Produces objects of
Abstract Interfaces
Python Objects

Figure 2.1: The components and interfaces of `nbodykit`. The main Python classes are Catalog, Mesh, and Algorithm objects, which are described in more detail in §2.2.3. Algorithm results can be *consistent*, where all processes hold the same data, or *distributed*, where data is spread out evenly across parallel processes.

We present some of the main classes and interfaces and how data flows through them in Figure 2.1. In the subsections to follow, we provide an overview of some of the components outlined in this figure.

Catalog

A Catalog is a Python object derived from a `CatalogSource` class that holds information about discrete objects[9] in a columnar format. Catalogs implement a *random-read interface*, which allows users to access arbitrary slices of the data while also taking advantage of the high throughput of a parallel file system. Often, users will initialize Catalog objects by reading data from a file on disk, using a `NumPy` array already stored in memory, or by generating simulated particles at runtime using one of `nbodykit`'s built-in classes.

Mesh

A Mesh is a Python object that computes a discrete representation of a continuous quantity on a uniform mesh. It is derived from a `MeshSource` class and provides a *paintable* interface, which refers to the process of "painting" the density field values onto the discrete mesh cells. When the user calls the `paint()` function, the mesh data is returned as a three-dimensional array. Mesh objects can be created directly from a Catalog via the `to_mesh()` function or by generating simulated fields directly on the mesh.

[9] Here, "object" can represent galaxies, simulation particles, mass elements, etc.

Algorithms

Algorithms are implemented as Python classes and interact with data by consuming Catalog and Mesh objects as input. The algorithm is executed when the user initializes the class, and the returned instance stores the results as attributes.

Serialization and Reproducibility

Most objects in `nbodykit` are serializable[10] via a `save()` function. Algorithm classes not only save the result of the algorithm but also save input parameters and meta-data. They typically implement both a `save()` and `load()` function, such that the algorithm result can be de-serialized into an object of the same type. The two main data containers, catalogs and meshes, can be serialized using `nbodykit`'s intrinsic format which relies on the massively parallel IO library `bigfile` (Feng 2017a). `nbodykit` includes support for reading these serialized results from disk back into Catalog or Mesh objects.

2.2.4 Parallelism

Data-based

`nbodykit` is fully parallelized using the Python bindings of the MPI standard available through `mpi4py`. The MPI standard allows processes running in parallel, each with their own memory, to exchange messages. This mechanism enables independent results to be computed by individual processes and then combined into a single result.

Both the Catalog and Mesh objects are *distributed* data containers, meaning that the data is spread out evenly across the available processes within an MPI communicator.[11] Nearly all algorithm calculations are performed on this distributed data, with final results computed via a reduce operation across all processes in the communicator. Rarely throughout the code base, data is instead gathered to a single root process, and operations are performed on this data before re-distributing the results to all processes. This only occurs when wall-clock time will not be a concern for most use cases and the additional complexity of a massively parallel implementation is not merited.

The distributed nature of the Catalog object is implemented by using the random-read interface to access different slices of the tabular data for different processes. The values of a Mesh object are stored internally on a three-dimensional `NumPy` array, which is distributed evenly across all processes. The domain of the 3D mesh is decomposed across parallel processes using the particle mesh library `pmesh`, which also provides an interface for computing parallel FFTs of the mesh data using `pfft-python`. The `pfft-python` software exhibits excellent scaling with the number of available processes, enabling high-resolution (large number

[10] Serialization (and its reverse, de-serialization) refers to the process of storing a Python object on disk in a format such that it can be reconstructed at a later time.

[11] The MPI communicator is responsible for managing the communication between a set of parallel processes.

of cells) mesh calculations.

Task-based

The analysis of LSS data often involves hundreds to thousands of repetitions of a single, less computationally expensive task. Examples include estimating the covariance matrix of a clustering statistic from a set of simulations and best-fit parameter estimation for a model. `nbodykit` implements a `TaskManager` utility to allow users to easily iterate over multiple tasks while executing in parallel. Users can specify the desired number of processes assigned to each task, and the `TaskManager` will iterate through the tasks, ensuring that all processes are being utilized.

2.3 Capabilities

In this section, we provide a more detailed overview of some of the main components of `nbodykit`. In particular, we describe how cosmology calculations are performed (§2.3.1), outline the available Catalog (§2.3.2) and Mesh (§2.3.3) classes, and provide details and references for the various algorithms currently implemented (§2.3.4).

2.3.1 Cosmology

The `nbodykit.cosmology` module includes functionality for representing cosmological parameter sets and computing various common theoretical quantities in LSS that depend on the background cosmological model. The underlying engine for these calculations is the `CLASS` Boltzmann solver (Blas et al. 2011; Lesgourgues 2011). We use the Python bindings of the `CLASS` C library provided by the `classylss` package. Comparing to the binding provided by the `CLASS` source code, `classylss` is a direct mapping of the `CLASS` object model to Python and integrates with the `NumPy` array protocol natively.

The main object in the module is the `Cosmology` class, which users can initialize by specifying a unique set of cosmological parameters (using the syntax of `CLASS`). This class represents the background cosmological model and contains methods to compute quantities that depend on the model. Most of the `CLASS` functionality is available through methods of the `Cosmology` object. Examples include distance as a function of redshift z, the Hubble parameter $H(z)$, the linear power spectrum, the nonlinear power spectrum, and the density and velocity transfer functions. Several `Cosmology` objects are provided for well-known parameter sets, including the *WMAP* 5, 7, and 9-year results (Komatsu et al. 2009, 2011; Hinshaw et al. 2013) and the *Planck* 2013 and 2015 results (Planck Collaboration et al. 2014, 2016a).

The `nbodykit.cosmology` module also includes classes to represent theoretical power spectra and correlation functions. The `LinearPower` class can compute the linear power spectrum as a function of redshift and wavenumber, using either the transfer function as computed by `CLASS` or the analytic approximations of Eisenstein & Hu (1999). The latter includes the so-called "no-wiggle" transfer function, which includes no BAO but the correct broadband features and is useful for quantifying the significance of potential BAO features. Similarly, we provide the `NonlinearPower` object to compute nonlinear power spectra, using the `Halofit` implementation in `CLASS` (Smith et al. 2003), which includes corrections from Takahashi et al. (2012). The `ZeldovichPower` class uses the linear power spectrum object to compute the power spectrum in the Zel'dovich approximation (tree-level Lagrangian perturbation theory). The implementation closely follows the appendices of Vlah et al. (2015) and relies on a Python implementation and generalization of the `FFTLog` algorithm[12] (Hamilton 2000). Finally, we also provide a `CorrelationFunction` object to compute theoretical correlation functions from any of the available power classes (using `FFTLog` to compute the Fourier transform).

We choose to use the `CLASS` software for the cosmological engine in `nbodykit` rather than the most likely alternative, the `astropy.cosmology` module. This allows `nbodykit` to leverage the full power of a Boltzmann solver for LSS calculations. We provide syntax compatibility between the `Cosmology` class and `astropy` when appropriate and provide functions to transform between the cosmology classes used by the two packages. However, we note that there are important differences between the two implementations. In particular, the treatment of massive neutrinos differs, with `astropy` using the approximations of Komatsu et al. (2011) rather than the direct calculations, as in `CLASS`.

2.3.2 Catalogs

In this section, we describe the two main ways that catalogs are created in `nbodykit`, as well as tools for cleaning and manipulating data stored in Catalog objects.

Reading Data from Disk

We provide support for loading data from disk into Catalog objects for several of the most common data storage formats in LSS data analysis. These formats include plaintext comma-separated value (CSV) data (via `pandas`, McKinney 2010), binary data stored in a columnar format, HDF5 data (via `h5py`, Collette & contributors 2017), FITS data (via `fitsio`, Sheldon 2017), and the `bigfile` data format. We also provide more specialized readers for particle data from the Tree-PM simulations of White (2002) and the legacy binary format of the GADGET simulations (Springel 2005). These Catalog objects use the `nbodykit.io` module, which includes several "file-like" classes for reading data from disk. These file-like objects implement a `read()` function that provides the random-read interface which returns a slice of the data for the requested columns. Users can easily support custom file formats by implementing their own subclass and `read()` interface.

Formats storing data on disk in a columnar format yield the best performance results, as the entirety of the data does not need to be parsed to yield the desired slice of the data on a given process. This is not true for the CSV storage format. We mitigate performance

[12] `https://github.com/eelregit/mcfit`

issues by implementing an enhanced version of the CSV parser in pandas that supports faster parallel random access. Our preferred IO format, bigfile, is massively parallel and stores data via a columnar format.

Finally, the Catalog object supports loading data from multiple files at once, providing a continuous view of the entirety of the data. This becomes particularly powerful when combined with the random-read interface, as arbitrary slices of the combined data can be accessed. For example, a single Catalog object can provide access to arbitrary slices of the output binary snapshots from an N-body simulation (stored over multiple files), often totaling 10-100 GB in size.

Generating Catalogs at Runtime

nbodykit includes several Catalog classes that generate simulated data at runtime. The simplest of these allows generating random columns of data in parallel using the numpy.random module. We also provide a UniformCatalog class that generates uniformly distributed particles in a box. These classes are useful for testing purposes, as well as for use as unclustered, synthetic data in clustering estimators.

nbodykit also includes functionality for generating more realistic approximations of large-scale structure. LogNormalCatalog generates a set of objects by Poisson sampling a log-normal density field and applies the ZelâĂŹdovich approximation to model nonlinear evolution (Coles & Jones 1991; Agrawal et al. 2017). The user can specify the input linear power spectrum and the desired output redshift of the catalog.

Catalog objects can also be created using the mock generation techniques of the Halotools software (Hearin et al. 2017) for populating halos with objects. Halotools includes functionality for populating halos via a wide range of techniques, including the halo occupation distribution (HOD), conditional luminosity function, and abundance matching methods. We refer the reader to Hearin et al. (2017) for further details. nbodykit supports using a generic Halotools model to populate a halo catalog. We also include built-in, specialized support for the HOD models of Zheng et al. (2007), Leauthaud et al. (2011), and Hearin et al. (2016).

Finally, the fastpm-python package implements an nbodykit Catalog object that generates particles via the FastPM approximate N-body simulation scheme (Feng et al. 2016). The FastPM library is massively parallel and exhibits excellent strong scaling with the number of available processes (see §2.6).

On-demand Data Cleaning

nbodykit uses the dask library (Team 2016) to represent the data columns of a Catalog object as dask array objects instead of using the more familiar NumPy array. The dask array has two key features that help users work interactively with data, and, in particular, large data sets. The first feature is *delayed evaluation*. When manipulating a dask array, operations are not evaluated immediately but instead stored in a task graph. Users can explicitly evaluate the dask array (returning a NumPy array) via a call to a compute()

function. Second, dask arrays are *chunked.* The array object is internally divided into many smaller arrays, and calculations are performed on these smaller “chunks.”

The delayed evaluation of dask arrays is particularly useful during the process of data cleaning, when users manipulate input data before feeding it into the analysis pipeline. Common examples of data cleaning include changing the coordinate system from galactic to Euclidean, converting between unit conventions, and applying masks. When using large data sets, the time to load the full data set into memory can be significant. This delay hinders data exploration and limits the interactive benefits of the Python language. dask arrays allow users to design data-cleaning pipelines on the fly. If the data format on disk supports random-read access, users can easily select and peek at a small subset of data without reading the full data set. This becomes especially useful when prototyping scientific models in an interactive environment, such as a Jupyter notebook.

The chunked nature of the dask array allows array computations to be performed on large data sets that do not fit into memory because the chunk size defines the amount of data loaded into memory at any given time. It effectively extends the maximum size of useable data sets from the size of memory to the size of the disk storage. This feature also simplifies the process of dealing with large data sets in interactive environments.

2.3.3 Meshes

Painting a Mesh

The Mesh object implements a paint() function, which is responsible for generating the field values on the mesh and returning an array-like object to the user. Meshes provide an equal treatment of configuration and Fourier space, and users can specify whether the painted array is defined in configuration or Fourier space. In the former case, a RealField is returned and in the latter, a ComplexField. These objects are implemented by the pmesh package and are subclasses of the NumPy ndarray class. They are related via a real-to-complex parallel FFT operation, implemented using pfft-python via the r2c() and c2r() functions.

The paint() function paints mass-weighted (or equivalently, number-weighted) quantities to the mesh. The field that is painted is

$$F(\boldsymbol{x}) = [1 + \delta'(\boldsymbol{x})]\, V(\boldsymbol{x}), \tag{2.1}$$

where $V(\boldsymbol{x})$ represents the field value painted to the mesh and $\delta'(\boldsymbol{x}) = n'(\boldsymbol{x})/\bar{n}' - 1$ is the weighted overdensity field. It is related to the unweighted number density as $n'(\boldsymbol{x}) = W(\boldsymbol{x})n(\boldsymbol{x})$, where $W(\boldsymbol{x})$ are the weights.

In nbodykit, users can control the behavior of both $V(\boldsymbol{x})$ and $W(\boldsymbol{x})$. In the default case, both quantities are unity, and the field painted to the mesh is $1 + \delta$. As an illustration, $V(\boldsymbol{x})$ can be specified as a velocity component to paint the momentum field (mass-weighted velocity). We also provide a mechanism by which users can further transform the painted field on the mesh. The apply() function can be used to apply a function to the mesh, either in configuration or Fourier space. Multiple functions can be applied to the mesh, and the operations are performed when `paint()` is called.

From Catalog to Mesh

All Catalog objects include a `to_mesh()` function which creates a Mesh object using the specified number of cells per mesh side. This function allows users to configure exactly how the catalog is interpolated onto the mesh. Users can choose from several different mass assignment windows, including the Cloud-In-Cell (CIC), Triangular Shaped Cloud (TSC), and Piecewise Cubic Spline (PCS) schemes (Hockney & Eastwood 1981). The Daubechies wavelet (Daubechies 1992) and its symmetric counterpart ("Symlets", see, e.g., `PyWavelets`[13]) are also available. By default, the CIC window is used. The interlacing technique (Hockney & Eastwood 1981; Sefusatti et al. 2016) can reduce the effects of aliasing in Fourier space. In this scheme, the Catalog object is interpolated onto two separate meshes separated by half of a cell size. When the fields are combined in Fourier space, the leading-order contribution to aliasing is eliminated.

Users can also configure whether or not the window is *compensated*, which divides the density field in Fourier space by (Hockney & Eastwood 1981)

$$W(\boldsymbol{k}) = \Pi_i \left[\mathrm{sinc}\left(\pi k_i / 2k_{\mathrm{N}}\right)\right]^p , \tag{2.2}$$

where $i \in \{x, y, z\}$, $p = 2, 3, 4$ for CIC, TSC, and PCS, respectively, and $\mathrm{sinc}(x) \equiv \sin(x)/x$. The Nyquist frequency of the mesh is given by $k_{\mathrm{N}} = \pi N/L$, where L is the box size, and N is the number of cells per box side.

We provide comparisons of the various interpolation windows and correction methods in this section. First, Figure 2.2 illustrates the effects of interlacing when using the CIC, TSC, and PCS schemes. This comparison is similar to the detailed analysis presented in Sefusatti et al. (2016). Second, we show the effectiveness of the wavelet windows at reducing aliasing in Figure 2.3. For both figures, we paint a `LogNormalCatalog` of 5×10^7 objects to a mesh of 512^3 cells in a box of side length 2500 h^{-1}Mpc. We compare the measured power spectrum to a "reference" power spectrum, computed using a mesh of 1024^3 cells and the PCS window. When using the CIC, TSC, and PCS windows, we de-convolve the interpolation window using equation 2.2, while we apply no such corrections when using wavelet-based windows.

Figure 2.2 confirms the results of Sefusatti et al. (2016)—the interlacing technique performs very well at reducing the effects of aliasing on the measured power spectrum. We achieve sub-percent accuracy up to the Nyquist frequency when combining interlacing with the CIC, TSC, and PCS windows. In general, higher-order windows perform better, with the PCS scheme achieving a precision of at least $\sim 10^{-5}$ up to the Nyquist frequency.

Figure 2.3 compares the performance of the Daubechies and Symlet wavelets to the CIC, TSC, and PCS windows. As in Figure 2.2, we plot the ratio of the power spectrum computed using meshes of size 512^3 and 1024^3 cells. We apply equation 2.2 for the CIC, TSC, and

[13] https://pywavelets.readthedocs.io

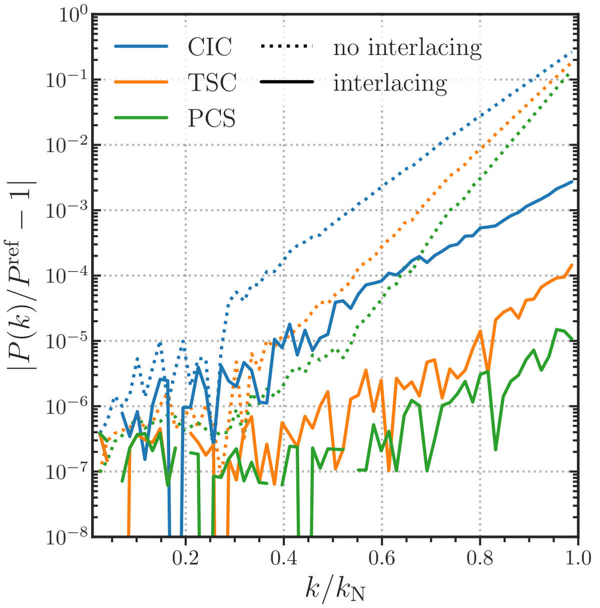


Figure 2.2: A comparison of the effects of interlacing when using the CIC, TSC, and PCS windows. We show the ratio of the power spectrum computed for a log-normal density field using a mesh with 512^3 cells to a reference power spectrum P^{ref}, computed using a mesh with 1024^3 cells. The ratio is shown as a function of wavenumber in units of the Nyquist frequency of the lower-resolution mesh. In all cases, the appropriate window compensation is performed using equation 2.2.

PCS windows but do not apply any corrections when using the wavelet windows. For this comparison, we do not use interlacing. We are able to confirm the results of Cui et al. (2008) and Yang et al. (2009), which claim 2% accuracy on the power spectrum up to $k \approx 0.7 k_{\mathrm{N}}$ when using the DB6 window without any additional corrections. However, the wavelet windows fail to match the precision achieved when using interlacing, even when using the largest wavelet size tested here ($a = 20$). Furthermore, the Daubechies windows introduce scale-dependence on large scales due to symmetry breaking (see the inset of Figure 2.3). The symmetric Symlet wavelets do not suffer from this issue but also cannot match the accuracy achieved when using interlacing.

Figure 2.3 also displays the relative speeds of each of the windows discussed in this section (bottom panel). These timing tests were performed using 64 processes on the NERSC Cori Phase I system. The wavelet windows are all significantly slower than the CIC, TSC, and PCS windows. The TSC and PCS methods are only marginally slower than the default CIC scheme, with slowdowns of ∼10% and ∼40%, respectively. The CIC, TSC, and PCS windows rely on optimized implementations in `pmesh`, while the wavelet windows use a slower lookup table implementation. Due to the precision of the interlacing technique and the relative speed of the TSC and PCS windows, we recommend using these options in most instances. However, it is generally best to determine the optimal set of parameters for a particular application by running convergence tests with different parameter configurations.

An Illustrative Example

We demonstrate the use of Mesh objects by example in Figure 2.4, which gives a short code snippet that creates a Mesh object from an existing Catalog, saves the configuration space density field to disk, and then reloads the data into memory. The snippet also demonstrates the `preview()` function, which can create a lower resolution projection of the full mesh field. This can be useful to quickly inspect mesh fields interactively, which would otherwise be difficult due to memory limitations. We show the preview of the density field from a log-normal catalog in the bottom panel of Figure 2.4, where the large-scale structure is clearly evident, even in the low-resolution projection.

2.3.4 Algorithms

The `nbodykit.algorithms` module includes parallel implementations of some of the most commonly used large-scale structure analysis algorithms. We take care to provide support for data sets from both observational surveys and N-body simulations. In this section, we provide an overview of the available functionality. The set of algorithms currently implemented is not meant to be exhaustive, but instead a solid foundation for LSS data analysis.

Power Spectra

For simulation boxes with periodic boundary conditions, the `FFTPower` algorithm measures the power directly from the square of the Fourier modes of the overdensity field. The 1D or 2D power spectrum, $P(k)$ or $P(k,\mu)$, can be computed, as well as the power spectrum multipoles $P_\ell(k)$. Here, μ represents the angle cosine between the pair separation vector and the line-of-sight. For observational data, in the form of right ascension (RA), declination (Dec), and redshift, the power spectrum multipoles of the density field can be computed using the `ConvolvedFFTPower` algorithm. The implementation uses the FFT-based estimator described in Hand et al. (2017b), which requires $2\ell + 1$ FFTs to compute a given multipole of order ℓ. This estimator improves the FFT-based estimator presented by Bianchi et al. (2015) and Scoccimarro (2015), building on the ideas of previous power spectrum estimators (Feldman et al. 1994; Yamamoto et al. 2006), and in particular, the treatment of the anisotropic 2PCF using spherical harmonics of Slepian & Eisenstein (2015a). We also provide the `ProjectedFFTPower` for computing the power spectrum of a field in a simulation box, projected along the specified axes. Such an observable is useful for e.g., Lyman-α or weak lensing data analysis. The correctness of these algorithms has been verified using independent implementations from within the Baryon Oscillation Spectroscopic Survey (BOSS) collaboration.

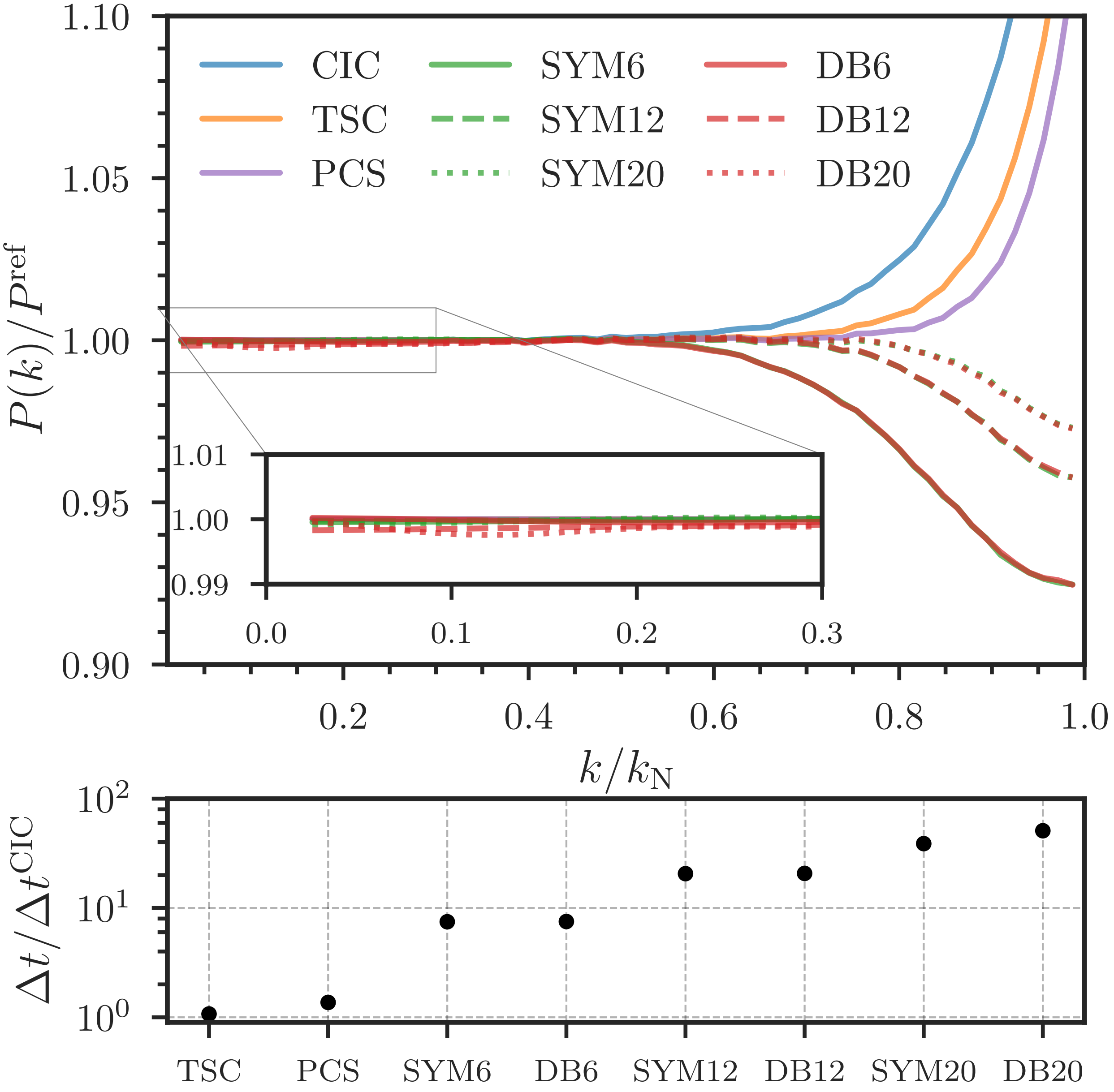


Figure 2.3: The performance of the Daubechies and Symlet wavelets in comparison to the CIC, TSC, and PCS windows. Wavelet windows of sizes $a = 6$, 12, and 20 are shown. *Top*: the ratio of the measured power to the reference power spectrum, as in Figure 2.2. Here, we apply no corrections when using the wavelet windows and apply equation 2.2 for the CIC, TSC, and PCS windows. No interlacing is used for this test. *Bottom*: the speed of each interpolation window, relative to the CIC window. Speeds were recorded when computing the power spectra in the top panel.

```
from nbodykit.lab import *
import matplotlib.pyplot as plt

# Initialize linear power spectrum with Planck 2015 cosmology
cosmo = cosmology.Planck15
Plin = cosmology.LinearPower(cosmo, redshift=0)

# Create a Catalog by sampling a log-normal density field
cat = LogNormalCatalog(Plin, nbar=3e-3, BoxSize=1380, Nmesh=256)

# Convert to a Mesh and use TSC painting
mesh = cat.to_mesh(Nmesh=256, window="tsc")

# Save the configuration-space Mesh
mesh.save("lognormal-mesh.bigfile", mode="real", dataset="Field")

# Preview a low-resolution projection of the density field
density = mesh.preview(Nmesh=64, axes=(0,1))
plt.imshow(density)

...

# Reload the Mesh from disk
mesh = BigFileMesh("lognormal-mesh.bigfile", dataset="Field")
```

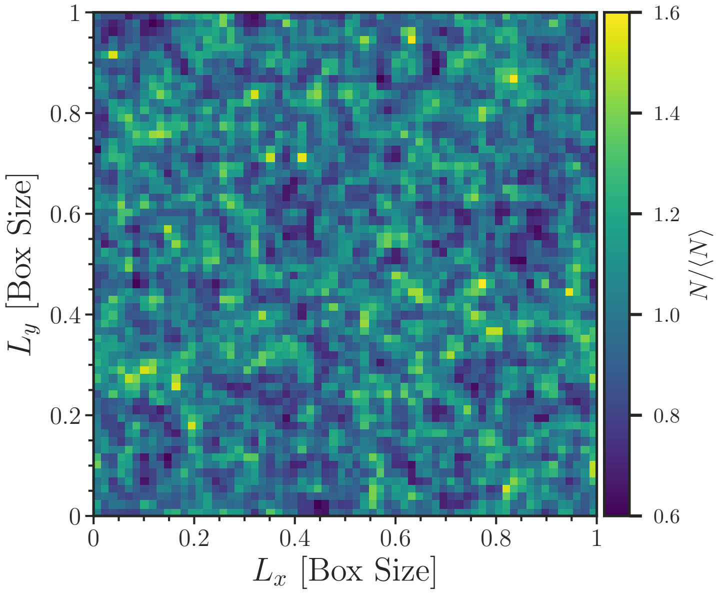


Figure 2.4: *Top*: an analysis pipeline illustrating the creation of a Mesh object from a Catalog, as well as how to serialize the painted mesh to disk and preview a low-resolution projection of the density field for inspection. *Bottom*: the two-dimensional, low-resolution preview of the painted density field $N/\langle N \rangle = 1 + \delta$.

2-Point Correlation Functions

`nbodykit` includes functionality for counting pairs of objects and computing their correlation function in configuration space. We leverage the blazing speed of the publicly available `Corrfunc` chaining mesh code for these calculations (Sinha & Garrison 2017). We adapt its highly optimized pair counting routines to perform calculations using MPI. We perform a domain decomposition on the input data such that the objects on a particular MPI rank are spatially confined to include all pairs within the maximum separation. For non-uniform density fields, the domain decomposition results in a particle load that is balanced across MPI ranks.[14] The relevant pair counting algorithms are `SimulationBoxPairCount` and `SurveyDataPairCount`. These classes can count pairs of objects as a function of the 3D separation r, the separation r and angle to the line-of-sight μ, the angular separation θ, and the projected distances perpendicular r_p and parallel π to the line-of-sight.

Users can compute the correlation function of a Catalog using the `SimulationBox2PCF` and `SurveyData2PCF` classes, which internally rely on the previously described pair counting classes. For data with periodic boundary conditions, we use analytic randoms to estimate the correlation function using the so-called "natural" estimator: $DD/RR - 1$. A Catalog object holding synthetic randoms can be supplied, in which case the Landy-Szalay estimator (Landy & Szalay 1993) is employed: $(DD - 2DR + RR)/RR$. The variations of the correlation function that can be computed by these two classes are as follows:

- as a function of three-dimensional separation, $\xi(r)$
- accounting for the angle to the line-of-sight, $\xi(r, \mu)$ and $\xi(r_p, \pi)$
- as a function of angular separation, $w(\theta)$
- projected over the line-of-sight separations, $w_p(r_p)$

The correctness of the pair counting and correlation function algorithms described here was independently verified using the `kdcount` and `Halotools` software.

3-Point Correlation Function

The `SimulationBox3PCF` and `SurveyData3PCF` classes compute the multipoles of the isotropic 3-point correlation function (3PCF) in configuration space. The algorithm follows the implementation described in Slepian & Eisenstein (2015a), which scales as (N^2), where N is the number of objects. Their improved estimator relies on a spherical harmonic decomposition to achieve a similar scaling with N as two-point clustering estimators. We note that the FFT-based implementation of this algorithm (presented in Slepian & Eisenstein 2016) and the anisotropic version described in Slepian & Eisenstein (2017) have not yet been implemented, although there are plans to do so in the future. We have verified the accuracy of the isotropic 3PCF classes against the implementation used in Slepian & Eisenstein (2015a).

[14] We thank Biwei Dai for the implementation of the load balancer.

An implementation of this algorithm including anisotropy written in C++ and optimized for HPC machines was recently presented in Friesen et al. (2017).

Grouping Methods

The FOF class implements the well-known Friends-of-Friends algorithm, which identifies clusters of points that are spatially less distant than a threshold linking length. It uses a parallel implementation of the algorithm described in Feng & Modi (2017), which utilizes KD-trees and the kdcount software. FOF groups can be identified as a function of three-dimensional or angular separation. We also provide functions for transforming the output of the FOF algorithm to a Catalog of halo objects (a HaloCatalog) in a manner compatible with the Halotools software.

nbodykit can also identify clusters of objects using a cylindrical rather than spherical geometry. We implement a parallel version of the algorithm described in Okumura et al. (2017) in the CylindricalGroups class. Our implementation relies on the neighbor querying capability of kdcount and the group-by methods of pandas.

Finally, the FiberCollisions class simulates the process of assigning spectroscopic fibers to objects in a fiber-fed redshift survey such as BOSS or eBOSS (Dawson et al. 2013, 2016). This procedure results in so-called “fiber collisions” when two objects are separated by an angular width on the sky that is smaller than the fiber size. We follow the procedure outlined in Guo et al. (2012) to assign fibers to an input catalog of objects. We identify angular FOF groups using a linking length equal to the fiber collision scale and assign fibers to the objects in such a manner as to minimize the number of objects that do not receive a fiber.

Miscellaneous

nbodykit also includes algorithms that generally serve a supporting role in other algorithms. The KDDensity class estimates a proxy density quantity for an input set of objects using the inverse cube of the distance to an object’s nearest neighbor. The RedshiftHistogram class computes the mean number density as a function of redshift, $n(z)$, from an input catalog of objects. We plan to generalize this algorithm to be a more universal histogram calculator that could, for example, compute mass or luminosity functions.

2.4 Development Workflow

2.4.1 Version Control

nbodykit is developed using the version control features of git,[15] and the code is hosted in a public repository on GitHub.[16] Major changes to the code base are performed using a *pull request* workflow, which provides a mechanism for developers to review changes before

[15] http://git-scm.com

[16] http://github.com/bccp/nbodykit

they are merged into the main source code. Users can contribute to `nbodykit` by first *forking* the main repository, making changes in this fork and submitting the changes to the main repository via a pull request. This workflow helps assure the overall quality of the code base and ensures that new changes are properly documented and tested. Bugs and new feature requests can be submitted as GitHub issues. Alternatively, users can send an email to `nbodykit-issues@fire.fundersclub.com`, which will automatically open an issue on GitHub. As `nbodykit` is intended as a community-based resource, we encourage user contributions and ideas for new functionality. We adopt a "mentoring" approach for new features and will gladly offer advice and guidance to new users who wish to contribute to `nbodykit` for the first time.

2.4.2 Automated Testing with MPI Support

`nbodykit` is extensively tested via hundreds of unit tests using the `runtests`[17] package (Feng & Hand 2017). As `mpi4py` does not provide a reusable framework for testing parallel applications, we have developed `runtests` to fill this gap in the development process. It extends the `py.test`[18] testing framework, adding several features. First, the test driver incrementally rebuilds and installs the Python package before running the test suite. Second, it adds MPI support by allowing users to specify the number of processes with which each test function should be executed. It also supports computing the testing coverage for parallel applications, where test coverage is defined as the percentage of the software covered by the test suite.

We execute the `nbodykit` test suite via the continuous integration (CI) service Travis,[19] using `runtests` to test both serial and parallel execution of the code. The test suite is currently executed on both Linux and Mac OS X operating systems and for Python versions 2.7, 3.5, and 3.6. Whenever a pull request is opened, the test suite is executed and the new changes will not be merged if the test suite fails. We also compute the testing coverage of the code base. Currently, `nbodykit` maintains a value of 95%. We use the Coveralls[20] service to ensure that new changes cannot be merged into the main repository if the testing coverage decreases.

2.4.3 Use on Personal and HPC Machines

`nbodykit` is compatible with both Python versions 2.7 and 3.x. For personal computing systems (Mac OS X and Linux), we provide binaries of `nbodykit` and its dependencies on the Berkeley Center for Cosmological Physics (BCCP) Anaconda channel.[21] `nbodykit` (and all of its dependencies) can be installed into an Anaconda environment using a simple command:

[17] https://github.com/rainwoodman/runtests

[18] http://pytest.org

[19] https://travis-ci.org

[20] https://coveralls.io

[21] https://anaconda.org/bccp

conda install -c bccp nbodykit. We ensure all packages on the BCCP channel are up-to-date using a nightly cron job hosted on the Travis CI service.

Supercomputing systems often require recompiling the dependencies of nbodykit using the machine-specific compilers and MPI configuration. For example, we use the "conda build" functionality of the Anaconda package to compile and update nbodykit and its dependencies nightly on the NERSC Cray supercomputers. The infrastructure for building nbodykit and its dependencies is publicly available on GitHub,[22] which users can re-use to setup nbodykit on HPC machines other than NERSC. However, we recommend that users first test if the default binaries on the BCCP channel are compatible with their supercomputing environment.

The remaining barrier to using nbodykit on HPC systems is the incompatibility of the Python launch system and the shared file systems of HPC machines. When launching an MPI application using Python, the file system will stall when all of the Python instances (can be thousands or more) query the file system for modules on the search path. This issue effectively prevents the use of Python applications on HPC machines.

nbodykit utilizes an open-source solution, denoted "python-mpi-bcast", to facilitate deploying Python applications on HPC machines (Feng & Hand 2016). This tool bundles and delivers runtime dependencies to the HPC computing nodes via an MPI broadcast operation, bypassing the file system bottleneck and allowing Python applications to launch at near-native speed. Users can modify their job scripts in a non-invasive manner to deploy our tool. Additional details and setup instructions can be found in Feng & Hand (2016). The tool is publicly available on GitHub.[23]

2.4.4 Documentation

Documentation for nbodykit is available on Read the Docs.[24] The documentation is generated using Sphinx[25] and includes comprehensive documentation of the nbodykit API. It also includes detailed walk-throughs of each of the main components of nbodykit.

We provide a set of recipes detailing a broad selection of the functionality available in nbodykit in the "Cookbook" section of the documentation. Ranging from simple tasks to more complex work flows, we hope that these recipes help users become acclimated to nbodykit as well as illustrate the power of nbodykit for LSS data analysis. The recipes are in the form of Jupyter notebooks. An interactive environment containing the recipe notebooks is available to users via the Binder service.[26] This allows new users to explore nbodykit without installing nbodykit on their own machine.

[22] https://github.com/bccp/conda-channel-bccp

[23] https://github.com/rainwoodman/python-mpi-bcast

[24] http://nbodykit.readthedocs.io

[25] http://www.sphinx-doc.org

[26] https://mybinder.org

```
from nbodykit.lab import *
from nbodykit import setup_logging
from fastpm.nbkit import FastPMCatalogSource

setup_logging()

# Setup initial conditions
cosmo = cosmology.Planck15
power = cosmology.LinearPower(cosmo, 0)
linear = LinearMesh(power, BoxSize=512, Nmesh=512)

# P(k) of initial field
r = FFTPower(linear, mode="1d")
r.save("linear-power.json")

# Run the FastPM particle mesh simulation
matter = FastPMCatalogSource(linear, Nsteps=10)

# Compute and save matter P(k,z=0)
r = FFTPower(matter, mode="1d", Nmesh=512)
r.save("matter-power.json")

# Run FOF to identify halo groups
fof = FOF(matter, linking_length=0.2, nmin=20)
halos = fof.to_halos(1e12, cosmo, 0.)

# Compute and save halo power P(k,z=0)
r = FFTPower(halos, mode="1d", Nmesh=512)
r.save("halo-power.json")

# Populate halos with galaxies
hod = halos.populate(Zheng07Model)

# Compute and save galaxy P(k,z=0)
r = FFTPower(hod, mode="1d", Nmesh=512)
r.save("galaxy-power.json")
```

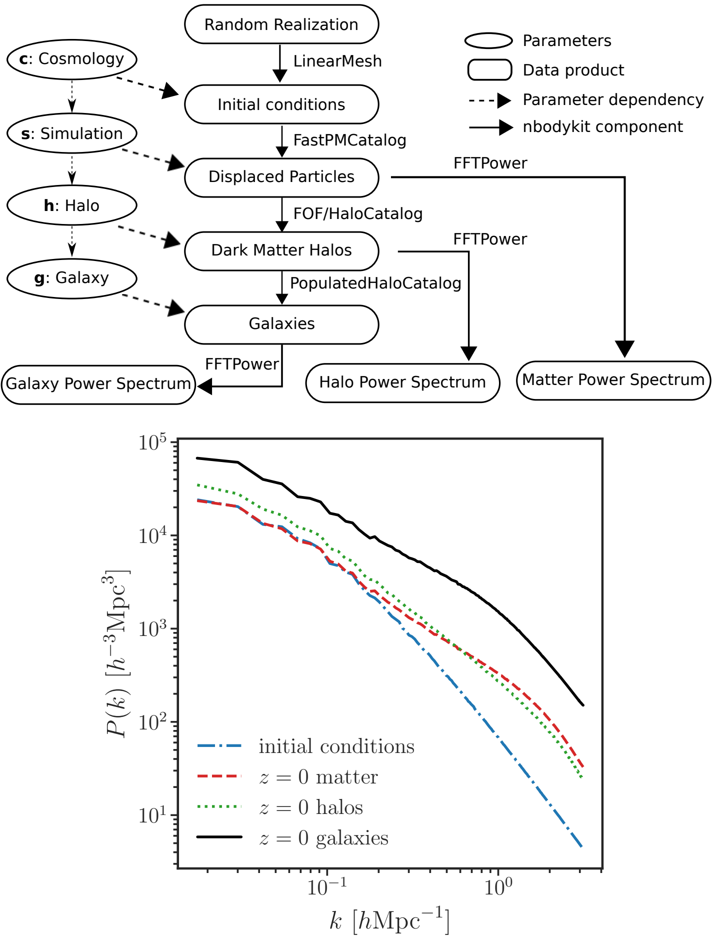


Figure 2.5: A galaxy clustering emulator, implemented with **nbodykit**. *Left*: the source code for the application, which evolves an initial Gaussian field to $z = 0$ using the FastPM simulation scheme, identifies FOF halos, populates those halos with galaxies, and records the power spectrum of each step. *Right, top*: the flow of data through the various components. *Right, bottom*: the resulting $P(k)$ measured for each step in the emulator. Performance benchmarks for this application are given in Figure 2.7.

2.5 In Action

In this section, we describe a realistic LSS application using `nbodykit`: a galaxy clustering emulator. The goal of the emulator is to produce the galaxy power spectrum from first principles, given a background cosmological model. The application combines several components of `nbodykit` to achieve this goal. The steps include:

- Initial conditions: the `LinearMesh` class creates a Gaussian realization of a density field in Fourier space from an input power spectrum.
- N-body simulation: the initial conditions are evolved forward to $z = 0$ using the FastPM quasi-N-body particle mesh scheme of Feng et al. (2016).
- Halo Identification: halos are identified from the matter field using the `FOF` grouping algorithm.
- Halo Population: halos are populated with galaxies using the HOD from Zheng et al. (2007) and the `Halotools` package.
- Clustering Estimation: $P(k)$ is computed for each of the above steps using the `FFTPower` algorithm.

We diagram the flow of data and parameters for these steps in the top right panel of Figure 2.5. We also show the source code for the application using `nbodykit`, which can be implemented using only ∼30 lines of code. We emphasize that with the component-based approach of `nbodykit`, the user is free to output and serialize any intermediate data products during the execution of the larger application, as we have done in this example for the power spectra of the initial, matter, and halo density fields. Finally, note that the source code in Figure 2.5 can be executed with an arbitrary number of MPI ranks. We discuss performance benchmarks for this application as a function of the number of MPI processes in the next section.

2.6 Performance Benchmarks

In this section, we present performance benchmarks for several `nbodykit` algorithms, as well as the emulator application discussed in Section 2.5. Tests are run on the NERSC Cori Phase I Haswell nodes, with 32 MPI cores per node. In Figure 2.6, we show the strong scaling results for the `FFTPower`, `ConvolvedFFTPower`, `SimulationBoxPairCount`, and `SimulationBox3PCF` algorithms. The benchmarks are performed for two different data configurations, meant to simulate the data sets of current and future surveys, denoted as "small" and "large", respectively. The "small" sample is modeled after the completed BOSS galaxy sample (Reid et al. 2016) and includes 10^6 galaxies in a cubic box of side length $L = 2500\ h^{-1}$Mpc. The "large" sample includes a factor of 10 more objects in a box of side

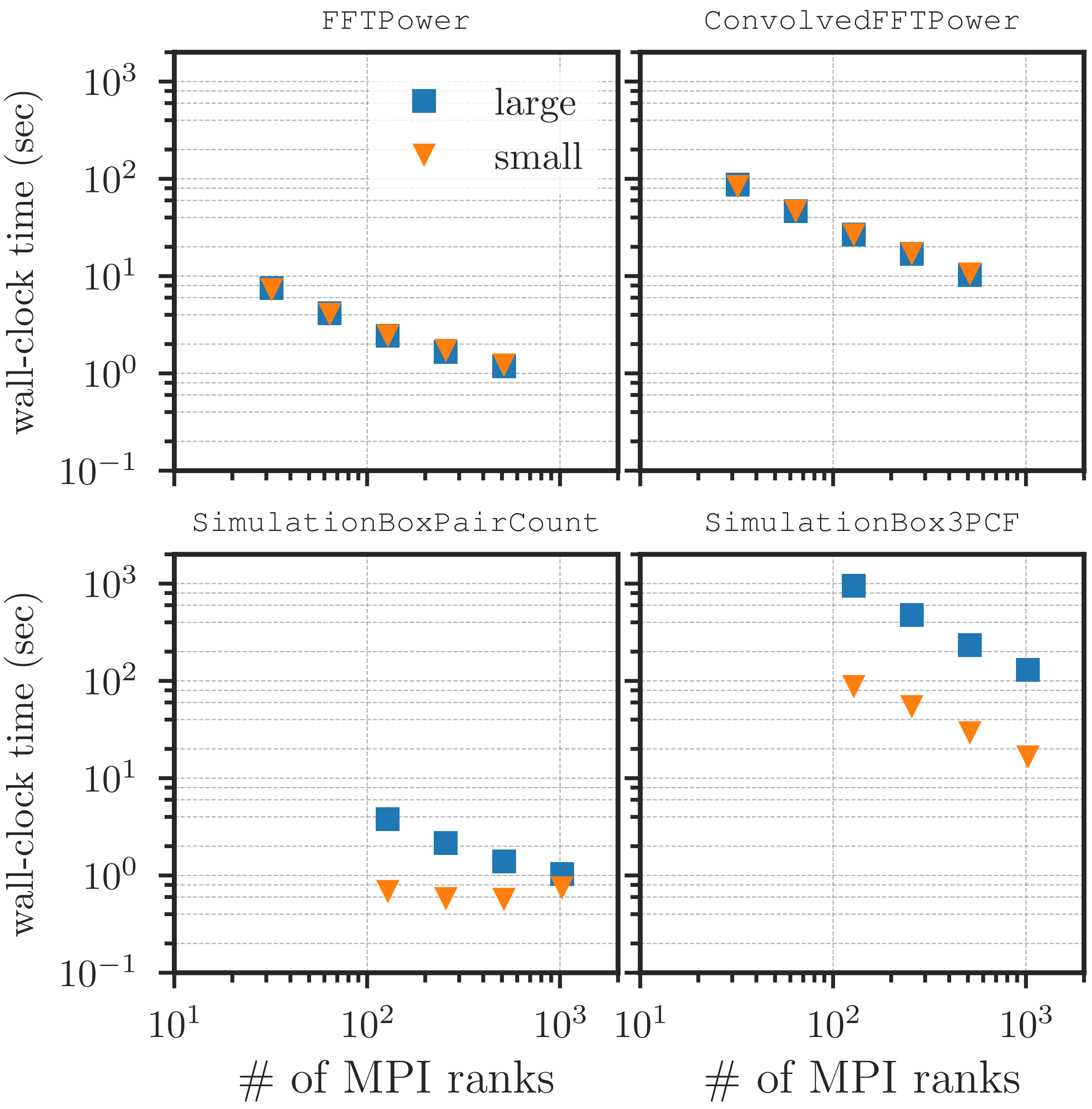


Figure 2.6: Performance benchmarks for four **nbodykit** algorithms for our "small" data set (10^6 objects) and our "large" data set (10^7 objects). The algorithms in the top row use FFT-based estimators to compute power spectra, while those in the bottom row of panels count pairs of objects in configuration space. The FFT-based algorithms take near-identical time for the large and small data sets due to the use of a 1024^3 mesh in both cases. The benchmarks were performed on the NERSC Cori Phase I Haswell nodes using 32 MPI ranks per node. See the text of Section 2.6 for further details on the test configurations.

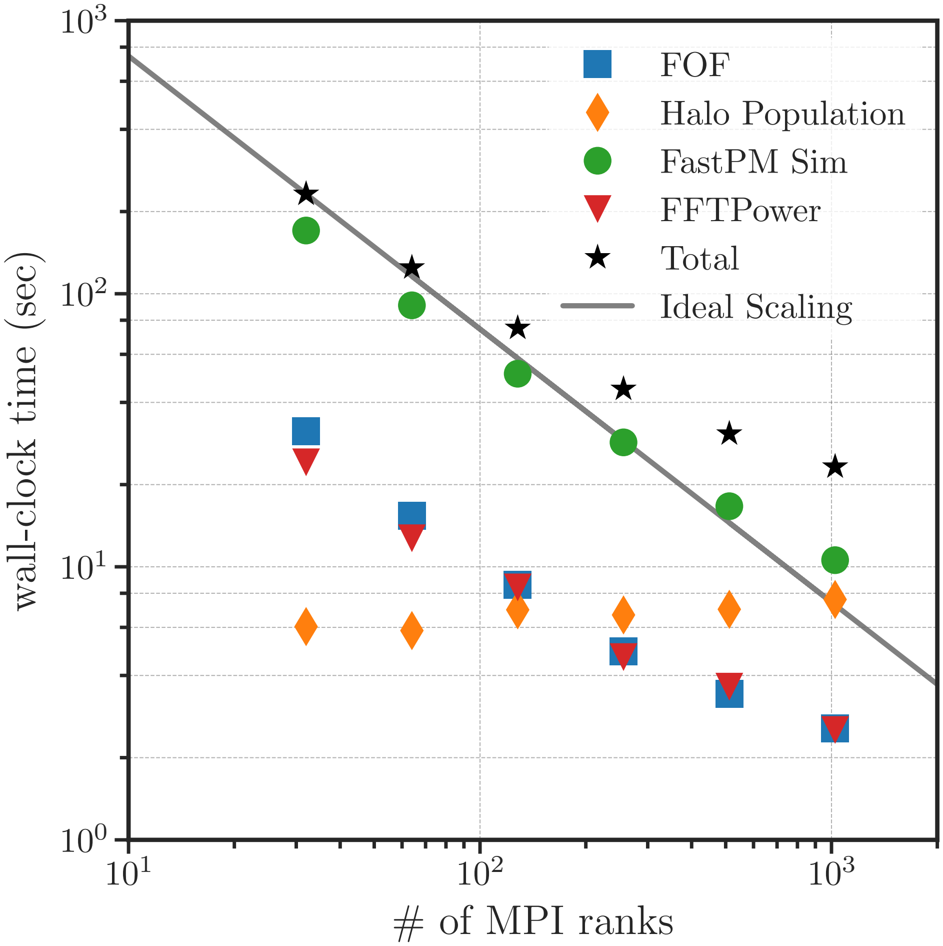


Figure 2.7: The wall-clock time as a function of the number of MPI ranks used for each step in the galaxy clustering emulator detailed in Figure 2.5. Overall, the application shows excellent scaling behavior, with deviations from the ideal scaling caused by the halo population step. This step does not currently have a massively parallel implementation and takes a roughly constant amount of time as more cores are used. The benchmarks were performed on the NERSC Cori Phase I Haswell nodes using 32 MPI ranks per node.

length $L = 5000\ h^{-1}$Mpc and is meant to represent data from future surveys such as DESI (DESI Collaboration et al. 2016a). We run four sets of benchmarking tests:

- `FFTPower`: compute $P(k, \mu)$ for 10 μ bins, using a mesh size of $N_{\mathrm{mesh}} = 1024$. This requires a single FFT operation.
- `ConvolvedFFTPower`: compute multipoles $P_\ell(k)$ for $\ell = 0$, 2, and 4 for survey data (RA, Dec, z), using a mesh size of $N_{\mathrm{mesh}} = 1024$. The algorithm requires $2\ell + 1$ FFT operations per multipole, and 15 in total for this test.
- `SimulationBoxPairCount`: count the number of pairs as a function of separation for 10 separation bins ranging from $r = 10\ h^{-1}$Mpc to $r = 150\ h^{-1}$Mpc and 100 μ bins.
- `SimulationBox3PCF`: compute the isotropic 3PCF multipoles for $\ell = 0, 1, ..., 10$ and 10 separation bins ranging from $r = 10\ h^{-1}$Mpc to $r = 150\ h^{-1}$Mpc.

In general, these four algorithms show excellent strong scaling with the number of MPI ranks. For the power spectrum algorithms (top row of Figure 2.6), the dominant calculation is the FFT operation, which has good scaling behavior. Because the FFT is the dominant time cost, we find nearly identical performances for the "small" and "large" samples. The wall-clock time for the `ConvolvedFFTPower` algorithm is roughly fifteen times that of the `FFTPower` algorithm, which is driven by the total number of FFTs that each algorithm computes. The pair-counting-based algorithms both take $\mathcal{O}(N^2)$ time to compute their results. However, the `SimulationBoxPairCount` algorithm relies on the highly optimized `Corrfunc` software, which is significantly faster than `SimulationBox3PCF`, which relies on `kdcount`. When using `SimulationBoxPairCount` on the "small" data set, we find that MPI communication costs are non-negligible due to the relatively small sample size, which hinders the scaling performance of the code.

We also present performance benchmarks for the emulator application described in Section 2.5. For this test, we run a FastPM particle mesh simulation with 512^3 total particles. The halo finder identifies roughly 225,000 dark matter halos that are then used to build a mock galaxy catalog. The wall-clock times for each step in the emulator are shown in Figure 2.7. We see that the dominant fraction of the wall-clock time is spent in the FastPM step, which shows excellent strong scaling behavior up to the number of cores we have tested. The implementation of the halo population step using `Halotools` is not massively parallel, and therefore, the time to solution for this step remains relatively constant as the number of cores increases. The wall-clock time for this step only becomes significant as the number of cores approaches ∼1024, and it would be worth investigating improving this step's scaling if users wish to run often at this scale. However, in our experience, we have not found that the time cost of this step justifies further efforts converting it to a massively parallel implementation.

We emphasize that for all benchmarks presented in this section, the number of MPI ranks can always be increased such that the time to solution is on the order of seconds.

This becomes important for realistic data analyses in LSS, which often require repeating an algorithm's calculation hundreds to thousands of times, e.g., while sampling a parameter space using Markov Chain Monte Carlo or optimization techniques. Due to the availability of large-scale computing resources and the scaling behavior of `nbodykit` demonstrated here, we believe that `nbodykit` will be able to meet the computational demands of future LSS data analyses.

2.7 Conclusions

We have presented the first public release of `nbodykit` (v0.3.0), a massively parallel Python toolkit for the analysis of large-scale structure data. Relying on the `mpi4py` binding of MPI, the package includes parallel implementations of a set of canonical algorithms in the field of large-scale structure, including two and three-point clustering estimators, halo identification and population tools, and quasi-N-body simulation schemes. The toolkit also includes a set of distributed data containers that support a variety of data formats common in astronomy, including CSV, FITS, HDF5, binary, and `bigfile` data. With these tools, we hope `nbodykit` can serve as a foundation for the community to build upon as we strive to meet the demands of future LSS data sets.

