OUR NON-GAUSSIAN UNIVERSE: HIGHER ORDER CORRELATION FUNCTIONS IN THE SLOAN DIGITIAL SKY SURVEY

by

Cameron Keith McBride

B.Sc. in Physics, Case Western Reserve University, 2003

B.A. in Chemistry, Case Western Reserve University, 2003

M.Sc. in Physics, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of
the Department of Physics & Astronomy in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF PHYSICS & ASTRONOMY

This dissertation was presented

by

Cameron Keith McBride

It was defended on

4 December 2009

and approved by

Andrew Connolly, University of Washington

Arthur Kosowsky, University of Pittsburgh

David Turnshek, University of Pittsburgh

Adam Lebovich, University of Pittsburgh

Rupert Croft, Carnegie Mellon University

Dissertation Advisors: Andrew Connolly, University of Washington,

Arthur Kosowsky, University of Pittsburgh

OUR NON-GAUSSIAN UNIVERSE: HIGHER ORDER CORRELATION FUNCTIONS IN THE SLOAN DIGITIAL SKY SURVEY

Cameron Keith McBride, PhD

University of Pittsburgh, 2010

Modern galaxy surveys, such as the Sloan Digital Sky Survey (SDSS), provide a wealth of information about large scale structure, galaxy evolution and cosmology. Even if initial density fluctuations were extremely Gaussian, gravitational collapse predicts the growth of non-Gaussianities in the galaxy distribution. Higher order clustering statistics, such as the three-point correlation function (3PCF), are necessary to probe the non-Gaussian structure and shape information in these distributions. We measure the clustering of spectroscopic galaxies in the SDSS Main Galaxy Sample, focusing on the shape or *configuration* dependence of the 3PCF in redshift and projected space. This work constitutes the largest observational dataset ever used to investigate the 3PCF, and the only known projected measurement for SDSS galaxies. The 3PCF exhibits extreme sensitivity to systematic effects such as sky completeness, binning scheme and insufficient error resolution. We show these systematics can dramatically affect our results, which are not consistently accounted for in comparable analyses. We find significant configuration dependence of the 3PCF on intermediate to large scales ($3 - 27\ h^{-1}$Mpc), in agreement with predictions from ΛCDM and disagreement with the hierarchical ansatz. Below 6 h^{-1}Mpc, the redshift space 3PCF shows reduced power and weak configuration dependence in comparison with projected measurements. Our results indicate that redshift distortions, and not galaxy bias, can make the 3PCF appear consistent with the hierarchical ansatz. Compared to the lower order 2PCF, the 3PCF shows a weaker dependence on luminosity with no significant dependence on scales above 9 h^{-1}Mpc. On scales less than 9 h^{-1}Mpc, the 3PCF shows a greater dependence on color than on luminosity.

We conclude that galaxies remain a biased tracer of the mass with a stronger bias associated with greater luminosity. Using a thorough error analysis in the linear regime ($9-27$ h^{-1}Mpc), we show bright galaxies ($M_r < -21.5$) are a biased realization of mass clustering at greater than 4.5σ in redshift space and 2.5σ in projected space. The strong degeneracy between linear and quadratic bias terms naturally explains the weak luminosity dependence of the 3PCF. Contrary to some claims, we find linear bias is sufficient to explain galaxy-mass bias of our samples.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This work would not have been possible without the support of my advisor, Andrew Connolly, who invested in me from the very first day. I would like to acknowledge Arthur Kosowsky, whose advice helped me to realize the context of graduate work and the scope of an academic career. My appreciation would be suspect if I did not include the significance of those that started me down this path, J. Chris Mihos and Glenn Starkman, as they taught me what a scientist is and sparked an intense interest for me to try and be one.

I am indebted to Jeffrey Gardner, Ryan Scranton and K. Simon Krughoff for their immense contribution in establishing the foundation of my graduate research. In addition, I am grateful to many in the SDSS collaboration for their patience and feedback that greatly aided my understanding of this work. I thank my academic siblings for their constructive comments and commiserations, specifically Ching-Wa, Sam, Jeremy and Niraj. There were many others whose friendship and support I found especially helpful in surviving graduate school, where the list includes Kiplin, Ramin, Ummi, Suman, Ben, Sandhya and especially Stephen. I do not want to even consider how people can successfully manage the process without Leyla, as I continually relied on her legendary shepherding skills. I want to thank David Turnshek for his candor and insistent emphasis on clear explanations, especially in this document. I am grateful for the guidance and advice that Andreas Berlind provided over the last couple years.

Personally, I want to thank my parents, Keith and Terry, along with my siblings: Evan, Brandon and Amber for supporting my curiosities and helping me to learn the importance of persistence. My Aunt Joyce, and her unwavering belief in my abilities, inspired continual forward progress. I want to express deep admiration of my wife, Cristy, and daughter, Ashlynne, for their support and patience that was necessary for me to complete this work.

The only thing I can offer to account for their sacrifice is my extreme gratitude in allowing me to pursue what I love.

1.0 INTRODUCTION

In the current paradigm of structure formation, we believe that an almost uniform distribution of mass at early times in the universe evolved through gravitational instability into the irregular and complex distribution that galaxies occupy today. The richness of structure we observe in galaxy redshift surveys or model with numerical simulations exhibit overdense knots, filaments and walls along with massive voids that all range from a few to hundreds of megaparsecs in size. Quantifying this "cosmic web" and understanding its evolution is the focus of research in *large-scale structure* (LSS; Peebles, 1980).

Structure formation does not try to describe the origin of the initial conditions of the universe, rather it builds off of an *inflationary cosmology* (Guth, 1981). As the initial hot and dense universe expands and cools, its dynamics enter a phase of *matter domination* (when the universe is roughly 30,000 years old) where the non-relativistic matter starts to dominate over free-streaming radiation pressure and small density perturbations can grow. However, matter and radiation remain coupled after this transition, thereby limiting the growth of structure due to competition between gravitational collapse and radiation pressure. About 330,000 years after inflation, the universe cools enough for atomic nuclei to capture electrons. This *epoch of recombination* causes a rapid transition from a hot plasma to a neutral, transparent universe. At this stage, the dynamics of gravitational collapse become decoupled from radiation and the growth of structure progresses unhindered.

The free steaming photons from recombination, observed today as the cosmic microwave background radiation (CMB), document the initial conditions of structure formation: small density fluctuations from an incredibly uniform spatial distribution that are well described by a Gaussian probability distribution function. Applications of gravitational perturbation theory and strong non-linear collapse in numerical simulations prove extraordinarily successful in describing the evolution of structure. We see strong support for *hierarchical structure formation*, where small objects form before larger ones (e.g. galaxies before galaxy clusters).

Gravitational dynamics are sensitive to cosmology and depend on the spatial curvature of the universe as well as the contents of the material in it. A combination of many observations including both LSS and CMB measurements support a standard model in accordance with current observations (see recent constraints in Komatsu et al., 2009). This concordance model of cosmology suggests a critically dense (spatially flat) universe. In addition to the baryonic matter that makes up stars, planets and people – the majority of mass exists in an as-of-yet undetected form, aptly called *dark matter*. To evolve into the structure we observe, the velocity of this matter must be well below relativistic speeds, and is therefore referred to *cold* dark matter (CDM). Finally, the expansion of the universe appears to be accelerating, a result of some form of *dark energy* that is currently consistent with a cosmological constant (denoted as Λ). This concordance model, referred to as ΛCDM, forms the basis of predicting LSS and the framework underlying galaxy formation and evolution.

We probe density perturbations today by studying the statistical properties of the distribution of galaxies which occupy highly overdense regions. We observe significant clumping and clustering in line with predictions from ΛCDM. However, there is a large conceptual hurdle between following the evolution of mass in gravitational collapse and that realized by galaxy positions. There is little reason to believe there exists a perfect one-to-one correspondence between mass overdensities and galaxy positions; complex galaxy formation processes such as merging and feedback should have significant contributions. This discrepancy between predicted "mass" relative to the observed "light" in galaxies is often described as *galaxy-mass bias*.

The parameterization of galaxy-mass bias allows us to use a two pronged approach to probe both cosmology and galaxy formation. On one side, we map the clustering of galaxies to that of the underlying mass distribution allowing us to understand and constrain cosmology. Alternatively, the parameterization of the bias itself encodes useful information concerning galaxy formation processes. This distills observational data from hundreds of thousands of galaxies into a significantly smaller and more manageable form, enabling powerful empirical constraints for theoretical models.

1.1 MOTIVATION

If the galaxy distribution was entirely Gaussian, clustering properties would be completely determined by the two-point correlation function (2PCF), or its Fourier space analog the power spectrum. Although analyses of the CMB suggest that the primordial mass fluctuations in our universe appear extremely Gaussian, we expect gravitational evolution between the early epoch observed by the CMB and today to produce non-Gaussian signatures in the galaxy distribution. The 2PCF remains only a limited view of the full distribution which cannot sufficiently probe non-Gaussian signals.

To investigate non-Gaussian structure, as well as shape information in galaxy distributions, we require higher order clustering statistics. In the hierarchy of n-point correlation functions, the three-point correlation function (3PCF) is the lowest order statistic to provide information on shape. For example, this enables probes of the triaxial nature of halos and extended filaments within the "cosmic web". Measurements of higher order moments allow a more complete picture of the galaxy distribution, breaking model degeneracies describing cosmology and galaxy properties.

Unfortunately, the information contained in higher order moments comes at a price. Their increased complexity make the measurements, modeling and interpretation difficult. Theoretically, non-linear contributions have significant non-trivial implications. Their calculation gets computationally challenging and efficient algorithms become critically important. They require larger and cleaner galaxy samples as they show more sensitivity to observational systematics than their lower order cousins (such as the 2PCF). As it was recently described: "the overlap between well understood theory and reliable measurements is in fact disquietingly small" (Szapudi, 2005). This work attempts to increase this overlap by leveraging the massive data available in a modern galaxy survey such as the SDSS.

1.2 OVERVIEW OF THESIS

In this thesis, we discuss efficient means to measure higher order correlation functions from large observational surveys, such as the Sloan Digital Sky Survey (SDSS), and how to make effective constraints using the data. We focus on the configuration dependence of the reduced 3PCF of galaxies in the SDSS, constraining parameters relevant to galaxy formation with implications for better understanding of cosmology. This work represents the largest sample of galaxy data ever analyzed with the reduced 3PCF, and the only investigation of the projected 3PCF for the SDSS.

In the rest of this chapter, we review the concepts relevant to this work. We describe statistical descriptions for quantifying large-scale structure (LSS) in §1.3, specifically defining the n-point correlation functions. The model to relate the clustering of galaxies to properties of the underlying mass, i.e. galaxy-mass bias, is reviewed in §1.4. We continue by presenting the effects of redshift distortions on clustering in §1.5 and a method to minimize these effects by projecting the correlation functions in §1.6. In §1.7, we review the methods of estimating correlation functions from data.

In Chapter 2, we describe our SDSS galaxy samples. We discuss their selection and completeness, and describe the parent catalog they are derived from (Blanton et al., 2005b). We cover the creation of "representative samples" (i.e. volume-limited subsets) characterized by different luminosities. We also discuss necessary information about our use of a dark matter N-body simulation and mock galaxy catalogs for comparison to observed galaxies.

In Chapter 3, we discuss the computational challenge of analyzing the increasingly massive observational data that is becoming available. We introduce a solution we developed applicable to a wide range of analyses: Ntropy, an efficient parallel framework to utilize massively parallel supercomputing facilities (thousands of processors) using efficient tree-based data structures. We present two applications implemented with this framework: (1) a friends-of-friends (Davis et al., 1985) group finder and (2) an n-point correlation calculator. We use the n-point application for our analysis of SDSS galaxies.

In Chapter 4, we present clustering measurements of SDSS DR6 galaxies, primarily focusing on the configuration dependence of the reduced 3PCF on scales between 3 and

27 h^{-1}Mpc. We use the full shape dependence at three different scales roughly corresponding to the non-linear ($3 - 9$ h^{-1}Mpc), quasi-linear ($6 - 18$ h^{-1}Mpc) and linear ($9 - 27$ h^{-1}Mpc) regimes. We find significant configuration dependence at all scales, similar to predictions from the canonical ΛCDM model and in contrast the lack of shape dependence proposed by the hierarchical ansatz. We investigate the luminosity and color dependence for three galaxy samples with different r-band magnitude ranges, where significant differences appear only on scales below 9 h^{-1}Mpc. We find a weak luminosity dependence, and a color dependence that changes with scale showing a more pronounced difference on our smallest scale configuration ($3 - 9$ h^{-1}Mpc). We analyze measurements and associated covariance matrices for the 3PCF, *both* in redshift and projected space. We note significant structure and correlation in the covariance that varies with galaxy sample luminosity. We discuss the effect of "super structures", such as the Sloan Great Wall (Gott et al., 2005), have on these clustering measurements. We show that a few structures dramatically affect measurements on the whole volume, but which structure dominates depends on the scale being measured.

In Chapter 5, we constrain galaxy-mass bias which quantifies the clustering difference between "light" (galaxies) and "mass" (as realized by dark matter particles from an N-body simulation). We utilize a modern analysis technique to incorporate the full covariance matrix and minimize inaccuracies due to noise. We investigate the structure of the covariance matrices by inspection of their eigenvectors, confirming that our errors are signal dominated and well resolved. We obtain fits for linear and quadratic bias on two galaxy samples and discuss the implications for cosmology. We show galaxy clustering can be described as a *biased* realization of the mass field, where the strength of the bias varies by luminosity.

In Chapter 6, we investigate systematic effects associated with our measurements, especially those not well studied for the 3PCF. We investigate the effect of sky completeness and binning, which can alter or mask expected signal in the 3PCF. We address the effectiveness of projected correlation functions in minimizing the impact of redshift distortions. Finally, we question the quality of error estimation by comparing cross validation estimation (i.e. jackknife re-sampling) with independent mock catalogs corresponding to our brightest galaxy sample. We justify that the choices we made in previous chapters should not affect our results and highlight the importance of systematics that are often overlooked in comparable analyses.

Chapter 7 summarizes our conclusions and presents additional discussion of our main results. Additionally, we briefly review future directions of this work.

On a practical note, all figures represent the work of the authors unless we explicitly mention otherwise.

1.3 QUANTIFYING LARGE-SCALE STRUCTURE

1.3.1 Defining the Correlation Function

We define the mass density as a function of position, $\rho(\vec{x})$, that has a *well known* average density of $\bar{\rho}$. We define the fractional overdensity about the mean at a local point as

$$\delta(\vec{x}) = \frac{\rho(\vec{x})}{\bar{\rho}} - 1 \; . \tag{1.1}$$

We note casting the density in terms of the overdensity effectively removes the first moment of the δ field, i.e. $\langle\delta(\vec{x})\rangle = 0$, where the $\langle\rangle$ denotes an ensemble average. The *two-point correlation function* (2PCF) can be defined in terms of δ values characterized by the separation of two positions, $r_{12} = |\vec{x}_1 - \vec{x}_2|$, which we write

$$\xi(r_{12}) = \langle\delta(\vec{x}_1)\delta(\vec{x}_2)\rangle \; . \tag{1.2}$$

Generally given a field of values, say discrete objects filling some volume, we might ask whether the objects *cluster* more than expected from a uniform random field. We cast this in terms of a probability. If we know the mean density of objects, $\bar{n}$, we relate the marginal probability, δP, of finding an object within some marginal volume, δV, such that $\delta P = \bar{n}\delta V$. If we then want to define the probability of finding another object within some distance r from the original object, we need to know an additional property of the distribution that relates *pairs*. If the field is *random* then there is no clustering nor correlation and the same formula holds independent of r. However, if the distribution exhibits some clustering, we can quantify it using the 2PCF. The *conditional* marginal probability (conditional since we already start from an object) can then be found by

$$\delta P = \bar{n}\delta V \left[1 + \xi(r)\right] \ . \tag{1.3}$$

The function $\xi(r)$ encodes the excess probability above (or equivalently below) random. For a random field, $\xi = 0$. Positive values denote increased clustering; there is a higher probability of finding a point a distance r over random. Negative values show an *anti-correlation*, i.e. less likely than random down to a limit that the probability becomes zero when $\xi = -1$. ξ remains unbounded for positive values.

We can relate ξ to the marginal probability, δP, of two positions each characterized by small volumes, δV_1 and δV_2, also parameterized by their separation, r_{12}, as

$$\delta P = \bar{n}^2 \delta V_1 \delta V_2 \left[1 + \xi(r_{12})\right] , \tag{1.4}$$

The difference between (1.3) and (1.4) is strictly the starting point. The conditional probability considers finding another object while sitting at the location of one. The latter formulation in (1.4) generalizes the relation for any two positions within the field, hence the dependence on $\bar{n}^2$.

1.3.2 Three-Point Correlation Function

A *Gaussian* field refers to any distribution that is fully described (statistically) by only its first and second moments (e.g. a mean and variance). For the δ field, the mean is zero and $\xi(r)$ successfully describes all clustering properties. A *non-Gaussian* field, basically every other possible distribution besides uniform and Gaussian, has non-trivial higher moments (i.e. non-zero higher order correlation functions). Higher order functions of the distribution can be similarly defined with respect to overdensity fluctuations, where the three-point correlation function (3PCF) is given by

$$\zeta(r_{12}, r_{23}, r_{31}) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2)\delta(\vec{x}_3) \rangle \ . \tag{1.5}$$

Instead of a single dependent variable, such as r_{12} in $\xi(r_{12})$, we see the 3PCF relies on three separations necessary to parameterize triplets. Further higher order correlation functions (greater than $n = 3$) require even more variables, resulting in a "combinatorial explosion" of parameters (Szapudi, 2005). Relating the 3PCF back to marginal probabilities, we write

$$\delta P = \bar{n}^3 \delta V_1 \delta V_2 \delta V_3 \left[1 + \xi_{12}\xi_{23} + \xi_{12}\xi_{31} + \xi_{23}\xi_{31} + \zeta_{123}\right] \ , \tag{1.6}$$

where we use a simplified notation with $\xi_{12} = \xi(r_{12})$, etc. We see the probability of finding triplets depends both on the 3PCF and the product of lower order 2PCFs. Formally, the probability relates to the *joint* moment of the mass overdensity δ field. The n-point correlation functions, which we denote for $n = 2$ and 3 with ξ and ζ, are the *connected* joint moments. Connected moments do not include "accidental associations" due to clustering in lower order connected moments. To visualize this, think of a Gaussian distribution where all *connected* joint moments for $n > 2$ are zero. However, the probability of finding a triplet is certainly not zero, and will depend on the clustering characterized in ξ.

1.3.3 Reduced Three-Point Correlation Function

The *hierarchical ansatz* posits that the 3PCF can be estimated by a cyclic combination of respective 2PCFs:

$$\zeta(r_{12}, r_{23}, r_{31}) \approx Q\left[\xi_{12}\xi_{23} + \xi_{12}\xi_{31} + \xi_{31}\xi_{23}\right], \tag{1.7}$$

where Q denotes a scaling constant to adjust the amplitude (Peebles, 1980). Initial measurements of 3PCF using angular surveys suggested that the hierarchical ansatz held at small scales with $Q \approx 1.3$ (Peebles, 1980). Current observations indicate that this hierarchical scaling does not hold for galaxies in the weakly non-linear regime where measurements of galaxies show both scale and configuration dependence.

What was originally called the hierarchical amplitude (Q) can be rewritten as a function, specifically

$$Q(r_{12}, r_{23}, r_{31}) = \frac{\zeta(r_{12}, r_{23}, r_{31})}{\xi_{12}\xi_{23} + \xi_{12}\xi_{31} + \xi_{31}\xi_{23}} \ . \tag{1.8}$$

This definition provides a useful normalization, and $Q(r_{12}, r_{23}, r_{31})$ is commonly referred to as the normalized or *reduced* 3PCF. As long as the 2PCF remains well above zero, i.e. the denominator in (1.8), the value of the function Q roughly equals unity regardless of scale. This form was later justified by gravitational perturbation theory, as the evolution of the

3PCF depends on quadratic terms in the equations of motion encapsulated in the square of the 2PCF (Bernardeau et al., 2002). An additional benefit of utilizing such a "ratio statistic" is that we expect Q to be insensitive to both time and cosmology. To leading order, Q only depends on the spectral index and triangle configuration (Bernardeau et al., 2002).

1.4 GALAXY-MASS BIAS

Galaxies might not perfectly trace the mass field. To account for differences between the observed light and mass, we can consider galaxies to be a *biased* realization of the ΛCDM evolved mass field. In the local bias model (Fry and Gaztanaga, 1993), the galaxy overdensity, δ_g, can be connected to the mass density, δ_m, by a non-linear Taylor series expansion:

$$\delta_g = \sum_k \frac{b_k}{k!}\delta_m^k \approx b_1\delta_m + \frac{b_2}{2}\delta_m^2 \; . \tag{1.9}$$

This relation describes the mapping between galaxy and mass overdensities by simple scalar values, to second order: the linear (b_1) and quadratic (b_2) bias.

With measurements on galaxy n-point correlation functions, the clustering of galaxies can be related to mass clustering via the bias parameters. The 2PCF can be used to constrain the linear bias by equating the correlation function between galaxies, ξ_g, to that of dark matter, ξ_m, such that

$$\xi_g(r) = b_1^2\,\xi_m(r) \; . \tag{1.10}$$

The 3PCF is the lowest order correlation function that is sensitive to the quadratic bias term (to leading order). The analog to (1.10) for the connected 3PCF is written

$$\zeta_g(r_{12}, r_{23}, r_{31}) = b_1^3\zeta_m(r_{12}, r_{23}, r_{31}) + b_1^2 b_2\left[\xi_{12}\xi_{23} + \xi_{12}\xi_{31} + \xi_{31}\xi_{23}\right] \; , \tag{1.11}$$

where $\xi_1 = \xi_m(r_1)$, etc. This simplifies for the reduced 3PCF which we write in terms of two scalar values $B = b_1$ and $C = b_2/b_1$ as follows:

$$Q_g(r_{12}, r_{23}, \theta) = \frac{1}{B}\left[\, Q_m(r_{12}, r_{23}, \theta) + C\, \right] \; . \tag{1.12}$$

We have changed notation slightly in (1.12), parameterizing r_{31} instead as θ, the opening angle between the two sides r_{12} and r_{23}, which still defines a unique triplet.

If there is any functional shape to Q_m, its shape can be affected by B (since it is a multiplicative factor), whereas C will only cause an offset. We can imagine that B and C are somewhat degenerate in this description, especially if there is no shape dependence in $Q(\theta)$ – as two parameters are used to describe a simple change in amplitude. However, when the 3PCF exhibits a shape dependence, the degeneracy is broken since B affects the shape and C is just an amplitude offset. Larger values of B will dampen the configuration dependence of $Q(\theta)$. Even with the degeneracy broken, the values of B and C will likely show a strong correlation.

1.5 REDSHIFT DISTORTIONS

The most accurate method to determine galactic distances is to determine a redshift from galaxy spectra. The redshift identifies the recession velocity (cz) which we relate to distance (d) using Hubble's law (Hubble, 1929). Any dynamical or *peculiar* velocity (v_{pec}) of a galaxy will induce an additional redshift or blueshift so that we really measure

$$cz = H_o d + v_{pec} \ , \tag{1.13}$$

where H_o denotes the value of the Hubble constant at the current epoch (today). The systematic error on distance due to the peculiar velocity is commonly referred to as a *redshift distortion*. We refer to the positions that include *distorted* distances as redshift space, which we denote with s as opposed to "real" space distance r.

Redshift distortions effectively couple the density and velocity fields, complicating the models needed to accurately describe observational galaxy samples. A nice review of linear theory implications is presented in Hamilton (1998), and we present their illustration of redshift distortion in Figure 1.1. There are two main effects that have been identified in galaxy distributions. In the linear regime, the peculiar velocities remain low and a small squashing effect can appear to slightly enhance large elongations perpendicular to our line-of-sight (Kaiser, 1987). In the strongly non-linear regions of gravitational collapse, we expect the relative velocities to be high, e.g. a galaxy falling into a galaxy cluster. The high peculiar velocity creates an extreme elongation along our line-of-sight; structures affectionately called "fingers-of-god"(FoG).

We show a slice of SDSS galaxies in Figure 1.2, where such FoG can be clearly seen. As as example of the magnitude of the effect, consider a typical galaxy cluster where the velocity dispersion might be $1000km/s$. If we assume a galaxy is falling into the cluster along our line-of-sight, the distortion distance due to this peculiar velocity would be $10h^{-1}$Mpc.

The redshift space distortion will significantly affect any statistical measure of clustering such as correlation functions. This breaks the isotropy of the galaxy distribution (it is no longer translational invariant, but maintains rotational symmetry). This results in a redshift space correlation function dependent on three variables, namely the redshift distances to the galaxies s_1 and s_2 as well as their redshift space separation s_{12} (Hamilton, 1998).

$$\xi(r_{12}) \rightarrow \xi(s_1, s_2, s_{12}) \tag{1.14}$$

We commonly use the spherically averaged, or monopole term, of the correlation function expressed as a function of just the redshift space separation: $\xi(s_{12})$. We show a comparison of this redshift space 2PCF with respect to real space measurement for a mock galaxy catalog in Figure 1.3. We see a severe reduction in power on small scales in the redshift space 2PCF as the FoG scatter pairs. At separations larger than 3 h^{-1}Mpc the redshift space $\xi(s)$ shows more power than real space.