In designing `nbodykit`, we have attempted to balance the requirements of both a scalable and interactive piece of software. Our ultimate goal was to produce a piece of software that is as usable in a Jupyter notebook environment as on an HPC machine. We have adopted a modular, component-based approach that should enable researchers to integrate `nbodykit` with their own software to build complicated applications. As an illustration of its power, we have discussed an implementation of a galaxy clustering emulator using `nbodykit`, which provides a complete forward model for the galaxy power spectrum, starting from an initial, Gaussian density field. We have also demonstrated that the toolkit shows excellent scaling behavior, presenting a set of performance benchmarks for the emulator as well as some of the more commonly used algorithms in `nbodykit`.

We have outlined our development workflow for producing a piece of reusable scientific software. `nbodykit` is open-source—we strongly believe in the idea of open science and have placed an emphasis on reproducibility when designing `nbodykit`. Designed for the LSS community, we hope that new users will find `nbodykit` useful in their own research and that the software can continue to grow and mature in new ways from community feedback and contributions. We are also strong believers in the necessity of unit testing and adequate documentation for open-source tools. We have attempted to meet these goals using the Travis automated testing service and the Read the Docs documentation hosting tools. Finally, we have aimed to make `nbodykit` widely available and easily installable. The package supports both Python versions 2 and 3, and binary distributions of `nbodykit` and its dependencies can be installed onto Mac OS X and Linux machines using the Anaconda package manager.

In the future, we hope to incorporate the expertise of new developers, from both the LSS and Python HPC communities. We expect the knowledge of both communities to be necessary in the data analysis of future surveys. The set of features currently implemented in nbodykit is not meant to be all-inclusive but rather a sampling of the more commonly used tools in the field. Most importantly, we hope that nbodykit provides a solid basis for the community to build upon. We warmly welcome feedback and contributions of all forms from the community. As open-source software, nbodykit was designed as a tool to help the LSS community, and we hope that community contributions can help maximize its benefits for its users.

Chapter 3

An optimal FFT-based anisotropic power spectrum estimator

Measurements of line-of-sight dependent clustering via the galaxy power spectrum's multipole moments constitute a powerful tool for testing theoretical models in large-scale structure. Recent work shows that this measurement, including a moving line-of-sight, can be accelerated using fast Fourier transforms (FFTs) by decomposing the Legendre polynomials into products of Cartesian vectors. In this chapter, we present a faster, optimal means of using FFTs for this measurement. We avoid redundancy present in the Cartesian decomposition by using a spherical harmonic decomposition of the Legendre polynomials. With this method, a given multipole of order ℓ requires only $2\ell+1$ FFTs rather than the $(\ell+1)(\ell+2)/2$ FFTs of the Cartesian approach. For the hexadecapole ($\ell = 4$), this translates to 40% fewer FFTs, with increased savings for higher ℓ. The reduction in wall-clock time enables the calculation of finely-binned wedges in $P(k, \mu)$, obtained by computing multipoles up to a large ℓ_{max} and combining them. This transformation has a number of advantages. We demonstrate that by using non-uniform bins in μ, we can isolate plane-of-sky (angular) systematics to a narrow bin at $\mu \simeq 0$ while eliminating the contamination from all other bins. We also show that the covariance matrix of clustering wedges binned uniformly in μ becomes ill-conditioned when combining multipoles up to large values of ℓ_{max}, but that the problem can be avoided with non-uniform binning. As an example, we present results using $\ell_{\mathrm{max}} = 16$, for which our procedure requires a factor of 3.4 fewer FFTs than the Cartesian method, while removing the first μ bin leads only to a 7% increase in statistical error on $f\sigma_8$, as compared to a 54% increase with $\ell_{\mathrm{max}} = 4$.

3.1 Introduction

The clustering of galaxies on the largest scales contains a significant amount of cosmological information. The baryon acoustic oscillation (BAO) feature on scales of $\sim$100 h^{-1}Mpc can be used as a standard ruler to gauge the Universe's expansion history and infer properties of dark energy (e.g., Wagner et al. 2008; Shoji et al. 2009). First detected in the 2-point correlation function (2PCF) by Eisenstein et al. (2005); Cole et al. (2005) and more recently in the 3-point function (3PCF) by Slepian et al. (2017), the BAO signal has provided percent-level measurements of the Hubble parameter $H(z)$ and angular diameter distance $D_A(z)$ (Alam et al. 2017). These analyses have measured both the characteristic BAO peak in configuration space (Ross et al. 2017; Vargas-Magaña et al. 2017) and the analogous wiggles in Fourier space (Beutler et al. 2017c; Gil-Marín et al. 2016a). Beyond the BAO, and on even larger scales, these clustering statistics also contain signatures of primordial non-Gaussianity, the deviation from Gaussian random field initial conditions in the early Universe (Creminelli et al. 2006; Desjacques & Seljak 2010).

Additional information can be extracted from these statistics by measuring the broadband clustering as a function of the angle to the line-of-sight (LOS). Although the underlying distribution of galaxies is assumed to be homogeneous and isotropic, observational effects such as the Alcock-Paczynski (AP; Alcock & Paczynski 1979) effect and redshift-space distortions (RSD; Kaiser 1987) introduce anisotropy into the measured clustering when a fiducial distance-redshift relation is used to translate redshifts into comoving coordinates. In particular, anisotropy around the line-of-sight is introduced by RSD because an object's redshift, used to infer the LOS coordinate, is sensitive to its peculiar velocity. Because this velocity is sourced by the large-scale gravity field, RSD measurements allow constraints on the growth rate of structure and can provide tests of general relativity (e.g., Guzzo et al. 2008). For galaxy pairs, RSD depends on the angle cosine μ between the pair separation $\boldsymbol{s}$ and the line-of-sight $\hat{\boldsymbol{n}}$. The clustering is typically measured as multipole moments of the 2-point correlation function, which gives the excess of pairs above random, or of the power spectrum, its Fourier-space analog. The Legendre polynomials form a complete basis and are equivalent to expanding in powers of μ. Parity demands that only even multipoles are non-zero. In linear theory, RSD generates only $\ell = 0, 2$, and 4 moments of the anisotropic power spectrum (Kaiser 1987) or correlation function (Hamilton 1992).

For wide-field galaxy surveys, only angle-averaged clustering, i.e., the monopole, can be measured accurately under the assumption of a single LOS to the entire survey. Under this assumption, it is straightforward to measure the clustering using a fast Fourier transform (FFT). What is more challenging is to define a clustering estimator for the higher-order multipoles that uses a line-of-sight that rotates to follow each galaxy pair's spatial or Fourier-space separation. Including the observer as a third vertex, the galaxy pair maps to a triangle, and more accurate line-of-sight choices are the angle bisector of this triangle or the vector from the observer to the separation midpoint. Less accurate but still better than a single LOS is taking the LOS to be the vector from observer to a single pair member, as first used in Yamamoto et al. (2006); Blake et al. (2011b). This latter method, often referred to as the local plane-parallel approximation, differs from angle bisector and midpoint methods at $\mathcal{O}(\theta^2)$, where θ is the angle the pair subtends; bisector and midpoint methods also differ from each other at $\mathcal{O}(\theta^2)$ (Slepian & Eisenstein 2015b). For the current generation of surveys, these wide-angle effects are not a significant source of error (Samushia et al. 2015; Yoo & Seljak 2015) but could become important for future surveys, especially for studies which focus primarily on large scales, i.e., primordial non-Gaussianity analyses. To address this, slight generalizations of the local plane parallel estimates for the multipoles can be combined to form the midpoint and bisector-based estimates (Slepian & Eisenstein 2015b).

Recently, Bianchi et al. (2015) and Scoccimarro (2015) showed that by using products of Cartesian coordinates as building blocks for the Legendre polynomials, one could evaluate the local plane-parallel method of Yamamoto et al. (2006) using FFTs, providing an enormous speed-up over the summation-based estimator. Around the same time, Slepian & Eisenstein (2016) demonstrated that FFTs could also be used for the anisotropic 2PCF by exploiting the spherical harmonic addition theorem to decompose the Legendre polynomials into spherical harmonics. In this chapter, we show that this spherical harmonic approach can also be used for the power spectrum multipoles. Importantly, the spherical harmonics are orthogonal to each other, whereas the Cartesian vectors used by Bianchi et al. (2015); Scoccimarro (2015) are not. Thus, the Bianchi et al. (2015); Scoccimarro (2015) implementation involves redundancy, requiring $(\ell+1)(\ell+2)/2 = \mathcal{O}(\ell^2)$ FFTs per multipole rather than the $2\ell+1 = \mathcal{O}(\ell)$ FFTs needed by our method. We emphasize that our algorithm scales linearly with ℓ whereas these previous works scaled with its square.

The additional speed-up provided by our implementation is not only useful for computing higher-order multipoles more quickly but also for the processing of a large number of mock catalogs for estimating covariance matrices. For example, the covariance matrix estimation of Alam et al. (2017) required evaluating clustering statistics for 3 separate redshift bins and 1000 mock catalogs. Furthermore, the calculation of higher-order multipoles is also useful for analyzing the clustering in wedges of μ (Kazin et al. 2012; Grieb et al. 2016). While there is little measurable signal in multipoles above the $\ell = 4$ hexadecapole, we show that the measurement of multipoles up to a large $\ell_{\rm max}$ allows the use of narrow μ bins. It also reduces the correlations between separate μ bins, allowing for easier theoretical modeling of the covariance of the clustering estimator. The use of narrow μ wedges becomes advantageous when measuring clustering contaminated by systematics in the plane of the sky, as is often the case for galaxy surveys, i.e., Pinol et al. (2017). Such a transverse systematic will contaminate all multipoles, but we demonstrate that the contamination can be effectively isolated to a narrow bin around $\mu \simeq 0$ when using wedges, with the width of the $\mu \simeq 0$ bin scaling as $(\ell_{\rm max}/2+1)^{-1}$. Non-uniform binning in μ can be chosen such that any artifacts of the systematic are eliminated for all bins beyond the first $\mu \simeq 0$ bin.

The chapter is laid out as follows. In §3.2.1, we first present the improved estimator of the power spectrum multipoles using a spherical harmonic expansion and demonstrate that it significantly outperforms the Cartesian decomposition method. This enables us to efficiently measure higher-order multipoles and then transform them into power spectrum wedges as shown in §3.2.2. We then discuss our implementation of the estimators in the publicly available large-scale structure analysis software `nbodykit` in §3.2.3. In §3.3, we develop a simple model for a systematic signal in the transverse ($\mu = 0$) direction and present a simple method to mitigate the contamination with a non-uniform binning scheme. We discuss the impact of survey window function on this method in §3.3.3. We show in §3.4.1 that the higher multipoles de-correlate the wedges even though they do not add additional signal. This means that one can reduce the information loss due to removal of the localized contamination by measuring more multipoles (§3.4.2). Finally, we conclude in §3.5.

3.2 Estimators

3.2.1 Multipoles

We begin by defining the weighted galaxy density field (Feldman et al. 1994),

$$F(\boldsymbol{r}) = \frac{w(\boldsymbol{r})}{I^{1/2}}\left[n(\boldsymbol{r}) - \alpha n_s(\boldsymbol{r})\right], \tag{3.1}$$

where n and n_s are the observed number density field for the galaxy catalog and synthetic catalog of random objects, respectively. The random catalog defines the expected mean density of the survey and also accounts for the angular mask and radial selection function. It contains no cosmological clustering signal. We allow for a general weighting scheme $w(\boldsymbol{r})$. The factor α normalizes the synthetic catalog to the number density of the galaxies. The field $F(\boldsymbol{r})$ is normalized by the factor of I, defined as $I \equiv \int d\boldsymbol{r}\; w^2\bar{n}^2(\boldsymbol{r})$. The estimator for the multipole moments of the power spectrum is (Feldman et al. 1994; Yamamoto et al. 2006)

$$\widehat{P}_\ell(k) = \frac{2\ell+1}{I}\int\frac{d\Omega_k}{4\pi}\left[\int \mathrm{d}\boldsymbol{r}_1\int \mathrm{d}\boldsymbol{r}_2\; F(\boldsymbol{r}_1)F(\boldsymbol{r}_2)e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)}\mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}_h) - P_\ell^{\mathrm{noise}}(\boldsymbol{k})\right], \tag{3.2}$$

where Ω_k represents the solid angle in Fourier space, $\boldsymbol{r}_h \equiv (\boldsymbol{r}_1+\boldsymbol{r}_2)/2$ is the line-of-sight to the mid-point of the pair of objects, and $\mathcal{L}_\ell$ is the Legendre polynomial of order ℓ. The shot noise P_ℓ^{noise} is

$$P_\ell^{\mathrm{noise}}(\boldsymbol{k}) = (1+\alpha)\int \mathrm{d}\boldsymbol{r}\;\bar{n}(\boldsymbol{r})w^2(\boldsymbol{r})\mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}), \tag{3.3}$$

and we assume that $P_\ell^{\mathrm{noise}} = 0$ for $\ell > 0$, as it is negligible relative to $\widehat{P}_\ell$. We then approximate the line-of-sight to the pair of objects with the line-of-sight to a single pair member, as $\mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}_h) \simeq \mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}_2)$. This approximation renders the integrals in equation 3.2 separable, yielding the so-called "Yamamoto estimator" (Yamamoto et al. 2006; Beutler et al. 2014b)

$$\widehat{P}_\ell^{\mathrm{yama}} = \frac{2\ell+1}{I}\int\frac{\mathrm{d}\Omega_k}{4\pi}\left[\int \mathrm{d}\boldsymbol{r}_1\; F(\boldsymbol{r}_1)e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1}\int \mathrm{d}\boldsymbol{r}_2\; F(\boldsymbol{r}_2)e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_2}\mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}_2) - P_\ell^{\mathrm{noise}}(\boldsymbol{k})\right]. \tag{3.4}$$

This approximate line-of-sight remains reasonably accurate over the typical range of scales considered in wide-field galaxy surveys, although it will eventually break down on very large scales (Yoo & Seljak 2015; Samushia et al. 2015; Slepian & Eisenstein 2016).

Recently, Bianchi et al. (2015) and Scoccimarro (2015) presented similar algorithms to accelerate the evaluation of equation 3.4 for the monopole, quadrupole, and hexadecapole ($\ell = 0, 2, 4$) using FFTs. By decomposing the dot product $\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}$ into its Cartesian components, they show that equation 3.4 for a given ℓ can be expressed as a sum over the Fourier transforms of the density field $F(\boldsymbol{r})$ weighted by products of Cartesian vectors. The $N\log N$ scaling of the FFT algorithm allows speed-ups of several orders of magnitude as compared to the naive summation implementation of equation 3.4. The implementation of Bianchi et al. (2015); Scoccimarro (2015) requires $(\ell+1)(\ell+2)/2$ FFTs to evaluate each $\hat{P}_\ell$, meaning $1+6+15 = 22$ FFTs for $\ell = 0, 2$, and 4. We note that Scoccimarro (2015) also defines a second estimator for $\ell > 2$ multipoles that requires fewer FFTs due to different choices regarding the line-of-sight and factorization of $\mathcal{L}_\ell(\mu)$. However, they note that this alternative estimator has larger cosmic variance than the estimator in equation 3.4 and as such, recommend against its use, despite the fewer required FFTs.

Rather than using a Cartesian decomposition, we use the spherical harmonic addition theorem (e.g., Arfken & Weber 2012, equation 16.57) to factor the Legendre polynomial into a product of spherical harmonics each depending on only a single unit vector:

$$\mathcal{L}_\ell(\hat{\boldsymbol{r}}_1 \cdot \hat{\boldsymbol{r}}_2) = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{r}}_1) Y^{\star}_{\ell m}(\hat{\boldsymbol{r}}_2). \tag{3.5}$$

This approach has recently been used by Slepian & Eisenstein (2016) to accelerate measuring the anisotropic 2PCF with the single-pair-member LOS estimator, as well as to accelerate the measurement of the 3PCF both with direct evaluations of the spherical harmonics (Slepian & Eisenstein 2015a) and using FFTs (Slepian & Eisenstein 2016). Slepian & Eisenstein (2017) further explores the use of spherical harmonics for the anisotropic 3PCF.

Using equation 3.5, the multipole estimator becomes

$$\widehat{P}_\ell(k) = \frac{2\ell+1}{I} \int \frac{\mathrm{d}\Omega_k}{4\pi} F_0(\boldsymbol{k}) F_\ell(-\boldsymbol{k}), \tag{3.6}$$

with

$$\begin{aligned} F_\ell(\boldsymbol{k}) &\equiv \int \mathrm{d}\boldsymbol{r}\ F(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \mathcal{L}_\ell(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{r}}), \\ &= \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{k}}) \int \mathrm{d}\boldsymbol{r}\ F(\boldsymbol{r}) Y^{*}_{\ell m}(\hat{\boldsymbol{r}}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}. \end{aligned} \tag{3.7}$$

The sum over m in equation 3.7 contains $2\ell+1$ terms, each of which can be computed using a FFT. Similar to Bianchi et al. (2015) and Scoccimarro (2015), we find that the multipole moments can be expressed as a sum of Fourier transforms of the weighted density field. The critical difference, however, is that by expanding the Legendre polynomial in terms of the orthonormal spherical harmonic basis we avoid redundant terms entering the summation for each multipole. For the purposes of memory efficiency, we evaluate equation 3.7 using a real-to-complex FFT and use the real form of the spherical harmonics, given by

$$Y_{\ell m}(\theta,\phi) \equiv \begin{cases} \sqrt{\dfrac{2\ell+1}{2\pi}\dfrac{(\ell-m)!}{(\ell+m)!}}\,\mathcal{L}_\ell^m(\cos\theta)\cos m\phi & m>0 \\ \sqrt{\dfrac{2\ell+1}{4\pi}}\,\mathcal{L}_\ell^m(\cos\theta) & m=0 \\ \sqrt{\dfrac{2\ell+1}{2\pi}\dfrac{(\ell-|m|)!}{(\ell+|m|)!}}\,\mathcal{L}_\ell^{|m|}(\cos\theta)\sin|m|\phi & m<0, \end{cases} \tag{3.8}$$

where $\mathcal{L}_\ell^m$ is the associated Legendre polynomial. The spherical harmonics can be expressed in terms of Cartesian vectors using equation 3.8 and the usual relations to transform from spherical to Cartesian coordinates. Thus, equations 3.6 and 3.7, combined with the spherical harmonic expressions in equation 3.8, enable computation of the multipole moments of the density field for arbitrary ℓ.

To compute each multipole, our implementation requires only $2\ell+1$ FFTs, as compared to $(\ell+1)(\ell+2)/2$ when using the Cartesian decomposition of Bianchi et al. (2015); Scoccimarro (2015). Often, we are concerned with computing all even-ℓ multipoles up to a given $\ell_{\rm max}$. For this case, our implementation requires a total of $(\ell_{\rm max}+2)(\ell_{\rm max}+1)/2 \sim \mathcal{O}(\ell_{\rm max}^2)$ FFTs, as compared to the total of $(\ell_{\rm max}+2)(\ell_{\rm max}+4)(2\ell_{\rm max}+3)/24 \sim \mathcal{O}(\ell_{\rm max}^3)$ for the Cartesian expansion. For example, for 9 multipoles ($\ell_{\rm max}=16$), our approach offers a factor of $525/153 \simeq 3.4$ improvement.

3.2.2 Wedges

The power spectrum can be expressed in terms of the multipole basis used in Section 3.2.1 as

$$P(k,\mu) = \sum_{\ell=0}^{\infty} P_\ell(k)\mathcal{L}_\ell(\mu), \tag{3.9}$$

where the power spectrum is parametrized by the amplitude k and the cosine of the angle to the line-of-sight μ. In linear theory (Kaiser 1987), only the $\ell=0,2,4$ multipoles contribute to the sum in equation 3.9, but nonlinear evolution generates nonzero moments for multipoles with $\ell>4$, albeit with diminishing importance as ℓ increases. In practice, we must truncate the sum in equation 3.9 at some $\ell_{\rm max}$. Thus, we define our estimator for clustering wedges, averaged over discrete k and μ bins, as

$$\widehat{P}(k_i,\mu_m) \equiv \sum_{\ell=0}^{\ell_{\rm max}} \widehat{P}_\ell(k)\bar{\mathcal{L}}_\ell(\mu_m,\mu_{m+1}), \tag{3.10}$$

where the multipole estimator $\widehat{P}_\ell$ can be evaluated using the implementation described in the previous section, and we have defined the mean Legendre polynomial across a wedge ranging from μ_m to μ_{m+1} as

$$\bar{\mathcal{L}}_\ell(\mu_m, \mu_{m+1}) = \frac{1}{\mu_{m+1} - \mu_m} \int_{\mu_m}^{\mu_{m+1}} \mathrm{d}\mu \, \mathcal{L}_\ell(\mu). \tag{3.11}$$

Here and throughout this chapter, hat denotes an estimator and subscripted k and μ indicate binned quantities. We assume uniform wavenumber bins and use k_i to denote the center of the i^{th} bin. We allow for non-uniform bins in μ, labeling the m^{th} wedge with μ_m to denote a bin ranging from $[\mu_m, \mu_{m+1}]$.

3.2.3 Implementation

We implement the multipole and wedge estimators as presented in sections 3.2.1 and 3.2.2 as part of the publicly available software toolkit `nbodykit` (Hand et al. 2017a).[1] As described in Chapter 2, our implementation is fully parallelized with Message Passing Interface (MPI) and uses a Python binding (Feng 2017d) of the `PFFT` software by Pippig (2013) to compute FFTs in parallel. We use the symbolic manipulation functionality available in the `SymPy` Python package (SymPy 2017) to compute the spherical harmonic expressions in equation 3.8 in terms of Cartesian vectors. This allows the user to specify the desired multipoles at runtime, enabling the code to be used to compute multipoles of arbitrary ℓ. Testing and development of the code was performed on the Cray XC-40 system Cori at the National Energy Research Supercomputing Center (NERSC), and the code exhibits strong scaling, with a roughly linear reduction in wall-clock time as the number of available processors increases. When computing all even multipoles up to $\ell_{\mathrm{max}} = 16$ (requiring in total 153 FFTs), our implementation takes roughly 90 seconds using 64 processors on Cori.

For the results presented in this chapter, we place the galaxies and random objects on a Cartesian grid using the Triangular Shaped Cloud (TSC) prescription to compute the density field $F(\boldsymbol{r})$ of equation 3.1. We use the interlaced grid technique of Sefusatti et al. (2016) to limit the effects of aliasing, and we correct for any artifacts of the TSC gridding using the correction factor of Jing (2005). The interlacing scheme allows computation of the FFTs on a 512^3 grid with accuracy comparable to the results when using a 1024^3 grid, but with a wall-clock time that is ~8 times smaller. When using interlacing, the catalog of galaxies is interpolated on to two meshes separated by half of the size of a grid cell. We sum these two density fields in Fourier space and inverse Fourier Transform back to configuration space. We then apply the spherical harmonic weightings of equation 3.8 to this combined density field and proceed with computing the terms in equation 3.7. The speed-up provided by interlacing is particularly powerful when computing large ℓ multipoles. When combined with TSC interpolation, we are able to measure power spectra up to the Nyquist frequency at $k \simeq 0.4\ h\mathrm{Mpc}^{-1}$ with fractional errors at the level of 10^{-3} (Sefusatti et al. 2016).

[1] `https://github.com/bccp/nbodykit`

3.3 Isolating transverse $\mu = 0$ systematics

As discussed above, cosmological information in the linear regime is limited to $\ell_{\max} = 4$, so one may question the value of algorithms that go to $\ell_{\max} > 4$. One reason is that in the nonlinear regime higher-order multipoles are generated, and their information can be used to constrain nonlinear RSD models. Another motivation is measurement contamination from systematics that are predominantly localized to some part of a clustering wedge. In this section, we present a method to isolate and potentially remove systematics from our clustering estimators, assuming that the systematic signal is dominant in the plane of the sky (i.e., angular), which is a common issue for galaxy surveys. The contamination in this case is confined to predominantly transverse $\mu = 0$ modes. We consider a toy model for the process of fiber assignment, a common issue for galaxy surveys where the physical process of assigning galaxy targets to spectrograph fibers leads to incomplete target selection and creates a systematic signal that must be accounted for. Our discussion is particularly relevant for the Dark Energy Spectrograph Instrument (DESI; Levi et al. 2013), as the process of fiber assignment has recently been shown in Pinol et al. (2017) to introduce a largely transverse systematic signal.

3.3.1 A toy model for fiber assignment

We model the effect of a plane-of-the-sky systematic by suppressing the observed power spectrum by a Dirac delta function at $\mu = 0$, as

$$P^{\mathrm{obs}}(k, \mu) = P(k, \mu) - P_c(k)\delta_{\mathrm{D}}(\mu), \tag{3.12}$$

where δ_{D} denotes a one-dimensional Dirac delta function, and $P_c(k)$ is the power spectrum of the contamination signal and describes the amplitude of the clustering suppression. Here, $P(k, \mu)$ is the true anisotropic power spectrum in the absence of systematics. In purely linear theory, $P(k, \mu)$ would be fully described by its $\ell = 0, 2,$ and 4 multipoles (Kaiser 1987).

The contamination signal is localized in μ but affects all observed multipoles, evident from the Legendre expansion of the Dirac delta function,

$$\delta_\ell = \frac{2\ell + 1}{2} \int_{-1}^{1} \mathrm{d}\mu \, \mathcal{L}_\ell(\mu)\delta_{\mathrm{D}}(\mu) = \frac{2\ell + 1}{2} \mathcal{L}_\ell(0). \tag{3.13}$$

In practice, we use only a finite number of multipoles, up to a desired $\ell_{\max}$, to reconstruct the two-dimensional power spectrum $P(k, \mu)$. We can define an estimator for the true power spectrum in the presence of a transverse systematic as

$$\begin{aligned}\widehat{P}(k,\mu) &= \widehat{P}^{\text{obs}}(k,\mu) + P_c(k)\sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{2}\mathcal{L}_\ell(0)\mathcal{L}_\ell(\mu), \\ &= \sum_{\ell=0}^{\ell_{\max}} \widehat{P}_\ell(k)\mathcal{L}_\ell(\mu) + P_c(k)\frac{\ell_{\max}+1}{2}\mathcal{L}_{\ell_{\max}+1}(0)\frac{\mathcal{L}_{\ell_{\max}+1}(\mu)}{\mu},\end{aligned} \tag{3.14}$$

where our estimator for the observed power $\widehat{P}^{\text{obs}}$ uses the measured multipoles $\widehat{P}_\ell$ up to $\ell_{\max}$, and we have used the Christoffel summation formula (Gradshteyn et al. 2007, equation 8.915.1),

$$\sum_{\ell=0}^{\ell_{\max}}(2\ell+1)\mathcal{L}_\ell(x)\mathcal{L}_\ell(y) = \frac{\ell_{\max}+1}{y-x}\left[\mathcal{L}_{\ell_{\max}}(x)\mathcal{L}_{\ell_{\max}+1}(y) - \mathcal{L}_{\ell_{\max}}(y)\mathcal{L}_{\ell_{\max}+1}(x)\right], \tag{3.15}$$

with $x = 0$ and $y = \mu$. Equation 3.14 demonstrates that the $\mu = 0$ contamination leaks into $\mu > 0$ modes because of the finite number of multipoles used to reconstruct $P(k,\mu)$ and that the angular dependence of this leakage is characterized by $\mathcal{L}_{\ell_{\max}+1}(\mu)/\mu$. We can describe the response of this leakage to the systematic signal as

$$R(\mu) \equiv \frac{\widehat{P}^{\text{obs}}(k,\mu) - \widehat{P}(k,\mu)}{P_c(k)} = -\frac{\ell_{\max}+1}{2\mu}\mathcal{L}_{\ell_{\max}}(0)\mathcal{L}_{\ell_{\max}+1}(\mu). \tag{3.16}$$

We show this response for various $\ell_{\max}$ values in Figure 3.1. While there is minimal signal in large ℓ multipoles, we can see from this figure that the utility of measuring higher-order multipoles is that it enables sharper reconstruction of the angular dependence of the contaminating signal. By increasing $\ell_{\max}$ we are able to increasingly localize the contamination around $\mu = 0$, with a width scaling as $\ell_{\max}^{-1}$.

The oscillatory structure of the response in Figure 3.1 suggests that we can employ a non-uniform binning in μ for our wedge estimator of Section 3.2.2 in order to localize the effect of the systematic to the first bin and cancel out the contamination in each of the other bins. If we desire to have as many wedge bins as number of observed multipoles (measuring even multipoles up to $\ell_{\max}$), then there will be $\ell_{\max}/2$ non-contaminated bins. The edges of these bins can be computed from the response in equation 3.16 as

$$\int_{\mu_i}^{\mu_{i+1}} d\mu\, \frac{\mathcal{L}_{\ell_{\max}+1}(\mu)}{\mu} \equiv 0, \quad i = 1, 2, \ldots, \ell_{\max}/2, \tag{3.17}$$

where μ_i specifies the left edge of the i^{th} bin, and we have assumed a total of $N_\mu = \ell_{\max}/2+1$ bins. By construction, we have $\mu_0 = 0$ and $\mu_{\ell_{\max}/2+1} = 1$. In this notation, the only contaminated bin is the first, ranging from $\mu_0 < \mu < \mu_1$. We show the non-uniform binning for $\ell_{\max} = 4$ and $\ell_{\max} = 16$ as the shaded regions in the left panel of Figure 3.2. Generically, the μ wedges first become wider and then significantly narrower ranging from $\mu = 0$ to $\mu = 1$.

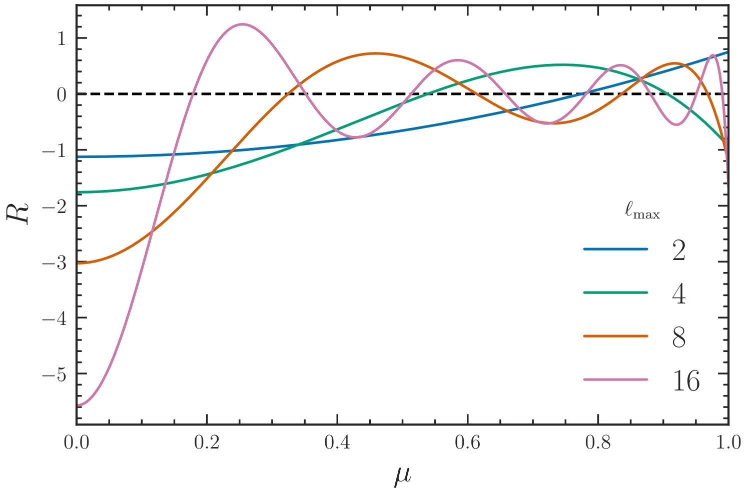


Figure 3.1: The leakage of a transverse $\mu = 0$ systematic into $\mu > 0$ power as a function of the maximum multipole used to reconstruct the observed power $P(k, \mu)$. We plot the response of this error, as given in equation 3.16. As multipoles are measured to larger $\ell_{\rm max}$, the contamination is better isolated around the origin, $\mu = 0$.

We also show the width of the first, contaminated bin, $\mu_1 - \mu_0$, in the right panel. The edge of the first bin closely follows the result in the uniform case, $\mu_1 - \mu_0 = (\ell_{\rm max}/2 + 1)^{-1}$. Larger $\ell_{\rm max}$ values clearly enable better isolation of the systematic signal in a narrow first bin, and in turn, create a larger μ range absent of any systematics.

3.3.2 Verification with simulations

We verify the utility of the non-uniform binning scheme discussed in §3.3.1 using simulated density fields. We generate uniformly clustered catalogs of discrete objects and simulate an example systematic signal by modulating the amplitude of the density field in the plane of the sky. We use a sinusoidal function for this modulation, which creates a large contaminating spike in Fourier space at a specific wavenumber, $k = k_c$. We perform this test for both periodic boxes and for mock catalogs where the geometry of the DR12 BOSS CMASS sample has been imposed (Alam et al. 2017; Reid et al. 2016). We denote these latter mocks as cutsky mocks. For the cubic boxes, we simply choose the z axis of the box to be the line-of-sight and modulate the amplitude of the density field in the (x,y) plane. For the cutsky mocks, which provide the angular and redshift coordinates of objects, we apply the sinusoidal variation as a function of right ascension and declination. We perform these tests for 50 cubic boxes of side length $L_{\rm box} = 2600\ h^{-1}$Mpc and for 84 cutsky catalogs and compute the average results to reduce noise.

We now compare our simulated results with the theoretical expectations from Section 3.3.1. Because the catalogs are uniformly clustered, the true signal is a constant shot noise that we can subtract from the results. We measure the clustering wedges in both uniform and

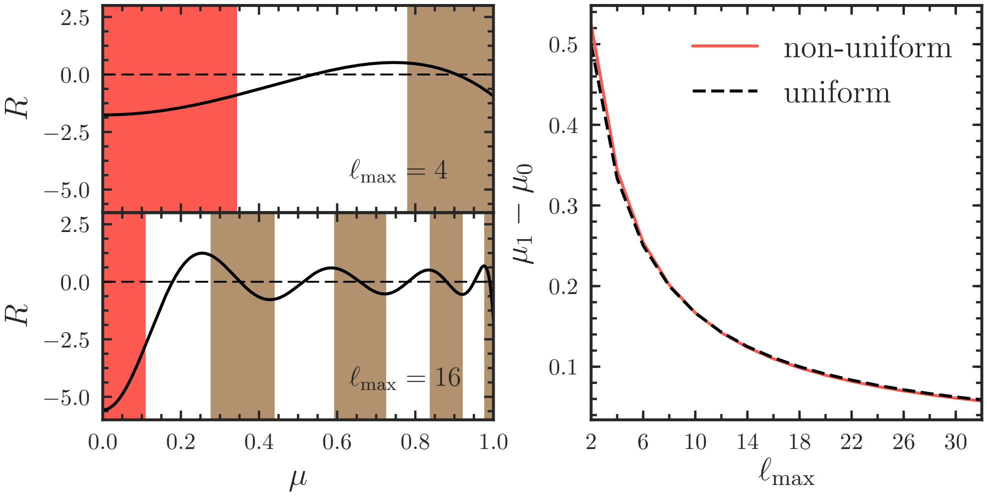


Figure 3.2: *Left*: the leakage of a transverse $\mu = 0$ systematic into $\mu > 0$ power (black) for $\ell_{\max} = 4$ (top) and $\ell_{\max} = 16$ (bottom). We show the appropriate non-uniform binning (shaded) that cancels the systematic in all but the first bin (red, shaded). *Right*: the width of the first $\mu \simeq 0$ bin, given by $\mu_1 - \mu_0$, for the cases of non-uniform (red) and uniform (black, dashed) bins.

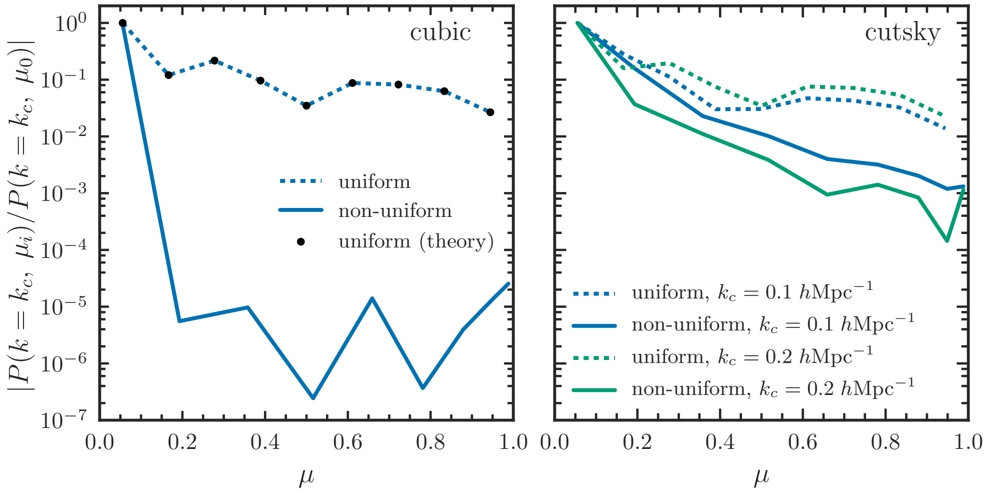


Figure 3.3: The amplitude of a contaminating spike in 9 $P(k, \mu)$ wedges relative to its amplitude in the first μ bin for cubic simulation boxes (left) and for cutsky mock catalogs with the BOSS DR12 selection function imposed (right). Wedges are computed from even multipoles up to $\ell_{\max} = 16$. The mock catalogs contain uniformly clustered objects with a density field modulated via a sinusoidal function in the plane of the sky, causing a large systematic spike in Fourier space at $k = k_c$. We show results for both uniform μ binning (dotted) and the non-uniform (solid) scheme discussed in §3.3.1.

non-uniform bins and compare the amplitude of the contaminating spike at $k = k_c$ for each wedge relative to its amplitude in the first μ bin. The wedges are computed using even multipoles up to $\ell_{\text{max}} = 16$, which results in 9 μ wedges. The left panel of Figure 3.3 shows the results for the cubic boxes. We obtain near-perfect cancellation of the systematic when using non-uniform bins, isolating the contamination to only the first bin at $\mu \simeq 0$. On the other hand, all wedges remain contaminated when using a uniform binning scheme. These results for uniform binning also agree well with our theoretical expectation (shown as black points), given the response in equation 3.16.

The removal of the systematic contamination using the cutsky catalogs, shown in the right panel of Figure 3.3, is not as prominent as in the cubic case. However, the non-uniform binning does reduce the amplitude of the systematic for all wedges, as compared to the uniform scheme, and this reduction is as large as an order of magnitude for most bins. We perform two separate tests for the cutsky mocks, introducing systematic spikes at $k_c = 0.1\ h\text{Mpc}^{-1}$ and at $k_c = 0.2\ h\text{Mpc}^{-1}$. We find varying levels of success in eliminating the systematic for these two cases, suggesting some unaccounted for k-dependence in the optimal binning scheme. It is likely that the survey geometry, which is not present in the cubic case, complicates the simple model discussed in Section 3.3.1. In the cutsky case, the estimator measures the power spectrum convolved with the window function. In particular, the systematic signal is also convolved with the window function, which mixes k and μ modes and invalidates our simple modeling assumptions in equation 3.14. We expect the window function to be isotropic and have less influence on small scales (large k); this is the trend we find in our results, as we find better cancellation of the systematic in the case of $k_c = 0.2\ h\text{Mpc}^{-1}$. We explore the effects of the window function on our non-uniform binning scheme in more detail in the next section.

3.3.3 A toy model for window function effects

Here, we outline a toy model to provide a qualitative understanding of the window function's impact on systematic removal. We show that the window function couples to the transverse systematic, effectively re-normalizing all of its coefficients in the Legendre basis and thus implying a different choice of non-uniform bin boundaries relative to the window-free case for systematic elimination.

We model the window function as a spherical top-hat in configuration space with radius R, so that

$$w(\boldsymbol{k}; R) = \frac{3 j_1(kR)}{kR}, \tag{3.18}$$

where j_1 is the spherical Bessel function of order one. The observed systematic is then convolved with the square of the window function as

$$P_c^{\text{win}}(k, \mu) = \left\{ w^2(k') \star P_c(k') \delta_{\text{D}}(\mu) \right\} (\boldsymbol{k}), \tag{3.19}$$

where star denotes convolution. We note that $P_c^{\rm win}(k, \mu)$ remains a function only of $|\boldsymbol{k}|$ and μ if the window function is isotropic, as in our toy model. We may evaluate this convolution using the Convolution Theorem, which gives

$$\begin{aligned}&\left\{w^2(k') \star P_c(k')\delta_{\rm D}(\mu)\right\}(k, \mu) = \\ &\qquad\qquad \mathrm{FT}\left\{\mathrm{FT}^{-1}\{w^2(k')\}(r)\ \mathrm{FT}^{-1}\{P_c(k')\delta_{\rm D}(\mu)\}(r)\right\}(k, \mu). \qquad (3.20)\end{aligned}$$

We first evaluate the inverse Fourier transform (FT) of $w^2(k)$. Applying the Convolution Theorem, the desired inverse FT is the convolution of two spherical top-hats, each of radius R with centers separated by r. The overlap integral is given by the volume $V_{\rm lens}(r; R)$ of the spherical lens enclosed by both spheres when they are separated by r (Weisstein 2017),

$$V_{\rm lens}(r; R) = \frac{\pi}{12}(4R + r)(2R - r)^2. \qquad (3.21)$$

This result gives the first term inside the outer curly brackets in equation 3.20. We now seek the second term, the inverse FT of the systematic. Writing the Delta function using its Legendre expansion (equation 3.13) and then expanding the Legendre polynomials into spherical harmonics using the spherical harmonic addition theorem, we find

$$P_c(k')\delta_{\rm D}(\mu) = P_c(k') \sum_{\ell=0}^{\infty} \delta_\ell \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{k}}') Y^*_{\ell m}(\hat{\boldsymbol{n}}), \qquad (3.22)$$

where δ_ℓ is defined in equation 3.13. The inverse FT can then be obtained by expanding the relevant exponential via the plane wave expansion into spherical Bessel functions and spherical harmonics (e.g., Arfken & Weber (2012), equation 16.52) and invoking orthogonality, leading to

$$\mathrm{FT}^{-1}\left\{P_c(k')\delta_{\rm D}(\mu)\right\}(r) = \sum_{\ell=0}^{\infty} S_\ell(r)\mathcal{L}_\ell(\mu), \qquad (3.23)$$

where $\mu = \hat{\boldsymbol{r}} \cdot \hat{\boldsymbol{n}}$ and

$$S_\ell(r) = \int \frac{k'^2 {\rm d}k'}{2\pi^2} j_\ell(k'r) P_c(k'). \qquad (3.24)$$

We now have both terms in the outer curly brackets of equation 3.20 and simply require their product's Fourier transform to obtain $P_c^{\rm win}(k, \mu)$, the systematic observed in the presence of the window function.

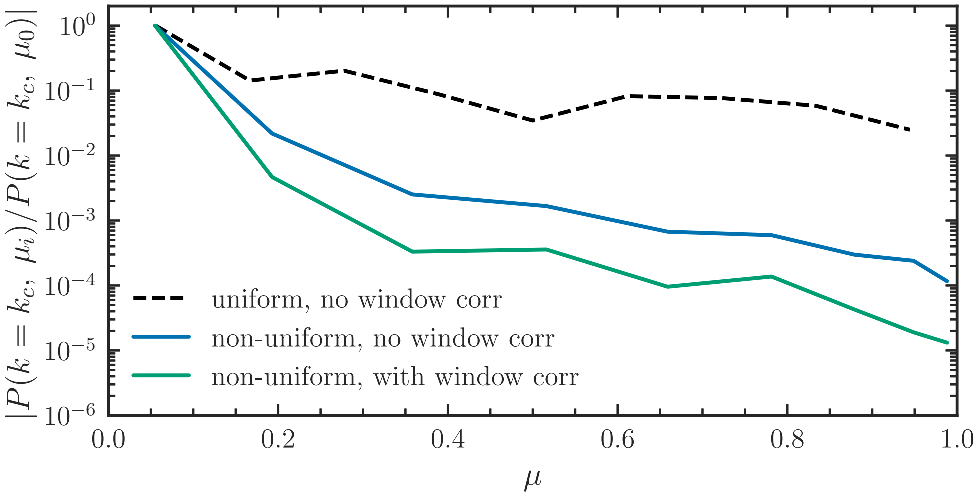


Figure 3.4: The amplitude of a systematic spike in 9 $P(k, \mu)$ wedges for a uniform clustering field with a toy-model selection function imposed as described in §3.3.3. When using a non-uniform binning scheme and accounting for window function effects, we can increase the success of the systematic removal by roughly an order of magnitude.

Expanding the Legendre polynomials of equation 3.23 into spherical harmonics using the addition theorem, again expanding the exponential via the plane wave expansion, and invoking orthogonality, we find

$$P_c^{\mathrm{win}}(k, \mu) = \sum_{\ell=0}^{\infty} \mathcal{L}_\ell(\mu)\delta_\ell \int r^2 \mathrm{d}r\ V_{\mathrm{lens}}(r; R) S_\ell(r) j_\ell(kr). \tag{3.25}$$

We pause to examine the limit where $R \to \infty$ and hence $V_{\mathrm{lens}}(r; R)$ is independent of r and can be taken outside the integral; this corresponds to a boundary-free survey. In this limit, the integral over r can be performed by substituting equation 3.24 and invoking the orthogonality relation for spherical Bessel functions and we recover that $P_c^{\mathrm{win}}(k, \mu) \to P_c(k)\delta_{\mathrm{D}}(\mu)$.

We see that in general, in the presence of an isotropic window function, the coefficients of the Legendre expansion of $P_c(k, \mu)$ change and are no longer given by the simple relation $\delta_\ell = (2\ell + 1)/2\ \mathcal{L}_\ell(0)$. Importantly, they now have k-dependence, as the window function mixes the purely isotropic systematic amplitude $P_c(k)$ with the μ-dependent Delta function. The non-uniform wedge boundaries of the previous section were set by the condition that for a given wedge, the sum of averaged Legendre polynomials weighted by the Delta function's coefficients would vanish. Here, we see that changing these coefficients simply means this criterion is satisfied for a different non-uniform binning scheme.

We use simulations to examine the effectiveness of our non-uniform binning scheme in the presence of this toy model window function. We apply a spherical top-hat window function of radius $R = 780\ h^{-1}$Mpc to a uniformly clustered density field in a cubic box of side length $L_{\mathrm{box}} = 2600\ h^{-1}$Mpc. As in previous sections, we model the systematic with a sinusoidal modulation of the density field in the (x, y) plane, assuming the z axis represents the line-of-sight. This modulation generates a spike in Fourier space at $k = k_c$, corresponding to the wavelength of the modulation. We once again use the measured multipoles to estimate wedges in non-uniform bins, and compare the results with and without accounting for the window function corrections in equation 3.25. We present this comparison in Figure 3.4, which shows that by accounting for the window function, an additional order of magnitude reduction in the systematic signal can be achieved in all but the first μ wedge. As in previous sections, we find that a uniform μ binning scheme performs worse than our non-uniform scheme, even when ignoring window function effects.

3.4 Statistical properties

In this section, we explore the covariance properties of the wedge estimator as a function of $\ell_{\max}$ and use a Fisher matrix formalism to describe the effect on the derived parameter constraints when using our non-uniform binning approach to mitigate systematics.

3.4.1 Covariance

Under the assumption of purely Gaussian statistics, the covariance of the power spectrum $P(k, \mu)$ averaged in bins of k and μ is (Grieb et al. 2016)

$$\mathrm{Cov}\big[P(k_i, \mu_m), P(k_j, \mu_n)\big] = \delta_{ij}\delta_{mn}\frac{2}{N_{k_i}\Delta\mu_m}\int \frac{4\pi k^2 \mathrm{d}k}{V_{k_i}}\frac{\mathrm{d}\mu}{\Delta\mu_m}\left[P(k, \mu) + \bar{n}^{-1}\right]^2, \tag{3.26}$$

where the number density of the sample considered is $\bar{n}$, the volume of the shell in k-space is $V_{k_i} = 4\pi[(k_i + \Delta k/2)^3 - (k_i - \Delta k/2)^3]/3$, and the number of modes in the i^{th} k bin is $N_{k_i} = 4\pi k_i^2 \Delta k V_s/(2\pi)^3$, where V_s is the volume of the sample considered. Under the assumption of Gaussian statistics, different clustering wedges are not correlated, as reflected by the Kronecker delta factor δ_{mn} in equation 3.26.

The computationally-efficient estimator presented in this chapter does not directly measure the quantity $P(k_i, \mu_m)$ that enters into equation 3.26. Rather, we reconstruct power spectrum wedges from a finite set of measured multipoles, up to a specified $\ell_{\max}$. Thus, the relevant quantity is the covariance of the multipoles averaged in k bins, which is given by

$$\mathrm{Cov}\big[\widehat{P}_\ell(k_i), \widehat{P}_{\ell'}(k_j)\big] = \delta_{ij}(2\ell + 1)(2\ell' + 1)\frac{2}{N_{k_i}}\int \mathrm{d}\mu \frac{2\pi k^2 \mathrm{d}k}{V_{k_i}}\mathcal{L}_\ell(\mu)\mathcal{L}_{\ell'}(\mu)\big[P(k, \mu) + \bar{n}^{-1}\big]^2, \tag{3.27}$$

where we see multipoles of different ℓ are correlated. From this covariance we can compute the covariance of the wedge estimator in equation 3.10 as

$$\mathrm{Cov}\big[\widehat{P}(k_i,\mu_m),\widehat{P}(k_j,\mu_n)\big] = \sum_{\ell=0}^{\ell_{\max}}\sum_{\ell'=0}^{\ell_{\max}} \bar{\mathcal{L}}_\ell(\mu_m)\bar{\mathcal{L}}_{\ell'}(\mu_n)\,\mathrm{Cov}\big[\widehat{P}_\ell(k_i),\widehat{P}_{\ell'}(k_j)\big], \tag{3.28}$$

where the mean Legendre polynomial across a wedge is given by equation 3.11.

The wedge covariance in equation 3.28 is difficult to further simplify analytically, but before comparing to simulations, we can make further progress using the simplifying assumption of linear theory. In this case, we can use the Kaiser model (Kaiser 1987)

$$P(k,\mu) = (1+\beta\mu^2)^2 b_1^2 P_\mathrm{r}(k), \tag{3.29}$$

where $\beta = f/b_1$ is the usual redshift-space distortion parameter, b_1 is the linear bias parameter, $P_\mathrm{r}(k)$ is the linear theory real-space power spectrum, and f is the logarithmic growth rate (Kaiser 1987). Now, we can separate the scale and angular dependence in equation 3.28. We leave the scale dependence implicit in our notation to focus on the angular subspace in order to improve clarity. With these assumptions, the wedge covariance becomes

$$\widehat{C}_{mn} \equiv \mathrm{Cov}\big[\widehat{P}(\mu_m),\widehat{P}(\mu_n)\big] = \frac{2\gamma_{mn}}{N_{k_i}}\overline{P_\mathrm{r}^2}, \tag{3.30}$$

where

$$\overline{P_\mathrm{r}^2} \equiv \int \frac{4\pi k^2 \mathrm{d}k}{V_{k_i}}\, P_\mathrm{r}^2(k), \tag{3.31}$$

and

$$\gamma_{mn} \equiv \sum_{\ell=0}^{\ell_{\max}}\sum_{\ell' 0}^{\ell_{\max}} (2\ell+1)(2\ell'+1)\bar{\mathcal{L}}_\ell(\mu_m)\bar{\mathcal{L}}_{\ell'}(\mu_n)\int\frac{\mathrm{d}\mu}{2}\mathcal{L}_\ell(\mu)\mathcal{L}_{\ell'}(\mu)(1+\beta\mu^2)^4. \tag{3.32}$$

From these equations, we see that in the simple Kaiser model, the correlation coefficient between between wedges μ_m and μ_n, defined as $\rho_{mn} = \widehat{C}_{mn}/(\widehat{C}_{mm}\widehat{C}_{nn})^{1/2}$ is independent of scale with the amplitude proportional to the quantity γ_{mn}.

We first compare our simple theoretical modeling to the wedge covariance measured from 990 independent Quick Particle Mesh (QPM) periodic simulations (White et al. 2014) at a redshift of $z = 0.55$ and with a box size of $L_\mathrm{box} = 2560\ h^{-1}$Mpc. These simulations were designed to mimic the clustering of the BOSS CMASS sample, with a linear bias of $b_1 \sim 2$ at $z \sim 0.5$. We estimate the clustering wedges using the measured multipoles up to a specified $\ell_{\max}$ and use the 990 realizations to estimate the covariance of the wedges. We show the resulting correlation matrix between separate μ wedges in Figure 3.5 and compare to the linear Kaiser result from equation 3.32. We perform this comparison using both non-uniform

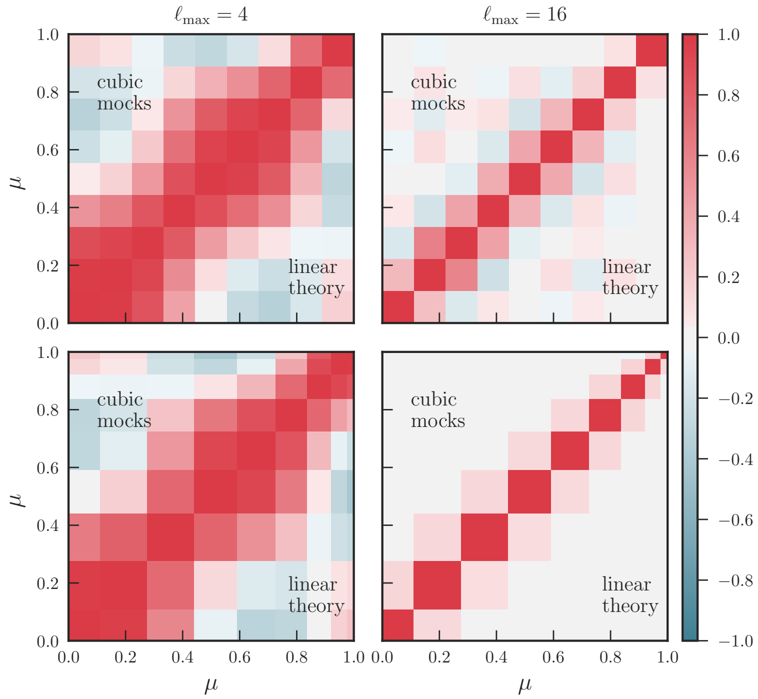


Figure 3.5: The correlation matrix between μ wedges measured from cubic box simulations (upper triangle) as compared to linear theory (lower triangle), when using uniform (top row) and non-uniform (bottom row) μ bins. For 9 μ wedges, we show results using $\ell_{\max} = 4$ (left column) and $\ell_{\max} = 16$ to estimate the wedges from the corresponding multipoles. We find excellent agreement between linear theory and the results measured from simulations.