As another alternative, we can assume the angle subtended between galaxies stays small, making the radial redshift distortions plane-parallel (for any given pair). We then cast the separation in terms of two coordinates perpendicular and parallel to our line-of-sight, respectively r_p and π, so

$$\xi(s_1, s_2, s_{12}) \approx \xi(r_p, \pi) \ . \tag{1.15}$$

This relates back to the redshift space separation, $s = \sqrt{r_p^2 + \pi^2}$.

We show the correlation function, $\xi(r_p, \pi)$, in Figure 1.4 for an SDSS galaxy sample. At small r_p, we note the signature of fingers-of-god in the correlation function. At large r_p we

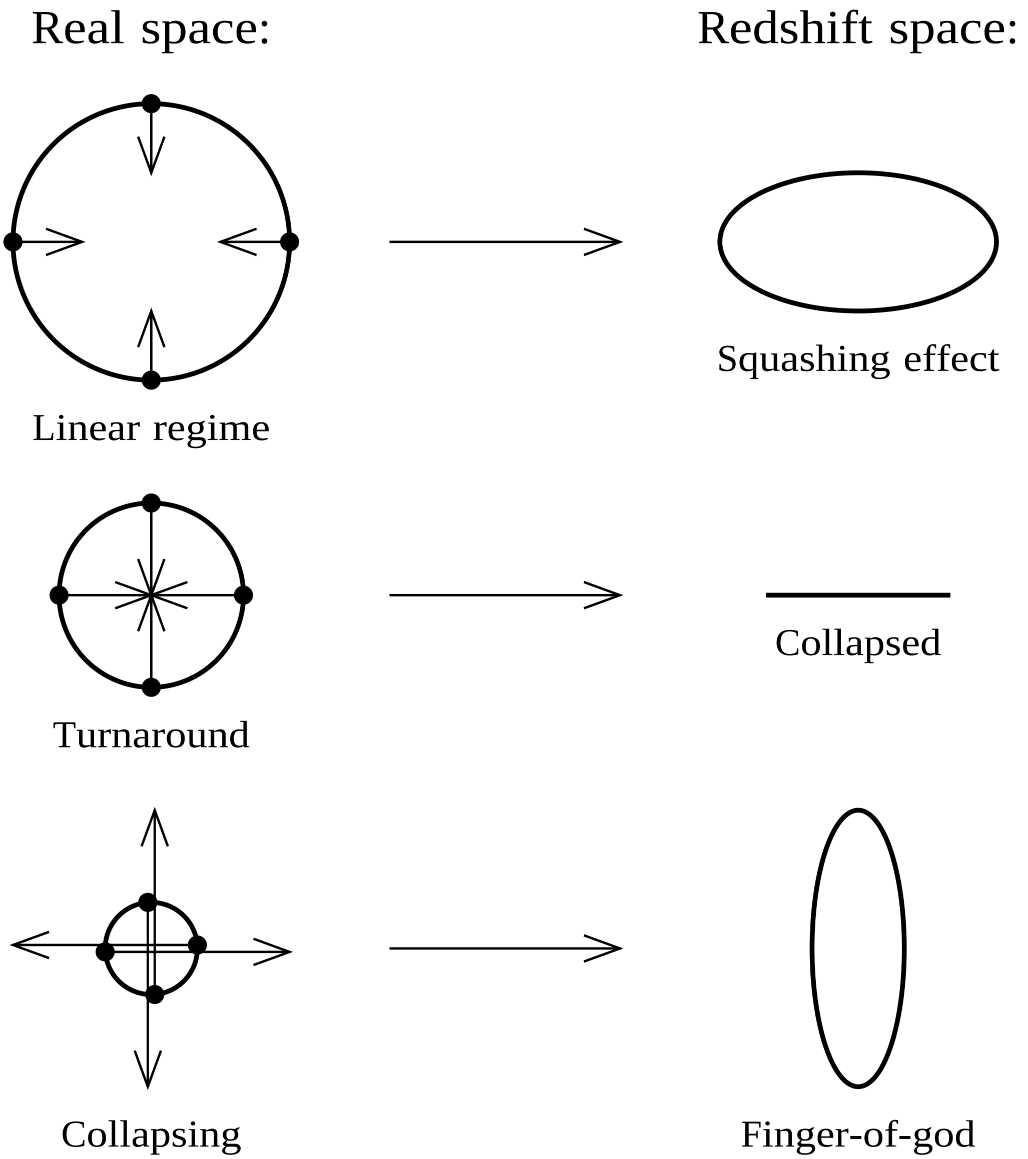


Figure 1.1 Illustrative description of redshift distortions. The shapes on the left represent shapes in real space, with points as “galaxies”. The arrows show the respective peculiar velocities which become entangled into the distance measurement and results in the redshift space shapes on the left. In the collapsing case, the strongly non-linear gravitational region actually appears to invert the structure (the closest “galaxy” to the observer appears the farthest away). *Image reproduced from Hamilton (1998)*.

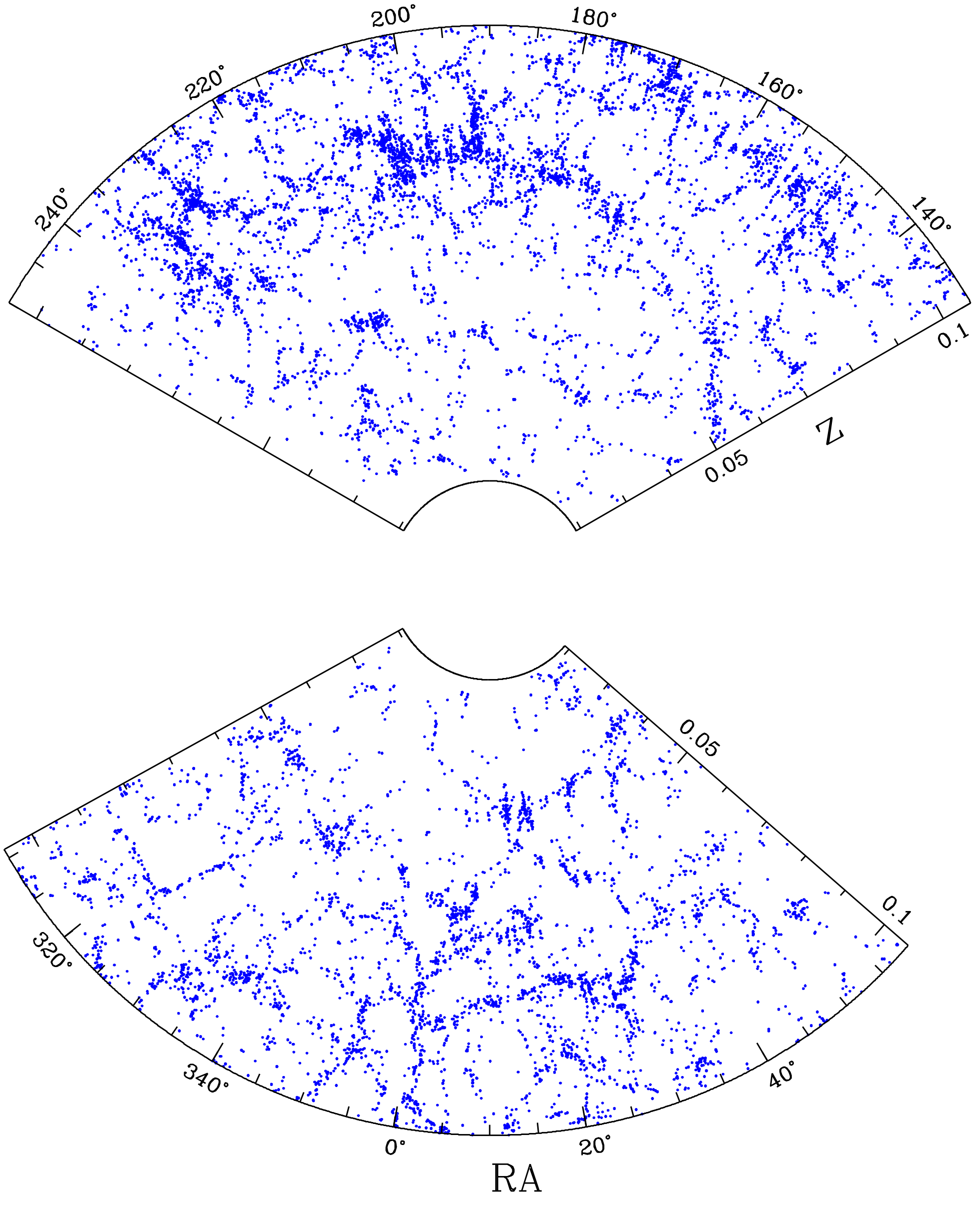


Figure 1.2 Slice of SDSS galaxy positions. Each point represents one galaxy, with the line-of-sight distance expressed as a redshift (denoted here with Z). Several finger-of-god structures can be seen, as well as the massive super structure in the top slice, often referred to as the Sloan Great Wall (Gott et al., 2005). *Image courtesy of A. Berlind.*

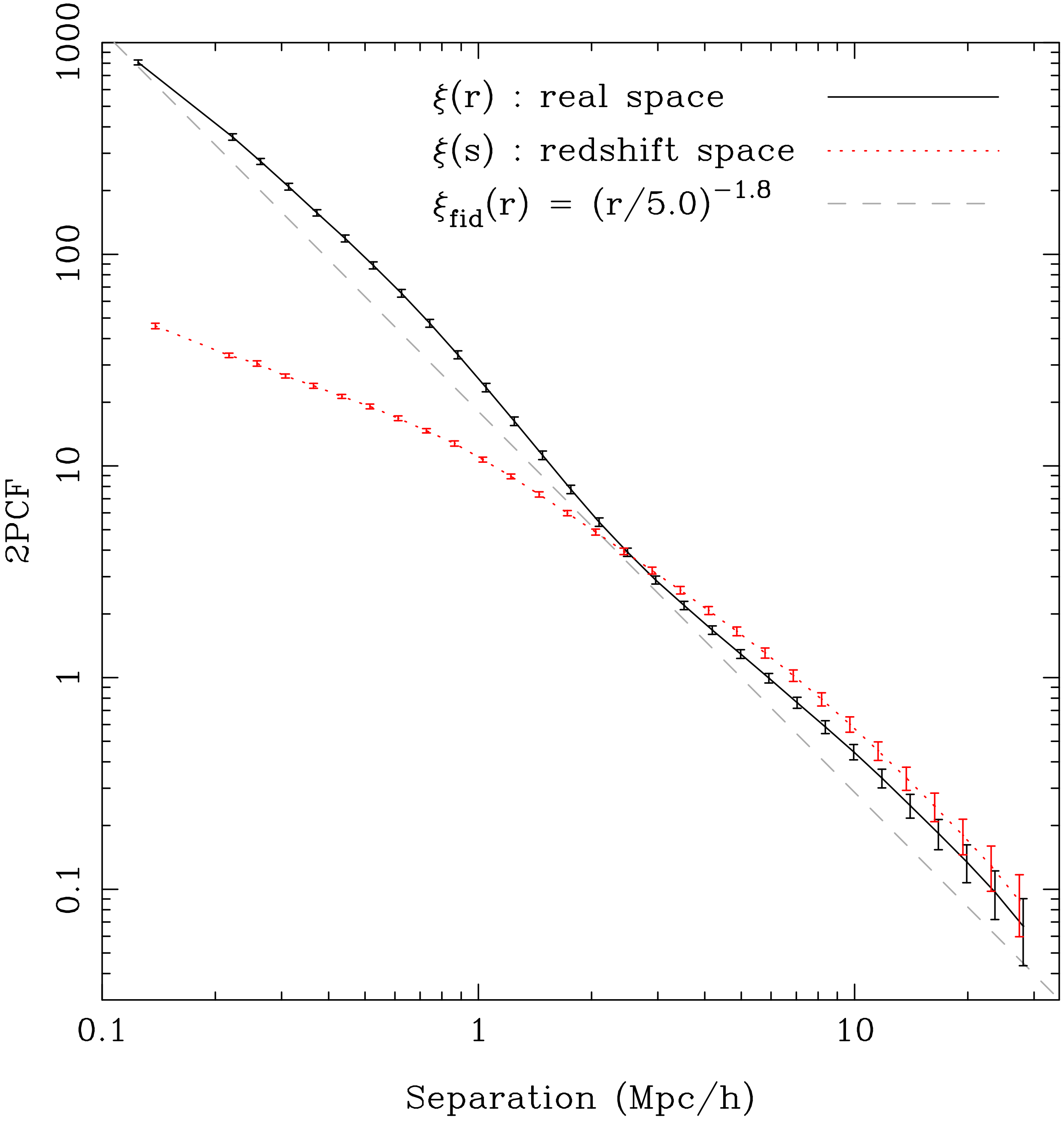


Figure 1.3 We show the 2PCF for a mock galaxy catalog we created from an N-body simulation, both with and without redshift space distortions. The error bars denote 1-σ uncertainties. For comparison, we include a *fiducial* power law model matching that used in Zehavi et al. (2005). Line-of-sight redshift distortions due to the peculiar velocities affect intrinsic clustering, illustrated by the difference in the solid black line (real space) and the dotted red line (redshift space).

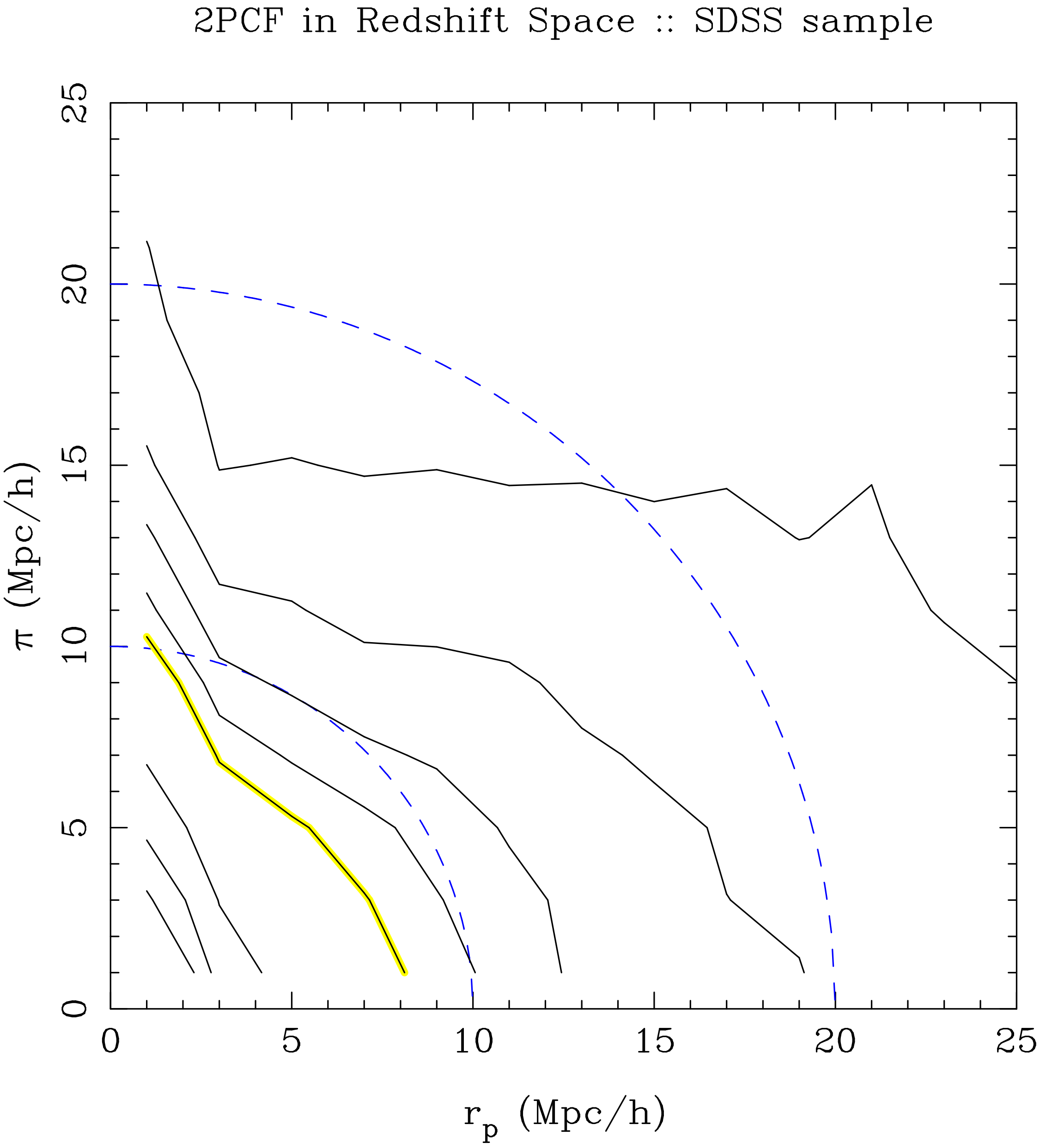


Figure 1.4 The 2PCF expressed in $r_p - \pi$ coordinates, perpendicular and parallel to our line-of-sight. The black lines are contours of a specific value of $\xi(r_p, \pi)$ with the yellow highlighted one corresponding to $\xi(r_p, \pi) = 1$. The blue (dot-dashed) semi-circles show a perfect isotropic correlation for comparison.

see the large scale Kaiser infall as a “squashing” of the contours. The symmetry of galaxy clustering is clearly broken by redshift distortions.

1.6 PROJECTED CORRELATION FUNCTIONS

Using the plane-parallel approximation of galaxies in redshift space, we notice that the π coordinate encapsulates the redshift distortion. We can integrate the redshift space 2PCF along π and introduce the projected two-point correlation function:

$$w_p(r_p) = 2\int_0^{\pi_{max}} \xi(r_p, \pi)\mathrm{d}\pi \ . \tag{1.16}$$

Physically, we project the 3D correlation function onto a 2D surface of the sky. Although we lose information from the dimensionality reduction of the projection, we achieve a statistical representation largely independent of redshift distortions. For small scales, where both the density and velocity fields are highly non-linear, redshift distortions become exceedingly difficult to model. The projected correlation function allows the high statistical significance of these small scale measurements to be robustly used with simple or even no modeling of the redshift distortions (we investigate its success later in Chapter 6).

Analogously, we define the projected 3PCF and its reduced form as:

$$\zeta_{proj}(r_{p12}, r_{p23}, r_{p31}) = \int\int \zeta(r_{p12}, r_{p23}, r_{p31}, \pi_{12}, \pi_{23})\mathrm{d}\pi_{12}\mathrm{d}\pi_{23} \tag{1.17}$$

$$Q_{proj}(r_{p12}, r_{p23}, r_{p31}) = \frac{\zeta_p(r_{p12}, r_{p23}, r_{p31})}{w_{p12}w_{p23} + w_{p12}w_{p31} + w_{p31}w_{p24}} \tag{1.18}$$

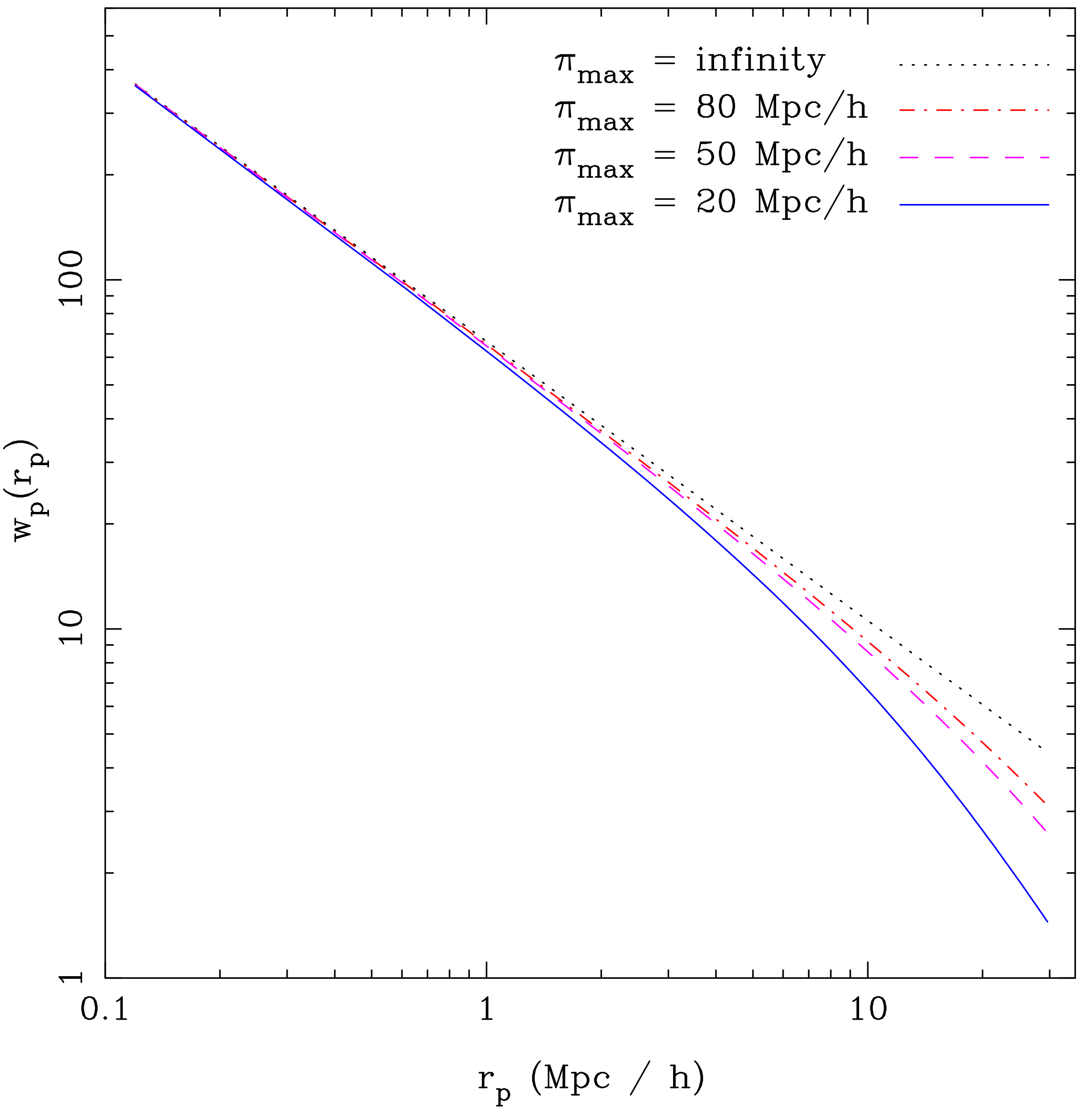


Figure 1.5 We show the *fiducial* projected 2PCF calculated by integrating the *fiducial* power-law model of the real space 2PCF: $\xi_{fid}(r) = \left(r/5\ h^{-1}\mathrm{Mpc}\right)^{-1.8}$. The dotted line depicts the full integration of the *fiducial* power-law in real space. We also show the effect of truncating the integration at π_{max} for three other values as noted in the legend.

1.7 ESTIMATING CORRELATION FUNCTIONS

To estimate correlation functions, we simply count pairs for the 2PCF, and triplets for the 3PCF. We convert the raw counts into estimates of clustering using the corresponding volume, which is typically done in Monte Carlo fashion by uniformly distributing points filling the same volume as the data. We refer to these latter distributions as *random catalogs* (reviewed in Szapudi, 2005).

Let us consider the 2PCF. From a finite data sample consisting of N points, we find all pairs that exist with separations between $(r - \Delta r/2)$ and $(r + \Delta r/2)$, where Δr denotes the bin-width. We refer to this count of data-data pairs as $dd(r)$. We normalize this raw count by the number of possible pairs, such that $DD(r) = \frac{dd(r)}{N(N-1)}$. We do the same using points from our random catalog finding the normalized random-random pairs $RR(r)$. The simplest representation of the 2PCF (Peebles, 1980) then becomes

$$\widehat{\xi}_{nat} = \frac{DD}{RR} - 1 \; , \tag{1.19}$$

where $\widehat{\xi}_{nat}$ signifies the *natural* estimator. Note, we have dropped the dependence on r only from our notation. In practice, the random catalogs typically contain many more points, often by factors of $5 - 50$, so the shot noise contribution (Poisson error) in $\widehat{\xi}$ are dominated by the data alone.

Unfortunately, we obtain galaxy data with many boundaries in irregular geometries. These “edges” produce significant contributions to estimates of correlation functions, deviating from the true underlying function of the data. New estimators were designed to optimally account for the *edge effects*, with two almost simultaneously presented: $\widehat{\xi}_{LS}$ by Landy and Szalay (1993) and $\widehat{\xi}_{Ham}$ by Hamilton (1993):

$$\widehat{\xi}_{LS} = \frac{DD - 2DR + RR}{RR} \; ; \tag{1.20}$$

$$\widehat{\xi}_{Ham} = \frac{DD\,RR}{DR^2} \; . \tag{1.21}$$

Here, DR corresponds to the normalized cross count of data-random pairs, such that $DR = \frac{dr}{N_D N_R}$ where N_D represents the number of data objects and N_R corresponds to that in random catalogs.