(bottom row) and uniform (top row) binning schemes, as well as for $\ell_{\max} = 4$ (left column) and $\ell_{\max} = 16$ (right column). In all cases, the number of wedges is fixed to $N_\mu = 9$. We find excellent agreement between a simple Kaiser model with $\beta = 0.35$ and the simulation results. As expected, we find the wedges to be significantly more correlated when using only three multipoles to reconstruct nine μ wedges, as is the case for $\ell_{\max} = 4$, than when using nine measured multipoles, as for $\ell_{\max} = 16$. Furthermore, in the case of $\ell_{\max} = 16$, we find that our non-uniform binning scheme achieves a significantly more diagonal covariance matrix between wedges, as seen in the right column of Figure 3.5. As the matrix becomes more diagonal, the covariance is better approximated by the Gaussian case, where the clustering wedges are fully independent.

We also compare our Kaiser modeling to a set of cutsky mock catalogs that include selection function effects, although we do not expect the simulation results to be well-described by this theoretical model in this case. We use a set of 84 mock catalogs which mimic the radial and angular selection functions of the BOSS DR12 CMASS sample (Reid et al. 2016; Alam et al. 2017). They model the true geometry, volume, and redshift distribution of the

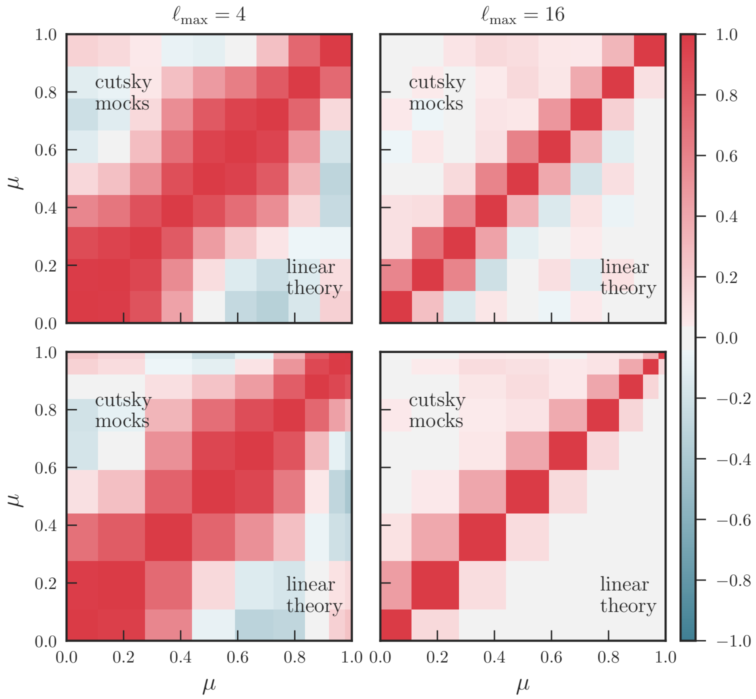


Figure 3.6: The correlation matrix between μ wedges measured from realistic cutsky mock catalogs (upper triangle) as compared to linear theory (lower triangle), when using uniform (top row) and non-uniform (bottom row) μ bins. For 9 μ wedges, we show results using $\ell_{\max} = 4$ (left column) and $\ell_{\max} = 16$ to estimate the wedges from the corresponding multipoles. While there are discrepancies introduced by the window function in comparison to the linear theory expectation, the general trends remain consistent with the periodic box results.

CMASS sample and were constructed from a set of seven independent, periodic box N-body simulations with the same cosmology and a side length of $L_{\mathrm{box}} = 2600\ h^{-1}\mathrm{Mpc}$. Each of the 84 mock catalogs is an independent realization, and the clustering of these cutsky catalogs is very similar overall to the BOSS CMASS sample at $z \sim 0.5$. As was done for the cubic box simulations, we compare simulation and theory for the correlation matrix for nine μ wedges using $\ell_{\max} = 4$ and $\ell_{\max} = 16$. These results are presented in Figure 3.6. As expected, the cutsky simulation results are not as well-described by the Kaiser model as in the cubic case due to window function effects. However, the general trends in the covariance are similar for the cutsky case as for the cubic case. Importantly, we once again find that using a higher $\ell_{\max}$ at fixed N_μ de-correlates the wedges and that the covariance is more diagonal when using our non-uniform binning scheme.

An additional disadvantage of using bins uniform in μ is that the wedge covariance matrix quickly becomes ill-conditioned for high $\ell_{\max}$. This can be seen in Figure 3.7, where the left panel shows the condition number of the γ_{mn} matrix as a function of $\ell_{\max}$ for both uniform

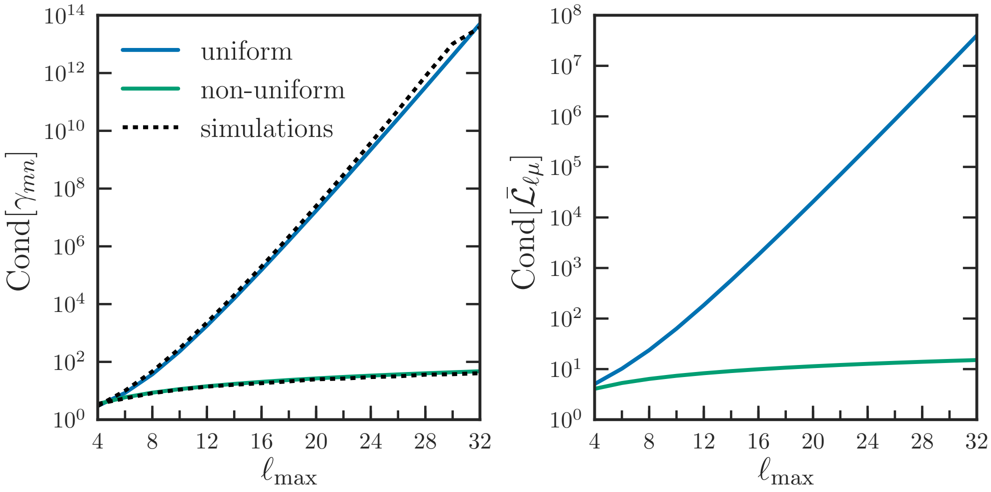


Figure 3.7: *Left*: the condition number of the covariance matrix γ_{mn}, assuming a linear Kaiser model with $\beta = 0.35$, as compared to the result computed from the 990 QPM simulations (black). The simulation results have been re-normalized to match the theoretical condition number at $\ell_{\max} = 4$. *Right*: the condition number of the matrix specifying the mean Legendre polynomial across each μ wedge $\bar{L}_{\ell\mu}$, as given by equation 3.11. We see that the wedge covariance matrix becomes ill-conditioned for large $\ell_{\max}$ values when using a uniform binning scheme, driven by the fact that the transformation matrix $\bar{\mathcal{L}}_{\mu\ell}$ also becomes ill-conditioned.

and non-uniform bins. Here, the condition number of a matrix $\mathbf{M}$ is defined as the ratio of its smallest to largest singular values, computed using Singular Value Decomposition (SVD) (e.g., Press et al. 1992). The SVD of a matrix is defined as $\boldsymbol{M} = \boldsymbol{U\Sigma V}^T$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with the singular values along the diagonal. We find similar trends for the condition number of the covariance matrix for our theoretical results assuming a linear Kaiser model with $\beta = 0.35$ and for results computed from the 990 QPM boxes. Both uniform and non-uniform binning result in a reasonable condition number for $\ell_{\max} = 4$, but the matrix in the uniform case becomes increasingly singular as $\ell_{\max}$ increases. In such a case, the inversion of the covariance, which is a necessary step of any likelihood analysis, becomes numerically unstable. This behavior at large $\ell_{\max}$ is largely driven by the $\bar{\mathcal{L}}_{\ell\mu}$ matrix, which defines the contribution of a multipole of order ℓ to a given μ wedge. The condition number of this matrix is shown in the right panel of Figure 3.7, and its behavior mirrors that of the full covariance matrix.

The functional form of $\mathcal{L}_\ell(\mu)$ can provide some insight into the large condition number of the covariance matrix when using uniformly spaced bins. The Legendre polynomial of order ℓ oscillates around zero, and the frequency of the oscillation increases with increasing μ. For large $\ell_{\max}$, there exist bins at $\mu \sim 1$ where the Legendre polynomial exhibits a positive/negative symmetry across the bin, and thus, the average value cancels very nearly to zero. This presents problems in equation 3.10, where our measured multipoles are weighted by the mean Legendre polynomial. These issues are mitigated by our non-uniform bins, which were constructed such that the width of the bins decreases as a function of μ, just as the Legendre polynomials oscillate more quickly. Thus, the bin cancellation is mostly avoided when non-uniform bins are used and the condition number of the resulting covariance matrix remains stable, even at large ℓ_{max}. Such a binning scheme becomes appealing for clustering analyses, even if systematic mitigation is not the primary goal.

3.4.2 Fisher information

We can evaluate the information content of our wedge estimator as a function of ℓ_{max} using the Fisher matrix formalism. As in Section 3.4.1, we assume a simple linear Kaiser model (equation 3.29), where the parameter vector of interest is $\boldsymbol{p} = (b_1\sigma_8, f\sigma_8)$. For clarity, we also suppress the k indexing here, as the μ and k dependence of the covariance is fully separable for the Kaiser model. Assuming a Gaussian likelihood function for the clustering wedge observables, we can express the Fisher matrix as

$$F_{ij} = \sum_{m=0}^{N_\mu - 1} \sum_{n=0}^{N_\mu - 1} \frac{\partial P(\mu_m)}{\partial p_i} \widehat{C}_{mn}^{-1} \frac{\partial P(\mu_n)}{\partial p_j}, \tag{3.33}$$

where N_μ is the number of (non-uniform) μ bins, $P(\mu_n)$ is the theoretical Kaiser model averaged over the μ_n wedge, and the covariance between the measured wedges $\widehat{C}_{mn}$ is given by equation 3.30. We can also use this formalism to quantify the cost of removing the first μ bin when using our non-uniform binning scheme. In this case, the Fisher matrix is given by

$$F_{ij}^{\mu\neq 0} = \sum_{m=1}^{N_\mu - 1} \sum_{n=1}^{N_\mu - 1} \frac{\partial P(\mu_m)}{\partial p_i} \widehat{C}_{mn}^{-1} \frac{\partial P(\mu_n)}{\partial p_j}, \tag{3.34}$$

where we have explicitly removed the contribution from the μ_0 wedge to the double sum in this equation.

We show the Fisher information for the auto-correlations of $b_1\sigma_8$ and $f\sigma_8$, as well as their cross correlation, as a function of ℓ_{max} in Figure 3.8. Results are computed for the non-uniform μ binning scheme presented in Section 3.3.1, assuming a value of $\beta = f/b_1 = 0.35$ for the Kaiser model. The left panel shows the information content when using all μ bins, and as expected, the information content saturates at $\ell_{\mathrm{max}} = 4$ because only the $\ell = 0, 2$, and 4 multipoles are non-zero in the Kaiser model. In the right panel of this figure, we show the Fisher information when we exclude the first μ bin from the analysis. In this case, the information on $b_1\sigma_8$ is partially lost, approximately proportional to the width of the missing wedge. However, the information on $f\sigma_8$ remains relatively unaffected by the missing wedge. The first wedge at $\mu \simeq 0$ is a prominent source of information on the amplitude of the power spectrum, as parametrized by $b_1\sigma_8$, but contains little information on the μ dependence of the clustering.

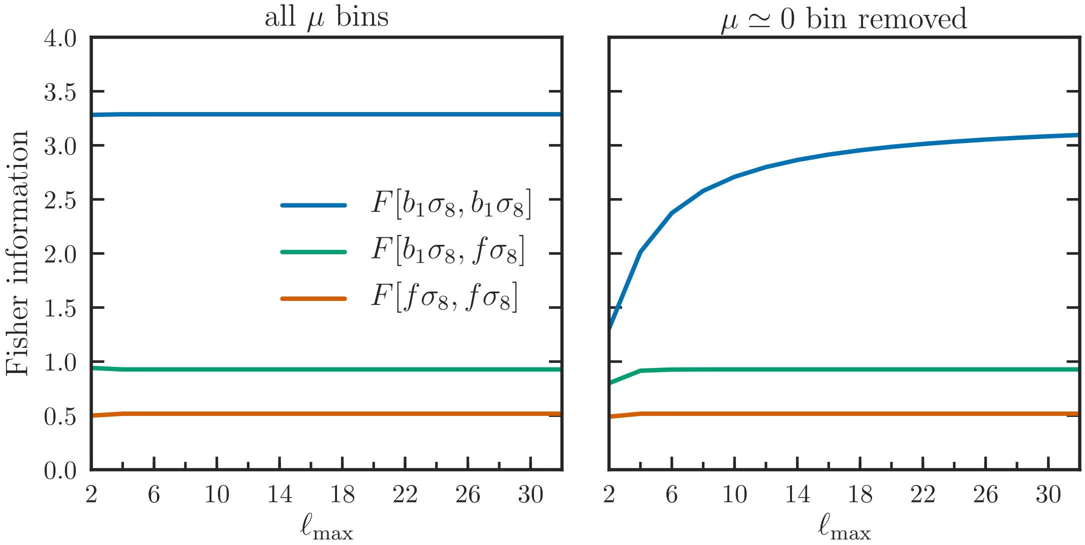


Figure 3.8: The Fisher information for the parameter vector $\boldsymbol{p} = (b_1\sigma_8, f\sigma_8)$ for our wedge estimator using non-uniform μ bins, in the case of the linear theory Kaiser model. We show results as a function of the maximum multipole used to reconstruct the clustering wedges, as well as the case when using all μ bins (left) and when excluding the first μ bin (right). A linear Kaiser model with $\beta = 0.35$ has been assumed.

The inverse of the Fisher matrix provides an estimate of the marginalized error on a given parameter, such that the error on the parameter A is given by $\sigma_A = (F^{-1})^{1/2}_{AA}$. Thus, we can use the Fisher formalism to evaluate the change in the parameter uncertainties when excluding the first $\mu \simeq 0$ wedge in the presence of a transverse systematic. We show this fractional change for $b_1\sigma_8$ and $f\sigma_8$ as a function of $\ell_{\max}$ in Figure 3.9, and we find the loss of constraining power drops rapidly with $\ell_{\max}$. For $\ell_{\max} = 16$, we find $\sim 7\%$ and $\sim 13\%$ increases in the uncertainties on $f\sigma_8$ and $b_1\sigma_8$, respectively, as compared to $\sim 54\%$ and $\sim 92\%$ for $\ell_{\max} = 4$. With a reasonably large choice for $\ell_{\max}$, we can exclude the contaminated $\mu \simeq 0$ bin with only marginal losses for the parameter constraints of interest.

3.5 Conclusions

In this chapter, we have presented an optimal estimator for the anisotropic power spectrum multipoles that is valid in the local plane-parallel approximation. Our implementation eliminates redundancy present in previous algorithms (Bianchi et al. 2015; Scoccimarro 2015). These works rely on a Cartesian decomposition of the Legendre basis to write the power spectrum estimator of Yamamoto et al. (2006) using fast Fourier transforms. We improve upon them by using a spherical harmonic decomposition of the Legendre polynomials, motivated by the approach of Slepian & Eisenstein (2016) for the anisotropic 2PCF. The method pre-

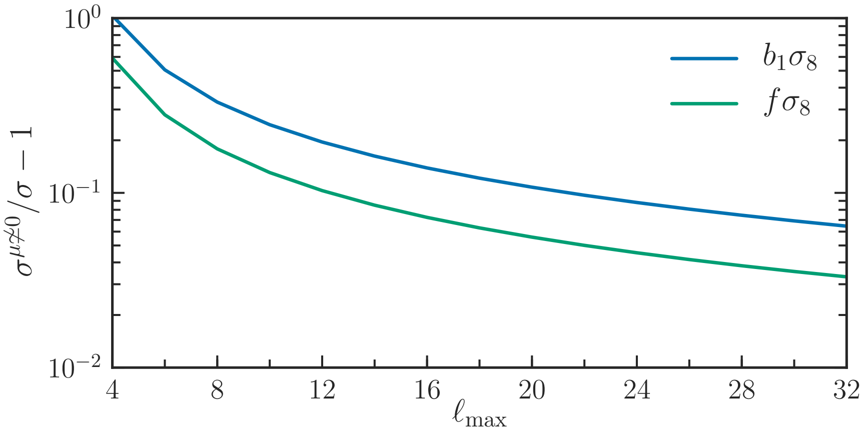


Figure 3.9: The fractional change in the uncertainties in $b_1\sigma_8$ and $f\sigma_8$ when using non-uniform μ wedges and excluding the first $\mu \simeq 0$ wedge from the analysis, as determined from the Fisher matrix. A linear Kaiser model with $\beta = 0.35$ has been assumed.

sented here is substantially faster than previous anisotropic power spectrum algorithms and renders calculation of multipoles to high $\ell_{\rm max}$ computationally feasible. For a given multipole of order ℓ, our method requires only $2\ell + 1$ FFTs rather than the $(\ell + 1)(\ell + 2)/2$ FFTs of the Cartesian approach. For the highest $\ell_{\rm max}$ used in this work, $\ell_{\rm max} = 16$, our approach is ~3.4 times faster than previous works, using 153 FFTs as opposed to 525.

Our estimator's significant reduction in wall-clock time allows construction of finely-binned wedges in $P(k, \mu)$ by combining multipoles up to high $\ell_{\rm max}$. We show that narrow μ bins are particularly advantageous for mitigating the effects of systematic contamination in the plane of the sky, as is often the case for galaxy surveys (Pinol et al. 2017, e.g.,). In the presence of such an angular systematic signal, we show that a non-uniform binning scheme in μ can effectively isolate the contamination to the first $\mu \simeq 0$ wedge and that the systematic contributions to all other bins can be eliminated. We have verified the effectiveness of our non-uniform bins on both periodic simulations and realistic mock catalogs that have a survey selection function. We have demonstrated with a toy model that a survey selection function mixes the k and μ dependence of the systematic signal, introducing k-dependence into the optimal non-uniform wedge boundaries. However, the systematic signal can still be reduced even when ignoring these effects. When analyzing galaxy survey data, knowledge of the window function and realistic simulations can be used to choose the optimal binning to reduce transverse systematics.

We have also explored the statistical properties of the wedge estimator as a function of the maximum measured multipole $\ell_{\rm max}$. We show using linear theory that the covariance of the wedge estimator quickly becomes ill-conditioned for large $\ell_{\rm max}$ when using uniform bins, and we verify this finding with simulations. Consequently, when using uniform bins the covariance inversion is numerically unstable, creating a significant barrier for any likelihood analysis. On the other hand, the non-uniform binning scheme described in this chapter remains well-conditioned for all $\ell_{\max}$ values, enabling its inversion and use in model fitting. We also show that at a fixed number of μ wedges, using larger values of $\ell_{\max}$ de-correlates separate wedges, and that the covariance matrix of wedges using non-uniform bins is more diagonal than in the uniform case. With a Fisher analysis assuming linear theory, we have demonstrated that the uncertainty on $f\sigma_8$ inflates by $\sim 7\%$ with $\ell_{\max} = 16$ when excluding the first μ wedge, assuming it is fully contaminated by systematics, as compared to a 54% increase with $\ell_{\max}$. Even larger choices for $\ell_{\max}$ can further reduce this increase and should be explored in more detail for future RSD analyses in the presence of transverse (angular) systematics.

We note that similar techniques as those presented in this chapter can be applied to clustering wedges in configuration space. However, the choice of optimal non-uniform bins to remove systematics is further complicated for a correlation function analysis, as the systematic signal is no longer localized to $\mu = 0$. Importantly, the optimal binning choice becomes a function of both the separation perpendicular and parallel to the line-of-sight, $r_\perp$ and $r_\parallel$, which introduces additional modeling complexity. Similar techniques in configuration space should be further explored to assess their effectiveness at minimizing the effects of angular systematics.

Finally, we also point out that, as shown in Slepian & Eisenstein (2015b) for the anisotropic 2PCF, slight generalizations of the local plane parallel multipole estimates can be combined to yield the separation midpoint or angle bisector method-based multipoles. This point is important because it enables midpoint and bisector-based multipoles to be obtained by FFTs. As the relevant geometry for anisotropic clustering is the same in Fourier space and configuration space, combining Slepian & Eisenstein (2015b) with the results of this work will enable estimation of midpoint or bisector-based multipoles to very high $\ell_{\max}$ with FFTs, relevant for properly handling wide-angle effects in next-generation surveys.

The improvements to the power spectrum estimator presented in this chapter will prove valuable for next generation redshift surveys such as DESI (Levi et al. 2013; DESI Collaboration et al. 2016a,b) and *Euclid* (Laureijs et al. 2011) both for the data measurement and for the covariance estimation, which requires analyzing a large number of mock catalogs. Given these surveys' large volumes and consequent high statistical precision, an unprecedented level of systematics control is required. The non-uniform clustering wedges described in this chapter will be important in this regard for DESI (recently described in Pinol et al. 2017). In the future, these methods should be developed and further tested on realistic end-to-end simulations of upcoming surveys.

Chapter 4

Extending the modeling of the anisotropic galaxy power spectrum to $k = 0.4\ h\mathrm{Mpc}^{-1}$

In this chapter, we present a model for the redshift-space power spectrum of galaxies and demonstrate its accuracy in describing the monopole, quadrupole, and hexadecapole of the galaxy density field down to scales of $k = 0.4\ h\mathrm{Mpc}^{-1}$. The model describes the clustering of galaxies in the context of a halo model and the clustering of the underlying halos in redshift space using a combination of Eulerian perturbation theory and N-body simulations. The modeling of redshift-space distortions is done using the so-called distribution function approach. The final model has 13 free parameters, and each parameter is physically motivated rather than a nuisance parameter, which allows the use of well-motivated priors. We account for the Finger-of-God effect from centrals and both isolated and non-isolated satellites rather than using a single velocity dispersion to describe the combined effect. We test and validate the accuracy of the model on several sets of high-fidelity N-body simulations, as well as realistic mock catalogs designed to simulate the BOSS DR12 CMASS data set. The suite of simulations covers a range of cosmologies and galaxy bias models, providing a rigorous test of the level of theoretical systematics present in the model. The level of bias in the recovered values of $f\sigma_8$ is found to be small. When including scales to $k = 0.4\ h\mathrm{Mpc}^{-1}$, we find 15-30% gains in the statistical precision of $f\sigma_8$ relative to $k = 0.2\ h\mathrm{Mpc}^{-1}$ and a roughly 10-15% improvement for the perpendicular Alcock-Paczynski parameter $\alpha_\perp$. Using the BOSS DR12 CMASS mocks as a benchmark for comparison, we estimate an uncertainty on $f\sigma_8$ that is $\sim$10-20% larger than other similar Fourier-space RSD models in the literature that use $k \leq 0.2\ h\mathrm{Mpc}^{-1}$, suggesting that these models likely have a too-limited parametrization.

4.1 Introduction

Galaxy redshift surveys measure the three-dimensional clustering of galaxies in the Universe, and over the past few decades, they have provided a wealth of cosmological information (Davis & Peebles 1983; Maddox et al. 1990; Tegmark et al. 2004; Cole et al. 2005; Eisenstein et al. 2005; Anderson et al. 2012, 2014a,b; Alam et al. 2017). In combination with other cosmological probes, such as observations of the cosmic microwave background, type-Ia supernova samples, and weak-lensing surveys, analyses of the large-scale structure (LSS) of the Universe have proven invaluable in establishing the current cosmological paradigm, the ΛCDM model, as well as measuring its parameters with ever-increasing precision.

Crucial to the success of galaxy surveys has been the ability to precisely and accurately measure the feature imprinted on the clustering of galaxies by baryon acoustic oscillations (BAO; see e.g., Bassett & Hlozek (2010) for a review) in the early Universe. The BAO signal can be used to provide constraints on the expansion history of the Universe and infer properties of dark energy (e.g., Wagner et al. (2008); Shoji et al. (2009)). The isotropic effect was first detected in the 2-point clustering in the SDSS (Eisenstein et al. 2005) and the 2dFGRS (Cole et al. 2005) and in the 3-point clustering in Slepian et al. (2017). Recent measurements of the anisotropic BAO signal, combined with the Alcock-Paczynski (AP; Alcock & Paczynski 1979) effect, have provided percent-level measurements of the Hubble parameter $H(z)$ and angular diameter distance $D_A(z)$ (Alam et al. 2017). Perhaps most encouragingly, the BAO signal is well-understood theoretically, with systematic effects on the distance scale expected to be sub-dominant for future generations of surveys (Eisenstein & White 2004; Seo & Eisenstein 2005; Angulo et al. 2008; Padmanabhan & White 2009; Mehta et al. 2011; Yoo et al. 2011; Slepian & Eisenstein 2015c).

Beyond the BAO signal, additional information is present in the clustering of galaxies through what are known as redshift-space distortions (RSD). The peculiar velocities of galaxies affect their measured redshifts through the Doppler effect, and in turn, these measured redshifts are used to infer the line-of-sight (LOS) position of those galaxies. The peculiar velocity field is sourced by the gravitational potential, and thus, an anisotropic signal containing information about the rate of structure growth in the Universe is imprinted on the clustering. Extracting information from RSD is inherently more difficult than with BAO, as it requires modeling of the full broadband shape of the clustering statistic and precise understanding of the anisotropy induced by RSD. The theoretical task is complicated by the fact that the well-understood, linear Kaiser model (Kaiser 1987) breaks down on relatively large scales, with various kinds of nonlinear effects complicating the theoretical modeling (e.g., Scoccimarro 2004; Okumura & Jing 2011; Jennings 2012; Kwan et al. 2012). Of particular importance is the large, nonlinear virial motions of satellite galaxies within halos, known as the Finger-of-God (FoG) effect (Jackson 1972). Since the statistical precision of clustering measurements is generally higher on smaller scales, where the effects of nonlinearities are worse, a direct limit on the amount of useable cosmological information is imposed due to theoretical uncertainties.

Despite these modeling challenges, RSD analyses have developed into one of the most popular and powerful cosmological probes today (Peacock et al. 2001; Hawkins et al. 2003; Tegmark et al. 2006; Guzzo et al. 2008; Yamamoto et al. 2008; Blake et al. 2011b; Beutler et al. 2012; Reid et al. 2012; Samushia et al. 2013; Chuang et al. 2013; Beutler et al. 2014a; Reid et al. 2014). Constraints on the growth rate of structure through measurements of the parameter combination $f(z)\sigma_8(z)$ can provide tests of General Relativity (e.g., Guzzo et al. 2008), as well as information about the properties of neutrinos (Lesgourgues & Pastor 2006; Beutler et al. 2014a) and tighter constraints on the expansion history through the AP effect (e.g., Shoji et al. 2009). Recent results from Data Release 12 (DR12) of the Baryon Oscillation Spectroscopic Survey (BOSS) (Sánchez et al. 2017; Beutler et al. 2017b; Grieb et al. 2017; Satpathy et al. 2017) have provided the tightest constraints to date on the growth rate of structure, with roughly 10% constraints on $f\sigma_8(z_\mathrm{eff})$ in 3 redshift bins centered at $z_\mathrm{eff} = 0.38$, 0.51, and 0.61.

To date, RSD analyses have generally either relied directly on the results of N-body simulations or on perturbative approaches to model the clustering of galaxies in the quasi-linear and nonlinear regimes. Both approaches have their pros and cons. For simulation-based analyses, e.g., Tinker et al. (2006); Hikage (2014); Reid et al. (2014); Guo et al. (2015b), the simulations represent the best possible description of nonlinearities, although individual simulations are expensive to run and, often, the relevant parameter space cannot be as sufficiently explored as one would like. On the other hand, modeling techniques relying on perturbation theory (PT), e.g., Beutler et al. (2017b); Grieb et al. (2016); Satpathy et al. (2017); Sánchez et al. (2017), are relatively fast to compute but will always break down on small enough scales and fail to fully capture non-perturbative features, such as the FoG effect from satellites. In either case, simulations play a crucial role in estimating the range of scales where a model remains accurate enough to recover cosmological parameters in an unbiased fashion.

In this chapter, we extend the work of Okumura et al. (2015), presenting a more accurate model for the redshift-space power spectrum of galaxies and extensively stress-testing this model on a suite of N-body simulations. We describe the galaxy clustering in the context of a halo model (Seljak 2000; Ma & Fry 2000; Peacock & Smith 2000; Cooray & Sheth 2002) and rely on a combination of Eulerian PT and N-body simulations to model the power spectrum of dark matter halos in redshift space. We use several sets of N-body simulations to validate our model, and we perform cosmological parameter analyses on realistic BOSS-like mock catalogs to verify both the accuracy and constraining power of the model. The model relies on the distribution function approach (Seljak & McDonald 2011; Okumura et al. 2012a,b; Vlah et al. 2012, 2013; Blazek et al. 2014) to map real-space statistics to redshift space. This formalism is different but complementary from other commonly used approaches in RSD analyses, such as the TNS model (Taruya et al. 2010) or the Gaussian streaming model (Reid & White 2011). We build upon the results presented in Okumura et al. (2015), which showed that the characterization of the redshift-space power spectrum of galaxies in terms of 1-halo and 2-halo correlations is accurate when compared against N-body simulations. We extend that work by improving the accuracy of the underlying model for the halo redshift-space power spectrum. The model is based on the PT results presented in Vlah et al. (2013),

Name	L_{box} [h^{-1}Mpc]	z_{box}	Ω_{m}	$\Omega_{\mathrm{b}}h^2$	h	n_{s}	σ_8
RunPB	1380	0.55	0.292	0.022	0.69	0.965	0.82
N-series; Challenge D,E	2600	0.562	0.286	0.02303	0.7	0.96	0.82
Challenge A,B,F,G	2500	0.5	0.30711	0.022045	0.6777	0.96	0.82
Challenge C	2500	0.441	0.27	0.02303	0.7	0.96	0.82

Table 4.1: The cosmological and simulation parameters for the various N-body simulations used in this chapter.

but uses simulation-based modeling for key terms. In particular, we develop and extend the Halo-Zel'dovich Perturbation Theory (HZPT) of Seljak & Vlah (2015), which relies on a combination of linear Lagrangian PT and simulation-based calibration. A Python software package `pyRSD` that implements the model described in this chapter is publicly available[1].

This chapter is organized as follows. Section 4.2 describes the set of simulations that we use to calibrate our model, as well as the test suite that we use for independent validation. We describe the power spectrum estimator, covariance matrix, and likelihood analysis used to perform parameter estimation in Section 4.3. In Section 4.4, we detail the power spectrum model, first reviewing the halo model formalism presented in Okumura et al. (2015) and then discussing several new modeling approaches for the redshift-space halo power spectrum. We assess the accuracy and performance of the model based on an independent test suite of simulations in Section 4.5. Finally, we discuss our results and future prospects in Section 4.6 and conclude in Section 4.7.

4.2 Simulations

We use several sets of N-body simulations for both calibrating and testing the model presented in this paper. The first set of simulations, described in Section 4.2.1, is used heavily in verifying individual components of the clustering model. Sections 4.2.2 and 4.2.3 describe an independent suite of high-fidelity simulations that we use to independently verify the accuracy and precision of the model. The relevant cosmological and simulation parameters for the mocks discussed in this section are summarized in Table 4.1.

4.2.1 RunPB

The main set of simulations used for calibration and testing purposes is the RunPB N-body simulation produced by Martin White with the TreePM N-body code of White (2002). These simulations have been used recently in a number of analyses (White 2014; Reid et al. 2014; Schmittfull et al. 2015a,b). The simulation set has 10 realizations of 2048^3 dark matter

[1] `https://github.com/nickhand/pyRSD`

particles in a cubic box of length $L = 1380\ h^{-1}\mathrm{Mpc}$. The cosmology is a flat ΛCDM model with $\Omega_\mathrm{b}h^2 = 0.022$, $\Omega_\mathrm{m} = 0.292$, $n_\mathrm{s} = 0.965$, $h = 0.69$, and $\sigma_8 = 0.82$.

For testing and calibration of the modeling of halo clustering, we use halo catalogs generated using a friends-of-friends (FOF) algorithm with a linking length of 0.168 times the mean particle separation to identify halos (Davis et al. 1985). We consider 8 halo mass bins (as a function of M_FOF) across 10 redshift outputs, ranging from $z = 0$ to $z = 1$. The redshift outputs considered are: $z \in \{0, 0.1, 0.25, 0.4, 0.5, 0.55, 0.65, 0.75, 1\}$. The (overlapping) halo mass bins range from $\log_{10} M_\mathrm{FOF} = 12.6$ to $\log_{10} M_\mathrm{FOF} = 14.4$ and are described in Table 4.2. For reference, Table 4.2 also gives the linear bias values at $z = 0.55$ and $z = 0$. The linear biases for each halo mass bin are determined from the ratio of the large-scale halo-matter cross power spectrum to the matter power spectrum at each redshift output.

We also rely heavily on a set of galaxy catalogs produced using halo occupation distribution (HOD) modeling from halo catalogs generated from the $z = 0.55$ RunPB realizations. The halo catalog production and the HOD modeling is the same as in Reid et al. (2014): halos are identified using a spherical overdensity (SO) algorithm and the HOD parameterization follows Zheng et al. (2005). In Reid et al. (2014), the RunPB simulations are denoted as the `MedRes` simulations. The HOD parameters used to generate the galaxy catalog used in this work are $\{\log_{10} M_\mathrm{min}, \sigma_{\log_{10} M}, \log_{10} M_1, \alpha, \log_{10} M_\mathrm{cut}\} = \{12.99, 0.308, 14.08, 0.824, 13.20\}$. These HOD parameters were chosen to reproduce the clustering of the BOSS CMASS sample (White et al. 2011), i.e., a large-scale linear bias of $b_1 \sim 2$ at $z \sim 0.5$.

4.2.2 N-series

The N-series cubic boxes are a set of realizations from a large-volume, high-resolution N-body simulation, used as part of a "mock challenge" testing procedure by the BOSS collaboration in preparation for publishing results as part of DR12 in Alam et al. (2017). Details of this mock challenge can be found in Tinker (2016). Briefly, the N-series suite consists of seven independent, periodic box realizations with the same cosmology, and a

bin	1	2	3	4	5	6	7	8
$\log_{10} M_\mathrm{FOF}^\mathrm{min}$	12.6	12.8	13.0	13.2	13.4	13.6	13.8	14.0
$\log_{10} M_\mathrm{FOF}^\mathrm{max}$	13.0	13.2	13.4	13.6	13.8	14.0	14.2	14.4
$b_1(z = 0.55)$	1.40	1.56	1.78	2.04	2.36	2.77	3.28	3.93
$b_1(z = 0)$	1.00	1.07	1.19	1.33	1.51	1.74	2.03	2.41

Table 4.2: The halo mass bins used when comparing results from the RunPB simulations to theoretical modeling of halo clustering. For each of the 10 redshift outputs ranging from $z = 0$ to $z = 1$, we consider 8 fixed halo mass bins. We give the corresponding large-scale, linear bias for each bin for two redshifts, $z = 0.55$ and $z = 0$.

side length of $L_{\text{box}} = 2600\ h^{-1}$Mpc at a redshift $z_{\text{box}} = 0.5$. The cosmology is given by: $\Omega_{\text{m}} = 0.286$, $\Omega_{\Lambda} = 0.714$, $\sigma_8 = 0.82$, $n_{\text{s}} = 0.96$, and $h = 0.7$. The N-body simulation was run using the GADGET2 code (Springel 2005), with sufficient mass and spatial resolution to resolve the halos that typical BOSS galaxies occupy. A single galaxy bias model was assumed, and HOD modeling was used to populate halos from the seven realizations with galaxies. The parameters of the HOD were chosen to reproduce the clustering of the BOSS CMASS sample (i.e., linear bias $b_1 \sim 2$ at $z \sim 0.5$).

An additional set of 84 mock catalogs were generated from the three orthogonal projections of each of the seven N-series cubic boxes using the `make_survey` software[2] (White et al. 2014). Denoted as the "cutsky" mocks, these mocks have the same angular and radial selection function as the NGC DR12 CMASS sample (Reid et al. 2016; Alam et al. 2017). They model the true geometry, volume, and redshift distribution of the CMASS NGC sample and provide a realistic simulation of the true BOSS data set. Each catalog is an independent realization, and these mocks were also used as part of the DR12 mock challenge.

4.2.3 Lettered Challenge Boxes

A second part of the BOSS DR12 mock challenge was performed on a suite of HOD galaxy samples constructed from a heterogeneous set of high-resolution N-body simulations. There are seven different HOD galaxy catalogs, constructed from large-volume periodic simulation boxes with varying cosmologies. The seven catalogs are labeled A through G. Several of the boxes are based on the Big MultiDark simulation (Riebe et al. 2013). The catalogs are constructed out of simulation boxes with a range of 3 underlying cosmologies. HOD models with varying parameters are applied to boxes with the same cosmology, changing the overall galaxy bias values by ±5%. The redshifts of the boxes range from $z = 0.441$ to $z = 0.562$.

These cases were designed to quantify the sensitivity of RSD models to the specifics of the galaxy bias model over a reasonable range of cosmologies, testing for any possible theoretical systematics. The cosmology and relevant simulation parameters for each of these boxes is given in Table 4.1. A comparison of the results from the mock challenge in context of the BOSS DR12 results is presented in Alam et al. (2017), and individual Fourier-space clustering results from the challenge are discussed in Beutler et al. (2017b); Grieb et al. (2017).

4.3 Analysis methods

In this work, we measure the 2-point clustering of galaxies as characterized by the power spectrum multipoles, defined in terms of the 2D anisotropic power spectrum as

$$P_\ell(k) = \frac{2\ell+1}{2} \int_{-1}^{1} d\mu P(k,\mu) \mathcal{L}_\ell(\mu), \tag{4.1}$$

[2] https://github.com/mockFactory/make_survey

where $\mathcal{L}_\ell$ is the Legendre polynomial of order ℓ. We estimate the multipoles from catalogs of discrete galaxies in periodic box N-body simulations and from more realistic, cutsky mock catalogs, which mimic real survey data. We estimate the continuous galaxy overdensity field using a Triangular Shaped Cloud interpolation scheme (see, e.g., Hockney & Eastwood 1981) to assign the galaxy positions to a 3D Cartesian grid.

For both periodic boxes and cutsky mocks, we employ Fast Fourier Transform (FFT) based estimators to compute the multipoles. In the case of simulation boxes with periodic boundary conditions, the line-of-sight is assigned to a specific box axis and the power spectrum $P(k, \mu)$ can be computed as the square of the Fourier modes of the overdensity field. The desired multipoles are then found by computing equation 4.1 as a discrete sum over $P(k, \mu)$. In the case of the cutsky mocks, we employ the FFT-based estimator described in Hand et al. (2017b), which modifies the FFT estimator presented by Bianchi et al. (2015) and Scoccimarro (2015). Building on the ideas of previous power spectrum estimators (Feldman et al. 1994; Yamamoto et al. 2006), this estimator uses a spherical harmonic decomposition to allow the use of FFTs to compute the higher-order multipoles, with 5 and 9 FFTs required to compute the quadrupole and hexadecapole, respectively. When computing FFTs, we ensure that the grid configuration is such that our desired maximum wavenumber is not greater than one-half of the Nyquist frequency of the grid, which should eliminate any aliasing effects on our measured power spectra (i.e., Sefusatti et al. 2016). The measured power spectra are estimated on a discrete k-grid which makes the angular distribution of Fourier modes irregular. This discreteness effect is especially important at low k and can be accounted for by modifying equation 4.1 as

$$P_\ell(k) = \frac{2\ell+1}{2} \int_{-1}^{1} d\mu P(k,\mu) \frac{N_\mathrm{modes}(k,\mu)}{N_\mathrm{bin}(k)} \mathcal{L}_\ell(\mu), \tag{4.2}$$

where $N(k, \mu)$ gives the total number of modes on the k-space grid, and the normalization is

$$N_\mathrm{bin}(k) = \int_{-1}^{1} d\mu N_\mathrm{modes}(k,\mu). \tag{4.3}$$

We account for this discreteness effect when comparing theoretical multipoles to simulation results by applying equation 4.2 to our model predictions. This procedure has been shown to sufficiently correct for this effect (Beutler et al. 2017b). For all power spectrum calculations, we use the publicly available software package `nbodykit`[3] (Hand et al. 2017a), which uses massively parallel implementations of these estimators for fast calculations optimized to run on high-performance computing machines. `nbodykit` is described in detail in Chapter 2.

We use the Markov chain Monte Carlo (MCMC) technique to derive the likelihood distributions of the model parameters described in detail in Section 4.4.4. We employ a modified version of the Python code `emcee`[4] (Foreman-Mackey et al. 2013) to explore the relevant

[3] https://github.com/bccp/nbodykit

[4] https://github.com/dfm/emcee

model parameter space. The data vector used in these fits is the concatenation of the monopole, quadrupole, and hexadecapole,

$$\mathcal{D} = [P_0(k), P_2(k), P_4(k)] \,, \tag{4.4}$$

where we have measured the multipoles from simulations as previously described. The inclusion of the hexadecapole $P_4(k)$ has been shown to offer significant improvements on RSD constraints, i.e., Beutler et al. (2017b); Grieb et al. (2017). In all fits, we use a bin spacing in wavenumber of $\Delta k = 0.005\ h\mathrm{Mpc}^{-1}$, and the maximum wavenumber included in the fits ranges from $k_\mathrm{max} = 0.2\ h\mathrm{Mpc}^{-1}$ to $k_\mathrm{max} = 0.4\ h\mathrm{Mpc}^{-1}$. The likelihood fits require an estimate of the covariance matrix, and we use the theoretical Gaussian covariance for the multipoles in Fourier space (i.e., Grieb et al. 2016). In the case of the cutsky mocks, we properly account for the redshift distribution and survey volume of the mock catalogs when computing the expected covariance, using e.g., Yamamoto et al. (2006).

Our choice for covariance matrix ignores non-Gaussian contributions produced by e.g., nonlinear structure growth, and in the case of the cutsky mocks, correlations induced by the window function due to the survey geometry. We have tested the impact of our choice for covariance matrix by comparing the parameter fits obtained when using a covariance matrix derived from a set of 1000 mock catalogs from the Quick Particle Mesh (QPM; White et al. 2014) simulations. Because the covariance derived from the QPM mocks uses a fixed cosmology, we focus our tests on the simulations described in Section 4.2 that have similar power spectra to that of the QPM mocks. This choice helps to minimize the impact of the cosmology dependence of the covariance matrix, allowing us to better gauge the effect of the non-Gaussian features of the covariance. While we do find variations in the best-fit parameters recovered when using the simulation-based covariance, the shifts are consistent with the derived errors, and we do not believe the use of analytic covariance matrices affects the conclusions of this chapter.

4.4 The power spectrum model

In this section, we present the model for the anisotropic clustering of galaxies in Fourier space, as characterized by the broadband, two-dimensional power spectrum. First, we connect the clustering of galaxies to the clustering of halos, reviewing the halo model formalism presented in Okumura et al. (2015) in §4.4.1. We describe our model for the redshift-space halo power spectrum and the various modeling improvements from past work in §4.4.2. In §4.4.3, we discuss how we account for various observational effects when modeling real galaxy survey data. Finally, we summarize the complete set of model parameters in §4.4.4.

4.4.1 Halo model formalism for galaxies

Our treatment of the clustering of galaxies is based upon the model presented in Okumura et al. (2015). The clustering of a given galaxy sample is considered within the context of a halo model (Seljak 2000; Peacock & Smith 2000; Ma & Fry 2000; Scoccimarro et al. 2001; Cooray & Sheth 2002), which allows one to separately consider contributions to the clustering arising from galaxies within the same halo and those from separate halos, known as the 1-halo and 2-halo terms, respectively. This formalism is ideal when accounting for the effects of satellite galaxies on the anisotropic power spectrum, where the radial distribution of satellites induces both 1-halo and 2-halo effects. We describe the relevant model details from Okumura et al. (2015), used in this work, below.

Galaxy sample decomposition

In redshift space, we can decompose contributions to the galaxy overdensity field δ_g^S into contributions from central and satellite galaxies as

$$\delta_g^S(\boldsymbol{k}) = (1-f_s)\delta_c^S(\boldsymbol{k}) + f_s\delta_s^S(\boldsymbol{k}), \tag{4.5}$$

where $f_s = N_s/N_g = 1 - N_c/N_g$ is the satellite fraction, N_c and N_s are the numbers of central and satellite galaxies, respectively, and $N_g = N_c + N_s$ is the total number of galaxies. It follows then that the power spectrum of the galaxy density field $(2\pi)^3 P_{gg}^S(\boldsymbol{k})\delta(\boldsymbol{k}+\boldsymbol{k}') \equiv \left\langle \delta_g^S(\boldsymbol{k})\delta_g^S(\boldsymbol{k}')\right\rangle$ can be expressed as

$$P_{gg}^S(\boldsymbol{k}) = (1-f_s)^2 P_{cc}^S(\boldsymbol{k}) + 2f_s(1-f_s)P_{cs}^S(\boldsymbol{k}) + f_s^2 P_{ss}^S(\boldsymbol{k}), \tag{4.6}$$

where P_{cc}^S, P_{cs}^S, and P_{ss}^S are the central auto power spectrum, the central-satellite cross power spectrum, and the satellite auto power spectrum in redshift space, respectively.

To fully separate 1-halo and 2-halo contributions to the power spectrum, we further decompose the central and satellite galaxy samples. We decompose the central galaxy density field into those centrals that do and do not have a satellite galaxy in the same halo, denoted as types "A" and "B" centrals, respectively. For the latter type, a 1-halo contribution will exist due to the central-satellite correlations inside the same halo. We use a similar decomposition for satellite galaxies, where we consider satellites that only have a single satellite in a halo (type "A") and those satellites that live in halos with more than one satellite (type "B"). The latter type will contribute a 1-halo term to the power spectrum, due to correlations between multiple satellites in the same halo.

With these galaxy sample definitions, we can express the central-satellite and satellite-satellite power spectra in terms of 1-halo and 2-halo correlations. Note that by construction, a halo can only have a single central galaxy, and thus, the centrals auto spectrum is a purely 2-halo contribution. The central-satellite cross power spectrum can be expressed as

$$\begin{aligned} P_{cs}^S(\boldsymbol{k}) &= (1-f_{c_B})P_{c_A s}^S(\boldsymbol{k}) + f_{c_B}P_{c_B s}^S(\boldsymbol{k}), \\ &= (1-f_{c_B})\left[(1-f_{s_B})P_{c_A s_A}^S + f_{s_B}P_{c_A s_B}^S\right] + f_{c_B}\left[(1-f_{s_B})P_{c_B s_A}^S + f_{s_B}P_{c_B s_B}^S\right], \end{aligned} \tag{4.7}$$

where $f_{c_B} = N_{c_B}/N_c$ is the fraction of centrals that have a satellite in the same halo, and $f_{s_B} = N_{s_B}/N_s$ is the fraction of satellites that live in halos with more than one satellite.

Because the sample c_B consists of central galaxies that have satellite galaxies inside the same halo, the term $P^S_{c_Bs}$ (and similarly, $P^S_{c_Bs_A}$ and $P^S_{c_Bs_B}$) contains a 1-halo contribution, so we write it as $P^S_{c_Bs} = P^{S,1h}_{c_Bs} + P^{S,2h}_{c_Bs}$. All other power spectra terms in equation 4.7 are purely 2-halo contributions.

Similarly, we can express the satellite auto power spectrum as

$$P^S_{ss}(\boldsymbol{k}) = (1 - f_{s_B})^2 P^S_{s_As_A}(\boldsymbol{k}) + 2 f_{s_B}(1 - f_{s_B}) P^S_{s_As_B}(\boldsymbol{k}) + f^2_{s_B} P^S_{s_Bs_B}(\boldsymbol{k}). \tag{4.8}$$

As in the case of $P^S_{c_Bs}$, the term $P^S_{s_Bs_B}$ includes both 1-halo and 2-halo contributions, which we can express as $P^S_{s_Bs_B} = P^{S,1h}_{s_Bs_B} + P^{S,2h}_{s_Bs_B}$. All other terms in equation 4.8 include only 2-halo contributions.

Combining the terms in equations 4.7 and 4.8, the galaxy power spectrum in redshift space is

$$P^S_{gg}(\boldsymbol{k}) = P^{S,1h}_{gg}(\boldsymbol{k}) + P^{S,2h}_{gg}(\boldsymbol{k}), \tag{4.9}$$

where the 2-halo contributions are given by

$$\begin{aligned} P^{S,2h}_{gg}(\boldsymbol{k}) = (1 - f_s)^2 P^S_{cc} & \\ &+ 2 f_s (1 - f_s) \left\{ (1 - f_{c_B}) \left[(1 - f_{s_B}) P^S_{c_As_A} + f_{s_B} P^S_{c_As_B} \right] \right\} \\ &+ 2 f_s (1 - f_s) \left\{ f_{c_B} \left[(1 - f_{s_B}) P^{S,2h}_{c_Bs_A} + f_{s_B} P^{S,2h}_{c_Bs_B} \right] \right\} \\ &+ f^2_s \left[(1 - f_{s_B})^2 P^S_{s_As_A} + 2 f_{s_B} (1 - f_{s_B}) P^S_{s_As_B} + f^2_{s_B} P^{S,2h}_{s_Bs_B} \right], \end{aligned} \tag{4.10}$$

and the 1-halo contributions are

$$\begin{aligned} P^{S,1h}_{gg}(\boldsymbol{k}) &= 2 f_s (1 - f_s) \left\{ f_{c_B} \left[(1 - f_{s_B}) P^{S,1h}_{c_Bs_A} + f_{s_B} P^{S,1h}_{c_Bs_B} \right] \right\} \\ &\quad + f^2_s f^2_{s_B} P^{S,2h}_{s_Bs_B}. \end{aligned} \tag{4.11}$$

Note that our halo model formalism is this section assumes that halos hosting satellite galaxies also host centrals. There is some observational evidence that this expectation may be violated in realistic data, e.g., Skibba et al. (2011). We can easily generalize the formalism presented in this section to include such effects, at the cost of introducing additional model parameters. However, past halo model studies (Guo et al. 2015b; Reid et al. 2014) have explored such generalizations and found the results insensitive to such model extensions. For this reason, we do not explicitly explore these generalizations in this work.

Modeling 1-halo and 2-halo terms in redshift space

In this subsection, we discuss how we model the 1-halo and 2-halo terms in redshift space that enter into equations 4.10 and 4.11. Our modeling of these terms in redshift space largely follows the work presented in Okumura et al. (2015); for completeness, we reproduce the relevant results from that work below.

Modeling complications arise due to the effects of the radial distribution of satellite galaxies inside halos on the galaxy power spectrum. In real space, correlations between galaxies on small scales give rise to the 1-halo term. In Fourier space, the 1-halo term manifests as a white noise-like term at low k, with departures from white noise at larger k due to the radial profile of satellites inside halos. As shown in Okumura et al. (2015), the deviations from white noise are small on the scales of interest for cosmological parameter inference ($k \lesssim 0.4\ h\mathrm{Mpc}^{-1}$). Thus, we treat all 1-halo terms in real space as independent of wavenumber. As discussed in Section 4.4.1, there are two sources of 1-halo terms: 1) the correlation between the c_B sample of centrals and satellites and 2) the auto-correlation between the s_B sample of satellites. We denote the real-space amplitude of these terms as $N_{c_B s}$ and $N_{s_B s_B}$, respectively.

In redshift space, satellite galaxies are spread out in the radial direction by their large virial velocities inside halos, an effect known as Fingers-of-God (Jackson 1972). Affecting both 1-halo and 2-halo correlations, the FoG effect is a fully nonlinear process, and it is not possible to accurately model it using perturbation theory. Quasi-linear perturbative approaches have been developed which use damping functions, i.e., a Gaussian or Lorentzian, to model the effect (Scoccimarro 2004; Taruya et al. 2010; Peacock & Dodds 1994; Park et al. 1994; Percival & White 2009). In previous studies, the effect is typically modeled with a single damping factor $G(k\mu;\sigma_v)$, with σ_v corresponding to the velocity dispersion of the full galaxy sample. In such a model, the redshift-space power spectrum of galaxies is modeled as $P^S_{gg}(k,\mu) = G^2(k\mu;\sigma_v)P^S_{hh}(k,\mu)$, where P^S_{hh} is the redshift-space halo power spectrum.

We separately model the FoG effect from each of the galaxy subsamples defined in Section 4.4.1. As demonstrated in Okumura et al. (2015), the FoG effect on the 1-halo and 2-halo terms from satellite galaxies can be accurately described using a damping function and the typical virial velocity associated with the halos hosting the satellites. The functional form of the damping function used in this work is

$$G(k\mu;\sigma_v) = \left(1 + k^2\mu^2\sigma_v^2/2\right)^{-2}, \tag{4.12}$$

which has a form slightly modified from the commonly used forms in the literature. As shown in Okumura et al. (2015), this damping function can accurately model the FoG effect from satellites over a wide range of scales, extending down to $k \sim 0.4\ h\mathrm{Mpc}^{-1}$. The dominant FoG effect arises from satellites, and we include velocity dispersion parameters for each of the satellite subsamples, σ_{v,s_A} and σ_{v,s_B}. Recent clustering analyses, e.g., Reid et al. (2014); Guo et al. (2015a), have also found evidence that central galaxies are not at rest with respect to the halo center, giving rise to an additional FoG effect (albeit smaller than the effect from satellites). To properly account for this possibility, we also include a velocity dispersion parameter associated with central galaxies, $\sigma_{v,c}$. We assume a single velocity dispersion for both the c_A and c_B galaxy samples.

There are five power spectra in equation 4.10 that include only 2-halo terms. Using the above FoG modeling, these terms become

$$P^s_{cc}(k,\mu) = G(k\mu;\sigma_{v,c})^2 P^S_{cc,h}(k,\mu) + \Sigma^2_{cc}, \tag{4.13}$$

$$P^S_{c_A s_A}(k,\mu) = G(k\mu;\sigma_{v,c}) G(k\mu;\sigma_{v,s_A}) P^S_{c_A s_A,h}(k,\mu), \tag{4.14}$$

$$P^S_{c_A s_B}(k,\mu) = G(k\mu;\sigma_{v,c}) G(k\mu;\sigma_{v,s_B}) P^S_{c_A s_B,h}(k,\mu), \tag{4.15}$$

$$P^S_{s_A s_A}(k,\mu) = G(k\mu;\sigma_{v,s_A})^2 P^S_{s_A s_A,h}(k,\mu) + \Sigma^2_{s_A s_A}, \tag{4.16}$$

$$P^S_{s_A s_B}(k,\mu) = G(k\mu;\sigma_{v,s_A}) G(k\mu;\sigma_{v,s_B}) P^S_{s_A s_B,h}(k,\mu), \tag{4.17}$$

where $P^S_{XX,h}$ represents the auto power spectrum of halos in which the galaxies of type X reside, Σ^2_{XX} is the shot noise contribution to the auto power from self pairs, and $P^S_{XY,h}$ is the cross spectrum of halos in which galaxies of types X and Y reside. Under the assumption of linear perturbation theory, $P^S_{XY,h}$ converges to the linear redshift-space power spectrum originally proposed by Kaiser (1987), $P^S_{XY,h}(k,\mu) = (b_{1,X} + f\mu^2)(b_{1,Y} + f\mu^2)P_L(k) + \delta^K_{XY}\Sigma^2_{XY}$, where $P_L(k)$ is the linear matter power spectrum, δ^K_{XY} is the Kronecker delta, Σ^2_{XX} is the shot noise contribution, which is non-zero only for auto spectra, and b_1 is the linear bias factor of the specified galaxy sample.