A detailed comparison between these and other estimators (Kerscher et al., 2000) found no distinguishable difference between $\widehat{\xi}_{LS}$ and $\widehat{\xi}_{Ham}$. They find that both of these minimum variance estimators optimally correct for edge effects and show stable estimates with the least number of randoms. We adopt $\widehat{\xi}_{LS}$ for our 2PCF estimates. The LS estimator was extended to a class of minimum variance estimators for all n-point correlation functions by Szapudi and Szalay (1998). We use this to estimate the 3PCF, specifically

$$\widehat{\zeta}_{SS} = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}\,. \tag{1.22}$$

The DDD denotes the normalized count of data-data-data triplets, DDR corresponds to data-data-random, etc. A brief comparison of three-point estimators shows $\widehat{\zeta}_{SS}$ performs favorably to alternatives (see appendix in Kayo et al., 2004).

2.0 DATA SAMPLES

In this chapter, we define the data samples we analyze throughout the thesis. We describe important aspects of the galaxy data obtained from the Sloan Digital Sky Survey (SDSS) in §2.1. We introduce the N-body simulations in §2.2 which we use to predict mass evolution through gravitational collapse in comparison to observational galaxy results. In addition, we detail the construction of mock galaxy catalogs in §2.3 which we later use to test our analysis and investigate the reliability of our estimated measurement errors in Chapter 6.

2.1 SLOAN DIGITAL SKY SURVEY

The SDSS (York et al., 2000) is an ambitious project based on a collaboration of over 25 institutions around the world. The data was obtained over eight years and managed as two sequential programs: SDSS-I (2000-2005) and SDSS-II (2005-2008). A third project, SDSS-III, is planned to continue operations until 2012, but we do not include any of these data in the current analysis. A dedicated 2.5 meter telescope at Apache Point Observatory in New Mexico obtained images and spectra covering nearly a quarter of the sky (see Gunn et al., 1998, for technical description). The 120 mega-pixel camera took imaging data in 5 different filters ($ugriz$) and a fiber spectrograph recorded spectra for over one million objects. The SDSS is arguably the most influential astronomical survey to date.

For our analysis, we use galaxy catalogs with accurate distances (obtained by measuring the spectroscopic redshift). In general, objects are identified from the multi-band imaging data and selected as targets for the fiber spectrograph according to a specific criteria detailed in Strauss et al. (2002). The algorithm which defines the Main galaxy sample selected about

90 galaxies per square degree which have a median redshift of 0.104 (Strauss et al., 2002). The Main galaxy sample has high completeness with an accurate statistical separation of starts and galaxies, preventing stellar contaminants in the galaxy samples.

The SDSS data is made available in almost yearly data releases. It is provided in a "scientifically digestible" reduced form as a product of a carefully developed software pipeline. We investigate systematics and perform initial measurements on a preliminary version of the 5th data release (DR5, Adelman-McCarthy et al., 2007) which we sometimes refer to as DR4+. We conduct our primary analysis of the three point correlation function on DR6 (Adelman-McCarthy et al., 2008). The final data product of SDSS-II was released to the public late in October 2008 as DR7 (Abazajian et al., 2009). Figure 2.1 depicts the main difference between the three releases: more available data in later releases due to larger angular coverage.

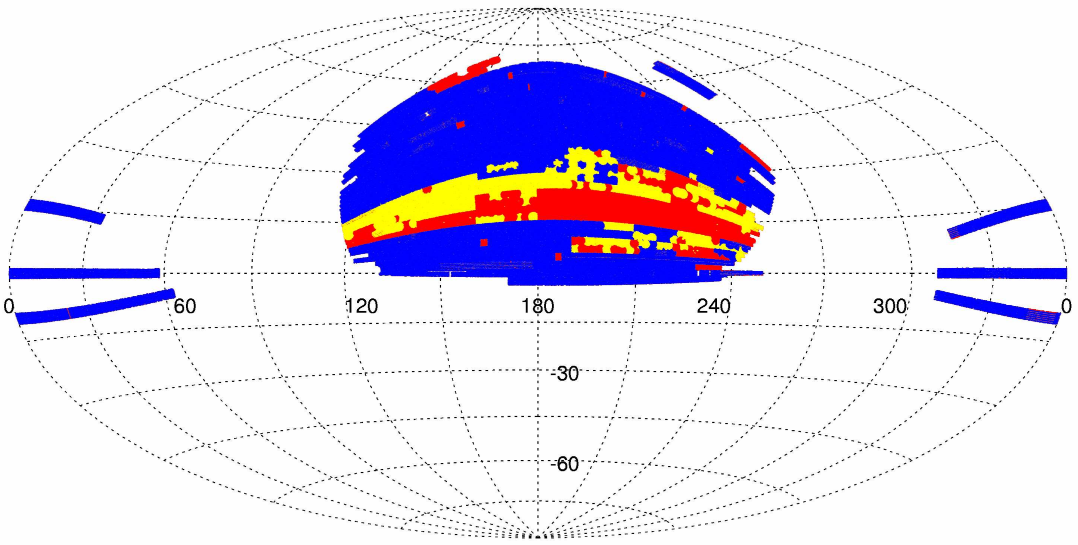


Figure 2.1 This figure is an Aitoff projection in equatorial coordinates of our specific SDSS spectroscopic galaxy samples from the NYU-VAGC. Our samples from the three latest data releases are shown: DR5 in blue, new regions to DR6 in yellow and new regions of DR7 in red.

2.1.1 Value-Added Galaxy Catalog

The SDSS galaxy samples are made more readily available for scientific analysis in the New York University Value-Added Galaxy Catalog (NYU-VAGC; Blanton et al., 2005b). This catalog focuses on spectroscopically targeted galaxies and provides detailed characterizations of the sample geometry and completeness as well as correcting for known systematics that are pertinent to large scale structure analyses. These include passive evolution corrections, K-corrections, and "fiber collision corrections" to correct for galaxies without measured redshifts due to galaxy pairs that are closer than fibers can be positions on the sky. We review these briefly here.

Statistical descriptions of the galaxy distribution, such as correlation functions, require a detailed account of the survey geometry. This is important to define the volume probed as well as define areas on the sky where objects could have been observed. The NYU-VAGC makes use of disjoint convex spherical polygons to store the geometry (Hamilton and Tegmark, 2004). This method is capable of accurately describing the complex shape of the SDSS footprint on the sky. Spectroscopic galaxy targets can be associated with a specific region that is described by these spherical polygons. A completeness is assigned for each region by taking the ratio of the number of successful redshifts to the number of identified targets.

The NYU-VAGC calculates additional quantities that are necessary for defining galaxy samples. A redshift is a measure of the recession velocity of the galaxy, which can be turned into an accurate distance measure for objects outside of the local group. However, the difference between the observed-frame wavelength bandpass to that of the actual rest-frame will also vary with distance. Basically, we observe a different part of the spectrum for distant objects (high redshift) as opposed to nearby objects. The "corrections" to account for this effect are referred to as K-corrections (see Oke and Sandage, 1968; Hogg et al., 2002). The absolute magnitude (M) can then be determined by using the apparent magnitude (m), the luminosity distance (D_L), and the K-correction (K):

$$M = m - 5\log\left(\frac{D_L}{10pc}\right) - K \; . \tag{2.1}$$

The K-correction is empirically determined by fitting *ugriz* magnitudes to carefully designed templates with the `kcorrect` code (Blanton et al., 2003b). They are calibrated at $z = 0.1$, which is close to the median redshift of the Main sample. The luminosity distance is calculated using the comoving distance element (for a concise review see Hogg, 1999) using a flat cosmology with $\Omega_{\mathrm{m}} = 0.3$, $\Omega_{\Lambda} = 0.7$, $H_o = h \times 100 km\ s^{-1}\ Mpc^{-1}$.

We expect the intrinsic brightness of galaxies to evolve with time. For the Main galaxy sample, a simple *passive evolution* model can characterize this as a function of redshift by examining the luminosity function and expected number density of objects. This correction to the absolute magnitude is determined and included in the NYU-VAGC (Blanton et al., 2005b) using a simple quadratic fit.

SDSS spectra are obtained by using 640 fibers per plate (one plate per observation). Hardware limitations restrict how close two fibers can physically be placed. Objects that are located within 55 $''$ "collide" and can not be simultaneously observed preventing redshifts for about 7% of the galaxy targets. A small number can be recovered in overlap regions between neighboring plates (a region covered by two different observations). Due to the significant clustering of galaxies, one must be careful to consider this effect which dramatically alters small scales. At large scales, a "fiber collision correction" can be applied to objects missing redshifts by assigning the exact redshift of the nearest angular neighbor to the object, which is incorporated into the NYU-VAGC. While this might appear dramatic, this method was carefully tested for clustering analyses by using simulated datasets (Zehavi et al., 2002) and found to be a sufficient correction for 5-6% of the sample (leaving only 1-2% of the total objects without redshifts).

2.1.2 Galaxy Sample Selection

To aid our understanding of galaxy clustering, we must analyze a "fair sample" of observational data. Galaxy surveys observe to an apparent brightness limit on the sky. The resulting flux-limited sample includes a larger volume of intrinsically brighter objects at farther distances with respect to fainter objects. This builds in an observational bias between brightness of the objects and total volume probed. This is typically handled in two ways. First, the radial distribution of the sample can be modeled and a radial selection function constructed to statistically weight the galaxy distribution and "correct" for the bias. The second option is to define a subset of the original sample by restricting the absolute magnitude range and redshift such that all objects within the brightness limits are observable through the entire redshift range. This subset is more "representative" of the universe although it is often of considerably smaller volume, hence referred to as a volume-limited sample.

There are benefits and drawbacks to both approaches. Clustering measurements based on the full flux-limited sample use the full statistical strength of all the available galaxies. However, the analysis can be sensitive to how well the radial selection function is modeled. This creates a chicken-or-the-egg problem: we must make significant assumptions about the galaxy distribution before actually measuring it. Volume-limited samples avoid the radial selection function by defining it to be the same over the entire redshift range. The drawback with using volume-limited samples is the smaller volume drastically reduces the data size and therefore the statistical strength of the measurements.

Given the size of data available with the SDSS, we choose to analyze volume-limited samples. Even the reduced volume of these samples can include over 100,000 objects (see Table 2.1). This avoids problems caused by the fact that the systematic effects in tabulating the radial selection function can dominate over statistical uncertainties in flux-limited samples.

We construct volume-limited samples by looking at the evolution corrected absolute r-band magnitude as a function of redshift. We then define a sample of objects by selecting bounding redshifts that correspond to a specific luminosity range. This is shown in Figure 2.2 for our SDSS DR6 samples. For each sample we tabulate the number of objects, volume and completeness corrected number density as shown in Table 2.1.

A common analytic form of the luminosity function, i.e. the number density of galaxies as a function of their intrinsic brightness, is given by Schechter (1976)

$$\phi(L)dL = \phi_* \left(\frac{L}{L_*}\right)^{\alpha} \exp\left(-\frac{L}{L_*}\right) d\left(\frac{L}{L_*}\right) \tag{2.2}$$

where $\phi(L)dL$ gives the number density of galaxies, parameterized by the normalization ϕ_*, a characteristic luminosity dividing the bright and faint ends L_* and the power-law slope

Specifics of SDSS galaxy samples

	Magnitude	Redshift	Volume $(Gpc/h)^3$	Number of Galaxies	Density $10^{-3}(Mpc/h)^{-3}$
BRIGHT	$M_r < -21.5$	0.010 to 0.210	0.1390	$37,875$	0.272
LSTAR	$-21.5 < M_r < -20.5$	0.053 to 0.138	0.0391	$106,824$	2.732
FAINT	$-20.5 < M_r < -19.5$	0.034 to 0.086	0.0098	$76,808$	7.849

Table 2.1 The redshift limits, volume, number of objects and sky completeness corrected number density (see text) are shown for the three galaxy samples constructed from the SDSS DR6 spectroscopic catalog. These are selected by cuts in redshift, z, and corrected (K-correction and passive evolution) absolute r-band magnitude, M_r, to create a volume-limited selection.

of the faint end in α. We plot this function with respect to the SDSS galaxy luminosity function (Blanton et al., 2003a) in Figure 2.3. We can see that fainter galaxies (below L_*) show a power law decrease in number density with increasing brightness. Brighter galaxies (well above L_*) become extremely rare exhibiting an exponential drop decrease in number density.

We define our galaxy samples in three luminosity ranges to investigate the luminosity dependence of our measurements. Observational datasets often have a characteristic luminosity, referred to at L_*, that represents the luminosity of the most prevalent galaxy in the survey. This is an observational "sweet spot" due to the competing effects of the galaxy luminosity function and limiting volume of observation.

Extremely faint galaxies can only be observed in a small volume and even though their expected density is high, the volume is too low to dominate the total galaxy count. At the very bright end, the exponential drop in the galaxy luminosity function is more rapid than the increasing volume and bright galaxies also become a minority contribution. For the SDSS Main galaxy sample, L_* corresponds to $M_r \approx -20.5$ (actually -20.42, see Blanton et al., 2003a). We choose two of our samples to be a bin of unit magnitude below L_* and above L_* to maximize the statistical strength while probing the luminosity dependence, respectively referred to as our FAINT and LSTAR samples. In addition, we investigate a brighter sample, called BRIGHT ($M_r < -21.5$), that covers the largest volume. Properties of these samples are detailed in Table 2.1.

We define our galaxy samples from two parent catalogs which are two versions of a flux-limited sample of SDSS DR6 available through the NYU-VAGC. The FAINT and LSTAR samples use the "safe" selection which conservatively applies an apparent magnitude restriction such that all galaxies are within $14.5 < m_r < 17.6$ (see blue data points in Figure 2.2). Our BRIGHT galaxy sample is defined from the "bright" flux-limited catalog (see black points in Figure 2.2), which does require members to be brighter than $m_r = 14.5$. The bright apparent magnitude cut was originally motivated to prevent contamination of fainter galaxy samples as an earlier version of the SDSS processing pipeline had more difficulties determining the apparent magnitude of near, bright galaxies.

Galaxy clustering measurements have been shown to vary strongly with color (Zehavi et al., 2005). We define red and blue sub-samples based on the $g - r$ color for two of our galaxy samples. There is a strong bi-modality between red and blue sub-samples that varies with absolute magnitude, as shown in Baldry et al. (2004). This can be seen in the color magnitude diagram (Figure 2.4), where the data points show a strong overdense linear structure, the "red sequence", which is predominantly populated by red elliptical galaxies and a less dense and more circular clump at bluer values, i.e. the "blue cloud". We adopt a simple linearly sloped color cut that depends on r-band magnitude to account for this effect, as described in Zehavi et al. (2005), such that the $g - r$ color limit is:

$$(g - r)_{lim} = -0.03M_r + 0.21 \ . \tag{2.3}$$

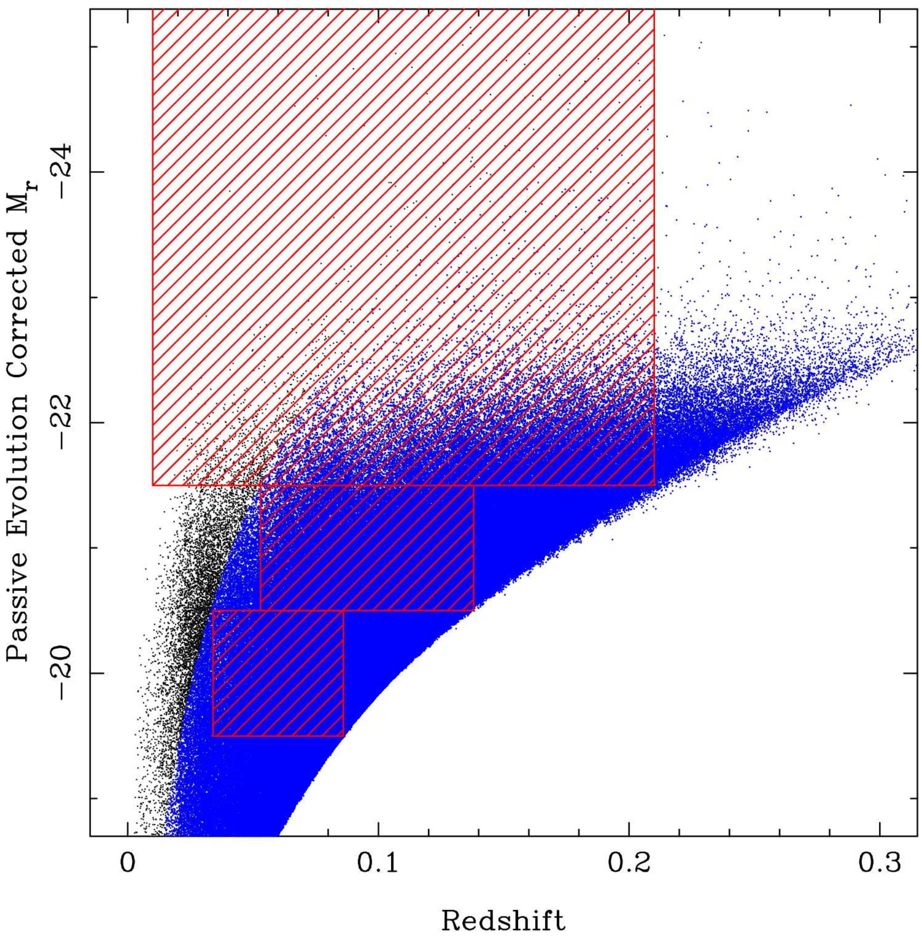


Figure 2.2 This figure depicts the sample selection for volume-limited samples from the SDSS DR6 release. We select the spectroscopic redshift limits where galaxies of all included magnitudes can be seen at both the inner and outer boundaries. The black points do not have a bright apparent magnitude cut, and is the base for our brightest volume-limited sample; the blue points show the so called "safe" catalog which we use to define our two fainter galaxy samples.

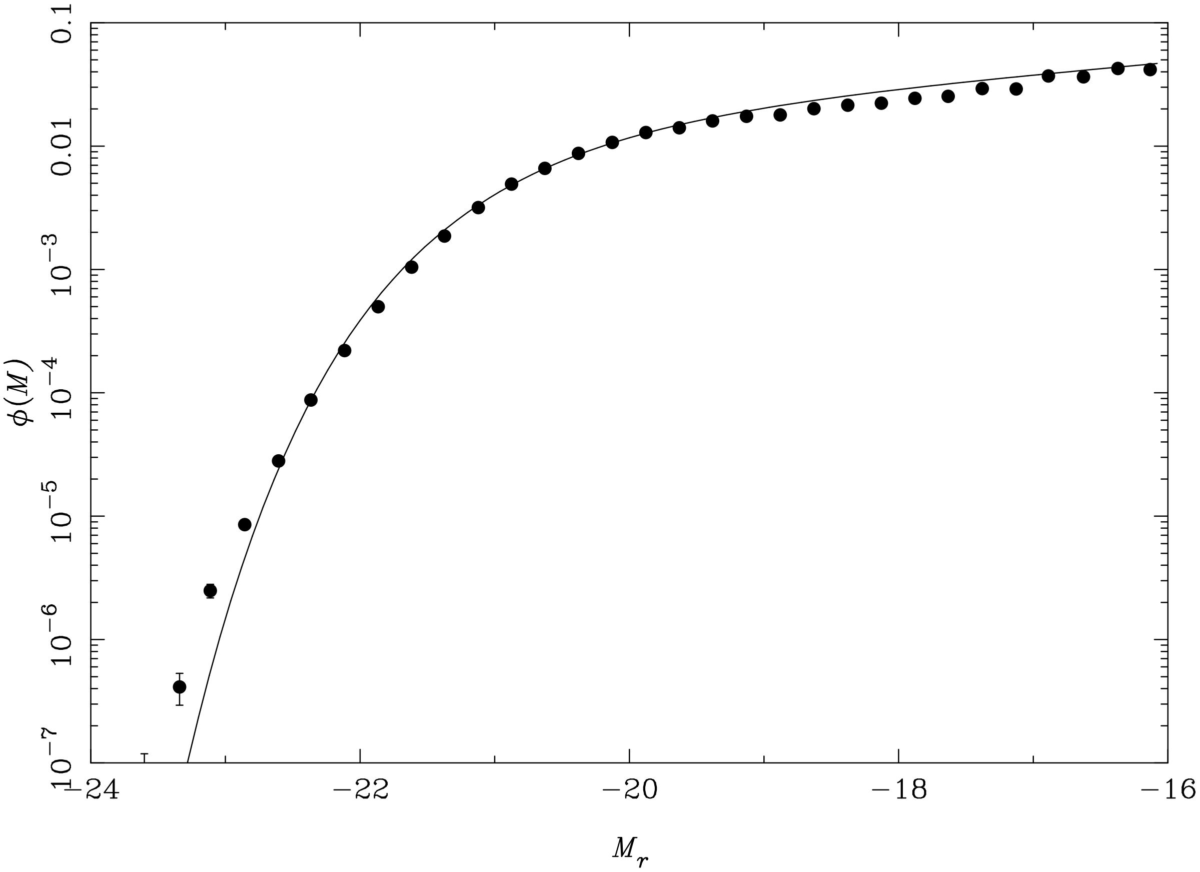


Figure 2.3 Luminosity function of SDSS Main galaxies from DR1 (symbols) with a best-fit Schechter function (line), taken from (Loveday, 2004). This luminosity function is characterized by magnitude (M) rather than luminosity (L), but is qualitatively the same as described in (2.2).

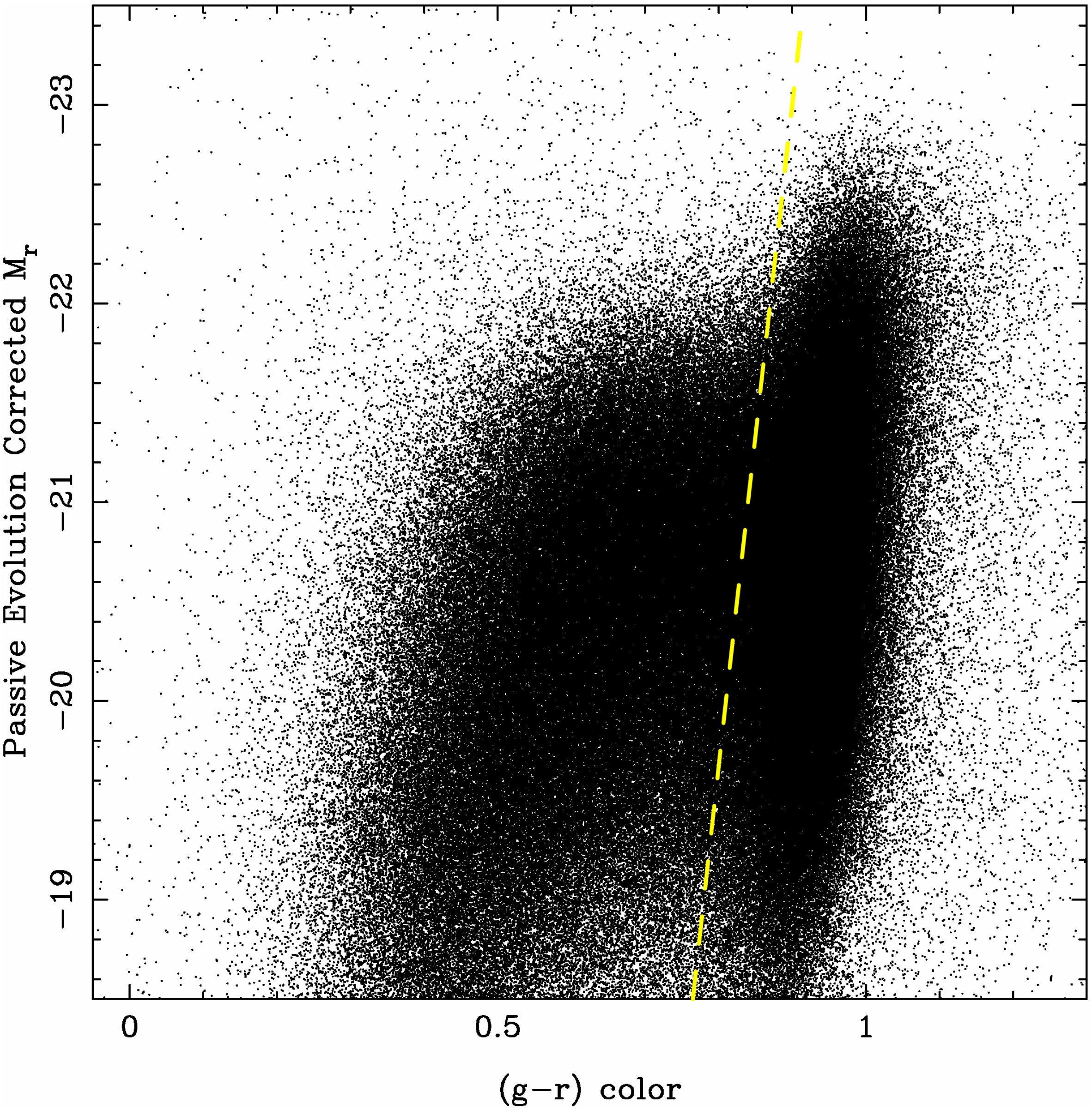


Figure 2.4 We show M_r magnitude as a function of color $(g - r)$ for the full DR6 galaxy sample. The "red sequence" can be seen the overdensity on the right which is redder (higher $g - r$ values) and slightly tilted. The "blue cloud" is the diffuse clump to the left (lower $g - r$ values). The dashed line shows the separation of the red and blue populations as a function of M_r as we discuss in the text.