The three power spectra that include both 1-halo and 2-halo terms can be expressed as

$$P^S_{c_B s_A}(k,\mu) = G(k\mu;\sigma_{v,c}) G(k\mu;\sigma_{v,s_A}) \left[P^S_{c_B s_A,h}(k,\mu) + N_{c_B s}\right], \tag{4.18}$$

$$P^S_{c_B s_B}(k,\mu) = G(k\mu;\sigma_{v,c}) G(k\mu;\sigma_{v,s_B}) \left[P^S_{c_B s_B,h}(k,\mu) + N_{c_B s}\right], \tag{4.19}$$

$$P^S_{s_B s_B}(k,\mu) = G(k\mu;\sigma_{v,s_B})^2 \left[P^S_{s_B s_B,h}(k,\mu) + N_{s_B s_B}\right] + \Sigma^2_{s_B s_B}, \tag{4.20}$$

where $N_{c_B s}$ is the 1-halo amplitude due to correlations between centrals and satellites in the same halo, and $N_{s_B s_B}$ is the 1-halo amplitude for satellites inside the same halo.

4.4.2 Halo clustering in redshift space

The remaining modeling unknown needed in equations 4.13–4.17 and 4.18–4.20 is the prescription for the redshift-space halo power spectrum, $P^S_{XY,h}(k,\mu)$. In this section, we describe our model for the halo power spectrum, which relies on a combination of perturbation theory and simulations. The model is based on the formalism presented in Vlah et al. (2013), with important differences and improvements discussed below.

Distribution function model for redshift-space distortions

Our model for the power spectrum of halos in redshift space relies on expressing the redshift-space halo density field in terms of moments of the distribution function (DF); the approach has been developed and tested in a previous series of papers (Seljak & McDonald 2011; Okumura et al. 2012a,b; Vlah et al. 2012, 2013; Blazek et al. 2014). If we consider halo samples X and Y, with linear biases $b_{1,X}$ and $b_{1,Y}$, the redshift-space power spectrum in the DF model can be expressed as a sum over mass-weighted velocity correlators

$$P^S_{XY,h}(k,\mu) = \sum_{L=0}^{\infty}\sum_{L'=0}^{\infty} \frac{(-1)^{L'}}{L!L'!}\left(\frac{ik\mu}{\mathcal{H}}\right)^{L+L'} P^{XY,h}_{LL'}(k,\mu), \tag{4.21}$$

where $\mathcal{H} = aH$ is the conformal Hubble parameter, and $P^{XY,h}_{LL'}$ is the power spectrum of the moments L and L' of the radial halo velocity field, weighted by the halo density field. These spectra are defined as

$$(2\pi)^3 P^{XY,h}_{LL'}(\boldsymbol{k})\delta_D(\boldsymbol{k}+\boldsymbol{k}') = \left\langle T^{X,L}_{\parallel}(\boldsymbol{k}) T^{Y,L'}_{\parallel}(\boldsymbol{k}')\right\rangle, \tag{4.22}$$

where $T^{X,L}_{\parallel}(\boldsymbol{k})$ is the Fourier transform of the corresponding halo velocity moment weighted by halo density,

$$T^{X,L}_{\parallel}(\boldsymbol{x}) = \left[1 + \delta^h_X(\boldsymbol{x})\right]\left(v^h_{\parallel,X}\right)^L, \tag{4.23}$$

where δ^h_X and $v^h_{\parallel,X}$ are the halo density and radial velocity fields for sample X, respectively. The velocity correlators defined in equation 4.22 have well-defined physical interpretations; for example, $P^{XX,h}_{00}$ represents the halo density auto power spectrum of sample X, whereas $P^{XX,h}_{01}$ is the cross-correlation of density and radial momentum for halo sample X. The DF approach naturally produces an expansion of $P^S_{XY,h}(k,\mu)$ in even powers of μ, with a finite number of correlators contributing at a given power of μ. For this work, we consider terms up to and including the μ^4 order in the expansion of equation 4.21.

To evaluate the halo velocity correlators in equation 4.21, we largely follow the results outlined in Vlah et al. (2012, 2013), where the correlators are evaluated using Eulerian perturbation theory. However, in order to increase the overall accuracy of the power spectrum model, our work differs from the results presented in Vlah et al. (2013) in several crucial areas. These differences will be discussed in the subsequent subsections of this section.

The modeling of halo bias

The spectra $P^{XY,h}_{LL'}(k,\mu)$ in equation 4.21 are defined with respect to the halo field, and a biasing model is needed to relate them to the correlators of the underlying dark matter density field. Following the results of Vlah et al. (2013), we use a nonlocal and nonlinear Eulerian biasing model, such that

$$\delta^h(\boldsymbol{x}) = b_1\delta(\boldsymbol{x}) + \frac{b_2}{2}(\delta^2(\boldsymbol{x}) - \langle\delta^2\rangle) + \frac{b_s}{2}(s^2(\boldsymbol{x}) - \langle s^2\rangle) + \frac{b_3}{6}\delta^3(\boldsymbol{x}), \tag{4.24}$$

where δ^h and δ are the halo and dark matter overdensity fields, respectively, b_i are the various bias parameters, and s^2 is the contracted tidal tensor contribution (see e.g., Baldauf et al. (2013) for the definition). We employ a modified version of equation 4.24 that uses four parameters: $\{b_1, b_2^{00}, b_2^{01}, b_s\}$. As discussed in Vlah et al. (2013), the spectra $P^{XX,h}_{00}$ and

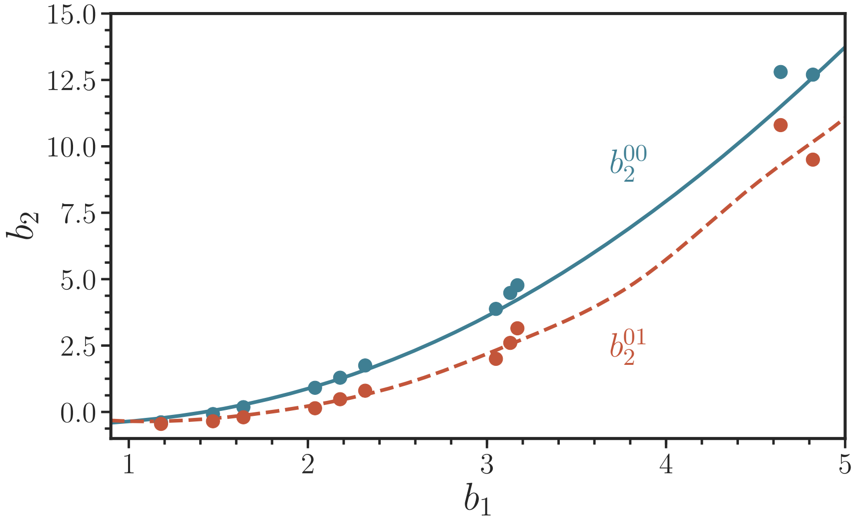


Figure 4.1: The dependence of the second-order nonlinear effective biases, b_2^{00} (blue, solid) and b_2^{01} (red, dashed), on the linear bias b_1 that is used in this work, as determined from the RunPB simulations. For comparison, the best-fit bias parameters from Vlah et al. (2013) are shown as circles.

$P_{01}^{XX,h}$ have distinct values for the quadratic, local bias b_2, resulting in the parameters b_2^{00} and b_2^{01}. This difference can be understood through the effects of the third-order, nonlocal bias $b_3^{\rm NL}$, which appears to be equally important to b_2 (McDonald & Roy 2009; Saito et al. 2014) (see further discussion in Vlah et al. (2013)). We would have obtained similar results using a biasing scheme with parameters: $\{b_1, b_2, b_s, b_3^{\rm NL}\}$. Note that the nonlinear and nonlocal biasing contributions have been demonstrated to improve the accuracy of theoretical models, e.g., Baldauf et al. (2013); Saito et al. (2014).

In summary, our biasing scheme has four bias parameters for each halo sample: the linear bias b_1, the two effective second-order biases, b_2^{00} and b_2^{01}, and the nonlocal tidal bias b_s. However, we treat the higher-order biases as functions of b_1, and thus, the only free bias parameter for each halo sample is the linear bias. In the case of the local Lagrangian bias model, we can predict the amplitude of the nonlocal tidal bias in terms of the linear bias (Chan et al. 2012; Baldauf et al. 2012)

$$b_s = -\frac{2}{7}(b_1 - 1). \tag{4.25}$$

As shown in Vlah et al. (2013), the tidal bias does not play a prominent role in the biasing model, but nonetheless, we include these terms in our model. The effective biases b_2^{00} and b_2^{01} have a roughly quadratic dependence on the linear bias b_1. Rather than freely varying these bias parameters, we treat them as a function of b_1, independent of redshift, and use simulations to determine the exact functional form of this dependence. We use the set of halo mass bins from the RunPB simulations described in Section 4.2.1 and use Gaussian

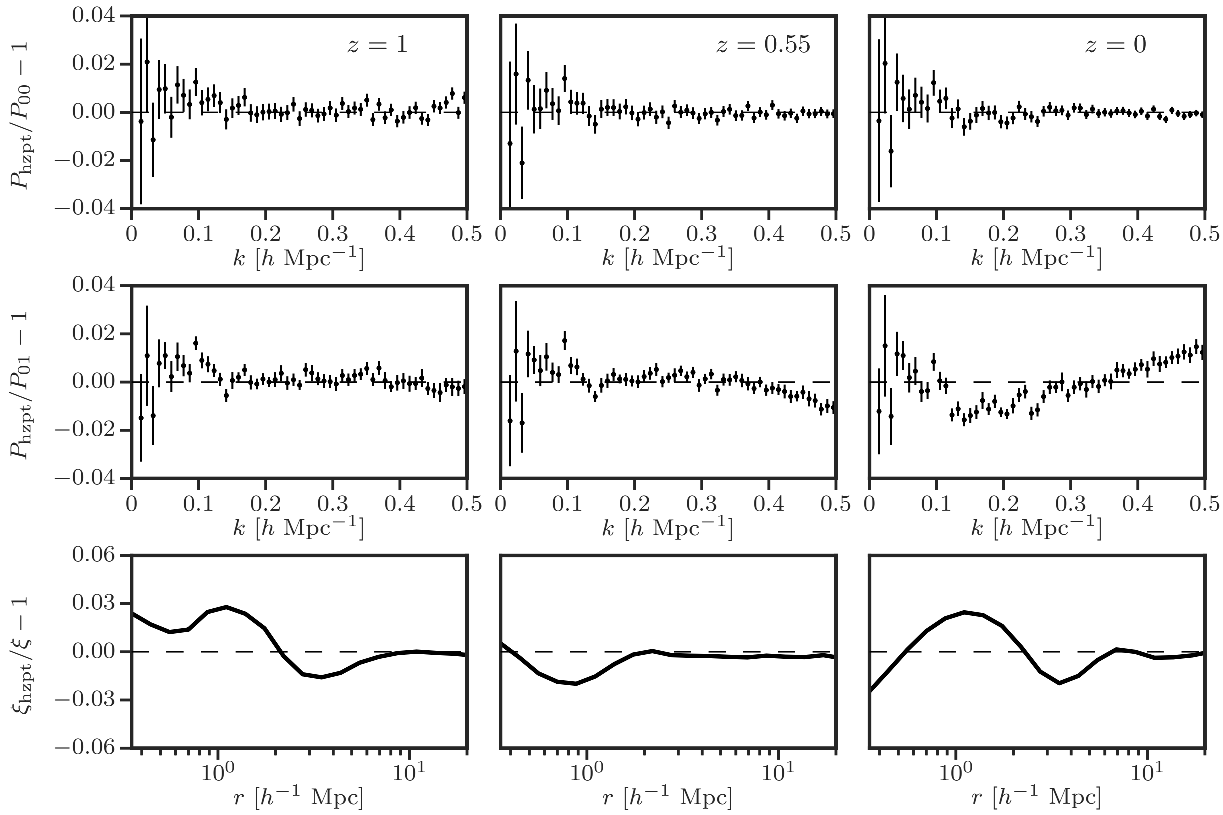


Figure 4.2: The accuracy of the dark matter HZPT modeling results used in this work, in comparison to the results from the RunPB simulation. We compare the dark matter power spectrum P_{00} (top), density – radial momentum cross power P_{01} (middle), and the small-scale correlation function ξ_{00} (bottom). We give results for three redshift outputs: $z = 1$ (left), $z = 0.55$ (center), and $z = 0$ (right). The updated HZPT model parameters are presented in Appendix B.1.

Process regression (see, e.g., Rasmussen & Williams 2006) to predict the functional form of $b_2^{00}(b_1)$ and $b_2^{01}(b_1)$. For this purpose, we use the public Gaussian Process package `george`[5] (Ambikasaran et al. 2015). The predictions for b_2^{00} and b_2^{01} used in this work, as determined from the RunPB simulations, are shown in Figure 4.1. We also show the best-fit bias parameters used in Vlah et al. (2013) for several redshifts, which are consistent with the results obtained from the RunPB simulations.

Improved modeling of dark matter correlators

We use the Halo-Zel'dovich Perturbation Theory (HZPT) approach of Seljak & Vlah (2015) to model the real-space dark matter density power spectrum $P_{00}(k)$ and the real-space cross-correlation of dark matter density and radial momentum $P_{01}(\boldsymbol{k})$. This differs from the

[5] https://github.com/dfm/george

results presented in Vlah et al. (2013), which uses standard perturbation theory (SPT) to evaluate these terms (which is known to break down at relatively large scales). Note that the density-momentum cross-correlation can be related to P_{00} through $P_{01}(\boldsymbol{k}) = \mu^2 dP_{00}/d\ln a$ (Vlah et al. 2012), so only an accurate model for P_{00} is required. The dark matter correlator $P_{01}(\boldsymbol{k})$ plays a crucial role in the modeling of the μ^2 angular dependence of the redshift-space power spectrum of halos.

The HZPT model connects the Zel'dovich approximation (Zel'dovich 1970; White 2014) with a Padé expansion for a 1-halo-like term that is determined from simulations using simple, physically motivated parameter scalings. The model for the dark matter power spectrum has been demonstrated to be accurate to 1% to $k \sim 1\ h\mathrm{Mpc}^{-1}$ (Seljak & Vlah 2015), and the Zel'dovich approximation performs sufficiently well when modeling BAO, relative to other modeling techniques (White 2014; Vlah et al. 2015).

We provide an update to the HZPT results presented in Seljak & Vlah (2015), using the dark matter RunPB simulations detailed in Section 4.2.1. We extend the analysis of Seljak & Vlah (2015) to include measurements of $P_{01}(\boldsymbol{k})$, as well as the small-scale matter correlation function. We also extend the redshift fitting range, using a set of 10 redshift outputs from the RunPB simulations, ranging from $z = 0$ to $z = 1$. We perform a global fit of the amplitude and redshift-dependence of the 5 parameters in the HZPT model using the $P_{00}(k)$ and $P_{01}(\boldsymbol{k})$ statistics over the range $k = 0.005 - 0.5\ h\mathrm{Mpc}^{-1}$, as well as the small-scale correlation function over the range $r = 0.3 - 25\ \mathrm{Mpc}/h$. Qualitatively, the results remain similar to those presented in Seljak & Vlah (2015), but the use of additional statistics (in particular, the small-scale correlation function) in the fit does allow some parameter degeneracies to be broken.

We review the HZPT model for P_{00} and P_{01} in Appendix B.1, and provide the updated best-fit model parameters. We also detail the necessary calculation of P_{01} in the Zel'dovich approximation in Appendix A. We show the accuracy of the HZPT model for the three statistics considered in Figure 4.2 for three redshift snapshots, $z = 0$, 0.55, and 1. It is evident that the 5-parameter HZPT model can provide a consistent picture of the power spectra to an accuracy of 1-2% over the range of scales considered in this work. Furthermore, keeping in mind that the inclusion of baryonic effects can effect the parameters R_1, R_{1h}, R_{2h} at the 5-10% level (van Daalen et al. 2011; Mohammed & Seljak 2014), the model used in this work performs reasonably well at modeling the notoriously difficult 1-halo to 2-halo regime of the correlation function.

We also extend the HZPT approach to model the dark matter radial momentum auto power spectrum, $P_{11}(\boldsymbol{k})$, which is important for modeling the μ^4 angular dependence of $P^S_{XY,h}(k,\mu)$. Specifically, we model the scalar component of $P_{11}[\mu^4]$ with the sum of a Zel'dovich term and Padé expression and model the vector contribution using 1-loop SPT (as was done in Vlah et al. 2012). The full model is given by the sum of the scalar and vector contributions

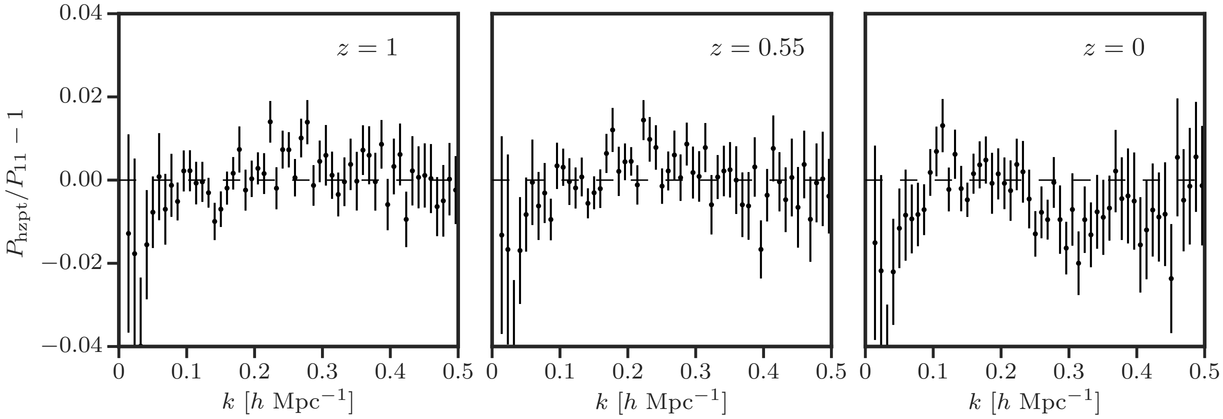


Figure 4.3: The accuracy of the HZPT model for the auto power spectrum of the dark matter radial momentum $P_{11}[\mu^4]$ in comparison to the results from the RunPB simulation. We give results for three redshift outputs: $z = 1$ (left), $z = 0.55$ (center), and $z = 0$ (right). The best-fit HZPT model parameters are presented in Appendix B.2.

$$\begin{aligned} P_{11}[\mu^4](k) &= P^S_{11,s}[\mu^4](k) + P^S_{11,v}[\mu^4](k), \\ &= P^{\rm zel}_{11,s}(k) + P^{BB}_{11}(k) - f^2 I_{31}(k), \end{aligned} \tag{4.26}$$

where the vector contribution $I_{31}(k)$ is defined in Vlah et al. (2012), and we discuss the Zel'dovich term $P^{\rm zel}_{11,s}$ in detail in Appendix A. We define the Padé term P^{BB}_{11} and give the best-fit parameters (fit using the RunPB simulations) in Appendix B.2. Figure 4.3 compares the accuracy of the model in equation 4.26 with the results from the RunPB simulations for three redshift snapshots. The figure shows the model to be accurate to 1-2% over the range of scales of interest.

Finally, we improve the modeling of the the dark matter density – velocity divergence cross power spectrum $P_{\delta\theta}(k)$ and the velocity divergence auto power spectrum $P_{\theta\theta}(k)$ in comparison to Vlah et al. (2013). An accurate model for $P_{\delta\theta}$ is needed to describe the μ^2 dependence of the halo density – momentum cross spectrum $P^{hh}_{01}(\boldsymbol{k})$, and the model for $P_{\theta\theta}$ enters into our model for the μ^4 dependence of the halo momentum auto spectrum $P^{hh}_{11}(\boldsymbol{k})$. Rather than using the 1-loop SPT expressions for these terms (as was done in Vlah et al. 2013), we use the fitting formula from Jennings (2012). While the 1-loop SPT expressions for $P_{\delta\theta}$ and $P_{\theta\theta}$ diverge from the truth at relatively large scales ($k \sim 0.1\ h$Mpc^{-1}), the model of Jennings (2012) achieves the necessary accuracy over the range of scales considered in this work.

Halo stochasticity

The μ^0 component of the redshift-space halo spectrum, $P^S_{XY,h}(k, \mu)$, in the DF model is the isotropic, real-space auto spectrum of the halo density field, $P^{hh}_{00}(k)$. For a complete description of this term, we must accurately model the contribution from the stochasticity of halos (Hamaus et al. 2010), defined for two separate halo mass bins (h and $\bar{h}$) as

$$\Lambda(k) = P_{00}^{h\bar{h}}(k) - \bar{b}_1(k)P_{00}^{hm}(k) - b_1(k)P_{00}^{\bar{h}m}(k) + b_1(k)\bar{b}_1(k)P_{00}(k), \tag{4.27}$$

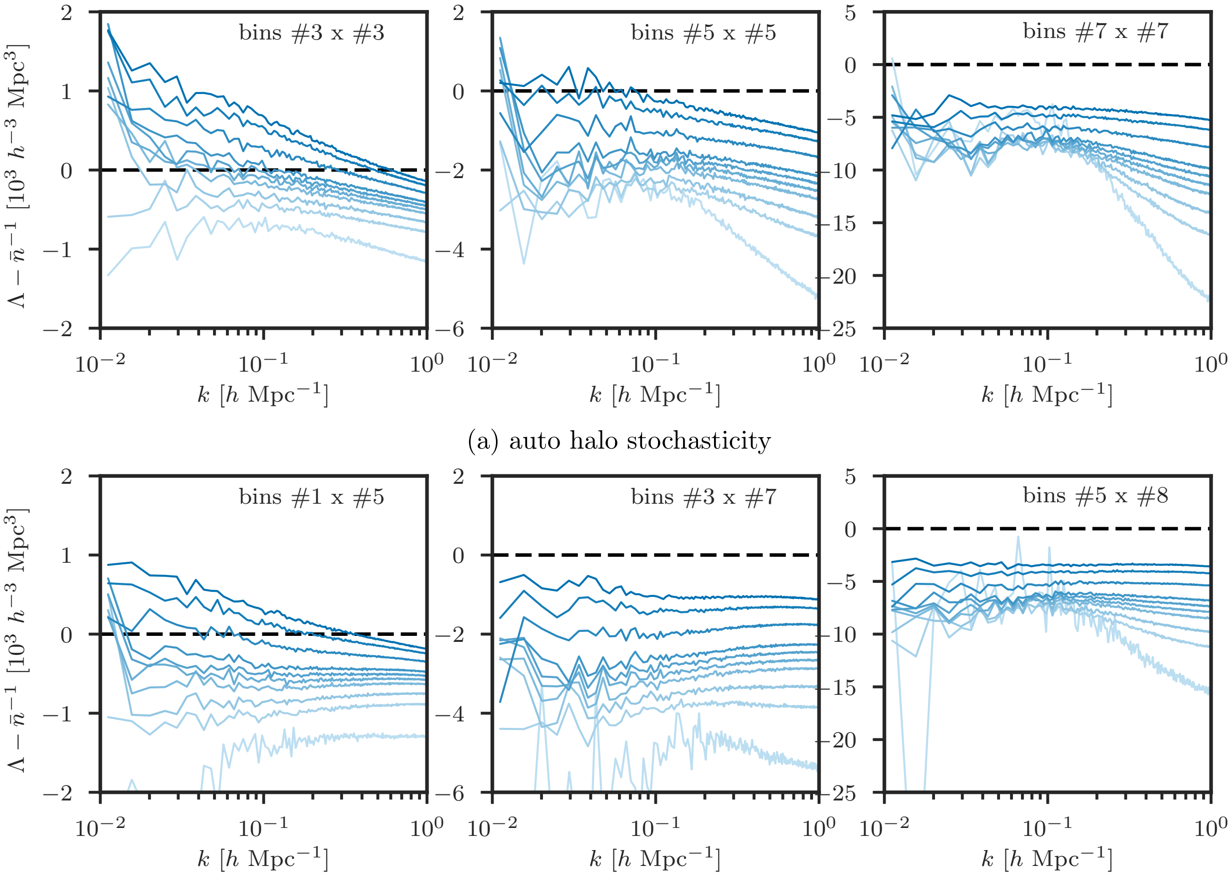
where P_{00}^{hm} and $P_{00}^{\bar{h}m}$ are the halo–matter cross power spectra for the halo mass bins h and $\bar{h}$, respectively, and P_{00} is the matter power spectrum (modeled using HZPT, as described in §4.4.2). The scale-dependent linear bias factors are defined as

$$b_1(k) \equiv \frac{P_{00}^{hm}(k)}{P_{00}(k)}, \qquad \bar{b}_1(k) \equiv \frac{P_{00}^{\bar{h}m}(k)}{P_{00}(k)}. \tag{4.28}$$

In the Poisson model, the leading-order term of the stochasticity is given by the Poisson shot noise, $\bar{n}^{-1}$, where $\bar{n}$ is the halo number density. However, there are significant deviations from this prediction that have complicated scale dependence. These corrections originate from two competing effects: first, the halo exclusion effects due to the finite size of halos and second, the nonlinear clustering of halos relative to dark matter (Baldauf et al. 2013; Vlah et al. 2013; Baldauf et al. 2016; Ginzburg et al. 2017). In the $k \to 0$ limit, the stochasticity behaves close to white noise, where halo exclusion lowers the stochasticity relative to the Poisson value and nonlinear clustering leads to a positive contribution. However, in the high-k limit, the stochasticity must approach the Poisson value, and these deviations vanish; thus, there exists a complicated scale dependence that is not yet well-understood theoretically.

We use the RunPB simulations at several redshift outputs and the halo mass bins defined in Table 4.2 to investigate the functional form of the halo stochasticity $\Lambda(k)$ as a function of mass and redshift. In Figure 4.4, we show the deviations of the halo stochasticity from the Poisson shot noise when considering the same halo mass bin and different halo mass bins. The trends are consistent with our theoretical understanding: as the average halo mass increases, the stochasticity becomes sub-Poissonian, sourced by halo exclusion effects, while positive contributions from nonlinear biasing become important for lower halo masses. As halos grow in time, the exclusion effects become more pronounced at lower redshifts (Baldauf et al. 2013; Vlah et al. 2013; Baldauf et al. 2016). However, the scale dependence and redshift scaling remains non-trivial, and there are significant differences in the scale dependence and amplitude when considering the cases of auto and cross mass bins. The stochasticity Λ eventually approaches the Poisson value, but only on very small scales ($k \sim 10$ hMpc^{-1}), which are not shown in Figure 4.4.

The halo stochasticity was studied in simulations in the context of the DF model in Vlah et al. (2013), and the results presented here agree with those findings. Vlah et al. (2013) employs a simple model with log scale dependence to model the auto stochasticity for several mass bins across three redshifts. We extend those results with finer resolution in both redshift and halo mass. In an attempt to capture as much complexity as possible, we treat the halo stochasticity results from the RunPB simulations as a training set and use Gaussian Process regression to predict the auto stochasticity $\Lambda(b_1, \sigma_8(z))$ and cross stochasticity $\Lambda(b_1, \bar{b}_1, \sigma_8(z))$, where we have parameterized the redshift dependence of the stochasticity



(a) auto halo stochasticity

(b) cross halo stochasticity

Figure 4.4: The deviation of the halo stochasticity $\Lambda(k)$, as defined in equation 4.27, from the Poisson shot noise, for the case of (a) the same halo mass bin and (b) different halo mass bins. Results are measured from the RunPB simulations for three separate combinations of bins. The average halo mass increases from left to right; see Table 4.2 for halo mass bin details. For each subplot, we show the results for 10 redshifts, ranging from $z = 1$ (dark) to $z = 0$ (light). Even when the mean halo mass is similar, the scale dependence and amplitude in the cases of auto and cross halo stochasticity can differ significantly.

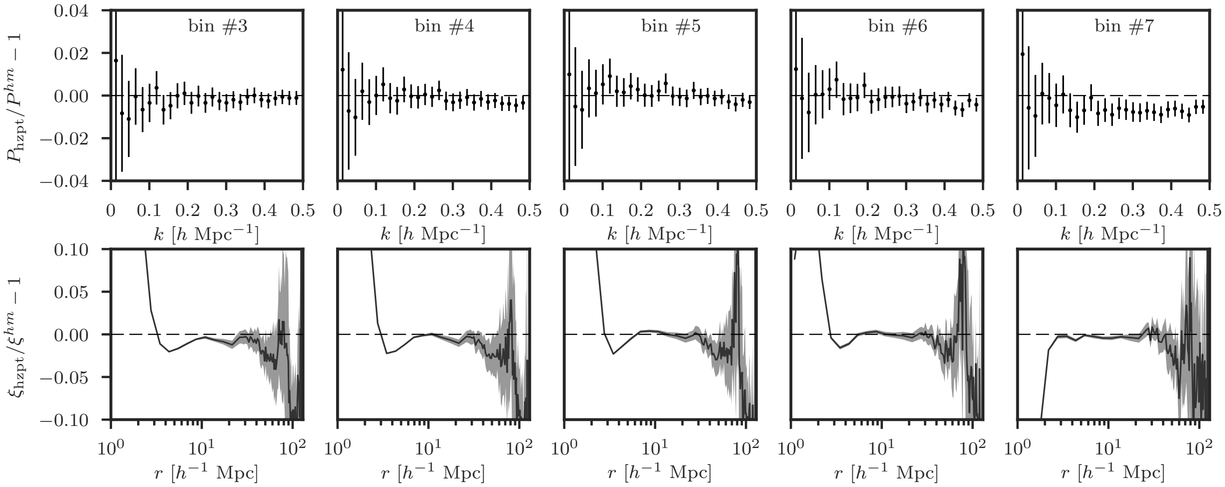


Figure 4.5: The accuracy of the HZPT model used in this work for the halo-matter cross-correlation, in comparison to the results from the RunPB simulation. We compare the cross power spectrum P_{hm} (top) and the correlation function ξ^{hm} (bottom) for 5 halo mass bins at $z = 0.55$ (see Table 4.2 for bin details). We show the measurement uncertainties as error bars for P^{hm} and as the grey shaded region for ξ^{hm}. The HZPT parameters have been fit using only $P^{hm}(k)$ from $0.005\ h\text{Mpc}^{-1} < k < 0.5\ h\text{Mpc}^{-1}$. The model is a good description of P^{hm} on these scales, as well as ξ^{hm} down to $r \sim 5$ Mpc/h, but fails once entering the 1-halo regime on small scales.

using the value of σ_8 at each redshift. The stochasticity of biased tracers is an active area of theoretical work, and we hope to incorporate analytic prescriptions for halo stochasticity based on existing work, i.e., Baldauf et al. (2013, 2016); Ginzburg et al. (2017), in to our model in the near future.

HZPT modeling for the halo-matter cross-correlation

The real-space halo-matter cross correlation $P^{hm}(k)$ plays a crucial role in accurately modeling the halo auto spectrum using equation 4.27. We develop a new model for the halo-matter power spectrum using HZPT and calibrate the model using the suite of halo mass bins from the RunPB simulations detailed in Table 4.2. To model the Zel'dovich term of the model, we employ a simple, linear bias model, such that the full HZPT model is given by

$$P^{hm}(k) = b_1 P_{00}^{\rm zel}(k) + P_{00}^{BB}(k, A_0, R, R_1, R_{1h}, R_{2h}), \tag{4.29}$$

where b_1 is the large-scale, linear bias of the halo field, $P_{00}^{\rm zel}$ is the matter density auto spectrum in the Zel'dovich approximation, and P_{00}^{BB} is a broadband Padé term, as given by equation B.1. To account for the biased nature of the halo field, we treat the HZPT model parameters, $\{A_0, R, R_1, R_{1h}, R_{2h}\}$, as a function of not only $\sigma_8(z)$ but also the linear bias b_1. We choose a simple power law functional form for the b_1 dependence, which performs well at modeling the bias dependence of P^{hm} over the range of scales of interest in this work. We perform a global fit across the 8 halo mass bins and 10 redshifts of the RunPB simulations to determine the best-fit HZPT model parameters. In our parameter fit, we have included the cross power spectrum P^{hm} on scales ranging from $k = 0.005\ h\text{Mpc}^{-1}$ to $k = 0.5\ h\text{Mpc}^{-1}$. The best-fit parameters are presented in Appendix B.3.

We show the accuracy of the halo-matter HZPT model in Figure 4.5 for several halo mass bins at $z = 0.55$. The trends evident at this redshift are consistent with the results from the full range of redshifts explored ($z = 0 - 1$). The model reproduces the cross power spectrum P^{hm} at the ~2% level, as well as the cross-correlation ξ^{hm} on scales $r \gtrsim 5$ Mpc/h. However, we see from the correlation function results on small scales that the model is unable to reproduce the clustering on scales within the 1-halo regime, where halo profile details become important. The model breakdown on these scales is due to our choice to use a power law dependence on b_1 for the HZPT parameters that are related to the halo profile. To better describe halo profiles and capture the effects of nonlinear and nonlocal bias terms, i.e., Saito et al. (2014), a more complicated functional form for the linear bias dependence is required. However, because Fourier-space statistics are the main concern of this work and the simplified model performs well when modeling the power spectrum on the scales of interest, we leave the investigation of improved small-scale modeling to future work.

4.4.3 Modeling observational effects

In this section we discuss several details that arise when modeling data from real galaxy surveys. In Section 4.4.3, we describe how we account for the geometric distortions of the clustering signal that arise when an inaccurate fiducial cosmology is assumed. Section 4.4.3 discusses how we treat the survey geometry and window function when modeling "cutsky" mocks, which have a realistic survey geometry imposed.

The Alcock-Paczynski effect

When analyzing data from galaxy surveys, we must transform observed angular positions and redshifts into physical coordinates, using a fiducial cosmological model to specify the relation between the redshift and the LOS distance (i.e., the Hubble parameter) and between the angular separation and the distance perpendicular to the LOS (i.e., the angular diameter distance). If the fiducial cosmology differs from the true cosmology, an anisotropic, geometric warping of the clustering signal is introduced. This distortion, known as the Alcock-Paczynski (AP) effect, (Alcock & Paczynski 1979) is distinct from RSD and can be used to measure cosmological parameters. The presence of the BAO feature at a fixed scale in the power spectrum helps distinguish the geometric AP effect and the dynamical RSD anisotropy, thus increasing the constraining power of full-shape modeling (Shoji et al. 2009; Ballinger et al. 1996).

The difference between the assumed and true cosmological models results in a rescaling of the wavenumbers transverse $k_\perp$ and parallel $k_\parallel$ to the LOS direction, such that

$$k'_\perp = q_\perp k_\perp \text{ and } k'_\parallel = q_\parallel k_\parallel, \tag{4.30}$$

where the primes denote quantities that are observed assuming the fiducial (and possibly incorrect) cosmology. The two distortion parameters $q_\perp$ and $q_\parallel$ are given by

$$q_\perp = \frac{D_A(z_{\text{eff}})}{D'_A(z_{\text{eff}})} \text{ and } q_\parallel = \frac{H'(z_{\text{eff}})}{H(z_{\text{eff}})}, \tag{4.31}$$

which are the ratios of the Hubble parameter and angular diameter distance in the fiducial and true cosmologies at the effective redshift of the survey. With these definitions, the theoretical prediction for the multipole power spectrum when including the AP effect can be expressed as

$$P_\ell(k') = \frac{2\ell+1}{2q_\perp q_\parallel^2} \int_{-1}^{1} d\mu \; P^S_{gg}\left[k(k',\mu'),\mu(\mu')\right] \mathcal{L}_\ell(\mu), \tag{4.32}$$

where $\mathcal{L}_\ell$ is the Legendre polynomial of order ℓ, and we use the model prediction of equation 4.9 for $P^S_{gg}[k(k',\mu'),\mu(\mu')]$. The true (k,μ) can be related to the observed (k',μ') via

$$k(k',\mu') = \frac{k'}{q_\perp}\left[1 + (\mu')^2\left(\frac{1}{F^2} - 1\right)\right]^{1/2}, \tag{4.33}$$

$$\mu(\mu') = \frac{\mu'}{F}\left[1 + (\mu')^2\left(\frac{1}{F^2} - 1\right),\right]^{-1/2} \tag{4.34}$$

where $F = q_\parallel / q_\perp$. The normalization scaling of the power spectrum with $q_\perp^{-1} q_\parallel^{-2}$ is due to the volume distortion between the two different cosmologies.

For comparison with BAO distance analyses, a second set of AP parameters is usually defined, given by

$$\alpha_\perp \equiv \frac{D_A(z_{\text{eff}})}{D'_A(z_{\text{eff}})}\frac{r'_d}{r_d} = q_\perp \frac{r'_d}{r_d}, \tag{4.35}$$

$$\alpha_\parallel \equiv \frac{H'(z_{\text{eff}})}{H(z_{\text{eff}})}\frac{r'_d}{r_d} = q_\parallel \frac{r'_d}{r_d}, \tag{4.36}$$

where we have defined $r_d \equiv r_s(z_d)$ as the sound horizon scale at the drag redshift z_d. BAO measurements are sensitive to the Hubble parameter and angular diameter distance relative to the sound horizon scale of a fixed "template" cosmology, and this second set of parameter definitions facilitates comparison of measurements using different template cosmological models.

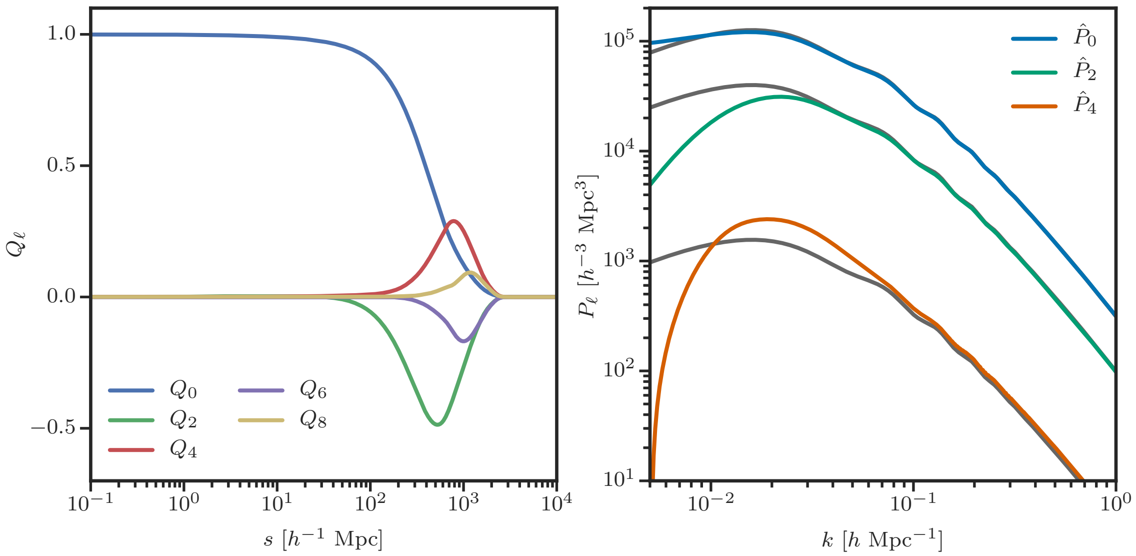


Figure 4.6: The window function multipoles in configuration space (left) and the effects of the window function on linear Kaiser power spectrum multipoles (right) for the DR12 CMASS NGC survey geometry. In the right panel, the solid grey lines show the original multipoles, while the colored lines correspond to the model after convolution with the window function, $\widehat{P}_\ell(k)$. The convolution procedure has large effects at small k, and we choose to use $k_{\rm min} = 0.02\ h{\rm Mpc}^{-1}$ in our data analysis to minimize these effects.

The survey geometry

When analyzing cutsky mock catalogs, we must account for the effects of the survey geometry when comparing our theoretical model to the measured power spectrum. We do this by convolving our theoretical model with the survey window function, rather than trying to remove the effect of the survey geometry from the data itself. Our window function treatment follows the method first presented in Wilson et al. (2017) and used in the analysis of BOSS DR12 data in Beutler et al. (2017b); Zhao et al. (2017).

Following Wilson et al. (2017), we compute the window function multipoles in configuration space using a pair counting algorithm and the catalog of random objects describing the survey geometry. We use the `Corrfunc` correlation function code (Sinha 2016) to compute the pair counts of the random catalog via

$$Q_\ell(s) \propto \int_{-1}^{1} d\mu RR(s,\mu)\mathcal{L}_\ell(\mu) \simeq \sum_i RR(s_i,\mu_i)\mathcal{L}_\ell(\mu_i), \tag{4.37}$$

where the normalization is such that $Q_0(s) \to 1$ for $s \ll 1$. The resulting multipoles Q_ℓ for the BOSS CMASS NGC sample are shown in the left panel of Figure 4.6. The Q_ℓ vanish for scales $\gtrsim 3000\ \ h^{-1}$Mpc, as these are the largest scales in the volume of the NGC. Note that on small scales, the clustering becomes isotropic, with the multipoles vanishing. In general, the contribution of the higher-order multipoles decreases as ℓ increases, which guarantees the convolution converges when including only the first few Q_ℓ. Here, we include multipoles up to and including Q_8, and have verified that the inclusion of Q_{10} does not affect our results.

With the measured Q_ℓ, we compute the convolved theoretical correlation function multipoles in configuration space as

$$\begin{aligned}
\widehat{\xi}_0(s) &= \xi_0 Q_0 + \frac{1}{5}\xi_2 Q_2 + \frac{1}{9}\xi_4 Q_4 + ... \\
\widehat{\xi}_2(s) &= \xi_0 Q_2 + \xi_2\left[Q_0 + \frac{2}{7}Q_2 + \frac{2}{7}Q_4\right] \\
&\quad + \xi_4\left[\frac{2}{7}Q_2\frac{100}{693}Q_4 + \frac{25}{143}Q_6\right] + ... \\
\widehat{\xi}_4(s) &= \xi_0 Q_4 + \xi_2\left[\frac{18}{35}Q_2 + \frac{20}{77}Q_4 + \frac{45}{143}Q_6\right] \\
&\quad + \xi_4\left[Q_0 + \frac{20}{77}Q_2 + \frac{162}{1001}Q_4 + \frac{20}{143}Q_6 + \frac{490}{2431}Q_8\right] + ..., \qquad (4.38)
\end{aligned}$$

where ξ_ℓ are the theoretical correlation function mutipoles, computed from the power spectrum multipoles via a 1D Hankel transform, evaluated using the `FFTLog` software (Hamilton 2000). We also perform the transformation from $\widehat{\xi}_\ell(s)$ to $\widehat{P}_\ell(k)$ using `FFTLog`.

The effects of the window function convolution can be seen in the right panel of Figure 4.6, where we illustrate the effects using linear Kaiser multipoles. The effects are most important on scales of order the survey size; for the NGC CMASS sample, the window function effects are only important on scales $k \lesssim 0.05\ h\mathrm{Mpc}^{-1}$. The impact of the survey geometry increases for the higher-order multipoles, with the anisotropy of the window function leading to non-trivial effects on our convolved model. In this work, we use a minimum wavenumber of $k_{\mathrm{min}} = 0.02\ h\mathrm{Mpc}^{-1}$ when comparing data and theory and have tested that the window function convolution has minimal impact on our parameter fitting analyses. However, as measurement errors decrease for future surveys, the window convolution will need to be carefully tested, given both the constraining power of the $\ell = 2$ and $\ell = 4$ multipoles and the larger convolution effects.

4.4.4 Model parametrization

Table 4.3 gives a summary of the parameters of the model described in this work. We give both the free parameters as well as the constrained parameters and the corresponding constraint expressions. There are 13 free parameters detailed in Table 4.3, and these parameters correspond to the parameter space used in our RSD analyses. The table also lists the assumed prior distribution for each parameter used during parameter estimation, which is either a flat (uniform) or Gaussian prior. We use physically motivated priors when possible and assume wide, flat priors on all cosmological parameters of interest. We describe the model parametrization in detail below.

Free Parameters	
Name [Unit]	Prior
$\alpha_\perp$	$\mathcal{U}(0.8, 1.2)$
$\alpha_\parallel$	$\mathcal{U}(0.8, 1.2)$
f	$\mathcal{U}(0.6, 1.0)$
$\sigma_8(z_{\text{eff}})$	$\mathcal{U}(0.3, 0.9)$
b_{1,c_A}	$\mathcal{U}(1.2, 2.5)$
f_s	$\mathcal{U}(0, 0.25)$
f_{s_B}	$\mathcal{U}(0, 1)$
$\langle N_{>1,s} \rangle$	$\mathcal{N}(2.4, 0.1)$
σ_c [h^{-1}Mpc]	$\mathcal{U}(0, 3)$
σ_{s_A} [h^{-1}Mpc]	$\mathcal{U}(2, 6)$
γ_{s_A}	$\mathcal{N}(1.45, 0.3)$
γ_{s_B}	$\mathcal{N}(2.05, 0.3)$
$f^{1h}_{s_B s_B}$	$\mathcal{N}(4, 1)$

Constrained Parameters	
Name [Unit]	Constraint Expression
b_{1,c_B}	equation C.7
b_{1,s_A}	$\gamma_{s_A} b_{1,c_A}$
b_{1,s_B}	$\gamma_{s_B} b_{1,c_A}$
f_{c_B}	equation C.2
σ_{s_B} [h^{-1} Mpc]	$\sigma_{s_A} \left[\sigma_v^{\text{model}}(b_{1,s_B}) / \sigma_v^{\text{model}}(b_{1,s_A})\right]$
$N_{c_B s}$ [h^{-3} Mpc^3]	equation C.12
$N_{s_B s_B}$ [h^{-3} Mpc^3]	equation C.16

Table 4.3: The parameter space of our full-shape RSD fits using the model described in this chapter. There are 13 free parameters (left) that are varied during the fitting process, with several additional parameters subject to constraint expressions (right). For all free parameters, we provide the prior used when fitting, either a normal prior $\mathcal{N}(\mu, \sigma)$ with mean μ and standard deviation σ, or a uniform prior $\mathcal{U}(a, b)$ with lower bound a and upper bound b. For a detailed description of the model parameters, see Section 4.4.4.

Cosmology parameters

The free parameters specifying the cosmology in our model are the AP distortion parameters, $\alpha_\parallel$ and $\alpha_\perp$, the growth rate f, and the amplitude of matter fluctuations σ_8, where both f and σ_8 are evaluated at the effective redshift of the sample, z_{eff}. During our fitting procedure, we vary f and σ_8 independently, although we only report results for the product $f\sigma_8$, which is the parameter combination most well-constrained by RSD analyses. The model requires a linear power spectrum in order to evaluate several perturbation theory integrals. These integrals are computationally costly (although see recent advances, Schmittfull & Vlah 2016; Schmittfull et al. 2016; McEwen et al. 2016), and for this reason, we do not vary any cosmological parameters affecting the shape of the linear power spectrum during parameter estimation. We evaluate the linear power spectrum using the fiducial cosmology and keep the shape fixed, allowing only the amplitude to vary through changes in σ_8. This choice assumes that the Planck (Planck Collaboration et al. 2016a) uncertainty for most of the parameters which define the shape of the power spectrum is much smaller than the uncertainty of our measurement and can be neglected. This has been shown to be a reasonable assumption for current data sets, i.e., Beutler et al. (2014b); Gil-Marín et al. (2016b).

Linear bias parameters

In the most general version of the model discussed in Section 4.4.1, we must specify linear bias parameters for each of the four galaxy subsamples: b_{1,c_A}, b_{1,c_B}, b_{1,s_A}, and b_{1,s_B}. As discussed in Section 4.4.2, the linear bias fully predicts the higher-order biasing parameters for a given sample. When varying the linear bias parameters of the s_A and s_B satellite samples, we enforce the expected ordering of the parameters: $b_{1,c_A} < b_{1,s_A} < b_{1,s_B}$. We use the relations $b_{1,s_A} = \gamma_{s_A} b_{1,c_A}$ and $b_{1,s_B} = \gamma_{s_B} b_{1,c_A}$ and choose to vary the parameters γ_{s_A} and γ_{s_B} instead. We use relatively wide Gaussian priors for these parameters centered on their expected fiducial values for a CMASS-like galaxy sample, $\gamma_{s_A} \sim 1.45$ and $\gamma_{s_B} \sim 2.05$ (see, e.g., Okumura et al. 2015, Table 1). For the linear bias of the c_B sample, we use the expected scaling of the bias with the biases of the satellite samples as given by equation C.7 and described in Appendix C.2.

Sample fractions, velocity dispersions, and 1-halo amplitudes

There are three sample fraction parameters: 1) f_s specifies the fraction of all galaxies that are satellites, 2) f_{c_B} gives the fraction of centrals that live in halos with a satellite, and 3) f_{s_B} defines the fraction of satellites that live in halos with multiple satellites. We must also specify the 1-halo amplitudes (assumed to be independent of k) that enter into equations 4.18–4.20. We denote the 1-halo amplitude due to correlations between centrals and satellites in the same halo as $N_{c_B s}$ and between satellites inside the same halo as $N_{s_B s_B}$. And, finally, we must specify the velocity dispersion parameters for each galaxy subsample in order to account for the FoG effect. We include a single velocity dispersion for centrals, σ_c, and parameters for each of the satellite subsamples, σ_{s_A} and σ_{s_B}. Thus, in the most general case, there are an additional 12 model parameters needed to fully evaluate our model, in addition to the 4 cosmological parameters.

We use dependencies between these parameters to shrink our fitting parameter space from the most general case (16 parameters) to its final size (13 parameters). In particular, we use the constraints outlined in Appendix C for the relative fraction of the c_B sample, f_{c_B}, (equation C.2) and for the 1-halo amplitudes, $N_{c_B s}$ and $N_{s_B s_B}$, (equations C.12 and C.16). In the former case, the constraint allows us to vary the parameter $\langle N_{>1,s} \rangle$, which is defined as the mean number of satellite galaxies in halos with more than one satellite. This parameter is typically centered on $\langle N_{>1,s} \rangle \sim 2.4$ for CMASS-like galaxy samples, with little variation around this center value. For the 1-halo amplitude $N_{s_B s_B}$, we vary a normalization parameter $f^{1h}_{s_B s_B}$ to account for uncertainty in the expected value, which should have a value of order unity. Finally, we do not vary the velocity dispersion of the s_B sample, σ_{s_B}, but rather use the physically motivated scaling with halo mass, $\sigma_v^2 \propto M^{2/3}$, and the halo bias – mass relation from Tinker et al. (2010). We do not use this model function $\sigma_v^{\mathrm{model}}(b_1)$ to predict the absolute amplitude of σ_{s_B}, but only the functional form. We always rescale the predicted value by the current value of σ_{s_A} (see Table 4.3 for details).

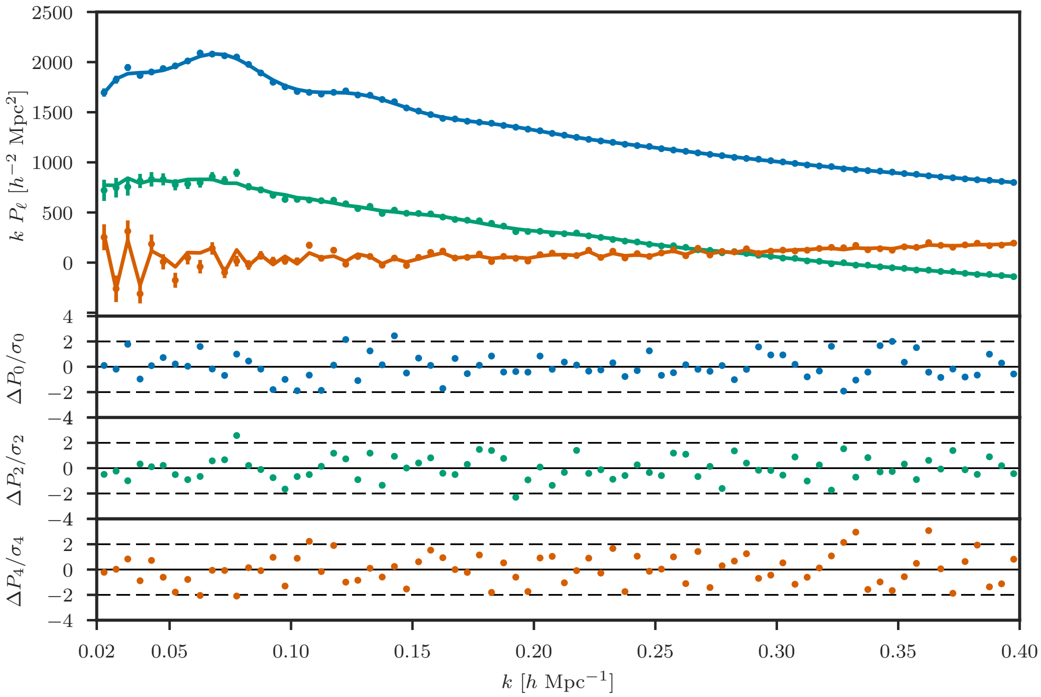


Figure 4.7: The best-fit monopole, quadrupole, and hexadecapole models (lines) as compared to the measurements (points) from the mean of 10 RunPB HOD galaxy realizations at $z = 0.55$, fit over the wavenumber range $k = 0.02 - 0.4\ h\mathrm{Mpc}^{-1}$. The lower panels show the model residuals for each multipole separately. The reduced chi-squared of the fit to all three multipoles is $\chi^2_{\mathrm{red}} = 1.12$. Note that the large variation from bin to bin in the hexadecapole is due to discrete binning effects.