2.2 HUBBLE VOLUME SIMULATION

We compare the SDSS galaxy distribution with the structure in cosmological N-body simulations. We use the *Hubble Volume* (HV) simulations (Colberg et al., 2000; Evrard et al., 2002) that were completed by the Virgo Consortium. We chose the simulation with ΛCDM cosmology: ($\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$, $H_o = 70km\ s^{-1}\ Mpc^{-1}$, $\sigma_8 = 0.9$). This HV simulation consists of 1000^3 particles in a box of $(3000h^{-1}\mathrm{Mpc})^3$ volume with a particle mass of $m_\mathrm{part} = 2.2 \times 10^{12} h^{-1} M_\odot$. To better compare with observational galaxy samples, we make use of a "light-cone" realization. Rather than have all positions at a single epoch, particles at farther distances from the origin correspond to earlier times in the gravitational evolution and hence trace the light-cone of observed galaxy samples. The light-cone output of the HV ΛCDM simulation was kindly provided by Gus Evrard and Jörg Colberg.

We apply observational constraints to the particles of the HV simulation. We filter particles to match the same angular footprint of the SDSS geometry (see Figure 2.1) as well as limiting the line-of-sight distance for each of the galaxy samples. This will exactly reproduce the volume of the galaxy sample. We also apply redshift distortions to the radial distance by using the particle's peculiar velocity (v_{pec}) to approximate the distortion distance,

$$d_{radial} = d_{comoving} + \frac{v_{pec}}{H_o}\ . \tag{2.4}$$

This is an approximation since we neglect evolution of the Hubble parameter and scale factor ($a = \frac{1}{1+z}$). At the low redshift of our galaxy samples, this is a fair approximation. Finally, we randomly downsample the number of dark matter particles to make the computational time of the analysis more manageable. To verify, we compare several downsampled realizations to a full distribution and noted the variation was well within expectations for Poisson sampling. We conclude that the measurements on dark matter distributions that we later use to compare to galaxy samples accurately represent the results of the full Hubble Volume simulation.

2.3 MOCK GALAXY CATALOGS

We use mock galaxy catalogs created to match the SDSS galaxy data, which are based on 49 independent N-body simulations. These were all evolved with the same cosmology, specifically $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, $h = 0.72$, $\sigma_8 = 0.9$ using different realizations with randomized phases where the initial conditions were generated from 2nd order Lagrangian perturbation theory (2LPT; Scoccimarro, 1998; Crocce et al., 2006). Please note that this cosmology differs slightly from that used in the observational galaxy samples, but this should not affect our main results (see Chapter 6). These simulations consist of 640^3 particles which we evolved using Gadget2 (Springel, 2005) from an initial redshift of $z_i = 49$ to the present epoch. The box side-length of 1280 h^{-1}Mpc was sufficiently large to match the geometry and cover the entire volume of our brightest SDSS sample.

The galaxy mocks were created by finding the number of galaxies that exist in each dark matter halo using a empirical description known as the halo occupation distribution (HOD; Berlind and Weinberg, 2002). The specific HOD model we use is described in detail in Tinker et al. (2005) with the parameters defined for $M_r < -21.5$ and $\sigma_8 = 0.9$ (Berlind, private communication). The halos were identified using a friends-of-friends algorithm (FoF; Davis et al., 1985) with a linking length of $b = 0.2$ in units of the mean interparticle separation. The least massive halos contained 33 particles, capable of representing the minimum halo mass necessary to host the faintest galaxies in the BRIGHT galaxy sample. Given the mass resolution, less massive halos that would host the fainter galaxies in the LSTAR and FAINT galaxy samples were not properly identified in these simulations. Therefore, these simulations can only be reliably used to for galaxy mocks corresponding to the BRIGHT galaxy sample.

3.0 COMPUTATION

In this chapter, we describe the computational difficulty associated with our investigation of the three-point correlation function and introduce code developed in collaboration with Jeffrey Gardner, Andrew Connolly and myself (see Gardner et al., 2007). We discuss the problem of analyzing the massive data available in current and future projects in §3.1. We present a general coding framework as a solution we developed for astrophysical contexts in §3.2. We further discuss two specific applications of our framework to address current scientific needs: (1) an n-point calculator in §3.3 and (2) a friends-of-friends groupfinder in §3.4. To be explicit about my involvement in code development: I focused on testing, debugging, result verification and minor development within the applications (along with their scientific use). The design and implementation of the *N*tropy framework is predominately the work of Jeff Gardner.

3.1 COMPUTATIONAL CHALLENGE

Over the past decade and continuing for the foreseeable future, we find ourselves fortunate enough to have a deluge of astrophysical data. Observational surveys are producing catalogs of unprecedented size. The Sloan Digital Sky Survey (SDSS) contains over 1 million spectra and more than 350 million unique objects in imaging data within the DR7 release. The large synoptic survey telescope (LSST) will produce time series photometry for over a billion objects, generating *terabytes* of data *every night*. On the theory side, the availability of supercomputing facilities with tens of thousands of processors makes simulations consisting of billions of particles commonplace. Obviously the situation is a good problem to have –

but how do we handle this data?

An appropriate answer is simple, but remains difficult to use in practice. To handle the increased data, we must employ larger, faster computing facilities and utilize better, more efficient algorithms. There are classes of problems where only one of the two solutions should suffice. For example, there are many problems that can be trivially divided up and run independently for each processor. Frameworks exist to make this approach highly effective, such as the Map-Reduce algorithm employed by Google (Dean and Ghemawat, 2004). In addition, sometimes an intractable calculation can become manageable by developing more efficient algorithms (Moore et al., 2001; Gray et al., 2004).

However, there remains a class of scientifically motivated inquiries that require interdependent calculations on data distributed simultaneously over many processors. Many of these applications can be addressed by efficiently sorting through the n-dimensional data partitioned through tree-based data structures. Examples include cluster finding, density estimation, object classification, n-point correlation functions, and N-body simulations of gravitational collapse (see additional discussion in Gardner et al., 2007). However, parallel codes often take many years to implement, debug and sufficiently test.

The use of n-point correlation functions represent an investigation that can be efficiently addressed using tree-based algorithms, and they are scientifically interesting for the massive amount of observational and simulated data that is becoming available. While calculations for the two-point correlation function (2PCF) remains tractable for most data with current workstations, the three-point correlation function (3PCF) is not, naively scaling as $O(N^3)$. Even when using "smart" algorithms (e.g. kd-tree based), which scales as $O(N^{1.4})$, a parallel approach is necessary. We illustrate the computational complexity of the 3PCF in Figure 3.1.

A typical researcher addressing his own scientific needs does not often have the time nor resources to develop complex and massively parallel codes. In addition, during the progression of the research many adjustments might be necessary to codify - in contrast to the workflow of the successful N-body shops. To address this, we introduce a novel framework, *N*tropy to aid rapid development of analysis codes for astronomical research that can utilize a tree-based approach. We use this framework to implement two applications: an n-point calculator and an astrophysical group finder.

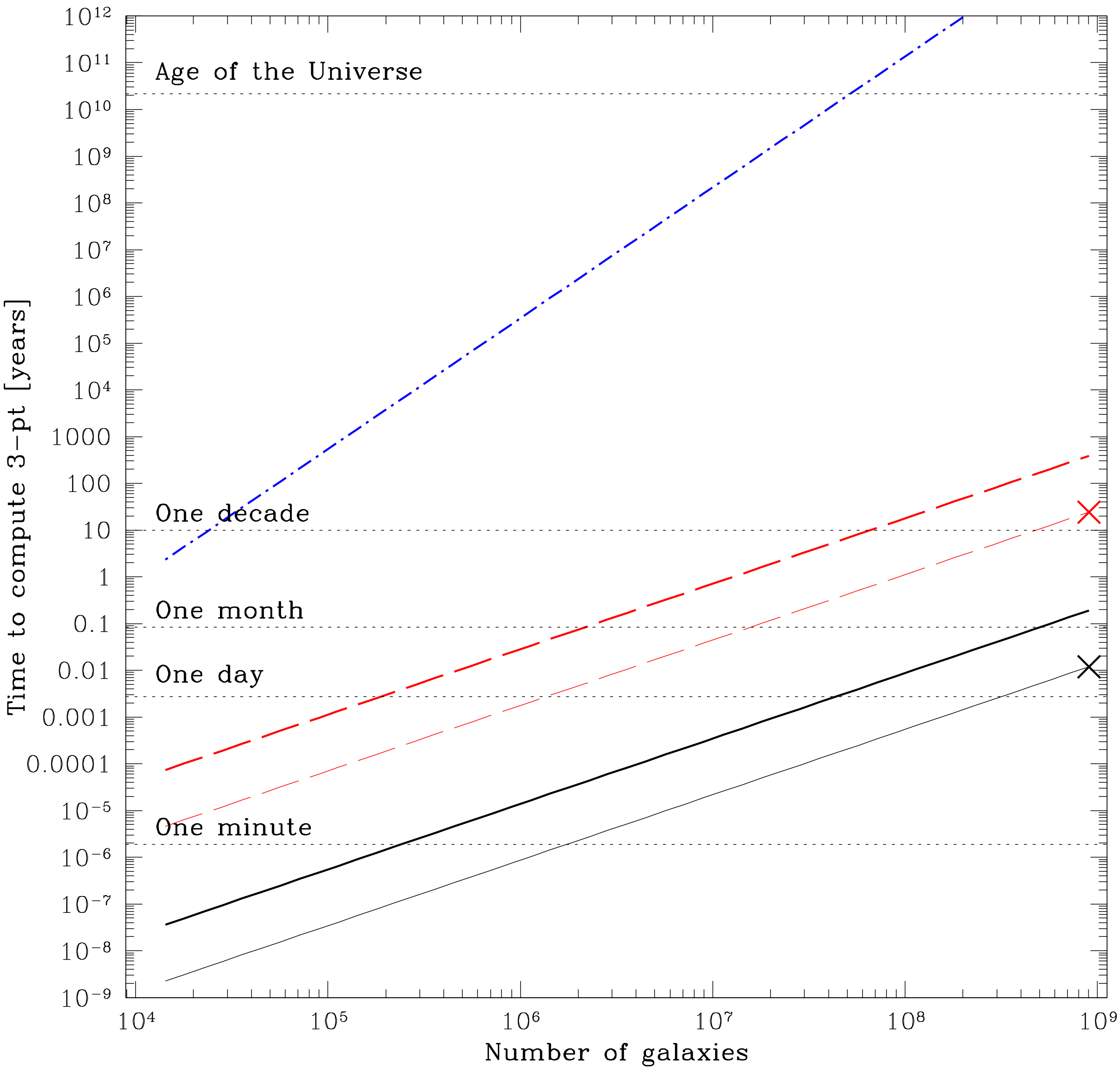


Figure 3.1 Wall-clock time to compute 3PCF as a function of the number of galaxies in dataset. The top blue dot-dashed line shows the naive algorithm of comparing every particle, which scales as $O(N^{2.8})$. The red long-dashed line shows calculation time using an efficient tree-based algorithm, such as *N*tropy's, that scales as $O(N^{1.4})$ (with the worst case being $\sim Q(N^{1.8})$). The black solid line shows the same efficient calculation using 2048 processors. Bold versions of the lines indicate current processing speeds, with non-bold lines denoting expected hardware improvements for the year 2012 (assuming a doubling in CPU capability every 18 months). The symbols on the right indicate the estimated time required to process 1 billion galaxies, the expected data size from LSST. This figure is reproduced from Gardner et al. (2007), which includes additional details of these quoted scalings.

3.2 *N*TROPY FRAMEWORK

The goal in developing a new programming framework is to enable research scientists to rapidly create analysis codes that utilized efficient and reliable methods with high parallel scalability. The hope is that our solution, *N*tropy, makes this possible even with little to no knowledge of parallel processing on the part of the researcher developing the code. The novel part of this work was assembling several existing parallelization strategies into one general framework to focus on requirements for a variety of astrophysical applications.

The parallel framework underlying *N*tropy originated in the development of a highly scalable gravitational N-body code, `PKDGRAV` (Stadel, 2001), which runs efficiently on many supercomputing platforms. This implementation of parallel kd-tree data structures is the result of over a decade of development, production use and optimization. We utilize this investment of time in code development and using it as a base to create a more general framework for similar yet subtly different analyses. Efficiently parallelizing tree-based data structures is a well studied topic – our solutions are not unique. The difficulty that we address is to extend and adapt solutions that were developed and tested for small numbers of processors (8-32) such that they perform well when scaled to thousands of processors.