4.5 Performance of the model

4.5.1 RunPB results

As a first test of the RSD model described in Section 4.4, we use a set of HOD galaxy catalogs constructed from 10 realizations of the $z = 0.55$ snapshot of the RunPB simulation, described previously in Section 4.2.1. The galaxy catalogs are made by populating halo catalogs according to a halo occupation distribution with parameters comparable to the BOSS CMASS sample. The halo catalogs are constructed in the manner described in detail in Reid et al. (2014). Briefly, the halo finder uses the spherical overdensity implementation of Tinker et al. (2008), using an overdensity of $\Delta_m = 200$ relative to the mean matter density ρ_m to define the halo virial radius. Central galaxies are not at rest with respect to the halo center-of-mass; they are assigned a velocity computed from the halo particles in the densest region of each halo (see Reid et al. (2014) for details). Note that the halo catalogs used here are not the same as the FOF halo catalogs described in Section 4.2.1, which we use to calibrate certain components of the RSD model. Differences in the halo finder algorithms lead to important differences in the clustering of the resulting galaxy catalogs. While we do not expect RSD fits to these galaxy catalogs to be a fully independent validation of the model, they do still provide a useful test of the accuracy of our model.

Using the model parameterization discussed in Section 4.4.4, we fit the mean of the measured monopole, quadrupole, and hexadecapole from 10 realizations at $z = 0.55$ as a function of the maximum wavenumber included in the fits, $k_\mathrm{max} = [0.2, 0.3, 0.4]$ $h\mathrm{Mpc}^{-1}$. The resulting best-fit model and residuals between measurements and theory are shown in Figure 4.7 for $k_\mathrm{max} = 0.4$ $h\mathrm{Mpc}^{-1}$. We are able to achieve excellent agreement between the model and simulation multipoles to scales of $k = 0.4$ $h\mathrm{Mpc}^{-1}$, well into the nonlinear clustering regime.

As a function of k_max, we report the mean and 1σ error for a subset of the model parameters in Table 4.4, as determined from the posterior distributions obtained via MCMC sampling. We also show the 2D posterior distributions for $f\sigma_8$, $\alpha_\parallel$, and $\alpha_\perp$ for each k_max value in Figure 4.8 (produced using Foreman-Mackey 2016). As expected, we obtain significant decreases in parameter uncertainties when including small-scale information in the fits. For the three cosmology parameters, $f\sigma_8$, $\alpha_\parallel$, and $\alpha_\perp$, we find decreases of 19%, 18%, and 18%, respectively, for $k_\mathrm{max} = 0.3$ $h\mathrm{Mpc}^{-1}$ and 38%, 24%, and 29% for $k_\mathrm{max} = 0.4$ $h\mathrm{Mpc}^{-1}$, relative to the fit using $k_\mathrm{max} = 0.2$ $h\mathrm{Mpc}^{-1}$. These decreases are roughly consistent with the expected scaling in the nonlinear regime, $\sigma \propto k_\mathrm{max}^{-1/2}$ (e.g., Blazek et al. 2014). For the AP parameters, we find more modest decreases in the uncertainty when ranging from $k_\mathrm{max} = 0.2$ $h\mathrm{Mpc}^{-1}$ to $k_\mathrm{max} = 0.4$ $h\mathrm{Mpc}^{-1}$. In particular, extending from $k_\mathrm{max} = 0.3$ $h\mathrm{Mpc}^{-1}$ to $k_\mathrm{max} = 0.4$ $h\mathrm{Mpc}^{-1}$ offers little improvement in the error on $\alpha_\parallel$. The constraining power for the AP parameters results from a combination of the BAO signal and information from the geometric distortion of the full broadband signal. As nearly all of the information from the BAO signal is present below $k = 0.2$ $h\mathrm{Mpc}^{-1}$, our more modest decreases in uncertainty for the AP parameters are consistent with our expectations.

We have central/satellite information for each galaxy in the RunPB catalogs and can assess the accuracy of the halo model decomposition described in Section 4.4.1. As seen in Table 4.4, we find a non-zero velocity dispersion for centrals with an amplitude $\sigma_c \sim 1$ $h^{-1}\mathrm{Mpc}$, which is consistent with the expected amplitude present in the underlying halo catalogs. The main discrepancy is that our recovered satellite fraction is significantly higher than the expected value. This difference is likely due to the fact that we rely on FOF halo catalogs for calibration of the halo clustering in the model, while we are fitting galaxy catalogs created from SO halo catalogs. The choice of halo finder alters the clustering on scales around the virial radius. A FOF halo finder tends to over-merge halos on these scales into a single halo, whereas a SO finder tends to preserve the multiple smaller halos. This well-known difference in halo finders would manifest itself as an increase in the satellite fraction and is consistent with our fitting results. While differences in halo finder algorithms are notoriously complex, we emphasize that we are able to absorb any discrepancies related to these simulation differences in our model and obtain unbiased results for our cosmological parameters.

We analyze the correlations between the posterior distributions to better understand the

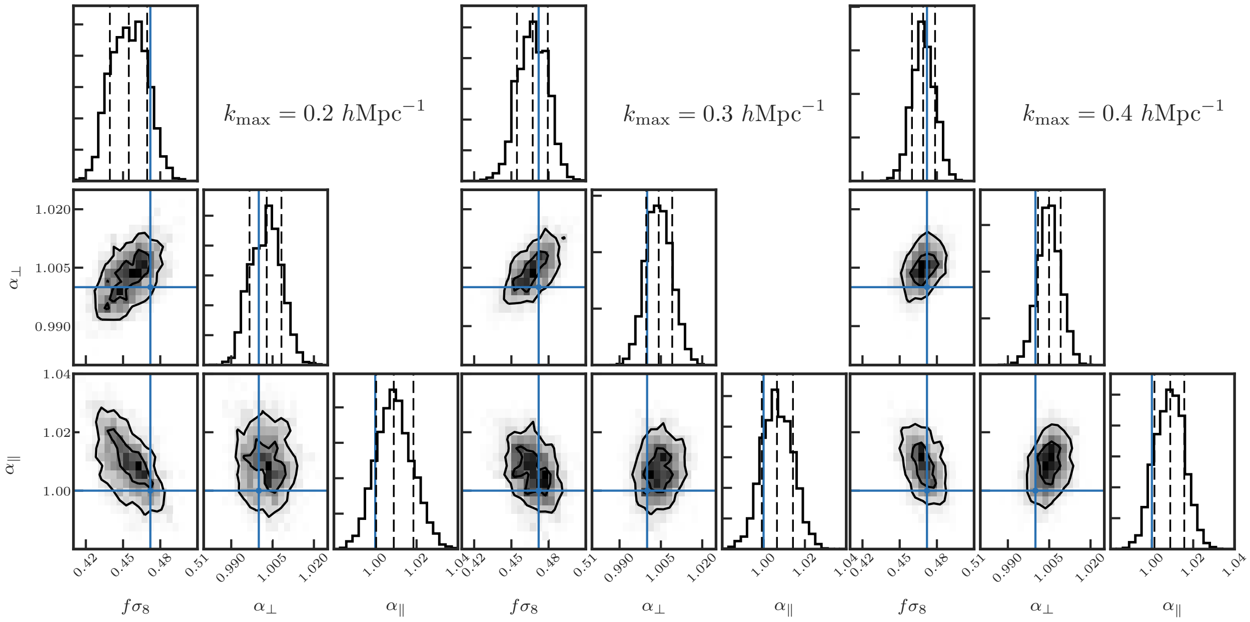


Figure 4.8: The 2D posterior distributions for the cosmology parameters $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ obtained from fitting the mean of 10 RunPB HOD galaxy realizations at $z = 0.55$. We show results when varying the maximum scale included in the fits: $k_{\rm max} = 0.2$ (left), 0.3 (center), 0.4 (right) $h{\rm Mpc}^{-1}$. The expected parameter values are marked with solid blue lines.

	$k_{\mathrm{max}} = 0.2\ h/\mathrm{Mpc}$	$k_{\mathrm{max}} = 0.3\ h/\mathrm{Mpc}$	$k_{\mathrm{max}} = 0.4\ h/\mathrm{Mpc}$	truth
$f\sigma_8$	$0.455\ ^{+0.015}_{-0.015}$	$0.467\ ^{+0.012}_{-0.013}$	$0.469\ ^{+0.010}_{-0.009}$	0.472
$f\sigma_8$ [fixed AP]	$0.457\ ^{+0.010}_{-0.010}$	$0.466\ ^{+0.009}_{-0.008}$	$0.468\ ^{+0.007}_{-0.007}$	0.472
$\alpha_\perp$	$1.003\ ^{+0.005}_{-0.006}$	$1.004\ ^{+0.005}_{-0.005}$	$1.005\ ^{+0.004}_{-0.004}$	1.000
$\alpha_\parallel$	$1.009\ ^{+0.010}_{-0.009}$	$1.006\ ^{+0.008}_{-0.007}$	$1.009\ ^{+0.007}_{-0.008}$	1.000
$b_1\sigma_8$	$1.266\ ^{+0.009}_{-0.009}$	$1.265\ ^{+0.008}_{-0.008}$	$1.268\ ^{+0.008}_{-0.008}$	1.272
f_s	$0.122\ ^{+0.019}_{-0.018}$	$0.143\ ^{+0.013}_{-0.013}$	$0.143\ ^{+0.008}_{-0.008}$	0.104
f_{c_B}	$0.104\ ^{+0.033}_{-0.030}$	$0.124\ ^{+0.022}_{-0.023}$	$0.122\ ^{+0.013}_{-0.015}$	0.089
f_{s_B}	$0.438\ ^{+0.203}_{-0.197}$	$0.438\ ^{+0.136}_{-0.123}$	$0.466\ ^{+0.081}_{-0.079}$	0.399
σ_c	$1.134\ ^{+0.214}_{-0.238}$	$0.906\ ^{+0.088}_{-0.111}$	$0.930\ ^{+0.062}_{-0.065}$	–
σ_{s_A}	$4.239\ ^{+0.476}_{-0.413}$	$3.737\ ^{+0.372}_{-0.464}$	$3.443\ ^{+0.278}_{-0.270}$	–
χ^2/d.o.f.	$113/(108-13) = 1.19$	$159/(168-13) = 1.03$	$241/(228-13) = 1.12$	

Table 4.4: Parameter constraints obtained when fitting the 13-parameter RSD model to $[P_0, P_2, P_4]$, as measured from the mean of the 10 RunPB galaxy catalogs at $z = 0.55$. We show results determined as a function of the maximum wavenumber included in the fits. Parameter posteriors are determined from MCMC sampling of the likelihood, assuming Gaussian covariance between multipoles.

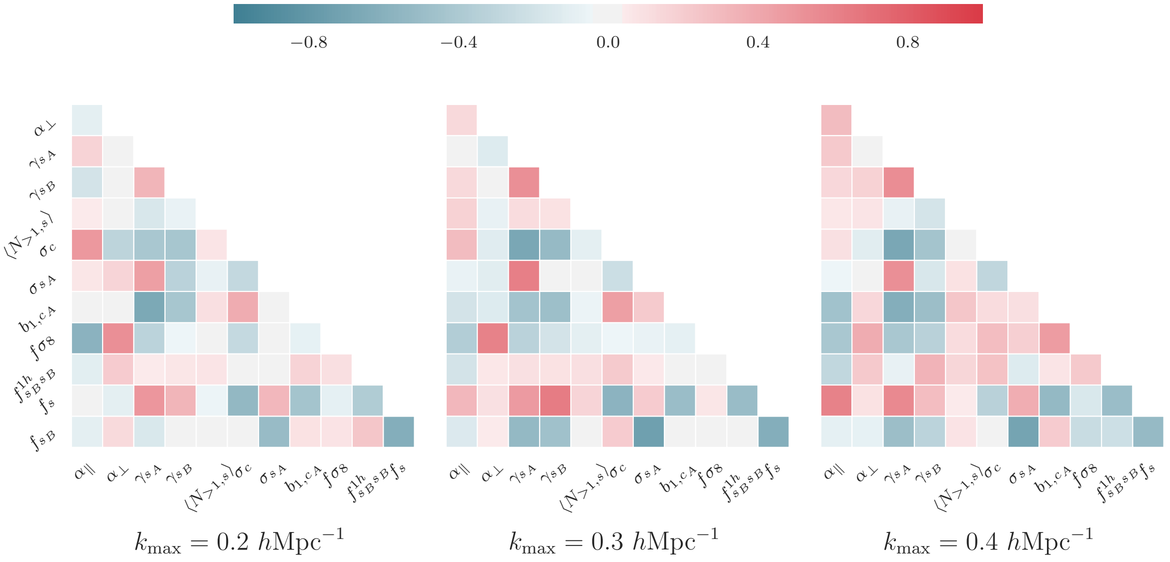
constraining power of our 13-parameter model. We show these parameter correlations for each fitting range in Figure 4.9. As expected, we find that the main parameter combination measuring the strength of RSD, $f\sigma_8$, is most correlated with the AP parameters $\alpha_\parallel$ and $\alpha_\perp$, which measure geometric distortions of the clustering signal. For each of the k_{max} fitting ranges, the correlation matrix for $(f\sigma_8, \alpha_\perp, \alpha_\parallel)$ is:

$$R^{0.2}[f\sigma_8, \alpha_\perp, \alpha_\parallel] = \begin{bmatrix} 1.000 & 0.536 & -0.583 \\ 0.536 & 1.000 & -0.094 \\ -0.583 & -0.094 & 1.000 \end{bmatrix}, \tag{4.39}$$

$$R^{0.3}[f\sigma_8, \alpha_\perp, \alpha_\parallel] = \begin{bmatrix} 1.000 & 0.605 & -0.361 \\ 0.605 & 1.000 & 0.133 \\ -0.361 & 0.133 & 1.000 \end{bmatrix}, \tag{4.40}$$

$$R^{0.4}[f\sigma_8, \alpha_\perp, \alpha_\parallel] = \begin{bmatrix} 1.000 & 0.377 & -0.418 \\ 0.377 & 1.000 & 0.292 \\ -0.418 & 0.292 & 1.000 \end{bmatrix}. \tag{4.41}$$

As we extend the maximum wavenumber included in our fits, small-scale information does help break degeneracies between $f\sigma_8$ and the AP parameters, reducing the correlation between $f\sigma_8$ and ($\alpha_\perp$, $\alpha_\parallel$). For comparison, Beutler et al. (2017b) reports a correlation between $f\sigma_8$ and $\alpha_\perp$ of 0.503 and $f\sigma_8$ and $\alpha_\parallel$ of 0.547 for the middle redshift bin for the



$k_{\max} = 0.2\ h\mathrm{Mpc}^{-1}$ $k_{\max} = 0.3\ h\mathrm{Mpc}^{-1}$ $k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$

Figure 4.9: Parameter correlations as measured from the posterior distributions when fitting the mean of the 10 RunPB galaxy catalogs. We show the correlations as a function of the maximum wavenumber included in the fit, illustrating the changes in parameter dependencies when fitting to smaller scales.

combined DR12 BOSS sample, where they have fit $[P_0,\ P_2]$ to $k_{\max} = 0.15\ h\mathrm{Mpc}^{-1}$ and P_4 to $k_{\max} = 0.1\ h\mathrm{Mpc}^{-1}$. This level of correlation is similar to our values obtained when fitting to $k_{\max} = 0.2\ h\mathrm{Mpc}^{-1}$, but we find a significant reduction in correlation fitting to $k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$. We can assess the freedom of our RSD modeling using the Fisher formalism, which predicts a correlation coefficient of unity between $\alpha_\parallel$ and $\alpha_\perp$ in the case where we perfectly understand RSD (Seo & Eisenstein 2003, 2007; Shoji et al. 2009). In the opposite limit, Fisher matrix calculations predict $r \sim -0.4$ (Seo & Eisenstein 2003, 2007; Beutler et al. 2017c) when only BAO information is used and RSD information is fully marginalized over. Thus, the correlation between $\alpha_\parallel$ and $\alpha_\perp$ provides a measure of the constraining power of our RSD parametrization, with the correlation decreasing from unity as additional freedom is introduced into the RSD model. Our results are consistent with this expectation, as we find the correlation increase for large $k_{\max}$. To model results only to $k_{\max} = 0.2\ h\mathrm{Mpc}^{-1}$, our model contains too much freedom, in comparison to the requirements of modeling to $k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$. Again for comparison, Beutler et al. (2017b) finds a correlation of $r = 0.257$ between $\alpha_\parallel$ and $\alpha_\perp$. Thus, our value of $r = 0.292$ indicates that we are able to recover a similar amount of information using our RSD model parametrization to $k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$.

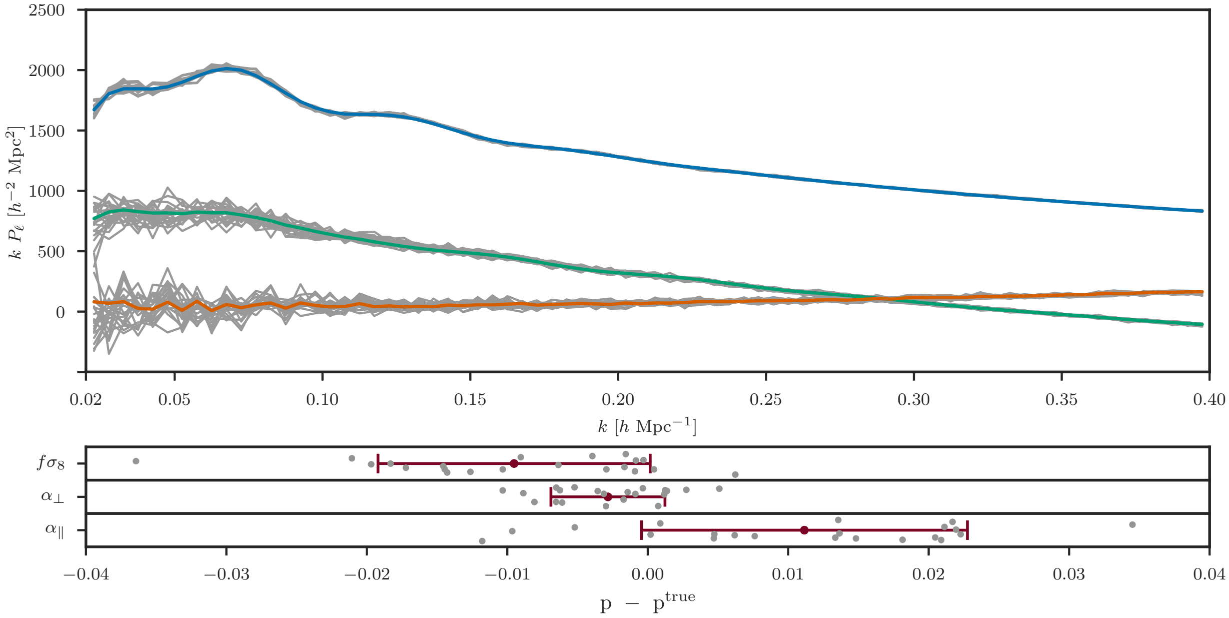


Figure 4.10: The mean of the best-fit monopole, quadrupole, and hexadecapole models (colored) as compared to the individual measurements (gray) from the 21 N-series cubic boxes at $z = 0.5$, fit over the wavenumber range $k = 0.02 - 0.4\ h\mathrm{Mpc}^{-1}$. The lower panels show the scatter in the recovered values for the 3 cosmology parameters, $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$, across the 21 boxes. The error bar shows the standard deviation of these results (not the error on the mean).

4.5.2 Independent tests on high resolution mocks

To fully assess the accuracy and precision of our RSD model, we perform independent tests using two sets of mocks based on high-fidelity, periodic N-body simulations. The first, described in §4.5.2, is a homogeneous set of 21 galaxy catalogs derived from 7 realizations of an N-body simulation with fixed cosmology and bias model. The second, described in §4.5.2, is a set of 7 heterogenous HOD galaxy catalogs where both the bias model and underlying cosmology varies from box to box. For details on the cosmology and simulation parameters for these mocks, see Table 4.1.

Cubic N-series results

Our first independent tests utilize the cubic N-series simulation, the large-volume ($L_{\mathrm{box}} = 2600\ h^{-1}$Mpc) periodic box simulations described in Section 4.2.2. We perform fits to the monopole, quadrupole, and hexadecapole from 21 HOD galaxy catalogs, constructed from 7 realizations at $z = 0.5$ and 3 orthogonal line-of-sight projections per box. The cosmology of these boxes is given in Table 4.1. As in Section 4.5.1, we perform fits to the data vector $[P_0, P_2, P_4]$ over a range of k_{max} values. The best-fitting parameters for each of the 21 catalogs are obtained by maximum a posterior (MAP) estimation using the LBFGS algorithm.

$k_{\rm max}$	$\Delta\langle\alpha_\parallel\rangle$	$S_{\alpha_\parallel}$	$\Delta\langle\alpha_\perp\rangle$	$S_{\alpha_\perp}$	$\Delta\langle f\sigma_8\rangle$	$S_{f\sigma_8}$	$\Delta\langle f\sigma_8\rangle$	$S_{f\sigma_8}$
[hMpc^{-1}]							fixed AP	
0.2	0.005	0.011	−0.002	0.005	−0.016	0.014	−0.010	0.013
0.3	0.007	0.011	−0.003	0.004	−0.009	0.013	−0.000	0.011
0.4	0.011	0.012	−0.003	0.004	−0.008	0.010	0.003	0.010

Table 4.5: The mean (with expected value subtracted) and standard deviation S of the best-fitting values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ found when fitting $[P_0, P_2, P_4]$ from the 21 cubic N-series catalogs. Results are reported as a function of the maximum wavenumber included in the fit. We also give results for $f\sigma_8$ when holding the AP parameters fixed to their true values.

Figure 4.10 shows the measured $\ell = 0$, 2, and 4 multipoles from the individual N-series catalogs, and we have over-plotted the mean of the best-fitting model from each fit using $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$. We report the mean (with the expected value subtracted) and standard deviation for the best-fitting $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ values from the 21 fits as a function of fitting range in Table 4.5. We also include the results for $f\sigma_8$ when holding the AP parameters fixed to their true values.

We find similar trends in our recovered cosmological parameters for the N-series boxes as for the RunPB results in §4.5.1. We obtain good fits to the measured $\ell = 0, 2, 4$ multipoles using our RSD model up to $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$. However, we do find some evidence for small systematic biases present in our RSD model, although it is difficult to properly assess the level of statistical significance with only seven fully independent realizations (clustering from boxes that vary only the line-of-sight projection are correlated). When using $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$, we find that are $f\sigma_8$ value is biased low by $\Delta f\sigma_8 = 0.008$ and $\alpha_\parallel$ is biased high by $\Delta\alpha_\parallel = 0.011$. These correspond to $\sim$0.8σ and $\sim$0.9σ shifts, respectively, relative to the box-to-box dispersion, as determined by the standard deviation of the 21 fits. Although it is important to note that, again, with only 7 independent realizations and 21 total fits, the standard deviation across the fits remains noisy. When fixing the AP parameters to their true values, we see a relatively large upwards shift in the mean $f\sigma_8$ value across the fits. As our model prefers a slightly larger $\alpha_\parallel$ value than expected, when its value is fixed to its correct value, the recovered value for $f\sigma_8$ shifts upwards, due to the anti-correlation between the parameters.

Lettered challenge box results

We perform additional tests of our model using a heterogeneous set of seven HOD galaxy catalogs, labeled A through G, which were constructed from high-fidelity cubic N-body simulations. These catalogs are described in detail in Section 4.2.3, and the cosmology and simulation parameters are reported in Table 4.1. The box size for these mocks is $\sim$2.5 h^{-1}Gpc; a single box has roughly 4 times the volume of the DR12 BOSS CMASS sample and 60% of the volume of the mean of the 10 RunPB realizations. They were designed to provide stringent stress-tests of full-shape RSD modeling analyses, and as such, they cover a range of redshifts ($z = [0.441, 0.5, 0.562]$), $f\sigma_8$ values, and galaxy bias models. As was done in previous sections, we compute fits to the monopole, quadrupole, and hexadecapole for each of the seven lettered challenge boxes, as a function of the maximum wavenumber included in the fits. We obtain full posterior distributions for each of our 13 model parameters using MCMC sampling. We report the recovered values for the cosmological parameters (with the expected value subtracted) and the 1σ parameter uncertainties for all seven boxes in Table 4.6. Figure 4.11 illustrates the fractional deviation of our recovered cosmology values from their reference values for each k_{max} value.

The recovered values show similar trends as a function of k_{max} as the results from the RunPB and cubic N-series results. We generally find $\sim 20 - 30\%$ improvements in the error on $f\sigma_8$ when extending the fit from $k_{\mathrm{max}} = 0.2\ h\mathrm{Mpc}^{-1}$ to $k_{\mathrm{max}} = 0.4\ h\mathrm{Mpc}^{-1}$. We find more modest decreases in the error for the AP parameters, with little improvement extending from $k_{\mathrm{max}} = 0.3\ h\mathrm{Mpc}^{-1}$ to $k_{\mathrm{max}} = 0.4\ h\mathrm{Mpc}^{-1}$. Within the expected 1σ uncertainty of each mock, we recover $f\sigma_8$ and $\alpha_\perp$ values consistent with the truth for all seven boxes, and the best-fitting values generally remain stable as a function of k_{max}. However, the recovered values for $\alpha_\parallel$ show a systematic positive bias for all boxes, relative to the truth, which can be most easily seen in Figure 4.11. This bias is present for each value of k_{max} used. It is difficult to assess the statistical significance of this potential bias, as several of the seven mocks are built on the same underlying N-body simulation, which correlates the derived parameters. Weighting each derived $\alpha_\parallel$ by the inverse uncertainty, we find a mean positive bias of $\Delta\alpha_\parallel = 0.02$, independent of k_{max}. This bias is slightly larger than was found for either the RunPB or cubic N-series mocks, where both results show a $\sim$0.01 positive bias in $\alpha_\parallel$.

We also include in Figure 4.11 and Table 4.6 the results for $f\sigma_8$ when fixing the AP parameters to their true values. As expected, we see substantial ($\sim$20-30%) error decreases since the correlation between $f\sigma_8$ and the AP parameters degrades constraints when $\alpha_\parallel$ and $\alpha_\perp$ are allowed to vary. Similar to previous results, we also find a systematic positive shift in the recovered $f\sigma_8$ values when holding $\alpha_\parallel$ and $\alpha_\perp$ fixed to their true values. This is expected due to the correlation between $f\sigma_8$ and $\alpha_\parallel$ and the systematic positive shift found for $\alpha_\parallel$.

4.5.3 Tests on realistic DR12 BOSS CMASS mocks

Finally, we test our RSD model using BOSS DR12 CMASS mock catalogs, using the 84 independent, N-series cutsky catalogs described in §4.2.2. This set of catalogs offers a chance to test the performance of our model in a realistic setting with a large enough number of catalogs to identify systematic biases up to the level of $\sqrt{84} = 9.16$ times smaller than the measurement uncertainty from a single mock. The 84 N-cutsky catalogs accurately model the geometry, volume, and redshift distribution of the DR12 CMASS NGC sample (Reid et al. 2016). We use the window function convolution procedure outlined in Section 4.4.3 to properly account for the effects of the selection function on the measured power spectrum multipoles. We measure the monopole, quadrupole, and hexadecapole for each of the 84

box	k_{max}	$\Delta\alpha_{\parallel}$	$\Delta\alpha_{\perp}$	$\Delta f\sigma_8$	$\Delta f\sigma_8$
	[hMpc^{-1}]				fixed AP
A	0.2	$0.031^{+0.012}_{-0.012}$	$-0.001^{+0.006}_{-0.007}$	$-0.026^{+0.018}_{-0.016}$	$-0.003^{+0.011}_{-0.012}$
	0.3	$0.029^{+0.011}_{-0.011}$	$0.001^{+0.007}_{-0.008}$	$-0.015^{+0.015}_{-0.014}$	$-0.001^{+0.011}_{-0.010}$
	0.4	$0.031^{+0.010}_{-0.013}$	$0.004^{+0.006}_{-0.006}$	$-0.009^{+0.013}_{-0.014}$	$0.005^{+0.008}_{-0.007}$
B	0.2	$0.029^{+0.012}_{-0.013}$	$-0.004^{+0.007}_{-0.007}$	$-0.010^{+0.018}_{-0.018}$	$0.009^{+0.014}_{-0.013}$
	0.3	$0.032^{+0.013}_{-0.012}$	$-0.003^{+0.008}_{-0.007}$	$-0.009^{+0.016}_{-0.018}$	$0.014^{+0.010}_{-0.009}$
	0.4	$0.032^{+0.012}_{-0.011}$	$-0.001^{+0.007}_{-0.007}$	$-0.004^{+0.015}_{-0.014}$	$0.019^{+0.010}_{-0.011}$
C	0.2	$0.045^{+0.014}_{-0.014}$	$-0.002^{+0.007}_{-0.007}$	$-0.035^{+0.020}_{-0.020}$	$-0.005^{+0.012}_{-0.014}$
	0.3	$0.048^{+0.011}_{-0.012}$	$-0.002^{+0.006}_{-0.007}$	$-0.030^{+0.016}_{-0.017}$	$0.002^{+0.011}_{-0.013}$
	0.4	$0.045^{+0.013}_{-0.013}$	$-0.000^{+0.006}_{-0.006}$	$-0.012^{+0.016}_{-0.014}$	$0.021^{+0.010}_{-0.010}$
D	0.2	$0.002^{+0.012}_{-0.011}$	$0.003^{+0.007}_{-0.006}$	$0.011^{+0.016}_{-0.018}$	$0.006^{+0.012}_{-0.012}$
	0.3	$0.002^{+0.010}_{-0.010}$	$0.002^{+0.005}_{-0.006}$	$0.011^{+0.017}_{-0.015}$	$0.012^{+0.010}_{-0.009}$
	0.4	$0.009^{+0.009}_{-0.008}$	$-0.001^{+0.005}_{-0.005}$	$0.013^{+0.011}_{-0.012}$	$0.017^{+0.009}_{-0.009}$
E	0.2	$0.000^{+0.013}_{-0.012}$	$0.006^{+0.006}_{-0.007}$	$0.015^{+0.018}_{-0.020}$	$0.007^{+0.010}_{-0.009}$
	0.3	$-0.001^{+0.010}_{-0.010}$	$0.002^{+0.006}_{-0.005}$	$0.008^{+0.016}_{-0.017}$	$0.005^{+0.009}_{-0.009}$
	0.4	$0.009^{+0.011}_{-0.011}$	$0.002^{+0.006}_{-0.006}$	$0.012^{+0.017}_{-0.015}$	$0.016^{+0.010}_{-0.009}$
F	0.2	$0.032^{+0.013}_{-0.012}$	$-0.002^{+0.007}_{-0.007}$	$-0.025^{+0.020}_{-0.019}$	$0.005^{+0.011}_{-0.012}$
	0.3	$0.034^{+0.011}_{-0.011}$	$0.003^{+0.006}_{-0.006}$	$-0.012^{+0.015}_{-0.017}$	$0.015^{+0.006}_{-0.006}$
	0.4	$0.015^{+0.010}_{-0.009}$	$0.008^{+0.006}_{-0.006}$	$0.006^{+0.014}_{-0.013}$	$0.009^{+0.005}_{-0.004}$
G	0.2	$0.014^{+0.010}_{-0.010}$	$-0.000^{+0.007}_{-0.007}$	$-0.022^{+0.017}_{-0.017}$	$-0.014^{+0.012}_{-0.012}$
	0.3	$0.012^{+0.009}_{-0.011}$	$0.001^{+0.005}_{-0.005}$	$-0.015^{+0.015}_{-0.017}$	$-0.004^{+0.010}_{-0.011}$
	0.4	$0.020^{+0.011}_{-0.011}$	$0.007^{+0.006}_{-0.006}$	$0.006^{+0.010}_{-0.010}$	$0.007^{+0.008}_{-0.007}$

Table 4.6: The best-fitting values for $f\sigma_8$, $\alpha_{\perp}$, and $\alpha_{\parallel}$ obtained when fitting our RSD model to the measured monopole, quadrupole, and hexadecapole from the 7 lettered challenge boxes. We report results as a function of the maximum wavenumber included in the fits. The 1σ uncertainties obtained via MCMC sampling are also shown.

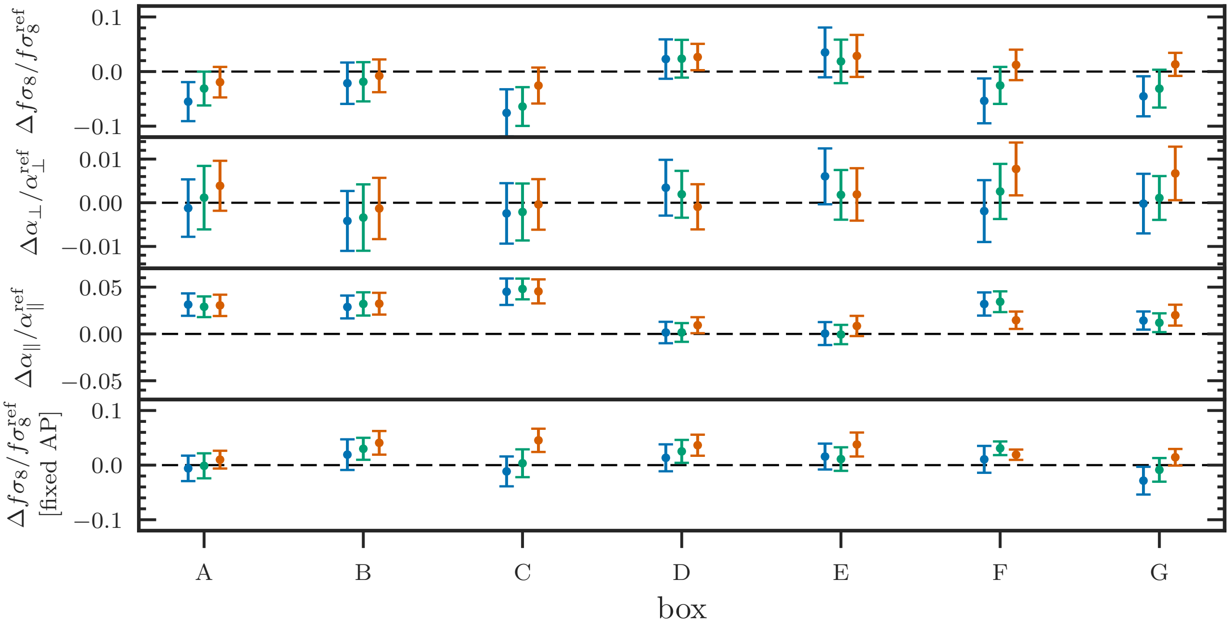


Figure 4.11: The fractional deviation of the best-fitting $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ values from their true values for each of the seven lettered challenge boxes. We also show the deviations for $f\sigma_8$ obtained when the AP parameters are fixed to their true values. For each box, we show results obtained using (from left to right) $k_{\rm max} = 0.2$ (blue), 0.3 (green), and 0.4 (orange) hMpc^{-1}. Error bars show the 1σ uncertainty as obtained via MCMC sampling.

catalogs and estimate the best-fitting model parameters using MAP optimization via the LBFGS algorithm. The power spectra have been measured using FKP weights with a value of $P_0 = 10^4\ h^{-3}\mathrm{Mpc}^3$. Similar to previous fits, we report parameter constraints as a function of the maximum wavenumber included in the fits. The minimum wavenumber included in the fits is $k_\mathrm{min} = 0.02\ h\mathrm{Mpc}^{-1}$, chosen to minimize any large-scale effects of the window function on our parameter fits.

We plot the best-fitting $\ell = 0$, 2, 4, and 6 theoretical models and the measured multipoles from a single catalog of the full 84 N-series cutsky test suite in Figure 4.12. Here, the best-fit model has been estimated using the data vector $[P_0, P_2, P_4]$ with $k_\mathrm{max} = 0.4\ h\mathrm{Mpc}^{-1}$. We also show the tetra-hexadecapole ($\ell = 6$) to illustrate that the model can accurately predict this higher-order multipole and that it contains little measurable signal. For this single mock, we find good agreement between theory and data, with a reduced chi-squared of $\chi^2_\mathrm{red} = 1.01$. The average value across all 84 mocks is $\langle \chi^2_\mathrm{red} \rangle = 1.08$.

We give the mean (with the expected value subtracted) and standard deviation of the best-fitting cosmological parameters from fits to each of the 84 cutsky mocks in Table 4.7. We also show the 1D histograms and 2D correlations of $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ for the individual fits in Figure 4.13, illustrating the constraining power of our model for these parameters as well as the correlations between the parameters. When fitting the monopole, quadrupole, and hexadecapole, we find good agreement between the mean of the recovered values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$. When including scales up to $k_\mathrm{max} = 0.4\ h\mathrm{Mpc}^{-1}$, we find modest mean biases of $\Delta\langle f\sigma_8\rangle = 0.005$, $\Delta\langle\alpha_\perp\rangle = -0.004$, and $\Delta\langle\alpha_\parallel\rangle = 0.004$, which represent 14%, 28%, and 17% of the expected mock-to-mock dispersion of each parameter, respectively. The statistical precision of the mean values due to the finite number of catalogs is $84^{-1/2} \simeq 0.1$ times the mock-to-mock dispersion. Thus, the results show evidence for a small bias in the derived $\alpha_\perp$ value and marginal evidence for small biases in $\alpha_\parallel$ and $f\sigma_8$. We also show results in Table 4.7 when fitting only the monopole and quadrupole in order to help quantify the impact of the hexadecapole on our final constraints. The mean best-fitting parameters remain consistent with the results obtained when fitting $[P_0,\ P_2,\ P_4]$, and we find the standard deviation of our best-fitting $f\sigma_8$ values inflates by roughly 30%, consistent with the findings of Beutler et al. (2017b). When fixing the AP parameters to their true values, we find that the hexadecapole adds negligible further information to our parameter constraints.

4.5.4 Comparison to published models

The set of 84 N-series mocks described in the previous section were utilized as part of the BOSS collaboration's internal RSD modeling tests in preparation for the DR12 parameter constraint analyses. This enables us to perform a direct comparison of our model with the main Fourier space RSD models used in the DR12 consensus results, which are described in the companion papers in Beutler et al. (2017b) and Grieb et al. (2017) and the main DR12 consensus paper in Alam et al. (2017). The model used in Grieb et al. (2017) was also applied to BOSS DR12 data in configuration space, with results presented in Sánchez et al. (2017). These analyses differ in a number of ways from ours. In particular, these

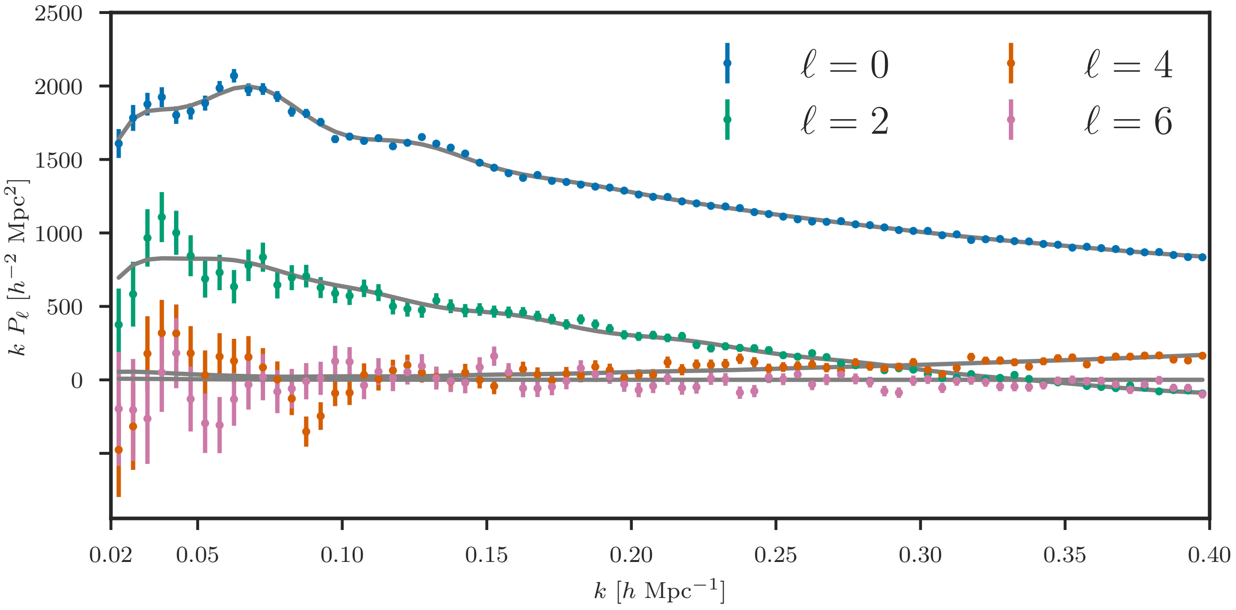


Figure 4.12: The best-fitting $\ell = 0$, 2, 4, and 6 theory (grey lines) and measurements (points with errors) from a single catalog of the N-series cutsky test suite, which accurately simulates the BOSS DR12 CMASS data set. The best-fit model has been estimated using the data vector $[P_0, P_2, P_4]$ while fitting over the wavenumber range $k = 0.02 - 0.4\ h\mathrm{Mpc}^{-1}$. We also show the tetra-hexadecapole ($\ell = 6$) to illustrate that the model can accurately predict this higher-order multipole and that it contains little measurable signal. The reduced chi-squared of the fit for this mock catalog is $\chi^2_{\mathrm{red}} = 1.01$. The average value across all 84 mocks is $\langle \chi^2_{\mathrm{red}} \rangle = 1.08$.

models have significantly fewer parameters (7-8 instead of 13) and use a smaller k_{max} value in their fits. We limit our fitting range to the same as those used in these works and directly compare the derived parameter constraints for the N-cutsky mocks in Table 4.8. For comparison, this table also includes results from fits using our model that include scales to $k_{\mathrm{max}} = 0.3\ h\mathrm{Mpc}^{-1}$ and $k_{\mathrm{max}} = 0.4\ h\mathrm{Mpc}^{-1}$, which goes beyond the scales used in Beutler et al. (2017b) and Grieb et al. (2017). When using comparable fitting ranges, we find that our model yields a standard deviation for $f\sigma_8$ that is larger by ~10% and ~20% as compared to the results recovered when using the models of Beutler et al. (2017b) and Grieb et al. (2017), respectively. We find comparable constraints on $f\sigma_8$ when extending our model to $k_{\mathrm{max}} = 0.3\ h\mathrm{Mpc}^{-1}$ and a modest 5-10% improvement when using $k_{\mathrm{max}} = 0.4\ h\mathrm{Mpc}^{-1}$.

For the AP parameters, we find a comparable constraint on $\alpha_\perp$ and a slighter worse constraint on $\alpha_\parallel$ as compared to the model of Beutler et al. (2017b). We find modest 5% and 10% reductions in the error on $\alpha_\parallel$ and $\alpha_\perp$ as compared to the model of Grieb et al. (2017). Extending the fits with our model to $k_{\mathrm{max}} = 0.4\ h\mathrm{Mpc}^{-1}$ does not provide much gain for the uncertainty of $\alpha_\parallel$, but we do find a roughly 20% reduction in the uncertainty of $\alpha_\perp$ as compared to the models of Beutler et al. (2017b) and Grieb et al. (2017). As

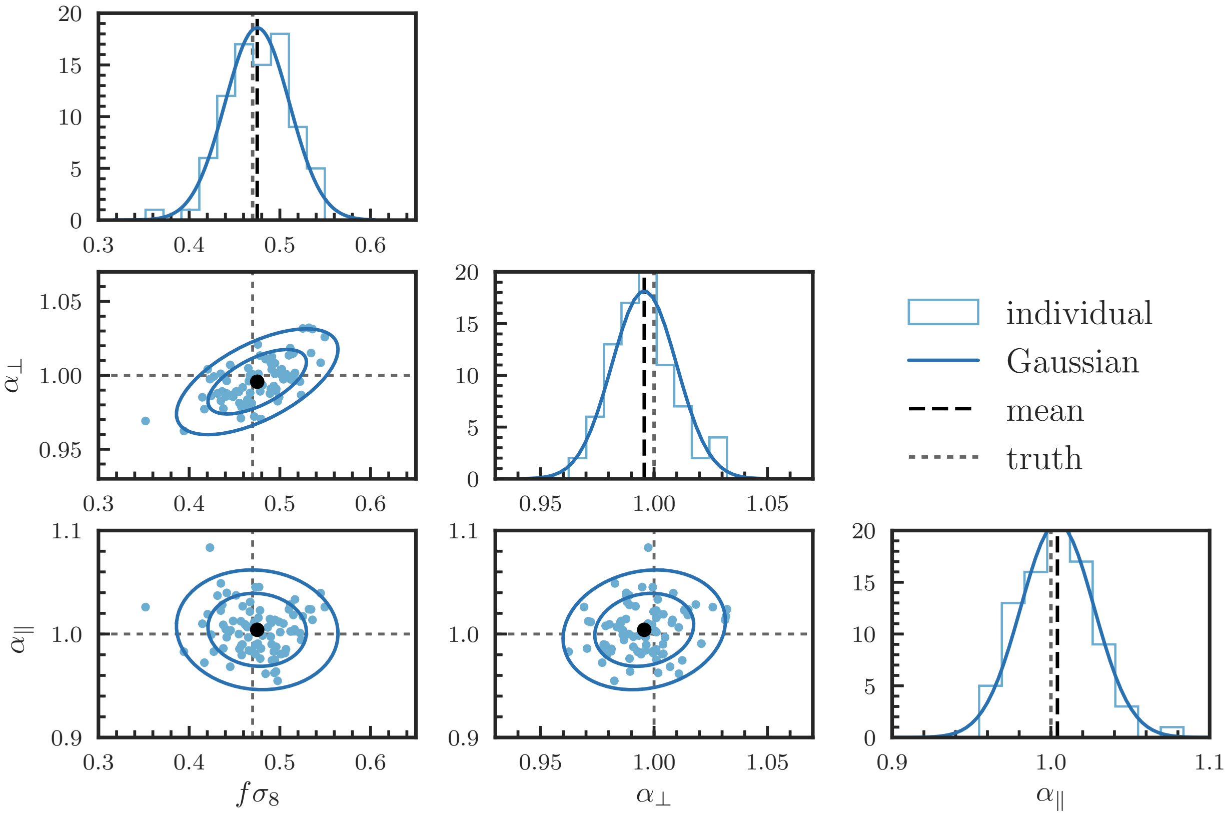


Figure 4.13: The best-fitting $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ parameters from fitting our RSD model to the measured $[P_0, P_2, P_4]$ multipoles from the 84 N-series cutsky mocks. We include wavenumbers in the range 0.02 $h\mathrm{Mpc}^{-1} \leq k \leq 0.4$ $h\mathrm{Mpc}^{-1}$. The diagonal panels show the histogram of the recovered parameters, with the mean best-fitting parameters indicated as black dashed lines and the true values as gray dotted lines. The panels below the diagonal show 2D plots with the 84 individual best-fitting parameters as blue dots and the mean as a filled circle. We also show a Gaussian fit to the marginalized parameter distributions in all panels.

statistics	$k_{\max}$	$\Delta\langle\alpha_\parallel\rangle$	$S_{\alpha_\parallel}$	$\Delta\langle\alpha_\perp\rangle$	$S_{\alpha_\perp}$	$\Delta\langle f\sigma_8\rangle$	$S_{f\sigma_8}$	$\Delta\langle f\sigma_8\rangle$	$S_{f\sigma_8}$
	[$h\text{Mpc}^{-1}$]							fixed AP	
$[P_0,\,P_2,\,P_4]$	0.2	0.007	0.024	-0.004	0.016	-0.020	0.041	-0.008	0.034
	0.3	0.007	0.025	-0.005	0.015	-0.008	0.039	0.005	0.030
	0.4	0.004	0.023	-0.004	0.014	0.005	0.036	0.013	0.027
$[P_0,\,P_2]$	0.2	-0.004	0.039	-0.001	0.019	-0.014	0.052	-0.013	0.035
	0.3	0.005	0.041	-0.004	0.019	-0.005	0.053	0.005	0.030
	0.4	0.012	0.036	-0.008	0.016	-0.010	0.040	0.007	0.025

Table 4.7: The mean and standard deviation of the best-fitting values for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ from fits to the 84 N-series cutsky catalogs. Results are reported as a function of the maximum wavenumber included in the fit. We show results obtained when including or excluding the hexadecapole from our fits in order to quantify the influence of the hexadecapole on our derived constraints.

	$k_{\max}$ for $[P_0, P_2, P_4]$	$\Delta\langle\alpha_\parallel\rangle$	$S_{\alpha_\parallel}$	$\Delta\langle\alpha_\perp\rangle$	$S_{\alpha_\perp}$	$\Delta\langle f\sigma_8\rangle$	$S_{f\sigma_8}$
Beutler et al. (2017b)	[0.15, 0.15, 0.1]	0.0049	0.0338	-0.0014	0.0180	-0.0049	0.0375
Grieb et al. (2017)	[0.2, 0.2, 0.2]	0.0089	0.0253	-0.0030	0.0175	0.0001	0.0383
This work	[0.15, 0.15, 0.1]	0.0003	0.0403	0.0011	0.0183	0.0043	0.0447
	[0.2, 0.2, 0.2]	0.0065	0.0239	-0.0041	0.0157	-0.0198	0.0409
	[0.3, 0.3, 0.3]	0.0074	0.0254	-0.0050	0.0152	-0.0077	0.0385
	[0.4, 0.4, 0.4]	0.0041	0.0231	-0.0043	0.0143	0.0050	0.0356

Table 4.8: The mean (with expected value subtracted) and standard deviation of the best-fitting cosmology parameters for the 84 N-cutsky mocks using the model in this work as well as the Fourier space models described in Beutler et al. (2017b) and Grieb et al. (2017). Results in all cases were computed using FKP weights with $P_0 = 10^4\ h^{-3}\text{Mpc}^3$.

seen in the results of Alam et al. (2017), the most powerful method for constraining the AP parameters, and thus $D_A(z)$ and $H(z)$, remains a BAO-only analysis that takes advantage of the additional statistical precision gained by the process of density field reconstruction. However, some additional constraining power can be gained from full-shape RSD analyses due the AP effect on sub-BAO scales. Here, the extra information provided by extending the modeling to $k_{\max} = 0.4\ h\text{Mpc}^{-1}$ aids the constraints on $D_A(z)$ and $H(z)$ and helps to de-correlate these parameters.

It is also instructive to compare our results for the N-cutsky mocks to the results published in Gil-Marín et al. (2016b), which fits the power spectrum monopole and quadrupole of the DR12 CMASS sample with $k_{\max} = 0.24\ h\text{Mpc}^{-1}$. The comparison yields similar conclusions as previously. In particular, Gil-Marín et al. (2016b) finds errors of $\sigma_{f\sigma_8} = 0.038$ and

$\sigma_{f\sigma_8} = 0.022$ when varying and fixing the AP parameters, respectively. These errors are both smaller than the uncertainties derived from our model by ~30% when fitting the monopole and quadrupole over similar wavenumber ranges. The constraints for the AP parameters in Gil-Marín et al. (2016b) are similarly smaller than those from our model by a comparable amount.

And, finally, it is worth noting that the $f\sigma_8$ constraints using the model in this chapter are not competitive with the 2.5% constraint on $f\sigma_8$ published in Reid et al. (2014), which remains the tightest measurement of $f\sigma_8$ in the literature to date. With fixed AP parameters, the work used a simulation-based analysis to model the small-scale correlation function of the DR10 CMASS sample well into the nonlinear regime, down to scales of $\sim 0.8\ h^{-1}$Mpc (smaller than the $R \sim k_{\rm max}^{-1} \sim 2.5\ h^{-1}$Mpc considered here). This work relied on simulations to accurately model the galaxy-halo connection whereas we use the analytic, halo model decomposition described in Section 4.4.2. Given the tight constraint on $f\sigma_8$ found by Reid et al. (2014), one might hope that Fourier space models could be similarly extended into the nonlinear regime and yield comparable increases in precision. However, while acknowledging the number of differences in the two analyses, we note that we do not find such large increases in precision in our measurement of $f\sigma_8$ when including small-scale information down to $k = 0.4\ h\text{Mpc}^{-1}$.

4.6 Discussion

The results presented in Section 4.5 provide tests of the RSD model presented in this work for a suite of simulations that span a wide range in both cosmology and galaxy bias models. Given the measurement uncertainties and the degrees of freedom in our model, we are able to achieve excellent agreement between the $\ell = 0, 2, 4$ multipoles measured from simulations and our best-fitting theory down to scales of $k = 0.4\ h\text{Mpc}^{-1}$. To quantify the impact of small-scale physics on our model, we perform fits for $k_{\rm max} = 0.2$, 0.3, and 0.4 $h\text{Mpc}^{-1}$. The results across the different sets of simulations indicate a positive systematic shift in the parallel AP parameter $\alpha_\parallel$ at the level of $0.01 - 0.02$ that is independent of the $k_{\rm max}$ value used. For fits using $k_{\rm max} = 0.4\ h\text{Mpc}^{-1}$, we find small biases at the level of ~0.005 for $f\sigma_8$ and $\alpha_\perp$. These deviations are small and can be effectively calibrated with simulations. The amplitude of the shifts is similar to the level of theoretical systematics present when using other RSD models in the literature, i.e., Alam et al. (2017). The positive bias in $\alpha_\parallel$ propagates into a small bias in $f\sigma_8$ when fixing the AP parameters to their expected values, due to the anti-correlation between $f\sigma_8$ and $\alpha_\parallel$. The exact amplitude of the bias in $\alpha_\parallel$ can be robustly estimated from a larger set of simulations than is considered in this work and the best-fitting $\alpha_\parallel$ value modified accordingly, while accounting for the systematic uncertainty in the error budget.

A primary goal of this work is to ensure that any model parameters that we introduce have physically meaningful values and are not just nuisance parameters. We attempt to capture the complex effects of satellite galaxies on the clustering signal in redshift space by considering separately the clustering of isolated satellites and those that live in halos with at least two satellites. This parametrization leads to a total of 13 model parameters, significantly more than other Fourier space RSD models in the literature, i.e., Beutler et al. (2017b); Grieb et al. (2017), which typically only have 7-8 parameters. In addition to differences in the treatment of RSD and perturbation theory choices, perhaps the most significant difference is the use of a single parameter to model the nonlinear FoG effect of the full galaxy sample, instead of separately modeling the effects for central and satellite subsamples, as is done in this work. They also typically float a constant, shot noise parameter, which is designed to absorb any potential deficiencies in the model. In some sense, these models are a limit of the more general parametrization considered in this work and are only valid over a certain range of scales and galaxy bias values.

As demonstrated in the analysis of Alam et al. (2017), the level of theoretical errors in $f\sigma_8$ measurements from full-shape RSD analyses ranges from ∼25-50% of the statistical precision for the three redshift bins considered for the completed BOSS DR12 sample. Another recent analysis (Beutler et al. 2017a) provides evidence for the possible shortcomings of the RSD model of Beutler et al. (2017b). The work extends the modeling to include the relative velocity effect of baryons and cold dark matter at decoupling but fails several null tests. The systematics situation is perhaps even more dire when considering the fact that the background cosmology model is essentially fixed by the Planck results (see Alam et al. 2017, Figure 11), indicating that a more relevant test of systematics should be done with the AP parameters fixed, often resulting in a ∼20-30% smaller error on $f\sigma_8$. This suggests that RSD analyses from full-shape modeling are already systematics dominated and will certainly be so for future galaxy surveys, without subsequent modeling improvements. While the model presented in this work has its own shortcomings, one such avenue for improvement is exploring more physically motivated model descriptions.

As discussed in Section 4.5.4, our parametrization leads to a derived uncertainty of $f\sigma_8$ that is roughly 10-20% larger than the constraint from the models of Beutler et al. (2017b); Grieb et al. (2017), which use fewer parameters. Each of our parameters has a physical motivation, and we apply reasonable priors based on these motivations when appropriate. Thus, we find no clear path to reduce the number of parameters in our model and do not believe that additional constraining power can be gained through the use of stronger priors. As such, RSD models in the literature are likely too-limited in their parametrization, with the uncertainty of $f\sigma_8$ underestimated by ∼10-20%. For a galaxy sample such as the BOSS CMASS sample with a satellite fraction $f_s \sim 0.1$, the clustering is dominated by the 2-halo correlations of centrals. However, we find the inclusion of parameters to properly treat the 1% effects of satellite-satellite correlations to be crucial to modeling the clustering down to scales of $k \sim 0.4\ h\mathrm{Mpc}^{-1}$. Using a Fisher analysis, we find similar errors on $f\sigma_8$ as found by the models of Beutler et al. (2017b); Grieb et al. (2017) for the N-cutsky mocks (see Table 4.8) when fixing the relative fraction of non-isolated satellites f_{s_B} and the central galaxy velocity dispersion σ_c. In this case, we only vary a single FoG velocity dispersion, as is the case for the models of Beutler et al. (2017b); Grieb et al. (2017), and fix the ∼1% contribution to the overall power spectrum from satellites living in halos containing multiple satellites.

Fully perturbative modeling approaches cannot accurately capture the effects of nonlinearities, i.e., the FoG effect from satellites, on small scales, and at some point, the modeling must become sensitive to the poorly understood physics of galaxy evolution. Presently, it is unclear how sensitive cosmological growth of structure measurements are to such small-scale physics. In particular, assembly bias remains a worrying potential systematic for galaxy clustering analyses (Zentner et al. 2014, 2016). The most promising avenue for including small-scale information in growth of structure analyses appears to be simulation-based modeling efforts. The most competitive constraint to date for $f\sigma_8$ published in Reid et al. (2014) uses a simulation-based model to describe the correlation function down to scales of $r \sim 0.8\,h^{-1}$Mpc. In order to achieve the desired accuracy for the RSD model presented in this work, we also find it necessary to include calibrations from simulations for key components of the model. The combination of perturbation theory with simulation-based calibration in our model likely limits the applicability of the model in comparison to a fully general, simulation-based approach. An emulator-based approach for the nonlinear clustering of galaxies in redshift space using the FastPM simulation method (Feng et al. 2016) is under active development.

An alternative approach for maximizing the constraining power of RSD analyses relies on limiting the effects of satellites on the modeling. These so-called halo reconstruction methods attempt to modify the measurement procedure to preferentially exclude satellites galaxies, thus measuring the clustering of the underlying halo density field, rather than the galaxy density field (Tegmark et al. 2006; Reid & Spergel 2009; Okumura et al. 2017). However, these methods often struggle to achieve a transformation accurate enough such that the added modeling complications from the transformation itself do not outweigh the benefits gained by removing satellites. The advantages include limiting the effects of nonlinearities, which simplifies the modeling and could allow use of models closer to purely linear theory. Reducing FoG effects raises the overall amplitude of the quadrupole and boosts the signal-to-noise of the measurement, although removing satellite galaxies does lower the overall bias, which reduces the constraining power of a given measurement.

As an illustration, we present our RSD constraints when fitting our 13 parameter model to the clustering of centrals and type A centrals (isolated centrals that have no satellites) from the RunPB simulations. We show the best-fit multipoles for these cases in comparison to the spectra of the full galaxy sample in Figure 4.14. As expected, the small-scale quadrupole shows a significant reduction in the effects of RSD, and we find a reduction in the linear bias due to the removal of the highly biased satellites. Corresponding parameter constraints for $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$ are presented in Table 4.9. We find the largest decreases in uncertainty when considering centrals only – the error on $f\sigma_8$ decreases by 31%, 19%, and 15% when fitting to $k_\mathrm{max} = 0.2, 0.3$, and $0.4\ h\mathrm{Mpc}^{-1}$, respectively. Similarly, we find decreases of 16%, 7%, and 25% for $\alpha_\parallel$ and 19%, 10%, and 0% for $\alpha_\perp$. While we find diminishing benefits to extending the fitting range from $k_\mathrm{max} = 0.2\ h\mathrm{Mpc}^{-1}$ to $k_\mathrm{max} = 0.4\ h\mathrm{Mpc}^{-1}$, the constraints using only centrals are in all cases better than when using all galaxies. Furthermore, fitting the clustering of only centrals to $k_\mathrm{max} = 0.2\ h\mathrm{Mpc}^{-1}$ is roughly as competitive in constraining $f\sigma_8$

as fitting the clustering of all galaxies to $k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$, and the latter is significantly more challenging to model than the former. While we recognize that this is certainly an idealized demonstration, we view halo reconstruction methods as an important area of future research for both their constraining power and simplified theoretical modeling.

Another avenue forward suggested in the literature relies on the use of nonlinear transformations of the density field, which can extend the range of scales that can be easily modeled, e.g., Neyrinck (2011); Wolk et al. (2015); Repp & Szapudi (2017). These methods have mostly been applied to dark matter to date, and since we observe galaxies, challenges exist in applying these methods to observational data sets. In addition, any nonlinear transform of the density field creates a velocity bias, even on large, linear scales, (e.g., Seljak 2012), which makes the linear RSD interpretation more difficult, further complicating the modeling procedure.

Recent RSD analyses of measurements of the void-galaxy cross-correlation (Hamaus et al. 2016, 2017; Hawken et al. 2017; Cai et al. 2016; Achitouv et al. 2017) similarly aim to simplify theoretical modeling by arguing that RSDs around voids can be described by linear theory. Initial studies have produced encouraging results, with constraints on the growth of structure competitive with those from state-of-the-art galaxy clustering studies. This approach to RSD includes its own set of modeling assumptions and challenges, including sensitivities to void profiles and sizes, and complicated velocity flows around and within voids. These methods are relatively new, and the understanding of these theoretical systematics have not yet been fully explored. Nevertheless, the method appears promising for measuring RSD in a manner complementary to galaxy clustering analyses.

4.7 Conclusion

We present a new model for the redshift-space power spectrum of galaxies and demonstrate its accuracy in modeling the monopole, quadrupole, and hexadecapole of the galaxy density field down to $k = 0.4\ h\mathrm{Mpc}^{-1}$ through a series of tests on high-fidelity N-body simulations. The model describes the clustering of galaxies in the context of a halo model, building upon the formalism presented in Okumura et al. (2015). We decompose galaxies into four subsamples: centrals with and without satellites and satellites with one or more neighboring satellite. We then model the clustering of the underlying halos in redshift space using a combination of Eulerian perturbation theory and N-body simulations. The modeling of RSD via the mapping from real space to redshift space is done using the so-called distribution function approach. In order to achieve sufficient accuracy in the modeling down to $k = 0.4\ h\mathrm{Mpc}^{-1}$, we utilize a set of simulations to calibrate the most important terms of the model. To this end, we extend the Halo-Zel'dovich Perturbation Theory of Seljak & Vlah (2015), which combines Lagrangian perturbation theory with physically motivated corrections calibrated from simulations. We improve the accuracy of this model for the dark matter density power spectrum and develop models for the dark matter velocity correlators P_{01} and P_{11}. Our final model has 13 free parameters, each of which is physically moti-

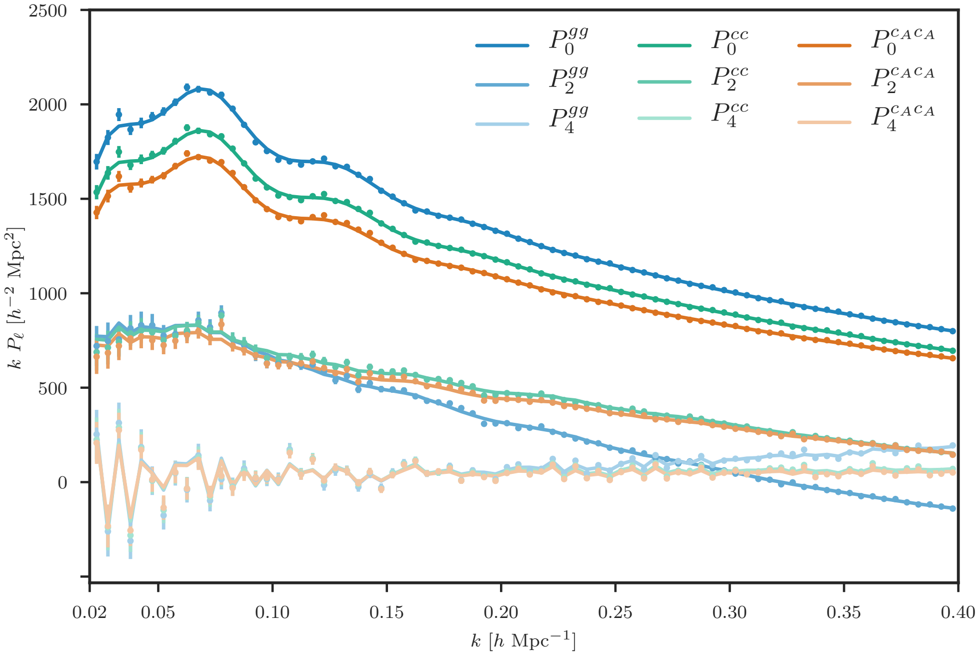


Figure 4.14: The best-fitting model and measured simulation points for the monopole (darkest shade), quadrupole, and hexadecapole (lightest shade) from the mean of 10 RunPB galaxy catalogs at $z = 0.55$ for all galaxies P_ℓ^{gg} (blue), all centrals P_ℓ^{cc} (green), and isolated centrals with no satellites in the same halo $P_\ell^{c_Ac_A}$ (orange). Linear biases for each sample are $b_{1,g} = 2.05$, $b_{1,c} = 1.93$, and $b_{1,c_A} = 1.84$.

$k_{\max}$		all galaxies	centrals only	type A centrals only
$k_{\max} = 0.2\ h\mathrm{Mpc}^{-1}$	$\Delta\alpha_{\parallel}$	$0.0090^{+0.0095}_{-0.0086}$	$0.0096^{+0.0076}_{-0.0076}$	$0.0101^{+0.0081}_{-0.0081}$
	$\Delta\alpha_{\perp}$	$0.0029^{+0.0054}_{-0.0063}$	$0.0023^{+0.0048}_{-0.0046}$	$0.0023^{+0.0047}_{-0.0047}$
	$\Delta f\sigma_8$	$-0.0174^{+0.0149}_{-0.0154}$	$-0.0042^{+0.0108}_{-0.0100}$	$-0.0089^{+0.0116}_{-0.0108}$
$k_{\max} = 0.3\ h\mathrm{Mpc}^{-1}$	$\Delta\alpha_{\parallel}$	$0.0065^{+0.0078}_{-0.0074}$	$0.0049^{+0.0072}_{-0.0070}$	$0.0038^{+0.0069}_{-0.0067}$
	$\Delta\alpha_{\perp}$	$0.0042^{+0.0049}_{-0.0046}$	$0.0044^{+0.0043}_{-0.0043}$	$0.0046^{+0.0045}_{-0.0046}$
	$\Delta f\sigma_8$	$-0.0048^{+0.0122}_{-0.0128}$	$0.0066^{+0.0100}_{-0.0104}$	$0.0031^{+0.0093}_{-0.0093}$
$k_{\max} = 0.4\ h\mathrm{Mpc}^{-1}$	$\Delta\alpha_{\parallel}$	$0.0089^{+0.0068}_{-0.0077}$	$0.0030^{+0.0055}_{-0.0054}$	$0.0012^{+0.0061}_{-0.0068}$
	$\Delta\alpha_{\perp}$	$0.0050^{+0.0042}_{-0.0040}$	$0.0029^{+0.0044}_{-0.0039}$	$0.0012^{+0.0044}_{-0.0041}$
	$\Delta f\sigma_8$	$-0.0031^{+0.0097}_{-0.0091}$	$0.0031^{+0.0085}_{-0.0076}$	$0.0059^{+0.0083}_{-0.0081}$

Table 4.9: The best-fit $f\sigma_8$, $\alpha_{\perp}$, and $\alpha_{\parallel}$ values and 1σ uncertainties obtained when fitting the monopole, quadrupole, and hexadecapole from the mean of 10 RunPB galaxy catalogs at $z = 0.55$ when including all galaxies, centrals only, and type A centrals only, which are isolated with no satellites in the same halo.

vated, as described in Table 4.3. The model accounts for the FoG effect from each of our galaxy subsamples, rather than using a single velocity dispersion to describe the combined effect. We account for the linear bias of each of the subsamples and describe the shot noise contributions to the power spectrum via the amplitude of the 1-halo galaxy correlations.

We fit our 13 parameter model to the monopole, quadrupole, and hexadecapole measured from several sets of simulations to test the accuracy and precision of the recovered parameters. These mock catalogs cover a range of cosmologies and galaxy bias models, providing stringent tests of our model. The test suite also includes realistic mock catalogs of the BOSS DR12 CMASS sample, which properly model the volume and selection effects of this data set. We perform fits as a function of the maximum wavenumber included in the fit, using $k_{\max} = 0.2, 0.3$, and $0.4\ h\mathrm{Mpc}^{-1}$. The results of these tests can be summarized as follows:

(i) Given the measurement covariance and degrees of freedom in the model, we find excellent agreement between our model and the measured $\ell = 0$, 2, and 4 multipoles from simulations down to scales of $k = 0.4\ h\mathrm{Mpc}^{-1}$.

(ii) A systematic shift in the best-fitting value of $\alpha_{\parallel}$ is identified at the level of $0.01 - 0.02$, independent of the $k_{\max}$ value used when fitting. Such a systematic shift can be calibrated from a large set of simulations and a correction applied to the best-fitting value.

(iii) The level of systematic bias in the parameters $f\sigma_8$ and $\alpha_{\perp}$ is found to be small, at the level of ∼0.005, which is similar to other published RSD models in the literature, i.e., Alam et al. (2017). However, considering that the Planck results essentially fix the background cosmology model, comparisons between this level of systematics and the error on $f\sigma_8$ for fixed AP parameters indicate that RSD analyses are nearly systematics dominated today. This will certainly be the case for the next generation of galaxy surveys, unless analyses are limited to the largest scales or substantial modeling improvements are made.

(iv) Using a set of BOSS DR12 CMASS mock catalogs as a benchmark for comparison, we estimate an uncertainty on $f\sigma_8$ that is $\sim$10-20% larger than when using the models of Beutler et al. (2017b); Grieb et al. (2017), when fitting over similar wavenumber ranges. With 5-6 fewer parameters, these models likely have a too-limited parametrization and are underestimating the resulting uncertainty of $f\sigma_8$.

(v) Extending the fitting range to $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$ provides 15-30% gains in the statistical precision of the $f\sigma_8$ constraint relative to $k_{\rm max} = 0.2\ h{\rm Mpc}^{-1}$. The gains are more modest when our model is compared to published models, which use a more limited parametrization; the error on $f\sigma_8$ is roughly 5-10% smaller with our model using $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$ than constraints found when using the models of Beutler et al. (2017b); Grieb et al. (2017) (with $k_{\rm max} \simeq 0.2\ h{\rm Mpc}^{-1}$) for the BOSS DR12 CMASS sample.

(vi) We find a $\sim$10-15% improvement in the constraint on $\alpha_\perp$ and only marginal gains for $\alpha_\parallel$ when extending from $k_{\rm max} = 0.2\ h{\rm Mpc}^{-1}$ to $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$. The constraint on $\alpha_\perp$ represents a 20% improvement relative to the results found when the published models of Beutler et al. (2017b); Grieb et al. (2017). This improvement will further help constrain and de-correlate the parameters $D_A(z)$ and $H(z)$ when combined with post-reconstruction BAO-only analyses.

Extending full-shape RSD modeling of galaxy clustering to smaller scales in both an accurate and precise manner remains a complicated endeavor. As we push into the nonlinear regime, there is no way to avoid the additional modeling complexity. As the number of free parameters necessarily becomes larger, simulations offer a good opportunity to place reasonable priors on model parameters. The question remains whether continuing to push to even smaller scales yields diminishing returns given the increased theoretical complexities. The results of this work suggest that further gains could be unlikely. We find relatively modest benefits when extending from $k_{\rm max} = 0.2\ h{\rm Mpc}^{-1}$ to $k_{\rm max} = 0.4\ h{\rm Mpc}^{-1}$, of order 30%, while the parameter errors increase by 10-20% when going from $\sim$7 parameters to 13 when using $k_{\rm max} < 0.2\ h{\rm Mpc}^{-1}$. Moreover, the parametrization of our model may not be sufficient to fully capture effects such as assembly bias, velocity bias, or other unknowns regarding the small-scale galaxy - halo connection that may become more important for future surveys. These findings suggest we may have exhausted the information content that can be reasonably extracted from RSD in the broadband power spectrum already. However, claims exist in the literature (e.g., Reid et al. 2014) that additional gains can be attained, and it is worth exploring the issue further.

There are several intriguing modeling approaches that can go beyond our analytic modeling approach and potentially improve the constraints further. Simulation-based approaches can leverage advances in high-performance computing to accurately model nonlinear clustering on small scales. Complementary approaches such as halo reconstruction or void-galaxy cross-correlation statistics can simplify modeling and reduce the need to include small-scale information by mitigating the complicated effects of nonlinear effects on the modeling procedure. It remains to be seen if these further advances will improve cosmological constraints, or whether with our model we have reached the limit due to the effects of nonlinear evolution and poorly known small-scale physics.

Applying models such as the one presented here will be necessary at the minimum to fully capitalize on the cosmological information contained in future galaxy surveys, such as the Hobby Eberly Telescope Dark Energy Experiment (Hill et al. 2008), the Dark Energy Spectroscopic Instrument (DESI) (Levi et al. 2013), the Subaru Prime Focus Spectrograph (Takada et al. 2014), and the ESA space mission *Euclid* (Laureijs et al. 2011). For example, we expect our model to perform well on the DESI emission line galaxy sample, which has a lower bias and higher satellite fraction as compared to the BOSS CMASS sample. Conversely, the gains when applying our model to next generation quasar samples would likely be more modest due to the high shot noise and the lower impact of one-halo correlations.

Chapter 5

Constraints on primordial non-Gaussianity from the clustering of eBOSS quasars

This chapter presents constraints on local primordial non-Gaussianity as parameterized by $f_{\rm NL}$ using the quasar sample from Data Release 14 (DR14) of the extended Baryon Oscillation Spectroscopic Survey (eBOSS). The DR14 data set contains 148,659 quasars covering more than 2000 square degrees of the sky. We measure and analyze the anisotropic clustering of the quasars in Fourier space, testing for the scale-dependent bias introduced by primordial non-Gaussianity on large scales. We derive a power spectrum estimator using optimal weights to account for the redshift evolution of $f_{\rm NL}$. As a baseline measurement, we analyze the eBOSS data set over the redshift range $0.8 \leq z \leq 2.2$ using a single redshift bin with an effective redshift of $z_{\rm eff} = 1.53$. We have verified and tested this analysis pipeline using a set of realistic mock catalogs that accurately model the angular selection function and redshift evolution of the DR14 sample. From this analysis, we find $f_{\rm NL} = 14^{+52}_{-73}$ and $f_{\rm NL} = 14^{+55}_{-48}$ when fitting the measured monopole and quadrupole from the NGC and SGC data sets, respectively. A joint fit to the entirety of the DR14 sample leads to a constraint of $f_{\rm NL} = -9^{+43}_{-46}$. We expect the optimal, redshift-weighted estimator derived in this work to improve upon this baseline constraint and plan to analyze the DR14 data using such weights in the near future.

5.1 Introduction

Measurements of the statistical properties of the late-time large-scale structure (LSS) of the Universe can provide insight into the physics that generated primordial density fluctuations. In particular, they offer the possibility to distinguish between different models of cosmic inflation using the fact these these models predict different levels of primordial non-Gaussianity (PNG), the deviation from Gaussian random field initial conditions. In this work, we focus on the local type of PNG, parameterized through the parameter $f_{\rm NL}$. Standard, single-field inflationary models predict an amplitude of $f_{\rm NL}$ that is unmeasurably small, and a detection of $|f_{\rm NL}| \gtrsim 1$ would robustly rule out this class of models (Maldacena 2003; Creminelli & Zaldarriaga 2004).

The current state-of-the-art constraint on PNG comes not from LSS data but from measurements of the bispectrum of the cosmic microwave background (CMB) by the *Planck* satellite, which has reported $f_{\rm NL} = 0.8 \pm 5.0$ (Planck Collaboration et al. 2016b). Unfortunately, the improvement in precision from CMB measurements is not expected to reach the level required to distinguish between inflationary models ($\sigma(f_{\rm NL}) \sim 1$) due to cosmic variance limitations (Baumann et al. 2009; Abazajian et al. 2016). However, forecasts for future LSS surveys, e.g., Doré et al. (2014); Yamauchi et al. (2014); Ferramacho et al. (2014); Ferraro & Smith (2015); Raccanelli et al. (2015); Camera et al. (2015); Alonso & Ferreira (2015); Tucci et al. (2016); de Putter & Doré (2017), indicate a strong potential for PNG constraints. Further gains can be made by surveys that observe multiple tracers, which are able to effectively remove noise from sample variance in their measurements (McDonald & Seljak 2009; Seljak 2009; Hamaus et al. 2011). The sensitivity to PNG originates from the distinctive scale-dependent signature that is imprinted on the clustering of biased tracers (e.g., galaxies or quasars) by local primordial non-Gaussianity (Dalal et al. 2008; Matarrese & Verde 2008; Slosar et al. 2008; Desjacques & Seljak 2010; Alvarez et al. 2014). The effect is proportional to the bias of the tracers themselves and scales as $f_{\rm NL}k^{-2}$; thus, it is most prominent on the largest scales probed by a survey.

The current best constraints from the analysis of large-scale structure data are comparable to those found by the *WMAP* CMB experiment, $\sigma(f_{\rm NL}) \simeq 20$ (Slosar et al. 2008; Ross et al. 2013; Giannantonio et al. 2014; Leistedt et al. 2014; Ho et al. 2015; Karagiannis et al. 2014; Bennett et al. 2013). The first such analysis by Slosar et al. (2008) combined a number of tracers from early SDSS releases to find $f_{\rm NL} = 28^{+23}_{-24}$. This analysis also demonstrated the constraining power of quasar data sets, finding $f_{\rm NL} = 8^{+26}_{-37}$ using only the SDSS photometric quasar sample. As quasars are highly biased and probe large volumes, they are ideal for measuring the PNG signal on large scales. More recent PNG studies have not achieved significant improvement on the constraint of Slosar et al. (2008), with contamination from systematics often hindering results (Pullen & Hirata 2013; Leistedt et al. 2013; Leistedt & Peiris 2014; Ho et al. 2015). Systematics control has spurred work on the use of cross-correlations in LSS PNG analyses, e.g., Giannantonio & Percival (2014); Schmittfull & Seljak (2017).

Data sets that probe large volumes offer the best chance to detect non-Gaussian biasing features on large scales, but they also complicate data analysis. For samples that span a wide redshift range, traditional analysis methods, such as using multiple, smaller redshift bins, become non-optimal. A proper treatment of the redshift evolution of the tracer bias and PNG signal is necessary to fully exploit the constraining power of a data set. Recent work has focused on using redshift weights to optimize LSS surveys for baryon acoustic oscillation (BAO) and redshift-space distortion (RSD) analyses (Zhu et al. 2015, 2016; Ruggeri et al. 2017). This weighting scheme was extended in Mueller et al. (2017) in order to optimize for

PNG constraints. In this work, we derive a redshift-weighted optimal quadratic estimator for the two-point clustering in Fourier space that yields optimal constraints for f_{NL}. The weights are similar to those derived by Mueller et al. (2017), but here, the clustering is estimated via a cross-correlation of two optimally weighted density fields. This differs from the method presented in Mueller et al. (2017), which advocates computing the auto-correlation of a single density field with tracer objects weighted by the square root of the total desired weight.

In this work, we use the Sloan Digital Sky Survey (SDSS) IV extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016) Data Release 14 quasar sample (DR14Q) to derive constraints on local primordial non-Gaussianity. This data set includes 148,659 quasars and spans a redshift range of $0.8 \leq z \leq 2.2$. This work represents an initial study of the DR14Q data set for constraining primordial non-Gaussianity and makes use of a novel technique for optimizing such measurements for data sets spanning wide redshift ranges. We expect many complementary analyses and approaches using this data set to improve upon the analysis methods presented here.

This chapter is organized as follows. In §5.2 we describe the eBOSS quasar sample described in this work. We present our new optimal estimator, which accounts for redshift evolution of f_{NL}, in §5.3. §5.4 outlines our analysis methods, including how we estimate the power spectrum multipoles of the data and the theoretical model used to estimate parameters. We present our constraints on f_{NL} in §5.5 and discuss and conclude in §5.6.

5.2 Data

In this section, we describe the eBOSS DR14Q sample and the synthetic mock catalogs used in our analysis.

5.2.1 eBOSS DR14Q sample

The extended Baryon Oscillation Spectroscopic Survey (Dawson et al. 2016) is part of the SDSS-IV experiment (Blanton et al. 2017). The eBOSS cosmology program relies on the same optical spectrographs (Smee et al. 2013) as the SDSS-III BOSS survey, installed on the 2.5 meter Sloan Foundation Telescope (Gunn et al. 2006) at the Apache Point Observatory in New Mexico. In addition to observing luminous red galaxies and emission line galaxies, eBOSS will observe and measure redshifts for ∼500,000 quasars across a volume of the Universe unexamined by previous spectroscopic surveys. First eBOSS cosmology results for the DR14Q sample were recently presented in Ata et al. (2018), which reported the first BAO distance measurement in the range $1 < z < 2$. The clustering properties of the eBOSS quasars have also been previously examined in Laurent et al. (2017); Rodríguez-Torres et al. (2017), although these works do not make use of the full DR14Q sample.

The imaging data, target selection, and catalog construction methods for the DR14Q sample used in this work are discussed in detail in Pâris et al. (2017). Targets are selected from the catalogs of the SDSS I/II surveys (York et al. 2000), released as part of SDSS DR7

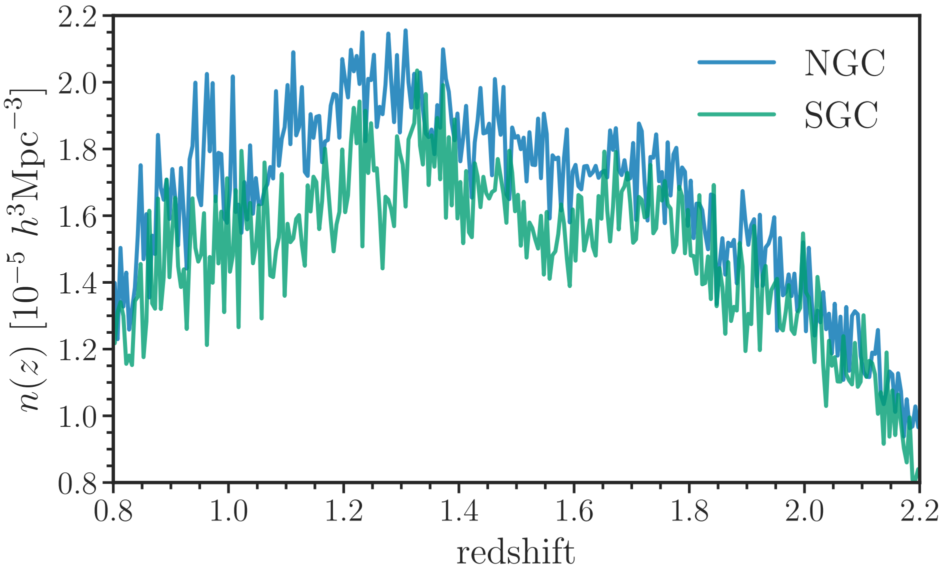


Figure 5.1: The mean density of quasars in the DR14Q sample as a function of redshift for the NGC (blue) and SGC (green) regions of the sky. The differences between the two regions are due to known discrepancies with the targeting efficiency.

Abazajian et al. (2009), and the SDSS-III survey (Eisenstein et al. 2011), released as part of SDSS DR8 (Aihara et al. 2011). eBOSS also makes use of several bands of the Wide Field Infrared Survey Explorer (WISE; Wright et al. 2010), as described in Myers et al. (2015). The target selection criteria for the DR14Q sample is presented in detail in Myers et al. (2007). Objects that satisfy this criteria and do not have a previously measured redshift are assigned a fiber as part of the eBOSS observations. Accurate redshift estimation is crucial for achieving the cosmology goals of eBOSS, which is particularly challenging for quasar spectra (Shen et al. 2016). As described in Pâris et al. (2017), the DR14Q sample contains three automated redshift estimates per object. In this work, we use the so-called "fiducial" redshift z_{fid}, which can be any of the three redshift estimates, depending on which one yields the lowest catastrophic failure rate (see Pâris et al. (2017) for further details).

The DR14Q sample contains 148,659 objects with spectroscopic redshifts in the range $0.8 \leq z \leq 2.2$. The observed objects are distributed in two separate angular regions in the North Galactic Cap (NGC) and South Galactic Cap (SGC). The effective areas of these regions are 1214.6 deg^2 and 899.3 deg^2, respectively. We show the observed number density as a function of redshift for the NGC and SGC regions in Figure 5.1. There are slight discrepancies in $n(z)$ between the two regions due to differences in targeting efficiency.

5.2.2 Completeness weights

Objects in the DR14Q sample are assigned weights to account for the incompleteness of the target selection process and other systematic effects that could potentially bias our clustering measurements. There are two main types of weights that we will discuss in this section: spectroscopic completion weights $w_{\rm spec}$ and systematic imaging-based weights $w_{\rm sys}$. The former accounts for the fact that a small percentage of targets do not receive a redshift while the latter set of weights corrects for systematics arising from photometric inhomogeneities in the targeting selection. When combining these two sets of weights, we take the total completeness weight as

$$w_{\rm c} = w_{\rm sys} \cdot w_{\rm spec}. \tag{5.1}$$

Spectroscopic weights

The first main cause of spectroscopic incompleteness in the DR14Q sample is *fiber collisions*. Fiber collisions result when a pair of quasars are separated by less than the 62″angular size of the SDSS spectrograph fiber, which prevents one of the objects from being observed. Missed observations are partially alleviated by the eBOSS tiling pattern, which naturally has overlapping tiles in regions with a higher density of targets on the sky, and thus, allows redshifts to be measured for objects separated by less than the 62″collision scale. Ultimately, 4% and 3% of the eBOSS quasar targets are fiber-collided objects that do not receive a spectroscopic observation in the NGC and SGC regions, respectively.

We account for the missing objects due to fiber collisions by up-weighting the nearest neighbor with a valid redshift and spectroscopic class. This procedure follows previous clustering analyses, e.g., as in BOSS (Anderson et al. 2014a; Reid et al. 2014). In practice, this is not a perfect correction, as a fraction of fiber collision pairs are projections and are not associated with the same dark matter halo. However, the nearest neighbor weighting scheme does preserve the large-scale bias of the clustering sample. As we are concerned only with the PNG signal on large scales, we leave exploration of more advanced fiber collision correction schemes, e.g., Hahn et al. (2017), for future work. We denote the weight used to correct for fiber collisions as a *close pair* weight, $w_{\rm cp}$. By default, its value is unity for all objects that are not involved in a fiber collision, and for the case of fiber collisions, it is equal to an integer with value greater than unity.

The second main cause of spectroscopic incompleteness is *redshift failures*, which refers to the subset of quasars that do not receive a valid automated redshift and are not visually inspected. The distribution of these objects is not uniform within the focal plane due to variations in detector efficiency. In past BOSS releases (Reid et al. 2016), redshift failures were an almost negligible fraction of the total number of objects, less than 1%. However, redshift determination for a quasar at $z \sim 1.5$ is more difficult than for an LRG at $z \sim 0.5$, and the DR14Q sample has a redshift failure rate of 3.4% and 3.6% in the NGC and SGC, respectively. With this increased rate, a more complex scheme than was used in previous BOSS analyses is required to adequately correct for the effect. This more complicated correction procedure was not used in the first eBOSS BAO analysis of Ata et al. (2018), but was implemented for the RSD analyses (Gil-Marín et al. in prep.). Here, we use a focal plane weight $w_{\rm foc}$ defined as

$$w_{\rm foc} = \left[1 - P_{\rm rf}(x_{\rm foc}, y_{\rm foc})\right]^{-1}\,, \tag{5.2}$$

where $P_{\rm rf}$ defines the probability of obtaining a redshift failure as a function of position in the focal plane. With this weight, quasars with measured redshifts that are observed in positions on the focal plane where $P_{\rm rf}$ is greater than zero will be up-weighted to account for the fact that, on average, targeted quasars are missing from the sample due to redshift failures. We refer the reader to Gil-Marín et al. (in prep.) for further details on the redshift failure weights. Finally, we assign the total spectroscopic completeness weight as the product of the fiber collision and redshift failure weights, $w_{\rm spec} = w_{\rm cp} \cdot w_{\rm foc}$.

Imaging weights

Each quasar in the DR14Q sample is also assigned a weight to mitigate photometric systematics, using the prescription studied in Laurent et al. (2017). The weights, denoted here as $w_{\rm sys}$, account for inhomogeneities in the quasar targeting selection related to the Galactic extinction and depth of the targeting image data. The weights used in this work have been utilized in previous eBOSS cosmology analyses (Ata et al. 2018; Gil-Marín et al. in prep.). They are described in detail in Section 3.4 of Ata et al. (2018), and we refer the reader to that work for further details.

5.2.3 Synthetic DR14Q catalogs

We make use of a set of mock catalogs designed to mimic the observational features of the DR14Q data set. The mocks are based on the Extended Zel'dovich (EZ) approximate N-body simulation scheme (Chuang et al. 2015). Throughout this work, we refer to this set of simulations as EZ mocks. In total, we utilize 1,000 independent realizations for each Galactic cap region. This large set of mocks enables us to estimate the covariance matrix of our clustering estimator. We also use the mocks to verify and test our analysis and parameter estimation pipelines.

The set of EZ mocks is generated following the methodology outlined in Chuang et al. (2015), matching both the angular footprint and redshift selection function of the DR14Q sample. Briefly, the EZ mock scheme relies on the Zel'dovich approximation to generate a density field, and implements nonlinear and halo biasing effects through the use of free parameters. These free parameters can be tuned to reproduce the two-point and three-point clustering of a desired data set. The method allows for the fast generation of a large number of mock catalogs without the computational cost of full N-body simulations, and it has been used extensively in previous BOSS cosmology analyses, e.g., Kitaura et al. (2016); Alam et al. (2017).

The EZ mock catalogs account for the redshift evolution of the eBOSS quasars by constructing a light-cone out of 7 redshift shells, generated from periodic boxes of side length

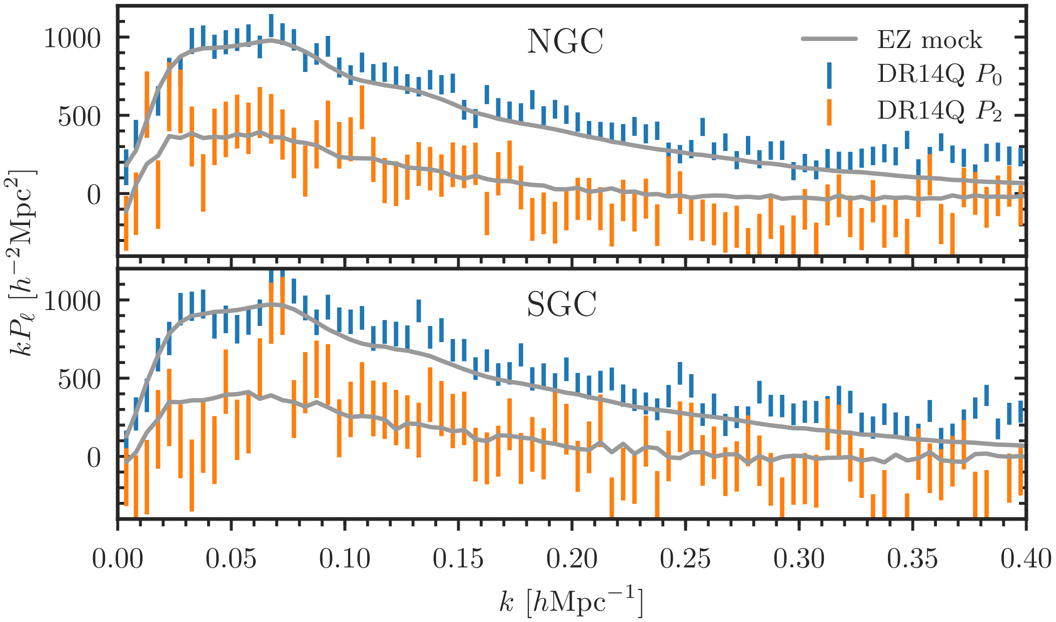


Figure 5.2: The measured power spectrum multipoles for the DR14Q sample, as compared to those measured from the mean of the 1,000 EZ mock catalogs. We show the comparison separately for the NGC (top) and SGC (bottom) data sets. Errors on the data measurements are computed from the variance of the 1,000 EZ mock measurements. We do not show error bars on the mean of the EZ mock multipoles (grey), as they are small compared to that of the data.

$L = 5000\ h^{-1}$Mpc at different redshifts. The free parameters of each box are calibrated independently, and the boxes are combined using the `make_survey` software (Carlson & White 2010). The background density field of the light cone mocks is continuous, as each of the boxes shares the same initial Gaussian density field. The NGC and SGC data sets are treated independently when deriving the best-fit internal EZ mock parameters. The cosmology of the EZ mocks is a flat, ΛCDM model with $\Omega_\mathrm{m} = 0.307115$, $\Omega_\mathrm{b} = 0.048206$, $h = 0.6777$, $\sigma_8 = 0.8255$, and $n_\mathrm{s} = 0.9611$. The mock catalogs also include the effects of fiber collisions and redshift failures (as discussed in Section 5.2.2). Each object in an EZ mock catalog has associated values for w_foc and w_cp. Fiber collisions are implemented by applying the tiling pattern to the mock data and removing pairs that fall within the collision scale that are not in overlapping tiles. Redshift failures are applied by statistically removing objects based on the position of the object in the focal plane, using the probability of a redshift failure $P_\mathrm{rf}(x_\mathrm{foc}, y_\mathrm{foc})$. We compare the power spectrum multipoles of the EZ mock catalogs and the DR14Q data set in Figure 5.2.

5.3 Theory

5.3.1 Local primordial non-Gaussianity

In this chapter, we focus on the local type of primordial non-Gaussianity, where the primordial potential is the sum of a random Gaussian field and its square,

$$\Phi = \phi + f_{\rm NL}\left(\phi^2 - \langle\phi^2\rangle\right), \tag{5.3}$$

where Φ is the primordial potential, ϕ is a Gaussian random field, and $f_{\rm NL}$ parametrizes the level of PNG present. The relation between Φ and the matter overdensity $\delta_{\rm m}$ is easiest to express in Fourier space, where it is given by $\delta_{\rm m}(k) = \alpha(k)\Phi(k)$, with

$$\alpha(k) = \frac{2c^2k^2T(k)D(z)}{3\Omega_{\rm m,0}H_0^2} \tag{5.4}$$

where $T(k)$ is the transfer function, c is the speed of light, $D(z)$ is the linear growth factor normalized to $(1+z)^{-1}$ in the matter-dominated era, $\Omega_{\rm m,0}$ is the present-day matter density parameter, and H_0 is the present-day Hubble parameter. We also define a related quantity $\widetilde{\alpha}$, which will be useful in the discussion to follow:

$$\widetilde{\alpha}(k) \equiv \frac{2\delta_{\rm c}}{\alpha(k)} = \frac{3\Omega_{\rm m,0}H_0^2\delta_{\rm c}}{c^2k^2T(k)D(z)}, \tag{5.5}$$

where $\delta_{\rm c} = 1.686$ is the critical density in the spherical collapse model.

As shown in Dalal et al. (2008); Slosar et al. (2008); Desjacques & Seljak (2010), local PNG as parameterized by $f_{\rm NL}$ introduces a scale-dependent halo bias, $\Delta b(k)$, given by

$$\Delta b(k) = 2(b-p)f_{\rm NL}\frac{\delta_{\rm c}}{\alpha(k)} = (b-p)f_{\rm NL}\widetilde{\alpha}(k), \tag{5.6}$$

where b is the bias of the sample, and p is a parameter that takes a value of 1 for a halo-mass-selected sample and 1.6 for a sample dominated by recent mergers (Slosar et al. 2008). It is argued in Slosar et al. (2008) that $p = 1.6$ is the more appropriate choice when studying quasars but we will keep our description in terms of a general p for now. At the linear order, and after adding redshift-space distortions (e.g., Kaiser 1987), the quasar overdensity is related to the matter overdensity in the presence of PNG as

$$\delta_{\rm QSO} = \left(b + f\mu^2 + \Delta b\right)\delta_{\rm m} \equiv \left(\widetilde{b} + \Delta b\right)\delta_{\rm m} \tag{5.7}$$

where $f = d\ln D/d\ln a$ is the logarithmic growth rate, and we have defined the convenient quantity $\widetilde{b} \equiv b + f\mu^2$, which accounts for both linear biasing and redshift-space distortions.

5.3.2 Optimal estimators in LSS

Our goal is to derive an estimator for the two-point clustering of a data set that yields the tightest constraint on $f_{\rm NL}$. We begin by describing the data, quasar positions, in terms of the pixelized overdensity $\delta_{\rm QSO}(\boldsymbol{r}_i)$, where $\boldsymbol{r}_i$ gives the pixel position. We will also need the mean density at a given pixel position, denoted as $\bar{n}(\boldsymbol{r}_i)$. An optimal analysis invariably requires inverse noise weighting of the data. For example, if $\bar{n}(\boldsymbol{r}_i) = 0$ then no data has been observed at that pixel and it should not be used for data analysis, suggesting that the noise should be infinite. An additional source of uncertainty is sample variance, which is caused by the finite number of measurable modes and is present even in the absence of noise.

When considering Gaussian statistics, the optimal inverse noise weighting of a data set has a well-defined solution, known as the optimal quadratic estimator (Tegmark 1997a; Bond et al. 2000), which weights the data inversely by the covariance matrix. If we collect our overdensity pixels into a vector $\boldsymbol{x}$, with $x_i = \delta_{\rm QSO}(\boldsymbol{r}_i)$, then its signal covariance matrix is S_{ij}, and the total covariance matrix is

$$C_{ij} = \langle x_i x_j \rangle = [V\bar{n}(\boldsymbol{r}_i)]^{-1}\delta^K_{ij} + S_{ij}, \tag{5.8}$$

where δ^K_{ij} is the Kronecker delta, V is the pixel volume, and we have assumed Poisson statistics for the noise term.

The optimal quadratic estimator for a parameter α is (Tegmark 1997b; Tegmark et al. 1997, 1998; Abramo et al. 2016)

$$\widehat{q}_\alpha = \frac{1}{2}\boldsymbol{x}^t C^{-1} C_{,\alpha} C^{-1}\boldsymbol{x} - \Delta q_\alpha, \tag{5.9}$$

where $C_{,\alpha}$ denotes the derivative of C with respect to α, and Δq_α subtracts the bias of the estimator.

The most difficult task is to compute $C^{-1}\boldsymbol{x}$. We shall simplify the calculation and use a diagonal form for the covariance matrix C. Suppose we want to determine the power spectrum at some k, where we expect the power to be approximated by a fiducial power spectrum $P_{\rm fid}$. If we assume that the power spectrum is locally flat (white noise) then its Fourier transfer would be a zero-lag correlation function determined by the amplitude of the power spectrum. This gives rise to a diagonal covariance matrix in configuration space:

$$C_{ij} = (P_{\rm fid} + \bar{n}^{-1})V^{-1}\delta^K_{ij}, \tag{5.10}$$

The fiducial power spectrum should in principle be varied with k, but this is usually not implemented. Here, we are concerned with the power on the largest scales, and the fiducial value will be of order $P_{\rm fid} \sim 3 \times 10^4\ h^{-3}{\rm Mpc}^3$.

We also need to evaluate the derivative $C_{,\alpha}$, where α is the parameter we wish to determine. Suppose we focus first on a single mode $\boldsymbol{k}$ with a volume $d\boldsymbol{k} = (2\pi)^3/V$. The Fourier transform of the power spectrum is the correlation function, which for this single mode gives

$C_{ij} = V^{-1}P(\boldsymbol{k})\exp[i\boldsymbol{k}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)]$. Its derivative with respect to $P(\boldsymbol{k})$ gives

$$\frac{dC_{ij}}{dP(\boldsymbol{k})} = V^{-1}e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j)}, \tag{5.11}$$

and the estimator of equation 5.9 for the power spectrum becomes

$$\widehat{P}(\boldsymbol{k}) = A\left|\sum_i e^{i\boldsymbol{k}\cdot\boldsymbol{r}_i}w_{\mathrm{FKP}}\right|^2, \tag{5.12}$$

where we have replaced the sum over pixels with a sum over discrete objects, such that $\delta_{\mathrm{QSO}}\bar{n}V = N_{\mathrm{QSO}}$, where N_{QSO} is the number of objects in the pixel (if the pixels are small enough, this can be viewed just as a sum over object positions $\boldsymbol{r}_i$). The weights w_{FKP} take the well-known form as first derived in Feldman et al. (1994), $w_{\mathrm{FKP}} = (1+\bar{n}P_{\mathrm{fid}})^{-1}$. We see that the operation in equation 5.12 is a Fourier transform, which can be computed rapidly using a fast Fourier transform (FFT) operation. The normalization A can be determined from performing the same operation on an unclustered catalog of synthetic objects, including FKP weights, and normalized to the total number of observed objects (Feldman et al. 1994; Yamamoto et al. 2006; Bianchi et al. 2015; Scoccimarro 2015; Hand et al. 2017b).

5.3.3 An optimal estimator for f_{NL}

Next, we consider instead the weighting scheme that yields an optimal constraint on f_{NL}. We explicitly account for redshift evolution by considering overdensity pixels as a function of time, $\boldsymbol{r} = \boldsymbol{r}(t)$. We begin by computing the signal covariance in the presence of PNG from equation 5.7,

$$\begin{aligned} S_{12} &= \langle\delta_{\mathrm{QSO}}(\boldsymbol{r}_1(t_1))\delta_{\mathrm{QSO}}(\boldsymbol{r}_2(t_2))\rangle & (5.13)\\ &= \left\langle\left[(\widetilde{b}_1+\Delta b_1)\delta_{\mathrm{m}}(\boldsymbol{r}_1)\right]\left[(\widetilde{b}_2+\Delta b_2)\delta_{\mathrm{m}}(\boldsymbol{r}_2)\right]\right\rangle, & (5.14)\end{aligned}$$

where we have defined the quantities $\widetilde{b}_1 = \widetilde{b}(t_1)$, $\boldsymbol{r}_1 = \boldsymbol{r}(t_1)$, $\Delta b_1 = \Delta b(t_1)$, and similar quantities at $t = t_2$. Evaluating the derivative of this expression at $f_{\mathrm{NL}} = 0$ yields

$$\left.\frac{dS_{12}}{df_{\mathrm{NL}}}\right|_{f_{\mathrm{NL}}=0} = \widetilde{b}_1(b_2-p)\widetilde{\alpha}_2\langle\delta_{\mathrm{m}}(\boldsymbol{r}_1)\delta_{\mathrm{m}}(\boldsymbol{r}_2)\rangle + \ 1\leftrightarrow 2, \tag{5.15}$$

where the second term is symmetric and can be computed via an exchange of indices. We can use the definition of the the power spectrum to express this equation as

$$\left.\frac{dS_{12}}{df_{\mathrm{NL}}}\right|_{f_{\mathrm{NL}}=0} = (b_1-p)b_2\int\frac{d\boldsymbol{k}}{2\pi^3}\widetilde{\alpha}_1(k)(1+\beta_2\mu_{r_2}^2)P_{\mathrm{m}}(k,t_1,t_2)e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)} + \ 1\leftrightarrow 2, \tag{5.16}$$

where $P_{\rm m}(k)$ the matter power spectrum, $\beta = f/b$ is the standard RSD parameter, and μ_{r_2} is the line-of-sight angle associated with position $\boldsymbol{r}_2$.

It is useful to parametrize some of the time dependencies in equation 5.16, using

$$P_{\rm m}(k,t_1,t_2) = P_{\rm m,0}(k)D(t_1)D(t_2), \tag{5.17}$$

$$\widetilde{\alpha}(k,t) = \frac{\widetilde{\alpha}_0(k)}{D(t)}, \tag{5.18}$$

where we have defined $P_{\rm m,0} \equiv P_{\rm m}(k,t_0)$ and $\widetilde{\alpha}_0 \equiv \widetilde{\alpha}(k,t_0)$. With these definitions, we can express the optimal estimator of equation 5.9 as a function of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ as

$$\begin{aligned}\widehat{q}_{f_{\rm NL}}(\boldsymbol{r}_1,\boldsymbol{r}_2) &= \frac{1}{2}C^{-1}\boldsymbol{x}^t\frac{dS_{12}}{df_{\rm NL}}\bigg|_{f_{\rm NL}=0} C^{-1}\boldsymbol{x} - \Delta q_{f_{\rm NL}} \\ &= \frac{1}{2}\frac{\delta_{\rm QSO}(\boldsymbol{r}_1)}{C}\Bigg[\int\frac{d\boldsymbol{k}}{2\pi^3}e^{i\boldsymbol{k}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)}P_{\rm m,0}(k)\widetilde{\alpha}_0(k)D(t_2)(1+\beta_2\mu^2_{\boldsymbol{r}_2})(b_1-p)b_2 \\ &+\ 1\leftrightarrow 2\Bigg]\frac{\delta_{\rm QSO}(\boldsymbol{r}_2)}{C} - \Delta q_{f_{\rm NL}}.\end{aligned} \tag{5.19}$$

And now, summing over $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, we obtain the estimator

$$\begin{aligned}\widehat{q}_{f_{\rm NL}} &= \frac{1}{2}\int\frac{d\boldsymbol{k}}{2\pi^3}P_{\rm m,0}(k)\widetilde{\alpha}_0(k) \\ &\Bigg\{\Bigg[\int d\boldsymbol{r}_1\ e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1}\frac{\delta_{\rm QSO}(\boldsymbol{r}_1)}{C}(b_1-p)\Bigg]\Bigg[\int d\boldsymbol{r}_2\ e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_2}\frac{\delta_{\rm QSO}(\boldsymbol{r}_2)}{C}b_2D(t_2)(1+\beta_2\mu^2_{\boldsymbol{r}_2}))\Bigg] \\ &+\ 1\leftrightarrow 2\Bigg\} - \Delta q_{f_{\rm NL}}.\end{aligned} \tag{5.20}$$

Note that in these equations the inverse noise weight factors of C^{-1} are identical to those discussed in Section 5.3.2, with the FKP weight being the near-optimal scheme. Thus, equation 5.20 suggests a power spectrum estimator of the following form to be near optimal:

$$\widehat{q}_{f_{\rm NL}} = \sum_k\left[\tilde{\delta}^b_{\rm QSO}(\boldsymbol{k})\tilde{\delta}^{b-p}_{\rm QSO}(-\boldsymbol{k})\right]P_{\rm m,0}(k)\widetilde{\alpha}_0(k) - \Delta q_{f_{\rm NL}}, \tag{5.21}$$

where we have introduced two weighted fields, $\delta^{b-p}_{\rm QSO} = (b(z){-}p)\delta_{\rm QSO}$ and $\delta^b_{\rm QSO} = D(z)b(z)\delta_{\rm QSO}$, and their FKP-weighted Fourier transforms as

$$\tilde{\delta}^{b-p}(\boldsymbol{k}) = \sum_i \left[b_i(z_i) - p\right] e^{i\boldsymbol{k}\cdot\boldsymbol{r}_i} w_{\rm FKP}(z_i), \tag{5.22}$$

$$\tilde{\delta}^{b}(\boldsymbol{k}) = \sum_i D(z_i) b(z_i) e^{i\boldsymbol{k}\cdot\boldsymbol{r}_i} w_{\rm FKP}(z_i), \tag{5.23}$$

where we have explicitly included the redshift dependency of the density field weights, and we sum over discrete objects labeled by index i. Note that in equation 5.23 we have taken a real-space limit, ignoring a factor of $(1+\beta\mu^2)$, which represents a sub-dominant contribution to the overall weighting scheme.

The last term in equation 5.21 provides signal weighting of the power spectrum, which increases towards low k due to the k^{-2} inside $\widetilde{\alpha}(k)$; it also contains the standard $P(k)$ weight term appropriate if we want to determine the overall amplitude of the power spectrum. Equation 5.21 also shows that an optimal analysis should weight one density field with bias $D(z)b(z)$, the second field with $b(z) - p$, and compute the cross-correlation. Finally, the estimator of equation 5.21 needs to be made unbiased by subtracting out the signal in the absence of any $f_{\rm NL}$ via the $\Delta q_{f_{\rm NL}}$ term.

In summary, an optimal analysis for constraining $f_{\rm NL}$ should apply a total weight to each quasar object that includes the standard FKP minimum variance weight as well as the redshift weight derived in this section. We shall define $w^X_{\rm tot}$ such that $w^X_{\rm tot} \in [w^b_{\rm tot}, w^{b-p}_{\rm tot}]$, and for a quasar with redshift z_i, we have

$$\begin{aligned} w^X_{\rm tot}(z_i) &= w^X_z(z_i) w_{\rm FKP}(z_i) \\ &= \begin{cases} [b(z_i) - p]\ \ w_{\rm FKP}(z_i) & \text{when } X = b-p \\ D(z_i)b(z_i)\ w_{\rm FKP}(z_i) & \text{when } X = b, \end{cases} \end{aligned} \tag{5.24}$$

and $p \in \{1, 1.6\}$.

5.4 Analysis Methods

5.4.1 Fiducial cosmology

Throughout our analysis, we assume the flat ΛCDM cosmology from Planck Collaboration et al. (2016a) as our fiducial background cosmology. The parameter set we use is $h = 0.6774$, $\Omega_{\rm b}h^2 = 0.0223$, $\Omega_{\rm c}h^2 = 0.1188$, $n_{\rm s} = 0.9667$, and $\sigma_8 = 0.8159$. We use this fiducial cosmology to convert observed quasar coordinates (right ascension, declination, and redshift) to Cartesian coordinates during the estimation of the power spectrum of the sample (see §5.4.2). The fiducial cosmology also determines the shape of the real-space matter power spectrum, which is used in our theoretical modeling (see §5.4.3).

5.4.2 Power spectrum estimation

We begin by defining the weighted quasar density field (Feldman et al. 1994)

$$F^X(\boldsymbol{r}) = \frac{w^X_{\rm tot}(\boldsymbol{r})}{A^{1/2}} \left[n'_{\rm QSO}(\boldsymbol{r}) - \alpha'_{\rm s} n_{\rm s}(\boldsymbol{r}) \right], \tag{5.25}$$

where $n'_{\rm QSO}$ and $n_{\rm s}$ are the number densities of the quasar sample and a synthetic catalog of random objects, respectively. The total weight $w^X_{\rm tot}$, with $X \in \{b, b-p\}$, is given in equation 5.24 and is applied to both the quasar and synthetic samples. The synthetic catalog contains unclustered objects—it is used to define the expected mean density of the survey, accounting for the radial and angular selection functions. The factor $\alpha'_{\rm s}$ gives the ratio of quasars to synthetic objects and properly normalizes the number density of the synthetic catalog. The field $F^X(\boldsymbol{r})$ is normalized by the factor of A, defined as

$$A = \int d\boldsymbol{r} \, \left[w_{\rm tot}(\boldsymbol{r}) n'_{\rm QSO}(\boldsymbol{r}) \right]^2. \tag{5.26}$$

In our notation, quantities marked with a prime ($\prime$) include the completeness weights, $w_{\rm c}$, specified in Section 5.2.2. The synthetic catalog defines our expected number density, and as such, does not require completeness weights. The synthetic sample has a number density $1/\alpha'_{\rm s}$ times more dense than the true sample. We assume that, on average, the relation $\langle n'_{\rm QSO}(\boldsymbol{r}) \rangle = \alpha'_{\rm s} \langle n_s(\boldsymbol{r}) \rangle$ holds true. We define $\alpha'_{\rm s}$ as $\alpha'_{\rm s} = N'_{\rm QSO}/N_{\rm s}$, where $N'_{\rm QSO} = \sum_{\rm QSO} w_c$, and $N_{\rm s}$ is the total number of objects in the synthetic catalog.

Now, the multipoles of the cross-correlation between the weighted density fields $F^b(\boldsymbol{r})$ and $F^{b-p}(\boldsymbol{r})$ can be estimated following Yamamoto et al. (2006), as

$$\widehat{P}_\ell = \frac{2\ell+1}{A} \int \frac{d\Omega_k}{4\pi} \left[\int d\boldsymbol{r}_1 \; F^b(\boldsymbol{r}_1) e^{i\boldsymbol{k}\cdot\boldsymbol{r}_1} \int d\boldsymbol{r}_2 \; F^{b-p}(\boldsymbol{r}_2) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_2} \mathcal{L}_\ell(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{r}}_2) \right] - S_\ell, \tag{5.27}$$

where we have introduced the shot noise contribution S_ℓ, defined as

$$S_\ell = A^{-1} \int d\boldsymbol{r} \; n'_{\rm QSO}(\boldsymbol{r}) (w_{\rm c}(\boldsymbol{r}) + \alpha'_{\rm s}) w^2_{\rm tot}(\boldsymbol{r}) \mathcal{L}_\ell(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{r}}), \tag{5.28}$$

which is only non-negligible relative to $\widehat{P}_\ell$ for the monopole $\ell = 0$. We assume $S_\ell = 0$ for $\ell > 0$.

We compute the normalization (equation 5.26) and shot noise (equation 5.28) as discrete sums over the quasar and synthetic catalogs. To do so, we make use of the following relation:

$$\int d\boldsymbol{r} \; n'_{\rm QSO}(\boldsymbol{r}) \ldots \longrightarrow \sum_i^{N_{\rm QSO}} w_{\rm c}(\boldsymbol{r}_i) \ldots \longrightarrow \alpha'_{\rm s} \sum_i^{N_{\rm s}} \ldots, \tag{5.29}$$

where the integral can be expressed equivalently as a sum over the quasar or synthetic catalogs. Thus, the normalization A can be computed as

$$A = \sum_i^{N_\mathrm{QSO}} n'_\mathrm{QSO}(\boldsymbol{r}_i) w_\mathrm{c}(\boldsymbol{r}_i) w^2_\mathrm{tot}(\boldsymbol{r}_i) \tag{5.30}$$

$$= \alpha'_\mathrm{s} \sum_i^{N_\mathrm{s}} n'_\mathrm{QSO}(\boldsymbol{r}_i) w^2_\mathrm{tot}(\boldsymbol{r}_i). \tag{5.31}$$

Note that while equations 5.30 and 5.31 are equivalent on average, in practice, we use the latter equation to estimate A due to the higher number density of the synthetic catalog. Similarly, we can express the shot noise contribution to the monopole (equation 5.28) as

$$S_0 = A^{-1} \left[\sum_i^{N_\mathrm{QSO}} w^2_\mathrm{c}(\boldsymbol{r}_i) w^2_\mathrm{tot}(\boldsymbol{r}_i) + \alpha'^2_\mathrm{s} \sum_i^{N_\mathrm{s}} w^2_\mathrm{tot}(\boldsymbol{r}_i) \right], \tag{5.32}$$

where the two terms compute the contributions to the shot noise from the quasar and synthetic catalogs, respectively. There is some uncertainty surrounding the impact of fiber collisions and completeness weights on the Poisson shot noise calculation of equation 5.32 (Beutler et al. 2014b, 2017b; Grieb et al. 2017). We choose to use the standard Poisson expression and vary a shot noise parameter while performing parameter estimation to account for any discrepancies (see Section 5.4.3).

Our implementation of equation 5.27 uses the FFT-based estimator of Hand et al. (2017b), which is described in detail in Chapter 3. This estimator builds upon similar estimators presented in Bianchi et al. (2015); Scoccimarro (2015), but reduces the number of FFTs required per multipole using a spherical harmonic decomposition. We calculate the power spectrum multipoles as

$$\widehat{P}_\ell(k) = \frac{2\ell+1}{A} \int \frac{d\Omega_k}{4\pi} F^b_0(\boldsymbol{k}) F^{b-p}_\ell(-\boldsymbol{k}), \tag{5.33}$$

with

$$\begin{aligned} F^X_\ell(\boldsymbol{k}) &\equiv \int d\boldsymbol{r}\ F^X(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \mathcal{L}_\ell(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{r}}), \\ &= \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{k}}) \int d\boldsymbol{r}\ F^X(\boldsymbol{r}) Y^*_{\ell m}(\hat{\boldsymbol{r}}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \end{aligned} \tag{5.34}$$

where $Y_{\ell m}$ are spherical harmonics. Note that equation 5.34 requires the calculation of $2\ell+1$ FFTs for a multipole of order ℓ.

To compute the FFTs required by our estimator, we estimate the overdensity field on a mesh of 1024^3 cells for the quasar and synthetic catalogs using a Triangular Shaped Cloud (TSC) interpolation scheme (see, e.g., Hockney & Eastwood 1981). When interpolating to the mesh, each quasar contributes a weight of $w_c w_{\rm tot}^X$ and each synthetic object a weight of $w_{\rm tot}^X$. When computing FKP weights as part of $w_{\rm tot}$, we use a fiducial power spectrum value of $P_0 = 3 \times 10^4\ h^{-3}{\rm Mpc}^3$, roughly equal to the expected power on the scales where PNG is prominent, $k \simeq 0.03\ h{\rm Mpc}^{-1}$. We use the interlaced grid technique of Sefusatti et al. (2016); Hockney & Eastwood (1981) to limit the effects of aliasing, and we correct for any artifacts of the TSC scheme by de-convolving the window in Fourier space. With the combination of TSC interpolation and interlacing, we are able to measure the power spectrum multipoles up to $k = 0.4\ h{\rm Mpc}^{-1}$ with fractional errors at the level of 10^{-3} (Sefusatti et al. 2016). To perform these operations, as well as estimate the power spectrum multipoles via equation 5.33, we utilize the massively parallel implementations available as part of the open-source Python toolkit `nbodykit` (Hand et al. 2017a) (see Chapter 2).

5.4.3 Modeling

The power spectrum model

We use linear theory to predict the quasar power spectrum in redshift space (Kaiser 1987)

$$P_{\rm QSO}(k,\mu) = G(k,\mu;\sigma_v)^2\ \left[b_{\rm tot}(k) + f\mu^2\right]^2 P_{\rm m}(k) + N_{\rm shot}, \tag{5.35}$$

where $P_{\rm m}$ is the real-space matter power spectrum, $N_{\rm shot}$ is a free parameter accounting for residual shot noise, and $b_{\rm tot}$ is the total quasar bias, including PNG, given by

$$b_{\rm tot}(k) = b_{\rm QSO} + \Delta b = b_{\rm QSO} + f_{\rm NL}(b_{\rm QSO} - p)\widetilde{\alpha}(k), \tag{5.36}$$

where $b_{\rm QSO}$ is the linear bias of the quasar sample, and $\widetilde{\alpha}$ is defined in equation 5.5. To account for damping of the power spectrum in redshift space, we include a Lorentzian damping function,

$$G(k,\mu;\sigma_v) = \left[1 + (k\mu\sigma_v)/2\right]^{-1}, \tag{5.37}$$

with a single free parameter σ_v, which represents the typical damping velocity dispersion. The physical motivation for the inclusion of $G(k,\mu)$ is the Finger-of-God effect in redshift space due to the virial motions of a quasar within its host dark matter halo (Jackson 1972). However, the damping term also accounts for errors in the spectroscopic redshift determination of the quasars (Dawson et al. 2016). The effect can be estimated for the DR14Q sample as $\sigma_z = 300$ km s^{-1} for $z < 1.5$ and $\sigma_z = [400(z-1.5) + 300]$ km s^{-1} for $z > 1.5$.

The multipoles of the power spectrum are then computed as

$$P_{{\rm QSO},\ell}(k) = \frac{2\ell+1}{2}\int_{-1}^{1} d\mu P_{\rm QSO}(k,\mu)\mathcal{L}_\ell(\mu), \tag{5.38}$$

We evaluate the linear, real-space matter power spectrum $P_{\rm m}(k)$ and the transfer function $T(k)$ using the `classylss` software (Hand & Feng 2017), which provides Python bindings of the `CLASS` CMB Boltzmann solver (Blas et al. 2011). We evaluate the linear power

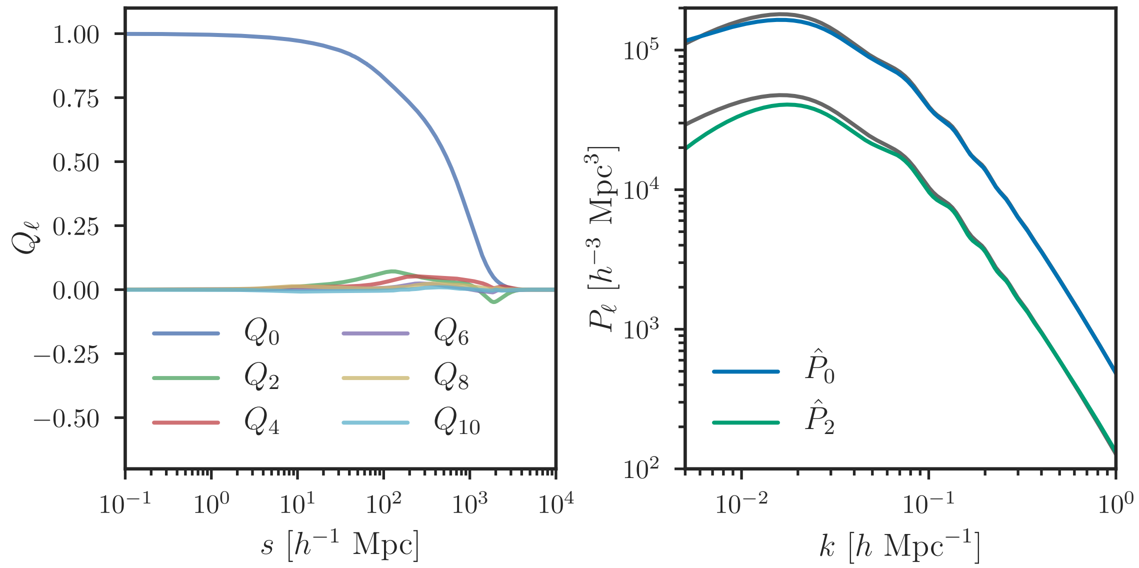


Figure 5.3: The window function multipoles in configuration space (left) and the effects of the window function on linear Kaiser power spectrum multipoles (right) for the eBOSS DR14Q NGC survey geometry. In the right panel, the solid grey lines show the original multipoles, while the colored lines correspond to the model after convolution with the window function, $\widehat{P}_\ell(k)$. The dominant consequence of the survey geometry is a reduction in power on large scales.

spectrum using the fiducial cosmology (§5.4.1) and keep the shape fixed during parameter estimation. This choice assumes that the uncertainty as determined by Planck Collaboration et al. (2016a) for most of the parameters which define the shape of the power spectrum is much smaller than the uncertainty of our measurement and can be neglected. This is a reasonable assumption given the expected constraining power of the DR14Q sample.

In summary, our power spectrum model includes four free parameters: the linear bias b_{QSO}, the damping velocity dispersion σ_v, the residual shot noise N_{shot}, and the PNG amplitude as parameterized by f_{NL}. We perform separate fits of the multipoles measured from the NGC and SGC sky regions, where we vary these four parameters. We also perform a joint fit to the combined data vector containing the monopole and quadrupole from both the NGC and SGC. In this case, we fit a single f_{NL} value while using separate b_{QSO}, σ_v, and N_{shot} values for each individual sky region.

The survey geometry

When comparing our theoretical model to the measurements from data, we must account for the effects of the DR14Q survey geometry on our measured multipoles. We do this by convolving our theoretical model with the survey window function, following the prescription first presented in Wilson et al. (2017) and used in a number of analyses since (Beutler et al. 2017b; Zhao et al. 2017; Hand et al. 2017c).

Following Wilson et al. (2017), we compute the pair counts of the synthetic catalog in configuration space using the `nbodykit` software and estimate the multipoles as

$$Q_\ell(s) \propto \int_{-1}^{1} d\mu RR(s,\mu)\mathcal{L}_\ell(\mu) \simeq \sum_i RR(s_i,\mu_i)\mathcal{L}_\ell(\mu_i), \tag{5.39}$$

where $RR(s,\mu)$ is the 2D pair counts, and the normalization of Q_ℓ is such that $Q_0(s) \to 1$ for $s \ll 1$. We show the multipoles Q_ℓ for the DR14Q NGC sample in the left panel of Figure 5.3. It is evident from this figure that anisotropic features of the survey geometry are sub-dominant.

Using the measured Q_ℓ, the convolved theoretical correlation function multipoles are calculated as

$$\begin{aligned}\widehat{\xi}_0(s) &= \xi_0 Q_0 + \frac{1}{5}\xi_2 Q_2 + \frac{1}{9}\xi_4 Q_4 + \ldots \\ \widehat{\xi}_2(s) &= \xi_0 Q_2 + \xi_2\left[Q_0 + \frac{2}{7}Q_2 + \frac{2}{7}Q_4\right] \\ &\quad + \xi_4\left[\frac{2}{7}Q_2\frac{100}{693}Q_4 + \frac{25}{143}Q_6\right] + \ldots\end{aligned} \tag{5.40}$$

where ξ_ℓ are the theoretical correlation function multipoles, computed from the power spectrum multipoles (equation 5.38) via a 1D Hankel transform, evaluated using the `FFTLog` software (Hamilton 2000). We show the effects of the window function convolution on linear Kaiser multipoles in the right panel of Figure 5.3. The dominant consequence of the window convolution is a reduction in power on large scales (small k) that becomes more prominent for larger values of ℓ.

5.4.4 Parameter estimation

We estimate the best-fit parameters of the model described in Section 5.4.3 using a likelihood analysis. We assume that the probability that our data vector $\boldsymbol{D}$ corresponds to a realization of our model $\boldsymbol{T}(\boldsymbol{\theta})$ is given by a multi-variate Gaussian of the form,

$$\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\Phi}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{D},\boldsymbol{\theta},\boldsymbol{\Phi})\right], \tag{5.41}$$

where $\boldsymbol{\theta}$ is our vector of model parameters, and χ^2 takes the quadratic form,

$$\chi^2(\boldsymbol{\theta}) = \sum_{ij}(D_i - T_i(\boldsymbol{\theta}))\Phi_{ij}(D_j - T_j(\boldsymbol{\theta})), \tag{5.42}$$

and $\boldsymbol{\Phi}$ is the inverse of the covariance matrix $\boldsymbol{C}$, often referred to as the precision matrix.

When performing our likelihood analysis, our data vector $\boldsymbol{D}$ consists of the monopole and quadrupole, measured using the procedure outlined in Section 5.4.2. We use linearly spaced

			NGC		SGC		Laurent et al. (2017)
$z_{\rm min}$	$z_{\rm max}$	$z_{\rm eff}$	$N_{\rm QSO}$	$b_{\rm QSO}$	$N_{\rm QSO}$	$b_{\rm QSO}$	$b_{\rm QSO}$
0.9	1.2	1.06	17191	1.63 ± 0.04	11013	1.71 ± 0.06	1.75 ± 0.08
1.2	1.5	1.35	21947	2.18 ± 0.04	14609	2.02 ± 0.05	2.06 ± 0.08
1.5	1.8	1.65	22127	2.36 ± 0.04	14793	2.52 ± 0.06	2.57 ± 0.09
1.8	2.2	1.99	23893	3.01 ± 0.05	16325	2.95 ± 0.07	3.03 ± 0.11
0.9	2.2	1.55	85158	2.33 ± 0.02	56740	2.33 ± 0.03	2.43 ± 0.05

Table 5.1: The best-fit linear bias parameters and 1σ errors obtained from fitting the power spectrum model of Section 5.4.3 to the monopole and quadrupole of the NGC and SGC DR14Q samples. We compare our results to those of Laurent et al. (2017), which reported bias values for the first year of eBOSS data (roughly half the size of the DR14Q sample).

bins of width $\Delta k = 0.005\ h\text{Mpc}^{-1}$. With the first bin separation at $k \sim 0.005\ h\text{Mpc}^{-1}$ and extending to $k_{\rm max} = 0.3\ h\text{Mpc}^{-1}$, we have a total of 120 data points in $\boldsymbol{D}$ (60 bins per multipole).

We estimate the covariance matrix of our data measurement using the 1,000 EZ mock realizations, described previously in Section 5.2.3. As the covariance is computed from a finite number of mock realizations, its inverse $\boldsymbol{\Phi}$ provides a biased estimate of the true precision matrix due to the skewed nature of the inverse Wishart distribution (Hartlap et al. 2007). To correct for this bias, we re-scale the precision matrix as

$$\boldsymbol{\Phi}' = \frac{N_{\rm mock} - n_b - 2}{N_{\rm mock} - 1} \boldsymbol{\Phi}. \tag{5.43}$$

When performing our likelihood analysis following equation 5.41, we use the rescaled precision matrix $\boldsymbol{\Phi}'$. In our analysis, we use $N_{\rm mocks} = 1000$ and $n_b = 120$, yielding a Hartlap factor of $\sim$0.88.

We find the best-fit model parameters using the LBFGS nonlinear minimization algorithm (Byrd et al. 1995). We verify that the minimization procedure converges by starting the algorithm from a number of different initialization states. We compute the full posterior distributions of the parameters of interest using the `emcee` software (Foreman-Mackey et al. 2013) to perform Markov Chain Monte Carlo (MCMC) sampling. We assume broad, uniform priors on all parameters of interest such that the priors serve only to bound the parameter values to the largest possible physically meaningful parameter space; they do not have an impact on our derived posterior distributions.

5.5 Results

5.5.1 Quasar bias as a function of redshift

We begin by measuring the linear bias of the DR14Q sample as a function of redshift. The bias of the quasars is expected to be a strong function of redshift, and it is important to understand if this evolution is consistent with our expectations, e.g., Croom et al. (2005); Ross et al. (2009); Laurent et al. (2017). We choose identical redshift bins as those used in Laurent et al. (2017), which reported quasar bias values for the first year of eBOSS data, corresponding to roughly half of the sample size of the DR14Q data set. We report our estimated linear bias parameters and their 1σ errors derived from MCMC sampling in Table 5.1. We measure the bias values by fitting our linear power spectrum model described in Section 5.4.3 to the monopole and quadrupole of the DR14Q sample. We fit over the range $10^{-3} < k < 0.3$ hMpc^{-1} and have verified that restricting to a smaller $k_{\rm max}$ value, i.e., only linear scales, does not change our results. Our values are roughly consistent with those of Laurent et al. (2017). Discrepancies can likely be attributed to the significantly larger sample size used here, as well as analysis differences (Laurent et al. (2017) use a configuration-space based analysis). In our analysis, we assume the redshift evolution of Laurent et al. (2017), given by

$$b_{\rm QSO}(z) = \alpha[(1+z)^2 - 6.565] + \beta, \tag{5.44}$$

with $\alpha = 0.278$ and $\beta = 2.393$. We use this equation to describe the redshift evolution of the bias when assigning weights to individual quasars using equation 5.24.

5.5.2 Unweighted results

In this section, we present constraints on $f_{\rm NL}$ using traditional analysis techniques. We treat the entirety of the DR14Q sample in a single redshift bin ranging from $0.8 \leq z \leq 2.2$ with an effective redshift of $z_{\rm eff} = 1.53$. We do not employ the redshift weighting scheme presented in Section 5.3.3. These results serve as a baseline measurement—an optimal weighting scheme should improve upon the $f_{\rm NL}$ constraints presented here. We first apply and test our parameter estimation methods using the 1,000 EZ mock catalogs described in Section 5.2.3 and then present the constraints measured from the DR14Q data sample.

EZ mock results

The set of EZ mock catalogs provides an opportunity to test our parameter estimation pipeline and modeling systematics. We have 1,000 independent realizations for both the NGC and SGC regions (2,000 in total). We perform fits to the measured monopole and quadrupole from each mock using a wavenumber range $10^{-3} < k < 0.3$ hMpc^{-1}. We summarize the resulting constraints in the top portion of Table 5.2. This table gives the mean values of the parameters across the 1,000 realizations as well as the asymmetric 1σ confidence

	Parameter Set	$f_{\rm NL}$	$b_{\rm QSO}$	σ_v	$N_{\rm shot}$
EZ Mock	NGC	63^{+42}_{-46}	$2.223^{+0.037}_{-0.037}$	$4.166^{+0.432}_{-0.385}$	-353^{+119}_{-110}
	SGC	37^{+49}_{-57}	$2.253^{+0.049}_{-0.050}$	$3.610^{+0.531}_{-0.487}$	-433^{+154}_{-156}
DR14Q Data	NGC	14^{+52}_{-73}	$2.269^{+0.041}_{-0.037}$	$4.850^{+0.362}_{-0.357}$	-54^{+98}_{-102}
	SGC	-27^{+55}_{-48}	$2.324^{+0.050}_{-0.049}$	$3.511^{+0.399}_{-0.416}$	-282^{+132}_{-141}
	NGC + SGC	-9^{+43}_{-46}			

Table 5.2: The best-fit model parameters as measured from the 1,000 EZ mock catalogs (top) and the DR14Q data set (bottom) for the NGC and SGC sky regions. We also give the constraint on $f_{\rm NL}$ when simultaneously fitting the DR14Q NGC and SGC samples. For these results, we treat the data sample as a single redshift bin ranging from $0.8 \leq z \leq 2.2$ with $z_{\rm eff} = 1.53$ and do not apply any redshift weights. Parameters are obtained by fitting the 4-parameter model of equation 5.35 to the monopole and quadrupole over the wavenumber range $10^{-3} < k < 0.3\ h{\rm Mpc}^{-1}$. The results from the EZ mocks give the mean values and 1σ uncertainties obtained from the distribution of the 1,000 fits, and DR14Q results are estimated using MCMC sampling.

level intervals, as determined from the distribution of best-fit values. We also show the one-dimensional histograms and two-dimensional correlations for the best-fit values in Figure 5.4. As expected, we find that $f_{\rm NL}$ is most strongly correlated with the linear bias $b_{\rm QSO}$. While centered near $f_{\rm NL} = 0$, the distribution of recovered $f_{\rm NL}$ values also shows significant non-Gaussian tails stretching to large positive and negative values.

We find the mean $f_{\rm NL}$ value to be $f_{\rm NL} = 63^{+42}_{-46}$ for the NGC and $f_{\rm NL} = 37^{+49}_{-57}$ for the SGC. The EZ mocks contain no PNG signal, and given the fact that these constraints are averaged over 1,000 mocks, we measure a statistically significant positive bias for both the NGC and SGC data sets. The most likely culprit for this bias is large-scale systematics contaminating the scale-dependent bias signature introduced by PNG. We have attempted to mitigate these issues by using the completeness weights discussed in Section 5.2.2. However, our findings here indicate that these weights may be insufficient for measuring $f_{\rm NL}$ robustly from the large-scale power spectrum. Further investigation of this potential large-scale contamination is required.

Results for the DR14Q sample

We perform an identical analysis on the DR14Q data set as discussed in the previous section when using the set of EZ mock catalogs. In particular, the results discussed here use a single redshift bin ranging from $0.8 \leq z \leq 2.2$ with $z_{\rm eff} = 1.53$, and we do not apply any redshift weights The parameter constraints are summarized in the bottom portion of Table 5.2. These constraints are obtained via MCMC sampling, and we report results when separately fitting the NGC and SGC samples. We also perform a joint fit to the combined data vector from both sky regions. In this case, we fit a single $f_{\rm NL}$ value while varying separate values for each of the other model parameters for the NGC and SGC multipoles. The resulting PNG constraint when jointly fitting the multipoles from the NGC and SGC sky regions is $f_{\rm NL} = -9^{+43}_{-46}$. We compare the best-fit theoretical model to the measured multipoles in Figure 5.5. Our 4-parameter model is able to accurately describe the multipoles over the range of scales considered, $10^{-3} < k < 0.3\ h{\rm Mpc}^{-1}$. The reduced chi-squared of the fits to the NGC and SGC multipoles is $\chi^2_{\rm red} = 0.92$ and $\chi^2_{\rm red} = 1.11$, respectively.

We show the resulting posterior distributions for each of the four model parameters in Figure 5.6.[1] This figure gives both the marginalized one-dimensional distributions as well as the joint, two-dimensional correlations between parameters. The results agree well with our expectations based on the results obtained from the EZ mock catalogs. We obtain slightly different $b_{\rm QSO}$ and σ_v values for the NGC and SGC data sets. We expect such discrepancies due to the slight differences in the targeting efficiency for these two samples (see Figure 5.1). Furthermore, given the spread of the parameters measured from the 1,000 EZ mocks, the DR14Q data appears to be a typical realization. Both the mean and 1σ uncertainties for the four model parameters measured from the DR14Q data are consistent with the results obtained from the EZ mock catalogs.

5.6 Conclusions and future work

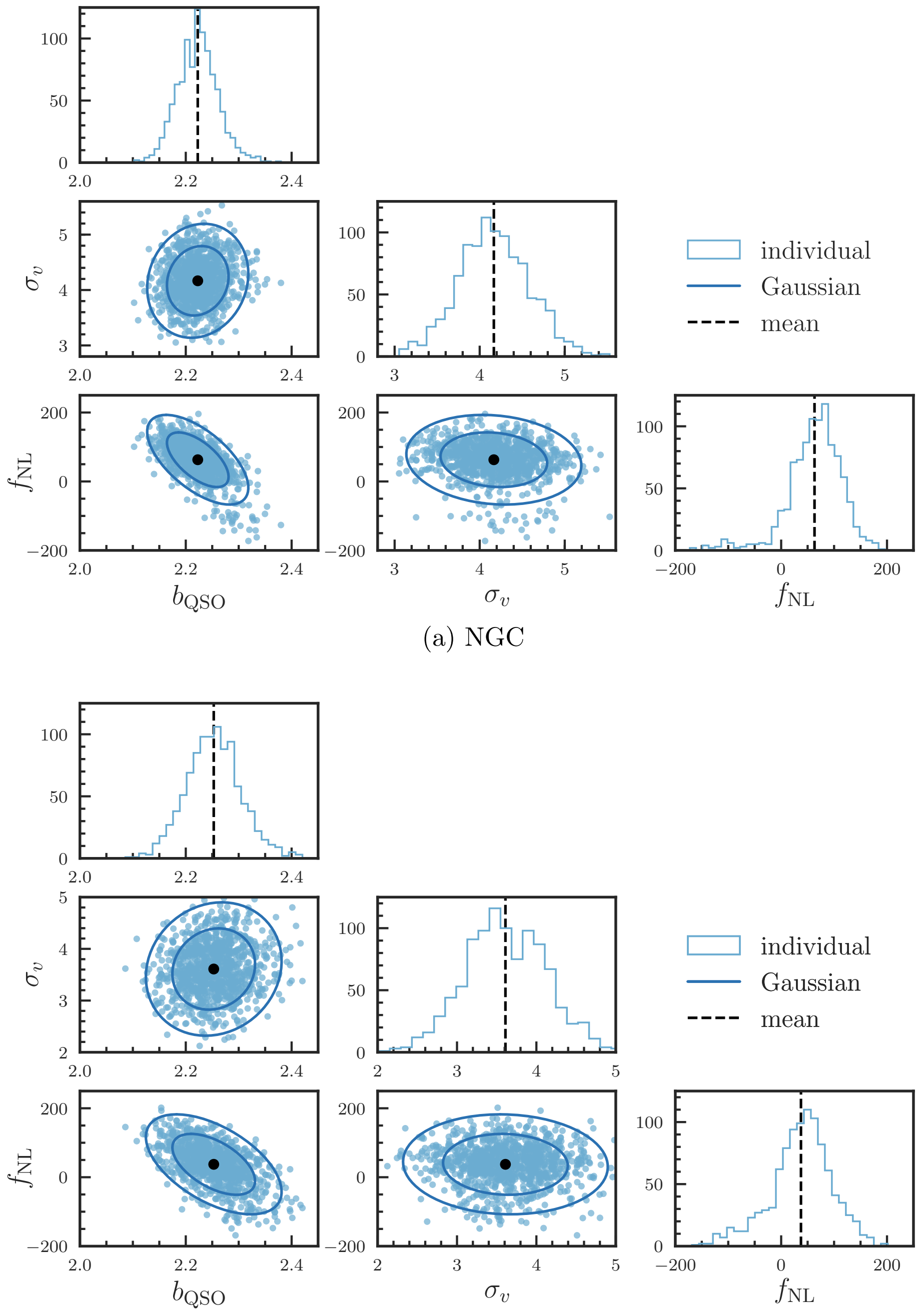
In this work, we have constrained local type primordial non-Gaussianity, as parameterized by $f_{\rm NL}$, using the eBOSS DR14 quasar sample. We have analyzed the quasar sample using traditional analysis techniques, considering the entirety of the data set in a single redshift bin ranging from $0.8 \leq z \leq 2.2$. Modeling the redshift-space power spectrum multipoles with a linear Kaiser model with Finger-of-God damping, we obtain a PNG constraint of $f_{\rm NL} = -9^{+43}_{-46}$ when jointly fitting the monopole and quadrupole measured from the NGC and SGC sky regions. We have tested and verified our analysis pipeline with a large set of realistic mock catalogs that accurately model the eBOSS angular selection function and redshift evolution. Our results obtained from the 1,000 mock catalogs indicate a potential positive bias in the recovered value of $f_{\rm NL}$. Such a bias is likely caused by large-scale contamination from systematics that is not fully corrected by our weighting scheme. Mitigation of systematics on the largest scales probed by a redshift survey is particularly challenging, and remains one of the most significant impediments to using large-scale structure measurements to probe PNG. Further investigation of any potential residual systematics on large scales in the eBOSS multipoles is required.

The constraint using a single redshift bin from this work is not competitive with the tightest published constraints on $f_{\rm NL}$ from large-scale structure data sets, e.g., $\sigma(f_{\rm NL}) \sim 20$. However, improvements can be made using more optimal analysis techniques. In Section 5.3.3, we derived an near-optimal power spectrum estimator for $f_{\rm NL}$ using weights to account for the evolution of the PNG signal as a function of redshift. An analysis including such weights

[1] Figure 5.6 was produced using the `ChainConsumer` software (Hinton 2016).

is currently underway and should yield improved constraints on $f_{\rm NL}$. Furthermore, eBOSS has observed quasars to higher redshift range than the maximum redshift of $z = 2.2$ used in this work. The PNG signal is larger at higher redshift where the quasars are more strongly biased. Consequently, the inclusion of all eBOSS quasars up to a redshift of $z \sim 3$ will further improve the $f_{\rm NL}$ constraint. The systematics associated with the quasars at higher redshift are currently not well-studied, and further analysis will be required to robustly utilize these objects to constrain $f_{\rm NL}$. Initial tests indicate that the PNG constraints could improve as much as ∼50% by extending the redshift range, so we believe additional systematics studies are well-justified.

The entirety of the eBOSS quasar sample will probe roughly three times the volume of the DR14 sample considered in this work. This additional volume will significantly enhance measurements of the large-scale quasar power spectrum and offer the possibility of competitive constraints on primordial non-Gaussianity. To fully exploit the constraining power of the complete data set, analysis techniques such as those developed in this work will need to be employed to optimally extract information across the enormous redshift range of the sample. Mitigation of large-scale systematics will be crucial for achieving these goals and constraining primordial non-Gaussianity in a robust manner from future large-scale structure data sets.



(a) NGC

(b) SGC

Figure 5.4: The distribution of the best-fit model parameters as measured from the 1,000 EZ mock catalogs for the NGC (top) and SGC (bottom) sky regions. We have fit the 4-parameter model of equation 5.35 to the monopole and quadrupole over the wavenumber range $10^{-3} < k < 0.3\ h\mathrm{Mpc}^{-1}$. For these results, we treat the data sample as a single redshift bin ranging from $0.8 \leq z \leq 2.2$ with $z_{\mathrm{eff}} = 1.53$ and do not apply any redshift weights. The diagonal plots give the 1D histogram of the best-fit values for each parameter and the off-diagonal plots show the 2D correlations between parameters.

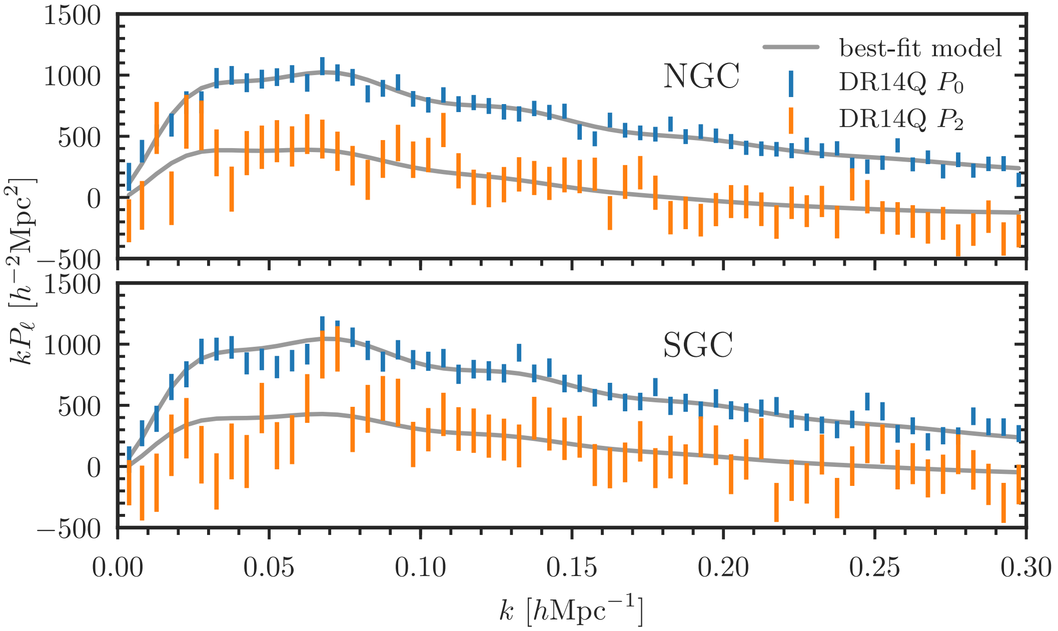


Figure 5.5: The measured power spectrum multipoles for the DR14Q sample (points with errors) and the best-fit theoretical model (grey lines). The model of equation 5.35 is able to accurately describe the monopole and quadrupole over the wavenumber range $10^{-3} < k < 0.3\ h\mathrm{Mpc}^{-1}$. The reduced chi-squared of the fits for the NGC and SGC data sets is $\chi^2_{\mathrm{red}} = 0.92$ and $\chi^2_{\mathrm{red}} = 1.11$, respectively.

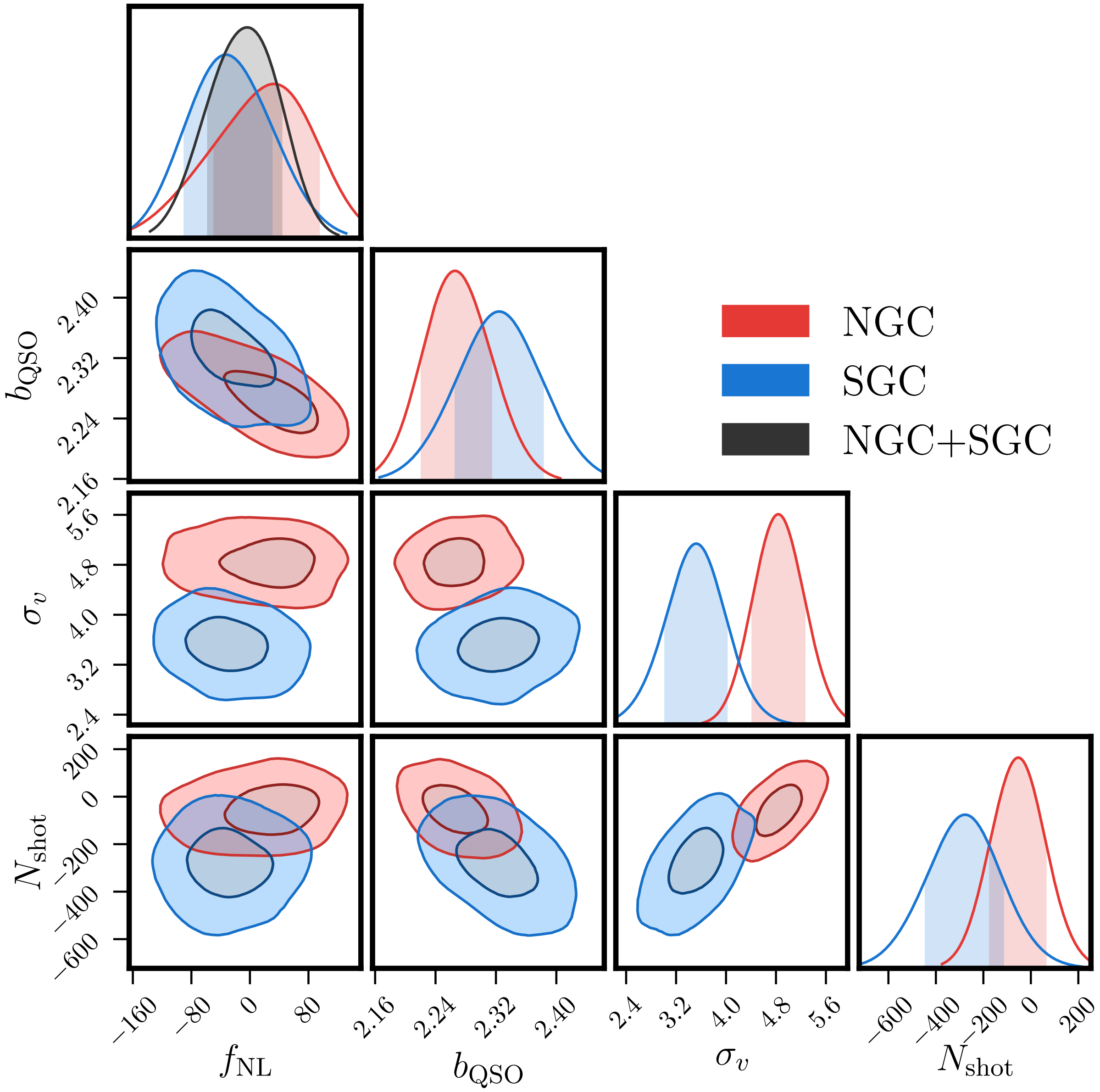


Figure 5.6: The posterior distributions for each of the four model parameters when fitting the DR14Q multipoles using a single redshift bin ranging from $0.8 \leq z \leq 2.2$. The marginalized posterior distributions are shown on the diagonal of the figure. Off-diagonal plots show joint two-dimensional posteriors. We show the constraints obtained when fitting the NGC (red) and SGC (blue) sky regions separately. We also give the 1D posterior for $f_{\rm NL}$ when jointly fitting the NGC and SGC multipoles (black). In this case, we obtain $f_{\rm NL} = -9^{+43}_{-46}$.

Appendix A

Velocity correlators in the Zel'dovich approximation

In this section, we use Lagrangian perturbation theory (LPT) to compute two dark matter velocity correlators that enter into the DF model: the density – radial momentum cross spectrum, P_{01}, and the radial momentum auto spectrum, P_{11}. We closely follow the notation of Vlah et al. (2015); see e.g., Vlah et al. (2015); Carlson et al. (2013); Matsubara (2008) and references therein for further review of Lagrangian perturbation theory.

A.1 P_{01} and P_{11} using LPT

Following the definitions of Vlah et al. (2013), the velocity correlators that we wish to compute are given by

$$\begin{aligned}(2\pi)^3 \widetilde{P}_{01}(\boldsymbol{k})\delta^D(\boldsymbol{k}+\boldsymbol{k}') &= \langle \delta(\boldsymbol{k}) | p_\parallel(\boldsymbol{k}')\rangle, \\ (2\pi)^3 \widetilde{P}_{11}(\boldsymbol{k})\delta^D(\boldsymbol{k}+\boldsymbol{k}') &= \langle p_\parallel(\boldsymbol{k}) | p_\parallel(\boldsymbol{k}')\rangle, \end{aligned} \tag{A.1}$$

where $p_\parallel$ is the momentum projected along the line-of-sight, i.e., $p_\parallel = \boldsymbol{p}\cdot\hat{z}$. The scalar component of the dark matter momentum (which correlates with density) can be computed using the continuity equation: $\dot{\delta}(\boldsymbol{k}) - i\boldsymbol{k}\cdot\boldsymbol{p} = 0$, where the dot in $\dot{\delta}$ represents the derivative with respect to conformal time τ. Using this equation, we can express the velocity correlators of interest as

$$\widetilde{P}_{01}(\boldsymbol{k}) = \frac{i\mu}{k} P_{\delta\dot{\delta}}(k), \tag{A.2}$$

$$\widetilde{P}_{11,s}(\boldsymbol{k}) = \frac{\mu^2}{k^2} P_{\dot{\delta}\dot{\delta}}(k), \tag{A.3}$$

where μ is defined as $k_\parallel/k$. Here, we explicitly note that $\widetilde{P}_{11,s}$ only includes scalar contributions, as only the scalar component of momentum enters into the continuity equation. The total contribution from these terms to the redshift-space power spectrum $P^S(k, \mu)$ is given by:

$$\begin{aligned} P_{01}(\boldsymbol{k}) &= 2\mathrm{Re}\left(\frac{-ik\mu}{\mathcal{H}}\right)\widetilde{P}_{01}(\boldsymbol{k}) = 2\frac{\mu^2}{\mathcal{H}}P_{\delta\dot{\delta}}(k), && \text{(A.4)} \\ P_{11,s}(\boldsymbol{k}) &= \left(\frac{k\mu}{\mathcal{H}}\right)^2\widetilde{P}_{11,s}(\boldsymbol{k}) = \frac{\mu^4}{\mathcal{H}^2}P_{\dot{\delta}\dot{\delta}}(k). && \text{(A.5)} \end{aligned}$$

These are the spectra that we wish to compute in the Zel'dovich approximation. In linear theory, these spectra are the anisotropic terms of the well-known Kaiser formula: $P_{01}(\boldsymbol{k}) = 2f\mu^2 P_L(k)$ and $P_{11,s}(\boldsymbol{k}) = f^2\mu^4 P_L(k)$ (Kaiser 1987).

We can compute δ and $\dot{\delta}$ using Lagrangian perturbation theory. In the Lagrangian clustering description, the overdensity field is given by

$$(2\pi)^3\delta^D(\boldsymbol{k}) + \delta(\boldsymbol{k}) = \int d^3q\ e^{i\boldsymbol{k}\cdot\boldsymbol{q}}\exp[i\boldsymbol{k}\cdot\Psi(\boldsymbol{q})], \tag{A.6}$$

where $\Psi(\boldsymbol{q})$ is the Lagrangian displacement field. The derivative of this equation with respect to conformal time is given by

$$\dot{\delta}(\boldsymbol{k}) = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}}\left(i\boldsymbol{k}\cdot\dot{\Psi}\right)\exp[i\boldsymbol{k}\cdot\Psi(\boldsymbol{q})]. \tag{A.7}$$

The quantity of interest for P_{01} is

$$\begin{aligned} (2\pi)^3 P_{\delta\dot{\delta}}(k)\delta^D(\boldsymbol{k}+\boldsymbol{k}') &= \langle\delta(\boldsymbol{k})|\dot{\delta}(\boldsymbol{k}')\rangle, \\ &= \int d^3q d^3q' e^{i\boldsymbol{k}\cdot\boldsymbol{q}+i\boldsymbol{k}'\cdot\boldsymbol{q}'}\left\langle\left(i\boldsymbol{k}'\cdot\dot{\Psi}'\right)e^{i\boldsymbol{k}\cdot\Psi+i\boldsymbol{k}'\cdot\Psi'}\right\rangle, && \text{(A.8)} \end{aligned}$$

where we have used the definition $\Psi' \equiv \Psi(\boldsymbol{q}')$. Similarly, for $P_{11,s}$ we need to compute

$$\begin{aligned} (2\pi)^3 P_{\dot{\delta}\dot{\delta}}(k)\delta^D(\boldsymbol{k}+\boldsymbol{k}') &= \langle\dot{\delta}(\boldsymbol{k})|\dot{\delta}(\boldsymbol{k}')\rangle, \\ &= \int d^3q d^3q' e^{i\boldsymbol{k}\cdot\boldsymbol{q}+i\boldsymbol{k}'\cdot\boldsymbol{q}'}\left\langle\left(i\boldsymbol{k}\cdot\dot{\Psi}\right)\left(i\boldsymbol{k}'\cdot\dot{\Psi}'\right)e^{i\boldsymbol{k}\cdot\Psi+i\boldsymbol{k}'\cdot\Psi'}\right\rangle. && \text{(A.9)} \end{aligned}$$

A.2 A generalized velocity generating function

To facilitate the calculation of equations A.6 and A.7, we introduce a generalized velocity generating function in this section. First, let us define the sum and difference of the displacement field $\Psi(\boldsymbol{q})$ defined at points $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ in space:

$$\Delta_i^- = \Psi_i(\boldsymbol{q}_2) - \Psi_i(\boldsymbol{q}_1), \quad \Delta_i^+ = \Psi_i(\boldsymbol{q}_2) + \Psi_i(\boldsymbol{q}_1). \tag{A.10}$$

Now we can define the generalized velocity generating function $\mathcal{G}$ as

$$(2\pi)^3\delta^D(\boldsymbol{k}) + \mathcal{G}(k,\gamma,\lambda) = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \left\langle e^{-i\boldsymbol{k}\cdot\Delta^- - i\gamma\boldsymbol{k}\cdot\dot{\Delta}^- - i\lambda\boldsymbol{k}\cdot\dot{\Delta}^+} \right\rangle. \tag{A.11}$$

Note that the case of $\gamma = \lambda = 0$ gives the well-known matter power spectrum $P_{\delta\delta}$ in the LPT formalism (Schneider & Bartelmann 1995). We define the following moments of $\mathcal{G}$:

$$\begin{aligned} G_{10}(k) &= \frac{d}{d\gamma}\mathcal{G}(k,\gamma,\lambda)\Big|_{\gamma=0,\lambda=0} = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \left\langle \left(i\boldsymbol{k}\cdot\dot{\Delta}^-\right) e^{-i\boldsymbol{k}\cdot\dot{\Delta}^-} \right\rangle, & \text{(A.12)} \\ G_{20}(k) &= \frac{d^2}{d\gamma^2}\mathcal{G}(k,\gamma,\lambda)\Big|_{\gamma=0,\lambda=0} = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \left\langle \left(i\boldsymbol{k}\cdot\dot{\Delta}^-\right)^2 e^{-i\boldsymbol{k}\cdot\dot{\Delta}^-}, \right\rangle, & \text{(A.13)} \\ G_{02}(k) &= \frac{d^2}{d\lambda^2}\mathcal{G}(k,\gamma,\lambda)\Big|_{\gamma=0,\lambda=0} = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \left\langle \left(i\boldsymbol{k}\cdot\dot{\Delta}^+\right)^2 e^{-i\boldsymbol{k}\cdot\dot{\Delta}^-}, \right\rangle. & \text{(A.14)} \end{aligned}$$

Substituting the definitions of Δ^+ and Δ^- into these equations yields the following relations,

$$G_{10}(k) = 2P_{\delta\dot{\delta}}(k), \tag{A.15}$$

$$G_{20}(k) - G_{02}(k) = 4P_{\dot{\delta}\dot{\delta}}(k). \tag{A.16}$$

We can evaluate $\mathcal{G}$ using the cumulant expansion theorem,

$$\left\langle e^{-iX}\right\rangle = \exp\left[\sum_{N=0}^{\infty}\frac{(-i)^N}{N!}\left\langle X^N\right\rangle_c\right], \tag{A.17}$$

where $X = \boldsymbol{k}\cdot\Delta^- + \gamma\boldsymbol{k}\cdot\dot{\Delta}^- + \lambda\boldsymbol{k}\cdot\dot{\Delta}^+$. In the Zel'dovich approximation (tree-level LPT), the displacement field remains Gaussian, so only the $N = 2$ term is non-zero in the above expansion. Thus, the quantity of interest is

$$\begin{aligned} \left\langle\left(\boldsymbol{k}\cdot\Delta^- + \gamma\boldsymbol{k}\cdot\dot{\Delta}^- + \lambda\boldsymbol{k}\cdot\dot{\Delta}^+\right)^2\right\rangle &= k_ik_j\left[A_{ij} + \gamma\dot{A}_{ij} + \gamma^2B^-_{ij} + \lambda^2B^+_{ij} + ...\right] \\ &\equiv k_ik_j\left[\mathcal{A}_{ij} + ...\right], & \text{(A.18)} \end{aligned}$$

where we have explicitly ignored terms that vanish upon taking the derivatives in the expressions for G_{10}, G_{20}, and G_{02}. The relevant definitions are

$$\begin{aligned} A_{ij}(k) &= \left\langle\Delta^-_i\Delta^-_j\right\rangle_c, & \text{(A.19)} \\ B^-_{ij}(k) &= \left\langle\dot{\Delta}^-_i\dot{\Delta}^-_j\right\rangle_c, & \text{(A.20)} \\ B^+_{ij}(k) &= \left\langle\dot{\Delta}^+_i\dot{\Delta}^+_j\right\rangle_c. & \text{(A.21)} \end{aligned}$$

Note that this definition of A_{ij} matches the notation used in the recent LPT work of Vlah et al. (2015); Carlson et al. (2013). Finally, using equations A.18, A.17, and A.11, the velocity generating function becomes

$$(2\pi)^3\delta^D(\boldsymbol{k}) + \mathcal{G}(k,\gamma,\lambda) = \int d^3q e^{i\boldsymbol{k}\cdot\boldsymbol{q}} \exp\left[-\frac{1}{2}k_ik_j\mathcal{A}_{ij}\right]. \tag{A.22}$$

A.3 The Zel'dovich approximation

In the Zel'dovich approximation, the displacement field and its time derivative are given

$$\Psi(\boldsymbol{k}) = i\boldsymbol{k}\delta_L(\boldsymbol{k})/k^2, \tag{A.23}$$

$$\dot{\Psi}(\boldsymbol{k}) = f\mathcal{H}\Psi(\boldsymbol{k}), \tag{A.24}$$

where δ_L is the linear overdensity field which scales in time as the linear growth function D, $\mathcal{H} = d\ln a/d\tau$ is the conformal Hubble parameter, and $f = d\ln D/d\ln a$ is the logarithmic growth rate.

With these relations, we can now compute the expressions for each term of equation A.18 (for a more in-depth discussion of this procedure, see Appendix B of Carlson et al. (2013)). The relevant expressions are:

$$A_{ij}(\boldsymbol{q}) = I^-_{ij}(\boldsymbol{q}), \tag{A.25}$$

$$\dot{A}_{ij}(\boldsymbol{q}) = 2f\mathcal{H}I^-_{ij}(\boldsymbol{q}), \tag{A.26}$$

$$B^-_{ij}(\boldsymbol{q}) = (f\mathcal{H})^2 I^-_{ij}(\boldsymbol{q}), \tag{A.27}$$

$$B^+_{ij}(\boldsymbol{q}) = (f\mathcal{H})^2 I^+_{ij}(\boldsymbol{q}), \tag{A.28}$$

where we have defined the integral

$$I^\pm_{ij}(\boldsymbol{q}) = 2\int \frac{d^3k}{(2\pi)^3}\left[1 \pm \cos(\boldsymbol{k}\cdot\boldsymbol{q})\right]\frac{k_ik_j}{k^4}P_L(k), \tag{A.29}$$

where $P_L(q)$ is the linear power spectrum. Here, I^-_{ij} is the same quantity that enters into the LPT calculation of the density auto power spectrum; for example, our expression is the same as equation A6 of Vlah et al. (2015) (restricting to tree-level).

Equation A.29 can be expressed in terms of two scalar functions as

$$X^\pm_{ij}(\boldsymbol{q}) = X^\pm(q)\delta^K_{ij} + Y^\pm(q)\hat{q}_i\hat{q}_j, \tag{A.30}$$

We can compute $X^\pm_{ij}$ and $Y^\pm_{ij}$ by performing the angular integration in equation A.29. To facilitate comparisons with previous work (e.g., Vlah et al. 2015; Carlson et al. 2013), we define

$$
\begin{aligned}
X_{ij}^{\pm}(q) &= 2\sigma^2 \pm \frac{1}{\pi^2}\int dk P_L(k)\frac{j_1(kq)}{kq} \equiv 2\sigma^2 \pm X_0(q), \\
X_{ij}^{-}(q) &= X(q), \\
Y_{ij}^{\pm}(q) &= \mp Y(q),
\end{aligned}
\tag{A.31}
$$

where the $\sigma^2 = 1/(6\pi^2)\int dqP(q)$ is the square of the linear displacement field dispersion, and the well-known Zel'dovich integrals $X(q)$ and $Y(q)$ are

$$
\begin{aligned}
X(q) &= \int \frac{dk}{2\pi^2}P_L(k)\left[\frac{2}{3} - 2\frac{j_1(kq)}{kq}\right], \\
Y(q) &= \int \frac{dk}{2\pi^2}P_L(k)\left[-2j_0(kq) + 6\frac{j_1(kq)}{kq}\right],
\end{aligned}
\tag{A.32}
$$

where j_n is the spherical Bessel function of order n.

With these integral expressions, we can now compute the relevant moments of $\mathcal{G}$ in order to evaluate P_{01} and $P_{11,s}$. First, for P_{01}, we have

$$
\begin{aligned}
P_{01}(k,\mu) = \frac{\mu^2}{\mathcal{H}}G_{10}(k), \\
= 2f\mu^2\int d^3qe^{ikq\bar{\mu}}\left[-\frac{1}{2}k^2\left(X+\bar{\mu}^2Y\right)\right]e^{-\frac{1}{2}k^2(X+\bar{\mu}^2Y)},
\end{aligned}
\tag{A.33}
$$

where we have introduced the angle between the given k-mode and separation vector $\bar{\mu} = \hat{q}\cdot\hat{k}$. Similarly, for $P_{11,s}$, we have

$$
\begin{aligned}
P_{11,s}(k,\mu) = \frac{\mu^4}{4\mathcal{H}^2}\left[G_{20}(k) - G_{02}(k)\right], \\
= \frac{1}{4}f^2\mu^4\int d^3qe^{ikq\bar{\mu}}k^2\left[-2X_0 + k^2X^2 + 2(k^2X-1)Y\bar{\mu}^2 + k^2Y^2\bar{\mu}^4\right]e^{-\frac{1}{2}k^2(X+\bar{\mu}^2Y)}.
\end{aligned}
\tag{A.34}
$$

Equations A.33 and A.34 represent the desired solution for P_{01} and $P_{11,s}$ in the Zel'dovich approximation. The angular integration over $\bar{\mu}$ in these expressions can be performed using the following expression (Schneider & Bartelmann 1995)

$$
\int_{-1}^{1} d\mu e^{iA\mu}e^{B\mu^2} = 2e^B\sum_{n=0}^{\infty}\left(-\frac{2B}{A}\right)^n j_n(A),
\tag{A.35}
$$

and the subsequent derivatives of this expression with respect to B yields

$$\int_{-1}^{1} d\mu \mu^2 e^{iA\mu} e^{B\mu^2} = 2e^B \sum_{n=0}^{\infty} \left(-\frac{2B}{A}\right)^n j_n(A) \left[1 + \frac{n}{B}\right], \tag{A.36}$$

$$\int_{-1}^{1} d\mu \mu^4 e^{iA\mu} e^{B\mu^2} = 2e^B \sum_{n=0}^{\infty} \left(-\frac{2B}{A}\right)^n j_n(A) \left[1 + \frac{n}{B^2}(n + 2B - 1)\right]. \tag{A.37}$$

With equations A.35, A.36, and A.37, we can compute the desired quantities in equations A.33 and A.34 as a quickly converging sum of one-dimensional integrals, where the one-dimensional integrals can be computed rapidly with the aid of software such as `FFTLog` (Hamilton 2000). Typically, the sum over n can be truncated at $n < 15$ for $k < 1$ hMpc^{-1}.

Appendix B

Improved HZPT modeling

In this section, we give the best-fit parameters for the updated HZPT modeling used in this work (as described in Section 4.4.2).

B.1 Dark matter correlators P_{00} and P_{01}

For the dark matter power spectrum P_{00}, we follow the parameterization of Seljak & Vlah (2015) and provide updated best-fit parameters. We use a Padé expansion with $n_{\max} = 2$, such that the broadband term is given by

$$P_{00}^{BB}(k) = A_0 \left(1 - \frac{1}{1 + k^2 R^2}\right) \frac{1 + (kR_1)^2}{1 + (kR_{1h})^2 + (kR_{2h})^4}, \tag{B.1}$$

where the free parameters of the model are given by: $\{A_0, R, R_1, R_{1h}, R_{2h}\}$. For these parameters, we find the best-fit parameters to be:

$$A_0 = 708 \left(\frac{\sigma_8(z)}{0.8}\right)^{3.65} \; [\, h^{-3}\mathrm{Mpc}^3], \tag{B.2}$$

$$R = 31.8 \left(\frac{\sigma_8(z)}{0.8}\right)^{0.13} \; [\, h^{-1}\mathrm{Mpc}], \tag{B.3}$$

$$R_1 = 3.24 \left(\frac{\sigma_8(z)}{0.8}\right)^{0.37} \; [\, h^{-1}\mathrm{Mpc}], \tag{B.4}$$

$$R_{1h} = 3.77 \left(\frac{\sigma_8(z)}{0.8}\right)^{-0.10} \; [\, h^{-1}\mathrm{Mpc}], \tag{B.5}$$

$$R_{2h} = 1.70 \left(\frac{\sigma_8(z)}{0.8}\right)^{0.42} \; [\, h^{-1}\mathrm{Mpc}]. \tag{B.6}$$

As first shown in Seljak & McDonald (2011) and discussed in Appendix A (see equation A.2), P_{01} is fully predicted from P_{00} through the relation

$$P_{01}(\boldsymbol{k}, a) = \mu^2 \frac{dP_{00}(k, a)}{d\ln a}, \tag{B.7}$$

where a is the scale factor. Thus, the appropriate time derivative of equation B.1, combined with the Zel'dovich expression for P_{01} discussed in detail in Appendix A amounts to a full model for $P_{01}(\boldsymbol{k})$, using the same 5 parameters defined in equations B.2-B.6.

We also include measurements of the small-scale dark matter correlation function when finding the best-fit parameters discussed in this section. For reference, we provide the full relation for $\xi_{BB}(r)$, the Fourier transform of equation B.1,

$$\begin{aligned}\xi_{BB}(r) = &-\frac{A_0 e^{-r/R}}{4\pi r R^2(1 - R_{1h}^2/R^2 + R_{2h}^4/R^4)} \\ &\times \Bigg[1 - R_1^2/R^2 \\ &+ A \exp\left[r\left\{R^{-1} - R_{2h}^{-2}\sqrt{(R_{1h}^2 - S)/2}\right\}\right] \\ &+ B \exp\left[r\left\{R^{-1} - R_{2h}^{-2}\sqrt{(R_{1h}^2 + S)/2}\right\}\right]\Bigg],\end{aligned} \tag{B.8}$$

where we have defined the following quantities:

$$S \equiv \sqrt{R_{1h}^4 - 4R_{2h}^4}, \tag{B.9}$$

$$\begin{aligned}A \equiv (2R_{2h}^4 S)^{-1}\Big[&R^2\left(-2R_{2h}^4 + R_1^2(R_{1h}^2 - S)\right) + R_{2h}^4(R_{1h}^2 - S) \\ &+ R_1^2(-R_{1h}^4 + 2R_{2h}^4 + R_{1h}^2 S)\Big],\end{aligned} \tag{B.10}$$

$$\begin{aligned}B \equiv -(2R_{2h}^4 S)^{-1}\Big[&R_{2h}^4(R_{1h}^2 + S) - R_1^2(R_{1h}^4 - 2R_{2h}^4 + R_{1h}^2 S) \\ &+ R^2\left(-2R_{2h}^4 + R_1^2(R_{1h}^2 + S)\right)\Big].\end{aligned} \tag{B.11}$$

B.2 Dark matter radial momentum power spectrum, P_{11}

We model the μ^4 term of the scalar component of the radial momentum auto power spectrum, $P_{11}[\mu^4]$, with a HZPT model, as the sum of a Zel'dovich term and a Padé sum

$$P_{11,s}[\mu^4](k) = P_{11,s}^{\mathrm{zel}}(k) + P_{11}^{BB}(k), \tag{B.12}$$

where $P_{11,s}^{\rm zel}$ is the Zel'dovich approximation expression for the radial momentum power spectrum discussed in detail in Appendix A. For $P_{11}^{BB}(k)$, we use a Padé sum of the form

$$P_{11}^{BB}(k) = A_0 \left(1 - \frac{1}{1+k^2R^2}\right) \frac{1}{1+(kR_{1h})^2}. \tag{B.13}$$

The redshift dependence of the parameters enters into the model through both $\sigma_8(z)$ and $f(z)$, where f is the logarithmic growth rate. The best-fit parameters used in this work are given by

$$A_0 = 659 \left(\frac{\sigma_8(z)}{0.8}\right)^{3.91} \left(\frac{f(z)}{0.5}\right)^{1.92} \; [\, h^{-3}{\rm Mpc}^3], \tag{B.14}$$

$$R = 19.0 \left(\frac{\sigma_8(z)}{0.8}\right)^{-0.37} \left(\frac{f(z)}{0.5}\right)^{-0.25} \; [\, h^{-1}{\rm Mpc}], \tag{B.15}$$

$$R_{1h} = 0.85 \left(\frac{\sigma_8(z)}{0.8}\right)^{-0.15} \left(\frac{f(z)}{0.5}\right)^{0.77} \; [\, h^{-1}{\rm Mpc}]. \tag{B.16}$$

Note that in the large-scale, linear perturbation regime, we have $P_{11,s}[\mu^4](k) = f^2 P_{\rm lin}$. As discussed in Seljak & Vlah (2015), the density auto spectrum in both SPT and the Zel'dovich approximation scales as the square of the linear power spectrum. Noting the additional factor of f^2 in the case of $P_{11,s}$, the low-k amplitude scalings predict $A_0 \propto f^2\sigma_8^4$; this result is close to the best-fit values found in equation B.14.

B.3 Halo-matter power spectrum, P^{hm}

The HZPT model for the halo-matter power spectrum, as discussed in Section 4.4.2, is

$$P^{hm}(k) = b_1 P_{00}^{\rm zel}(k) + P_{00}^{\rm BB}(k, A_0, R, R_1, R_{1h}, R_{2h}), \tag{B.17}$$

where $P_{00}^{\rm BB}$ is the broadband Padé term, as given by equation B.1. The best-fit parameters for the Padé term used in this work are

$$A_0 = 752\ b_1^{1.66} \left(\frac{\sigma_8(z)}{0.8}\right)^{3.65} \quad [\ h^{-3}\mathrm{Mpc}^3], \tag{B.18}$$

$$R = 16.9\ b_1^{-0.12} \left(\frac{\sigma_8(z)}{0.8}\right)^{-1.07} \quad [\ h^{-1}\mathrm{Mpc}], \tag{B.19}$$

$$R_1 = 5.19\ b_1^{-0.57} \left(\frac{\sigma_8(z)}{0.8}\right)^{0.16} \quad [\ h^{-1}\mathrm{Mpc}], \tag{B.20}$$

$$R_{1h} = 8.25\ b_1^{-0.84} \left(\frac{\sigma_8(z)}{0.8}\right)^{-0.13} \quad [\ h^{-1}\mathrm{Mpc}], \tag{B.21}$$

$$R_{2h} = 3.05\ b_1^{-1.03} \left(\frac{\sigma_8(z)}{0.8}\right)^{-0.36} \quad [\ h^{-1}\mathrm{Mpc}]. \tag{B.22}$$

Appendix C

Relation between model parameters in the halo model

In this section, we describe the relations between parameters of our model in the context of the halo model, as discussed in Section 4.4.1. We apply previous analyses of clustering in the halo model, i.e., Berlind & Weinberg (2002); Zheng (2004); Hikage & Yamamoto (2013); Abramo et al. (2015), to the specific notation used in our model. In particular, we are able to constrain the relative fraction (§C.1) and the linear bias (§C.2) for the sample of centrals with satellites in the same halo. We also derive expressions for the 1-halo amplitudes, $N_{c_B s}$ and $N_{s_B s_B}$, in terms of other model parameters using the halo model in Section C.3.

C.1 The fraction of centrals with satellites

The relative fraction for the c_B sample f_{c_B}, which gives the fraction of central galaxies that live in halos with at least one satellite galaxy, can be related to the other galaxy sample fractions. The number of galaxies in the c_B sample is equal to the number of centrals with only one satellite plus the number of centrals with greater than one satellite. Assuming each halo has exactly one central galaxy, we can express this as

$$f_{c_B} = \frac{N_{c_B}}{N_c} = \frac{N_{s_A}}{N_c} + \frac{1}{\langle N_{>1,s} \rangle} \frac{N_{s_B}}{N_c}, \tag{C.1}$$

where we have defined $\langle N_{>1,s} \rangle$ to be the mean number of satellites galaxies in halos with greater than one satellite. This parameter normalizes the number of s_B galaxies to the number of centrals, such that $\langle N_{>1,s} \rangle^{-1} N_{s_B}$ gives the number of centrals with greater than one satellite in the same halo. For a HOD similar to the BOSS CMASS galaxy sample, we typically have $\langle N_{>1,s} \rangle \sim 2.4$.

Using the definitions $f_s = N_s / N_g$ and $f_{s_B} = N_{s_B} / N_s$, and noting that $N_s = N_{s_A} + N_{s_B}$ and $N_g = N_c + N_s$, we can simplify equation C.1 as

$$f_{c_B} = \frac{f_s}{1-f_s}\left[1 + f_{s_B}\left(\langle N_{>1,s}\rangle^{-1} - 1\right)\right]. \tag{C.2}$$

C.2 The linear bias of centrals with satellites

Using the halo model, we can express the bias of a specific galaxy sample as an integral over the halo mass function, weighted by bias

$$b_X = \frac{1}{\bar{n}_X}\int d\ln M \frac{d\bar{n}_h}{d\ln M}\bar{N}_x(M)b(M)u(k|M), \tag{C.3}$$

where $\bar{n}_x$ is the mean number density of the sample, $d\bar{n}_h/d\ln M$ is the halo mass function, $\bar{N}_x$ gives the mean halo occupation for the sample as a function of halo mass, $b(M)$ is the halo bias – mass relation, and $u(k|M)$ describes the halo profile in Fourier space.

For the sample of central galaxies with satellites in the same halo (denoted as c_B), we are able to express the mean occupation $\bar{N}_{c_B}$ in terms of quantities defined for the two satellite samples, s_A and s_B. In particular, we can write

$$\bar{N}_{c_B} = \bar{N}_{s_A} + \langle N_{>1,s}\rangle^{-1}\bar{N}_{s_B}, \tag{C.4}$$

where $\bar{N}_{s_A}$ is the occupation of satellites with only a single satellite in a halo, and $\bar{N}_{s_B}$ is the occupation of satellites with multiple satellites in the same halo. Here, we have defined $\langle N_{>1,s}\rangle$ to be the mean number of satellite galaxies in halos with greater than one satellite. Using equations C.3 and C.4, we can relate the linear biases as

$$\bar{n}_{c_B}b_{1,c_B} = \bar{b}_{s_A} + \langle N_{>1,s}\rangle^{-1}\bar{n}_{s_B}b_{1,s_B}. \tag{C.5}$$

We can relate the number density of individual samples to the total galaxy number density $\bar{n}_g$ as

$$\begin{aligned}\bar{n}_{c_B} &= f_{c_B}(1-f_s)\bar{n}_g,\\ \bar{n}_{s_A} &= f_s(1-f_{s_B})\bar{n}_g,\\ \bar{n}_{s_B} &= f_s f_{s_B}\bar{n}_g.\end{aligned}$$

Finally, we obtain the expression for b_{1,c_B}

$$b_{1,c_B} = \frac{(1-f_{s_B})f_s}{f_{c_B}(1-f_s)}b_{1,s_A} + \frac{f_{s_B}f_s}{\langle N_{>1,s}\rangle f_{c_B}(1-f_s)}b_{1,s_B}. \tag{C.6}$$

Using the expression for f_{c_B} from equation C.2, we can simplify this equation as

$$b_{1,c_B} = \frac{1-f_{s_B}}{1+f_{s_B}(\langle N_{>1,s}\rangle^{-1}-1)}b_{1,s_A} + \frac{f_{s_B}}{\langle N_{>1,s}\rangle(1-f_{s_B})+f_{s_B}}b_{1,s_B}. \tag{C.7}$$

Note that, as expected, the weights in this linear combination, $b_{1,c_B} = w_1 b_{1,s_A} + w_2 b_{1,s_B}$, sum to unity such that $w_1 + w_2 = 1$.

C.3 1-halo term amplitudes

In this section, we express the 1-halo amplitudes $N_{c_B s}$ and $N_{s_B s_B}$ in terms of other model parameters using a description of the shot noise in terms of pair counts of galaxies. Generically, we can write the shot noise of galaxies as

$$P^{\rm shot} = V \frac{\sum_{\rm halos} N_i^2}{\left(\sum_{\rm halos} N_i\right)^2} = V \frac{\sum_{\rm halos} N_i^2}{N_g^2}, \tag{C.8}$$

where V is the volume of the survey, N_i represents the number of galaxies in the $i^{\rm th}$ halo, N_g is the total number of galaxies, and we sum over all halos. Note that in the limit of a single object per halo, this simplifies to the usual expression for the Poisson shot noise, $P^{\rm shot} = V N_g / N_g^2 = \bar{n}_g^{-1}$, where $\bar{n}_g = V/N_g$ is the number density of the galaxy sample.

We can decompose the sum in the numerator of equation C.8 as

$$\begin{aligned}
\sum_{\rm halos} N_i^2 &= N_g + \sum_{\rm halos} N_i(N_i - 1), \\
&= N_g + \sum_{{\rm halos}, N=2} N_i(N_i - 1) + \sum_{{\rm halos}, N=3} N_i(N_i - 1) + \ldots, \\
&= N_g + 2N^{\rm halos}_{N=2} + 6N^{\rm halos}_{N=3} + \ldots, \\
&= N_g + \sum_{{\rm halos}, j=2}^{{\rm halos}, j=\infty} j(j-1) N^{\rm halos}_{N=j},
\end{aligned} \tag{C.9}$$

where $N^{\rm halos}_{N=j}$ is the total number of halos with exactly j galaxies in the halo.

To mirror our definitions of galaxy subsamples, we can decompose the sum over halos with greater than one galaxy member in equation C.9 into the contributions from central - satellite pairs and those between only satellites. For the former case, we can consider the number of pairs between centrals and satellites as

$$N^{\rm pairs}_{cs} = 2 \sum_{\rm halos} N_{s,i} = 2N_s = 2 f_s N_g, \tag{C.10}$$

where $N_{s,i}$ is the number of satellite galaxies in the $i^{\rm th}$ halo. And then using equation C.8, the total contribution of this term to the shot noise is

$$P^{1h}_{c_B s} = \frac{V}{N_g^2} N^{\rm pairs}_{cs} = \frac{2 f_s}{\bar{n}_g}, \tag{C.11}$$

and using the fact that $P^{1h}_{c_B s} = 2 f_s (1 - f_s) f_{c_B} N_{c_B s}$, we have

$$N_{c_B s} = \frac{1}{\bar{n}_g} \left[(1 - f_s) f_{c_B} \right]^{-1}, \tag{C.12}$$

where $\bar{n}_g$ is the number density of the full galaxy sample.

Similarly, we can consider the contribution to equation C.9 from the correlations between satellites. The contribution to the shot noise from satellite-satellite pairs is

$$P^{\rm shot}_{ss} = V \frac{\sum_{N_{s,i} > 1} N_{s,i} (N_{s,i} - 1)}{N_g^2}, \tag{C.13}$$

$$= \frac{V}{N_g^2} \left\langle N_{s,i} (N_{s,i} - 1) \right\rangle_{>1,s} N^{\rm halos}_{>1,s}, \tag{C.14}$$

where the quantity $\langle N_{s,i}(N_{s,i} - 1) \rangle_{>1,s}$ is averaged over halos with greater than one satellite, and $N^{\rm halos}_{>1,s}$ is the total number of halos that have more than one satellite. We can express the latter quantity as

$$N^{\rm halos}_{>1,s} = N_g \left[f_{c_B} (1 - f_s) - f_s (1 - f_{s_B}) \right], \tag{C.15}$$

where the first term represents the total number of halos with at least one satellite, and the second term is the number of halos with exactly one satellite. Here, we have explicitly assumed that every halo has exactly one central galaxy.

Using the fact that $P^{1h}_{s_B s_B} = f_s^2 f_{s_B}^2 N_{s_B s_B}$, the 1-halo amplitude becomes

$$N_{s_B s_B} = \frac{f^{1h}_{s_B s_B}}{\bar{n}_g f_s^2 f_{s_B}^2} \left[f_{c_B} (1 - f_s) - f_s (1 - f_{s_B}) \right], \tag{C.16}$$

where we have defined a normalization nuisance parameter $f^{1h}_{s_B s_B}$, which allows for variations in the unknown quantity $\langle N_{s,i}(N_{s,i} - 1) \rangle_{>1,s}$. Typically, for a CMASS-like galaxy sample, we find $f^{1h}_{s_B s_B} \sim 4$. For comparison, if $N_{s,i} = 2$ (3) for all halos with greater than one satellite, then $f^{1h}_{s_B s_B} = 2$ (6).

Bibliography

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Abazajian, K. N., Adshead, P., Ahmed, Z., et al. 2016, ArXiv e-prints, arXiv:1610.02743

Abramo, L. R., Balmès, I., Lacasa, F., & Lima, M. 2015, MNRAS, 454, 2844

Abramo, L. R., Secco, L. F., & Loureiro, A. 2016, MNRAS, 455, 3871

Achitouv, I., Blake, C., Carter, P., Koda, J., & Beutler, F. 2017, Phys. Rev. D, 95, 083502

Agrawal, A., Makiya, R., Chiang, C.-T., et al. 2017, J. Cosmology Astropart. Phys., 10, 003

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29

Alam, S., Ata, M., Bailey, S., et al. 2017, MNRAS, 470, 2617

Albrecht, A., & Steinhardt, P. J. 1982, Physical Review Letters, 48, 1220

Alcock, C., & Paczynski, B. 1979, Nature, 281, 358

Alonso, D., & Ferreira, P. G. 2015, Phys. Rev. D, 92, 063525

Alvarez, M., Baldauf, T., Bond, J. R., et al. 2014, ArXiv e-prints, arXiv:1412.4671

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2015, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, arXiv:1403.6015 [math.NA]

Anderson, L., Aubourg, E., Bailey, S., et al. 2012, MNRAS, 427, 3435

Anderson, L., Aubourg, É., Bailey, S., et al. 2014a, MNRAS, 441, 24

Anderson, L., Aubourg, E., Bailey, S., et al. 2014b, MNRAS, 439, 83

Angulo, R. E., Baugh, C. M., Frenk, C. S., & Lacey, C. G. 2008, MNRAS, 383, 755

Arfken, G. B., & Weber, H. J. 2012, Mathematical Methods for Physicists, Seventh Edition: A Comprehensive Guide, 7th edn. (Academic Press)

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Ata, M., Baumgarten, F., Bautista, J., et al. 2018, MNRAS, 473, 4773

Baldauf, T., Codis, S., Desjacques, V., & Pichon, C. 2016, MNRAS, 456, 3985

Baldauf, T., Seljak, U., Desjacques, V., & McDonald, P. 2012, Phys. Rev. D, 86, 083540

Baldauf, T., Seljak, U., Smith, R. E., Hamaus, N., & Desjacques, V. 2013, Phys. Rev. D, 88, 083507

Ballinger, W. E., Peacock, J. A., & Heavens, A. F. 1996, MNRAS, 282, 877

Bardeen, J. M. 1980, Phys. Rev. D, 22, 1882

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Bassett, B., & Hlozek, R. 2010, Baryon acoustic oscillations, ed. P. Ruiz-Lapuente, 246

Baumann, D., Jackson, M. G., Adshead, P., et al. 2009, in American Institute of Physics

Conference Series, Vol. 1141, American Institute of Physics Conference Series, ed. S. Dodelson, D. Baumann, A. Cooray, J. Dunkley, A. Fraisse, M. G. Jackson, A. Kogut, L. Krauss, M. Zaldarriaga, & K. Smith, 10

Baumgart, D. J., & Fry, J. N. 1991, ApJ, 375, 25

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, Phys. Rep., 367, 1

Betoule, M., Kessler, R., Guy, J., et al. 2014, A&A, 568, A22

Beutler, F., Seljak, U., & Vlah, Z. 2017a, MNRAS, 470, 2723

Beutler, F., Blake, C., Colless, M., et al. 2011, MNRAS, 416, 3017

—. 2012, MNRAS, 423, 3430

Beutler, F., Saito, S., Brownstein, J. R., et al. 2014a, MNRAS, 444, 3501

Beutler, F., Saito, S., Seo, H.-J., et al. 2014b, MNRAS, 443, 1065

Beutler, F., Seo, H.-J., Saito, S., et al. 2017b, MNRAS, 466, 2242

Beutler, F., Seo, H.-J., Ross, A. J., et al. 2017c, MNRAS, 464, 3409

Bianchi, D., Gil-Marín, H., Ruggeri, R., & Percival, W. J. 2015, MNRAS, 453, L11

Blake, C., & Glazebrook, K. 2003, ApJ, 594, 665

Blake, C., Kazin, E. A., Beutler, F., et al. 2011a, MNRAS, 418, 1707

Blake, C., Brough, S., Colless, M., et al. 2011b, MNRAS, 415, 2876

—. 2012, MNRAS, 425, 405

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28

Blas, D., Lesgourgues, J., & Tram, T. 2011, J. Cosmology Astropart. Phys., 7, 034

Blazek, J., Seljak, U., Vlah, Z., & Okumura, T. 2014, J. Cosmology Astropart. Phys., 4, 001

Bond, J. R., Jaffe, A. H., & Knox, L. 2000, ApJ, 533, 19

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. 1995, SIAM J. Sci. Comput., 16, 1190

Cai, Y.-C., Taylor, A., Peacock, J. A., & Padilla, N. 2016, MNRAS, 462, 2465

Camera, S., Santos, M. G., & Maartens, R. 2015, MNRAS, 448, 1035

Carlson, J., Reid, B., & White, M. 2013, MNRAS, 429, 1674

Carlson, J., & White, M. 2010, ApJS, 190, 311

Castorina, E., & White, M. 2017, ArXiv e-prints, arXiv:1709.09730

Chan, K. C., Scoccimarro, R., & Sheth, R. K. 2012, Phys. Rev. D, 85, 083509

Chevallier, M., & Polarski, D. 2001, International Journal of Modern Physics D, 10, 213

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., & Yepes, G. 2015, MNRAS, 446, 2621

Chuang, C.-H., Prada, F., Cuesta, A. J., et al. 2013, MNRAS, 433, 3559

Coil, A. L. 2013, The Large-Scale Structure of the Universe, ed. T. D. Oswalt & W. C. Keel, 387

Cole, S., Fisher, K. B., & Weinberg, D. H. 1995, MNRAS, 275, 515

Cole, S., & Kaiser, N. 1989, MNRAS, 237, 1127

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505

Coles, P., & Jones, B. 1991, MNRAS, 248, 1

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Collette, A., & contributors. 2017, HDF5 for Python

Contreras, C., Blake, C., Poole, G. B., et al. 2013, MNRAS, 430, 924

Cooray, A., & Sheth, R. 2002, Phys. Rep., 372, 1

Creminelli, P., Nicolis, A., Senatore, L., Tegmark, M., & Zaldarriaga, M. 2006, Journal of Cosmology and Astro-Particle Physics, 5, 4

Creminelli, P., & Zaldarriaga, M. 2004, J. Cosmology Astropart. Phys., 10, 006

Crocce, M., & Scoccimarro, R. 2006a, Phys. Rev. D, 73, 063520

—. 2006b, Phys. Rev. D, 73, 063519

Croom, S. M., Boyle, B. J., Shanks, T., et al. 2005, MNRAS, 356, 415

Cui, W., Liu, L., Yang, X., et al. 2008, ApJ, 687, 738

Dalal, N., Doré, O., Huterer, D., & Shirokov, A. 2008, Phys. Rev. D, 77, 123514

Dalcín, L., Paz, R., Storti, M., & DâĂŹElía, J. 2008, Journal of Parallel and Distributed Computing, 5, 655

Dalcin, L. D., Paz, R. R., Kler, P. A., & Cosimo, A. 2011, Advances in Water Resources, 34, 1124 , new Computational Methods and Software Tools

Das, S., Louis, T., Nolta, M. R., et al. 2014, J. Cosmology Astropart. Phys., 4, 014

Daubechies, I., ed. 1992, Ten lectures on wavelets

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Davis, M., Groth, E. J., & Peebles, P. J. E. 1977, ApJ, 212, L107

Davis, M., & Peebles, P. J. E. 1977, ApJS, 34, 425

—. 1983, ApJ, 267, 465

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44

de Jong, R. S., Barden, S., Bellido-Tirado, O., et al. 2014, in Proc. SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, 91470M

de la Torre, S., Guzzo, L., Peacock, J. A., et al. 2013, A&A, 557, A54

de Putter, R., & Doré, O. 2017, Phys. Rev. D, 95, 123513

Delubac, T., Bautista, J. E., Busca, N. G., et al. 2015, A&A, 574, A59

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016a, ArXiv e-prints, arXiv:1611.00036 [astro-ph.IM]

—. 2016b, ArXiv e-prints, arXiv:1611.00037 [astro-ph.IM]

Desjacques, V., Jeong, D., & Schmidt, F. 2016, ArXiv e-prints, arXiv:1611.09787

Desjacques, V., & Seljak, U. 2010, Classical and Quantum Gravity, 27, 124011

Diehl, H. T., Abbott, T. M. C., Annis, J., et al. 2014, in Proc. SPIE, Vol. 9149, Observatory Operations: Strategies, Processes, and Systems V, 91490V

Ding, Z., Seo, H.-J., Vlah, Z., et al. 2017, ArXiv e-prints, arXiv:1708.01297

Dodelson, S. 2003, Modern cosmology

Doré, O., Bock, J., Ashby, M., et al. 2014, ArXiv e-prints, arXiv:1412.4872

Efstathiou, G., Sutherland, W. J., & Maddox, S. J. 1990, Nature, 348, 705

Eisenstein, D., & White, M. 2004, Phys. Rev. D, 70, 103523

Eisenstein, D. J., & Hu, W. 1999, ApJ, 511, 5

Eisenstein, D. J., Hu, W., & Tegmark, M. 1998, ApJ, 504, L57

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23

Feng, Y. 2017a, bigfile

—. 2017b, fastpm-python

—. 2017c, kdcount

—. 2017d, pfft-python

—. 2017e, pmesh

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273

Feng, Y., & Hand, N. 2016, in Proceedings of the 15th Python in Science Conference, ed. Sebastian Benthall & Scott Rostrup, 137

Feng, Y., & Hand, N. 2017, runtests

Feng, Y., & Modi, C. 2017, Astronomy and Computing, 20, 44

Ferramacho, L. D., Santos, M. G., Jarvis, M. J., & Camera, S. 2014, MNRAS, 442, 2511

Ferraro, S., & Smith, K. M. 2015, Phys. Rev. D, 91, 043506

Font-Ribera, A., Kirkby, D., Busca, N., et al. 2014, J. Cosmology Astropart. Phys., 5, 027

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 24

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Freedman, W. L., Madore, B. F., Scowcroft, V., et al. 2012, ApJ, 758, 24

Friesen, B., Patwary, M. M. A., Austin, B., et al. 2017, in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '17 (New York, NY, USA: ACM), 20:1

Fry, J. N., & Gaztanaga, E. 1993, ApJ, 413, 447

George, E. M., Reichardt, C. L., Aird, K. A., et al. 2015, ApJ, 799, 177

Giannantonio, T., & Percival, W. J. 2014, MNRAS, 441, L16

Giannantonio, T., Ross, A. J., Percival, W. J., et al. 2014, Phys. Rev. D, 89, 023511

Gil-Marín, H., Percival, W. J., Cuesta, A. J., et al. 2016a, MNRAS, 460, 4210

Gil-Marín, H., Percival, W. J., Brownstein, J. R., et al. 2016b, MNRAS, 460, 4188

Gil-Marín, H., Guy, J., Burtin, E., et al. in prep., MNRAS

Ginzburg, D., Desjacques, V., & Chan, K. C. 2017, Phys. Rev. D, 96, 083528

Gradshteyn, I. S., Ryzhik, I. M., Jeffrey, A., & Zwillinger, D. 2007, Table of Integrals, Series, and Products

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., & Dalla Vecchia, C. 2016, MNRAS, 457, 1577

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., et al. 2017, MNRAS, arXiv:1607.03143

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332

Guo, H., Zehavi, I., & Zheng, Z. 2012, ApJ, 756, 127

Guo, H., Zheng, Z., Zehavi, I., et al. 2015a, MNRAS, 453, 4368

—. 2015b, MNRAS, 446, 578

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Guzzo, L., Pierleoni, M., Meneux, B., et al. 2008, Nature, 451, 541

Hahn, C., Scoccimarro, R., Blanton, M. R., Tinker, J. L., & Rodríguez-Torres, S. A. 2017, MNRAS, 467, 1940

Hamaus, N., Cousinou, M.-C., Pisani, A., et al. 2017, J. Cosmology Astropart. Phys., 7, 014

Hamaus, N., Pisani, A., Sutter, P. M., et al. 2016, Physical Review Letters, 117, 091302

Hamaus, N., Seljak, U., & Desjacques, V. 2011, Phys. Rev. D, 84, 083509

Hamaus, N., Seljak, U., Desjacques, V., Smith, R. E., & Baldauf, T. 2010, Phys. Rev. D, 82, 043515

Hamilton, A. J. S. 1992, ApJ, 385, L5

—. 1993, ApJ, 406, L47

—. 2000, MNRAS, 312, 257

Hand, N., & Feng, Y. 2017, classylss

Hand, N., Feng, Y., Beutler, F., et al. 2017a, ApJ

Hand, N., Li, Y., Slepian, Z., & Seljak, U. 2017b, J. Cosmology Astropart. Phys., 7, 002

Hand, N., Seljak, U., Beutler, F., & Vlah, Z. 2017c, J. Cosmology Astropart. Phys., 10, 009

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Hawken, A. J., Granett, B. R., Iovino, A., et al. 2017, A&A, 607, A54

Hawkins, E., Maddox, S., Cole, S., et al. 2003, MNRAS, 346, 78

Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., & Tollerud, E. 2016, MNRAS, 460, 2552

Hearin, A. P., Campbell, D., Tollerud, E., et al. 2017, AJ, 154, 190

Heymans, C., Grocutt, E., Heavens, A., et al. 2013, MNRAS, 432, 2433

Hikage, C. 2014, MNRAS, 441, L21

Hikage, C., & Yamamoto, K. 2013, J. Cosmology Astropart. Phys., 8, 019

Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki, 115

Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19

Hinton, S. 2016, The Journal of Open Source Software, 45

Ho, S., Agarwal, N., Myers, A. D., et al. 2015, J. Cosmology Astropart. Phys., 5, 040

Hockney, R. W., & Eastwood, J. W. 1981, Computer Simulation Using Particles

Hubble, E. 1934, ApJ, 79, 8

Hubble, E. P. 1926, ApJ, 64

Huchra, J., Davis, M., Latham, D., & Tonry, J. 1983, ApJS, 52, 89

Jackson, J. C. 1972, MNRAS, 156, 1P

Jennings, E. 2012, MNRAS, 427, L25

Jing, Y. P. 2005, ApJ, 620, 559

Jones, E., Oliphant, T., Peterson, P., et al. 2001–2017, SciPy: Open source scientific tools for Python

Kaiser, N. 1984, ApJ, 284, L9

—. 1987, MNRAS, 227, 1

Karagiannis, D., Shanks, T., & Ross, N. P. 2014, MNRAS, 441, 486

Kazin, E. A., Sánchez, A. G., & Blanton, M. R. 2012, MNRAS, 419, 3223

Kazin, E. A., Koda, J., Blake, C., et al. 2014, MNRAS, 441, 3524

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, MNRAS, 456, 4156

Komatsu, E., Dunkley, J., Nolta, M. R., et al. 2009, ApJS, 180, 330

Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18

Krauss, L. M., & Turner, M. S. 1995, General Relativity and Gravitation, 27, 1137

Kwan, J., Lewis, G. F., & Linder, E. V. 2012, ApJ, 748, 78

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints, arXiv:1110.3193 [astro-ph.CO]

Laurent, P., Eftekharzadeh, S., Le Goff, J.-M., et al. 2017, J. Cosmology Astropart. Phys., 7, 017

Leauthaud, A., Tinker, J., Behroozi, P. S., Busha, M. T., & Wechsler, R. H. 2011, ApJ, 738, 45

Leistedt, B., & Peiris, H. V. 2014, MNRAS, 444, 2

Leistedt, B., Peiris, H. V., Mortlock, D. J., Benoit-Lévy, A., & Pontzen, A. 2013, MNRAS, 435, 1857

Leistedt, B., Peiris, H. V., & Roth, N. 2014, Physical Review Letters, 113, 221301

Lesgourgues, J. 2011, ArXiv e-prints, arXiv:1104.2932 [astro-ph.IM]

Lesgourgues, J., & Pastor, S. 2006, Phys. Rep., 429, 307

Levi, M., Bebek, C., Beers, T., et al. 2013, ArXiv e-prints, arXiv:1308.0847 [astro-ph.CO]

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473

Linde, A. D. 1982, Physics Letters B, 108, 389

Linder, E. V. 2003, Physical Review Letters, 90, 091301

Loveday, J., Efstathiou, G., Maddox, S. J., & Peterson, B. A. 1996, ApJ, 468, 1

Ma, C.-P., & Fry, J. N. 2000, ApJ, 543, 503

Maddox, S. J., Efstathiou, G., Sutherland, W. J., & Loveday, J. 1990, MNRAS, 242, 43P

Maldacena, J. 2003, Journal of High Energy Physics, 5, 013

Matarrese, S., & Verde, L. 2008, ApJ, 677, L77

Matsubara, T. 2008, Phys. Rev. D, 78, 083519

McDonald, P., & Roy, A. 2009, J. Cosmology Astropart. Phys., 8, 020

McDonald, P., & Seljak, U. 2009, J. Cosmology Astropart. Phys., 10, 007

McEwen, J. E., Fang, X., Hirata, C. M., & Blazek, J. A. 2016, J. Cosmology Astropart. Phys., 9, 015

McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51

Mehta, K. T., Seo, H.-J., Eckel, J., et al. 2011, ApJ, 734, 94

Merz, H., Pen, U.-L., & Trac, H. 2005, New A, 10, 393

Modi, C., Castorina, E., & Seljak, U. 2017, MNRAS, 472, 3959

Mohammed, I., & Seljak, U. 2014, MNRAS, 445, 3382

Momcheva, I., & Tollerud, E. 2015, ArXiv e-prints, arXiv:1507.03989 [astro-ph.IM]

Moore, A. W., Connolly, A. J., Genovese, C., et al. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann, 71

Mueller, E.-M., Percival, W., Linder, E., et al. 2016, ArXiv e-prints, arXiv:1612.00812

Mueller, E.-M., Percival, W. J., & Ruggeri, R. 2017, ArXiv e-prints, arXiv:1702.05088

Myers, A. D., Brunner, R. J., Nichol, R. C., et al. 2007, ApJ, 658, 85

Myers, A. D., Palanque-Delabrouille, N., Prakash, A., et al. 2015, ApJS, 221, 27

Neyrinck, M. C. 2011, ApJ, 742, 91

NSF. 2017, NSF Committee on Software Infrastructure for Heterogeneous Computing

Oka, A., Saito, S., Nishimichi, T., Taruya, A., & Yamamoto, K. 2014, MNRAS, 439, 2515

Okumura, T., Hand, N., Seljak, U., Vlah, Z., & Desjacques, V. 2015, Phys. Rev. D, 92, 103516

Okumura, T., & Jing, Y. P. 2011, ApJ, 726, 5

Okumura, T., Seljak, U., & Desjacques, V. 2012a, J. Cosmology Astropart. Phys., 11, 014

Okumura, T., Seljak, U., McDonald, P., & Desjacques, V. 2012b, J. Cosmology Astropart. Phys., 2, 010

Okumura, T., Takada, M., More, S., & Masaki, S. 2017, MNRAS, 469, 459

Ostriker, J. P., & Steinhardt, P. J. 1995, Nature, 377, 600

Padmanabhan, N., & White, M. 2009, Phys. Rev. D, 80, 063508

Pâris, I., Petitjean, P., Ross, N. P., et al. 2017, A&A, 597, A79

Park, C., Gott, III, J. R., & da Costa, L. N. 1992, ApJ, 392, L51

Park, C., Vogeley, M. S., Geller, M. J., & Huchra, J. P. 1994, ApJ, 431, 569

Peacock, J. A. 1999, Cosmological Physics, 704

Peacock, J. A., & Dodds, S. J. 1994, MNRAS, 267, 1020

Peacock, J. A., & Smith, R. E. 2000, MNRAS, 318, 1144

Peacock, J. A., Cole, S., Norberg, P., et al. 2001, Nature, 410, 169

Peebles, P. J. E. 1973, ApJ, 185, 413

—. 1980, The large-scale structure of the universe

Percival, W. J., & White, M. 2009, MNRAS, 393, 297

Percival, W. J., Baugh, C. M., Bland-Hawthorn, J., et al. 2001, MNRAS, 327, 1297

Perez, F., & Granger, B. E. 2007, Computing in Science Engineering, 9, 21

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565

Pinol, L., Cahn, R. N., Hand, N., Seljak, U., & White, M. 2017, J. Cosmology Astropart. Phys., 4, 008

Pippig, M. 2013, SIAM Journal on Scientific Computing, 35, C213

Pippig, M. 2013, SIAM Journal on Scientific Computing, 35, C213

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A16

—. 2016a, A&A, 594, A13

—. 2016b, A&A, 594, A17

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing

Pullen, A. R., & Hirata, C. M. 2013, PASP, 125, 705

Raccanelli, A., Doré, O., & Dalal, N. 2015, J. Cosmology Astropart. Phys., 8, 034

Rasmussen, C., & Williams, C. 2006, Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning (Cambridge, MA, USA: MIT Press), 248

Rees, M. J. 1985, MNRAS, 213, 75P

Reid, B., Ho, S., Padmanabhan, N., et al. 2016, MNRAS, 455, 1553

Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2014, MNRAS, 444, 476

Reid, B. A., & Spergel, D. N. 2009, ApJ, 698, 143

Reid, B. A., & White, M. 2011, MNRAS, 417, 1913

Reid, B. A., Samushia, L., White, M., et al. 2012, MNRAS, 426, 2719

Repp, A., & Szapudi, I. 2017, MNRAS, 464, L21

Riebe, K., Partl, A. M., Enke, H., et al. 2013, Astronomische Nachrichten, 334, 691

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Riess, A. G., Macri, L. M., Hoffmann, S. L., et al. 2016, ApJ, 826, 56

Rodríguez-Torres, S. A., Comparat, J., Prada, F., et al. 2017, MNRAS, 468, 728

Ross, A. J., Samushia, L., Howlett, C., et al. 2015, MNRAS, 449, 835

Ross, A. J., Percival, W. J., Carnero, A., et al. 2013, MNRAS, 428, 1116

Ross, A. J., Beutler, F., Chuang, C.-H., et al. 2017, MNRAS, 464, 1168

Ross, N. P., Shen, Y., Strauss, M. A., et al. 2009, ApJ, 697, 1634

Ruggeri, R., Percival, W. J., Gil-Marín, H., et al. 2017, MNRAS, 464, 2698

Saito, S., Baldauf, T., Vlah, Z., et al. 2014, Phys. Rev. D, 90, 123522

Samushia, L., Branchini, E., & Percival, W. J. 2015, MNRAS, 452, 3704

Samushia, L., Reid, B. A., White, M., et al. 2013, MNRAS, 429, 1514

Sánchez, A. G., Baugh, C. M., Percival, W. J., et al. 2006, MNRAS, 366, 189

Sánchez, A. G., Scoccimarro, R., Crocce, M., et al. 2017, MNRAS, 464, 1640

Satpathy, S., Alam, S., Ho, S., et al. 2017, MNRAS, 469, 1369

Schmittfull, M., Baldauf, T., & Seljak, U. 2015a, Phys. Rev. D, 91, 043530

Schmittfull, M., Baldauf, T., & Zaldarriaga, M. 2017, Phys. Rev. D, 96, 023505

Schmittfull, M., Feng, Y., Beutler, F., Sherwin, B., & Chu, M. Y. 2015b, Phys. Rev. D, 92, 123522

Schmittfull, M., & Seljak, U. 2017, ArXiv e-prints, arXiv:1710.09465

Schmittfull, M., & Vlah, Z. 2016, Phys. Rev. D, 94, 103530

Schmittfull, M., Vlah, Z., & McDonald, P. 2016, Phys. Rev. D, 93, 103528

Schneider, P., & Bartelmann, M. 1995, MNRAS, 273, 475

Scoccimarro, R. 2004, Phys. Rev. D, 70, 083007

—. 2015, Phys. Rev. D, 92, 083532

Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20

Sefusatti, E., Crocce, M., Scoccimarro, R., & Couchman, H. M. P. 2016, MNRAS, 460, 3624

Seljak, U. 2000, MNRAS, 318, 203

—. 2009, Physical Review Letters, 102, 021302

—. 2012, J. Cosmology Astropart. Phys., 3, 004

Seljak, U., & McDonald, P. 2011, J. Cosmology Astropart. Phys., 11, 039

Seljak, U., & Vlah, Z. 2015, Phys. Rev. D, 91, 123516

Seo, H.-J., & Eisenstein, D. J. 2003, ApJ, 598, 720

—. 2005, ApJ, 633, 575

—. 2007, ApJ, 665, 14

Shane, C., & Wirtanen, C. 1967, The distribution of galaxies (University of California)

Sheldon, E. 2017, A python package for FITS input/output wrapping cfitsio

Shen, Y., Brandt, W. N., Richards, G. T., et al. 2016, ApJ, 831, 7

Shoji, M., Jeong, D., & Komatsu, E. 2009, ApJ, 693, 1404

Sinha, M. 2016, Corrfunc: Corrfunc-1.1.0

Sinha, M., & Garrison, L. 2017, Corrfunc: Blazing fast correlation functions on the CPU, Astrophysics Source Code Library, ascl:1703.003

Skibba, R. A., van den Bosch, F. C., Yang, X., et al. 2011, MNRAS, 410, 417

Slepian, Z., & Eisenstein, D. J. 2015a, MNRAS, 454, 4142

—. 2015b, ArXiv e-prints, arXiv:1510.04809

—. 2015c, MNRAS, 448, 9

—. 2016, MNRAS, 455, L31

—. 2017, ArXiv e-prints, arXiv:1709.10150

Slepian, Z., Eisenstein, D. J., Brownstein, J. R., et al. 2017, MNRAS, 469, 1738

Slosar, A., Hirata, C., Seljak, U., Ho, S., & Padmanabhan, N. 2008, J. Cosmology Astropart. Phys., 8, 031

Smee, S. A., Gunn, J. E., Uomoto, A., et al. 2013, AJ, 146, 32

Smith, R. E., Peacock, J. A., Jenkins, A., et al. 2003, MNRAS, 341, 1311

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints, arXiv:1503.03757 [astro-ph.IM]

Springel, V. 2005, MNRAS, 364, 1105

Springel, V., Yoshida, N., & White, S. D. M. 2001, New A, 6, 79

Suzuki, N., Rubin, D., Lidman, C., et al. 2012, ApJ, 746, 85

SymPy. 2017, sympy: a Python library for symbolic mathematics, http://www.sympy.org

Tadros, H., Ballinger, W. E., Taylor, A. N., et al. 1999, MNRAS, 305, 527

Takada, M., Ellis, R. S., Chiba, M., et al. 2014, PASJ, 66, R1

Takahashi, R., Sato, M., Nishimichi, T., Taruya, A., & Oguri, M. 2012, ApJ, 761, 152

Taruya, A., Bernardeau, F., Nishimichi, T., & Codis, S. 2012, Phys. Rev. D, 86, 103528

Taruya, A., Nishimichi, T., & Saito, S. 2010, Phys. Rev. D, 82, 063522

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, J. Cosmology Astropart. Phys., 6, 036

Team, D. D. 2016, Dask: Library for dynamic task scheduling

Tegmark, M. 1997a, Phys. Rev. D, 55, 5895

—. 1997b, Physical Review Letters, 79, 3806

Tegmark, M., Hamilton, A. J. S., Strauss, M. A., Vogeley, M. S., & Szalay, A. S. 1998, ApJ, 499, 555

Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22

Tegmark, M., Blanton, M. R., Strauss, M. A., et al. 2004, ApJ, 606, 702

Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, Phys. Rev. D, 74, 123507

Thomas, K., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, 20th International Conference on Electronic Publishing, ELPUB

Tinker, J. 2016

Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878

Tinker, J. L., Weinberg, D. H., & Zheng, Z. 2006, MNRAS, 368, 85

Tonry, J., & Davis, M. 1979, AJ, 84, 1511

Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2017, ArXiv e-prints, arXiv:1708.01538

Tucci, M., Desjacques, V., & Kunz, M. 2016, MNRAS, 463, 2046

Turk, M. J., Smith, B. D., Oishi, J. S., et al. 2011, ApJS, 192, 9

van Daalen, M. P., Schaye, J., Booth, C. M., & Dalla Vecchia, C. 2011, MNRAS, 415, 3649

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engineering, 13, 22

Vargas-Magaña, M., Ho, S., Fromenteau, S., & Cuesta, A. J. 2017, MNRAS

Vlah, Z., Seljak, U., & Baldauf, T. 2015, Phys. Rev. D, 91, 023508

Vlah, Z., Seljak, U., McDonald, P., Okumura, T., & Baldauf, T. 2012, J. Cosmology Astropart. Phys., 11, 009

Vlah, Z., Seljak, U., Okumura, T., & Desjacques, V. 2013, J. Cosmology Astropart. Phys., 10, 053

Vogeley, M. S., Park, C., Geller, M. J., & Huchra, J. P. 1992, ApJ, 391, L5

Wagner, C., Müller, V., & Steinmetz, M. 2008, A&A, 487, 63

Waters, D., Di Matteo, T., Feng, Y., Wilkins, S. M., & Croft, R. A. C. 2016, MNRAS, 463, 3520

Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., et al. 2013, Phys. Rep., 530, 87

Weisstein, E. 2017, Sphere-Sphere Intersection, http://mathworld.wolfram.com/Sphere-SphereIntersection.html

White, M. 2002, ApJS, 143, 241

—. 2014, MNRAS, 439, 3630

White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594

White, M., Blanton, M., Bolton, A., et al. 2011, ApJ, 728, 126

Wilson, M. J., Peacock, J. A., Taylor, A. N., & de la Torre, S. 2017, MNRAS, 464, 3121

Wolk, M., Carron, J., & Szapudi, I. 2015, MNRAS, 454, 560

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Yamamoto, K., Nakamichi, M., Kamino, A., Bassett, B. A., & Nishioka, H. 2006, PASJ, 58, 93

Yamamoto, K., Sato, T., & Hütsi, G. 2008, Progress of Theoretical Physics, 120, 609

Yamauchi, D., Takahashi, K., & Oguri, M. 2014, Phys. Rev. D, 90, 083520

Yang, Y.-B., Feng, L.-L., Pan, J., & Yang, X.-H. 2009, Research in Astronomy and Astrophysics, 9, 227

Yoo, J., Dalal, N., & Seljak, U. 2011, J. Cosmology Astropart. Phys., 7, 018

Yoo, J., & Seljak, U. 2015, MNRAS, 447, 1789

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Zel'dovich, Y. B. 1970, A&A, 5, 84

Zentner, A. R., Hearin, A., van den Bosch, F. C., Lange, J. U., & Villarreal, A. 2016, ArXiv e-prints, arXiv:1606.07817

Zentner, A. R., Hearin, A. P., & van den Bosch, F. C. 2014, MNRAS, 443, 3044

Zhao, G.-B., Wang, Y., Saito, S., et al. 2017, MNRAS, 466, 762

Zheng, Z. 2004, ApJ, 610, 61

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791

Zhu, F., Padmanabhan, N., & White, M. 2015, MNRAS, 451, 236

Zhu, F., Padmanabhan, N., White, M., Ross, A. J., & Zhao, G. 2016, MNRAS, 461, 2867

Zwicky, F., Herzog, E., Wild, P., Karpowicz, M., & Kowal, C. T. 1961, Catalogue of galaxies and of clusters of galaxies, Vol. I