*N*tropy employs several data management techniques to make this scalability possible, such as caching of interprocessor data transfers, intelligent partitioning of the high-level tree nodes, and dynamic workload management (Gardner et al., 2007). We include the capability to perform parallel data read-in with a simplified interface. Writing and debugging these capabilities from scratch is difficult and time consuming. However, a developer using *N*tropy gets all of them "for free". In addition, the design of *N*tropy does not require a parallel product – the efficiency of some *N*tropy applications have shown as good or better performance than competing serial applications (Gardner et al., 2007).

3.3 *N*TROPY: *N*-POINT APPLICATION

The flagship application of the *N*tropy framework was chosen to be the n-point calculator. If we can solve this problem efficiently (see Figure 3.1), other applications could be easily managed.

The n-point application employs an exact counting scheme, developed using *marked kd-trees*. The algorithm was developed by Moore et al. (2001) and Gray et al. (2004), but our application remains a completely independent implementation using a parallel approach within the *N*tropy framework. Their algorithm is implemented in a publicly available serial version, `npt`, which has been used in many astrophysical studies including the redshift space 3PCF (Nichol et al., 2006; Kulkarni et al., 2007).

As part of our implementation, we add the capability to do *projected* calculations in addition to the *spatial* n-point calculations. We provide the ability to search for data configurations parameterized in projected separation (r_p) and line-of-sight distance (π). This enables accurate calculations of *projected* correlation functions – an important complement to the redshift based analyses available from spatial searches, which we described in Chapter 1.

3.3.1 Verification

The validity of our n-point calculator is of paramount importance. We carefully scrutinized results by separately implementing naive calculators for spatial and projected counts in both the 2PCF and 3PCF and checking for identical pair and triplet counts. In addition to internal independent confirmation, we verified consistency with external codes used by different research groups. For the spatial 3PCF, we reproduced exact results as those by `npt`, used for studies of the redshift space 3PCF (Nichol et al., 2006). We checked the projected 2PCF by comparing to measurements generated by Zehavi et al. (2005). As an example, we show a comparison of the r_p-π diagram for a subset of SDSS galaxies between our `ntropy-npoint` and the code used by Zehavi et al. (2005) in Figure 3.2.

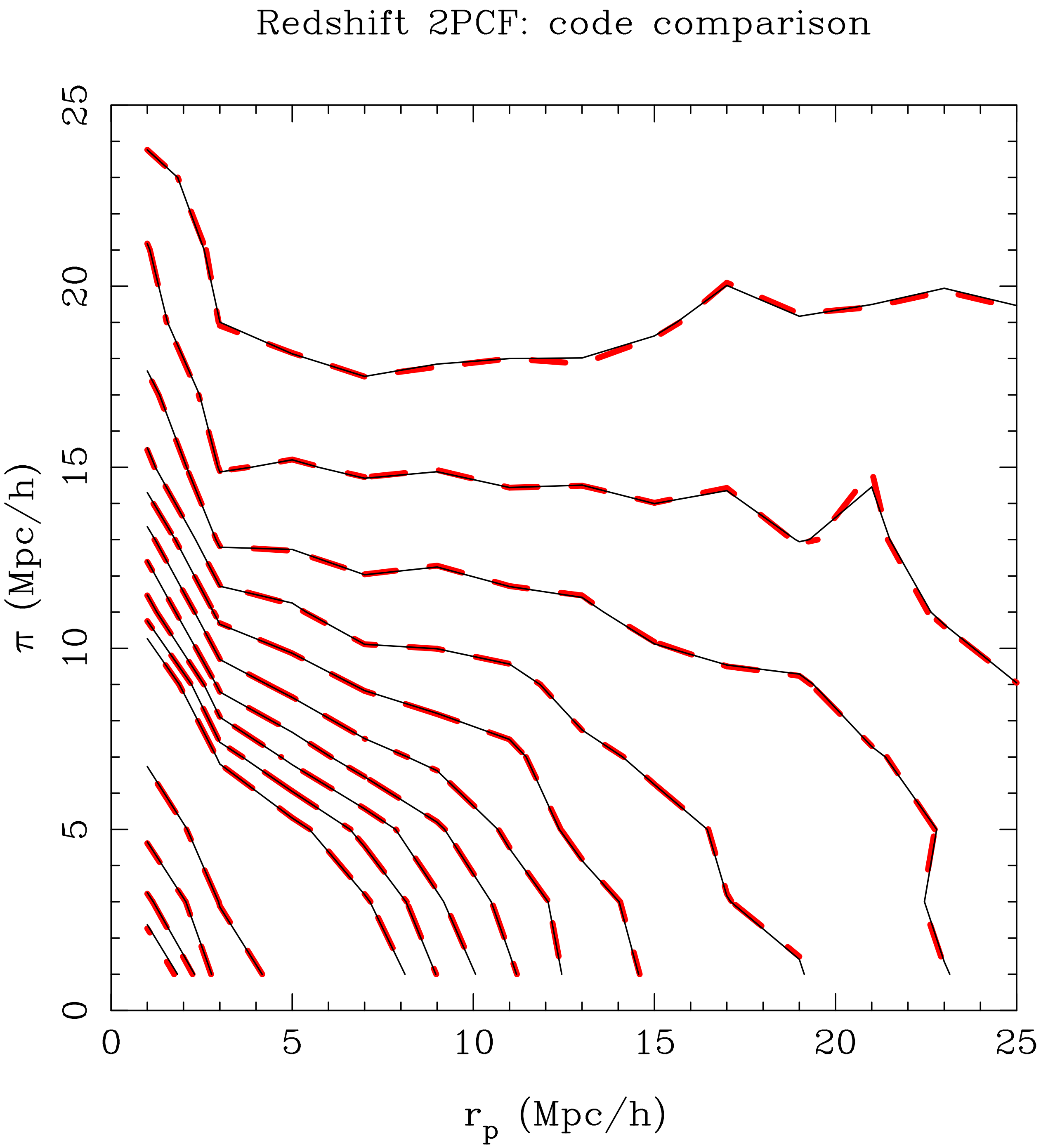


Figure 3.2 The r_p-π plot of the two-point correlation function (2PCF) for a subset of SDSS data. The thin solid black contours show a calculation using *N*tropy. The thick dashed red contours are results based the code used by Zehavi et al. (2005). No significant differences were found.

3.3.2 Performance

The analysis we present in this thesis on the 3PCF in the SDSS utilized the n-point application on massively parallel machines. We utilized hundreds of processors at a time and required over *3 decades* of CPU years to complete the n-point calculations (over $300,000$ CPU hours). This would not have been possible without an efficient parallel code available through the *N*tropy framework.

Many of the computational features developed for *N*tropy were initiated based on performance metrics of calculating the n-point functions on observational data. As an example of the complexity, we discuss the need for dynamic load balancing, i.e. for the code itself to divide the workload among processors in an adaptive manner as it runs. This turned out to be surprisingly critical piece to achieve parallel scalability. *N*tropy originated from an N-body code, which "guessed" at the best division between CPUs based on spatial positions and was simplified since the typical input was a periodic cube of a set size. Since much of the computation was local for gravity calculation, this worked very well – the kd-tree was well balanced and neighbor searches were highly efficient. With the irregular geometry of observations (see the SDSS footprint in Figure 2.1), the kd-tree became imbalanced and the division of labor for multiple processors could no longer be accurately pre-calculated. We show the dramatic loss of parallel scalability for a modest number of processors in the 3PCF in Figure 3.3, due primarily to irregular geometry and inherit clustering. We note the mock data, even though clustered like observational galaxies, scales well due to being distributed in regular box geometry.

The inefficiency of poor workload management was exacerbated by the larger computational burden of a three point calculation compared with a gravity calculation for a given data size. Naively, triplet counting scales as $O(N^3)$ and gravity determination as $O(N^2)$, which become $O(N^{1.8})$ and $O(N \log N)$, respectively, for tree-based methods. With the implementation of dynamic load balancing and increased data overlap with the interprocessor cache, the 3PCF calculation shows excellent scalability to *thousands* of processors (see Figure 3.4).

In addition to parallel performance, we found our implementation of the spatial n-point calculator ran 6 to 30 times faster in serial than a widely used alternative implementation

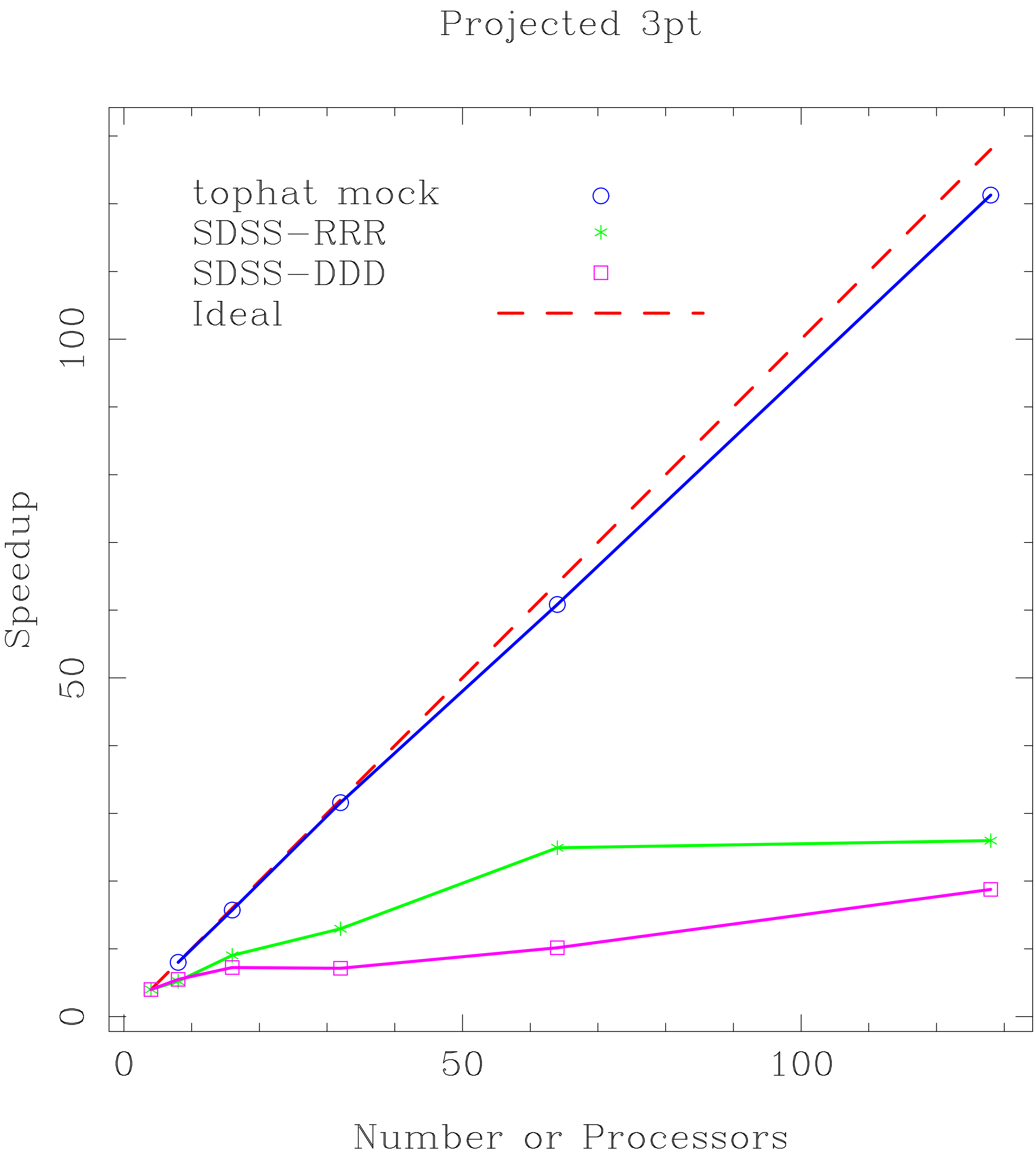


Figure 3.3 We demonstrate the effect of data geometry on an early version of the *N*tropy *n*-point calculator, before dynamic load balancing was implemented. The *speedup* is defined as the factor of improvement in runtime, as a function of using N additional processors t_N/t_1. The red dashed line shows ideal scaling where each CPU divides up the work perfectly. The blue circles denote a 3PCF calculation using a mock catalog in a periodic box, the geometry typically used in cosmological N-body simulations. The green stars (SDSS-RRR) denote smooth random catalogs distributed in the SDSS footprint and show a rapid decrease in efficiency since *N*tropy cannot correctly guess how to divide the data with the irregular geometry. Finally, the magenta squares (SDSS-DDD) show the same SDSS geometry but using the clumped distribution of observational galaxies which further unbalance the kd-tree and degrade parallel performance.

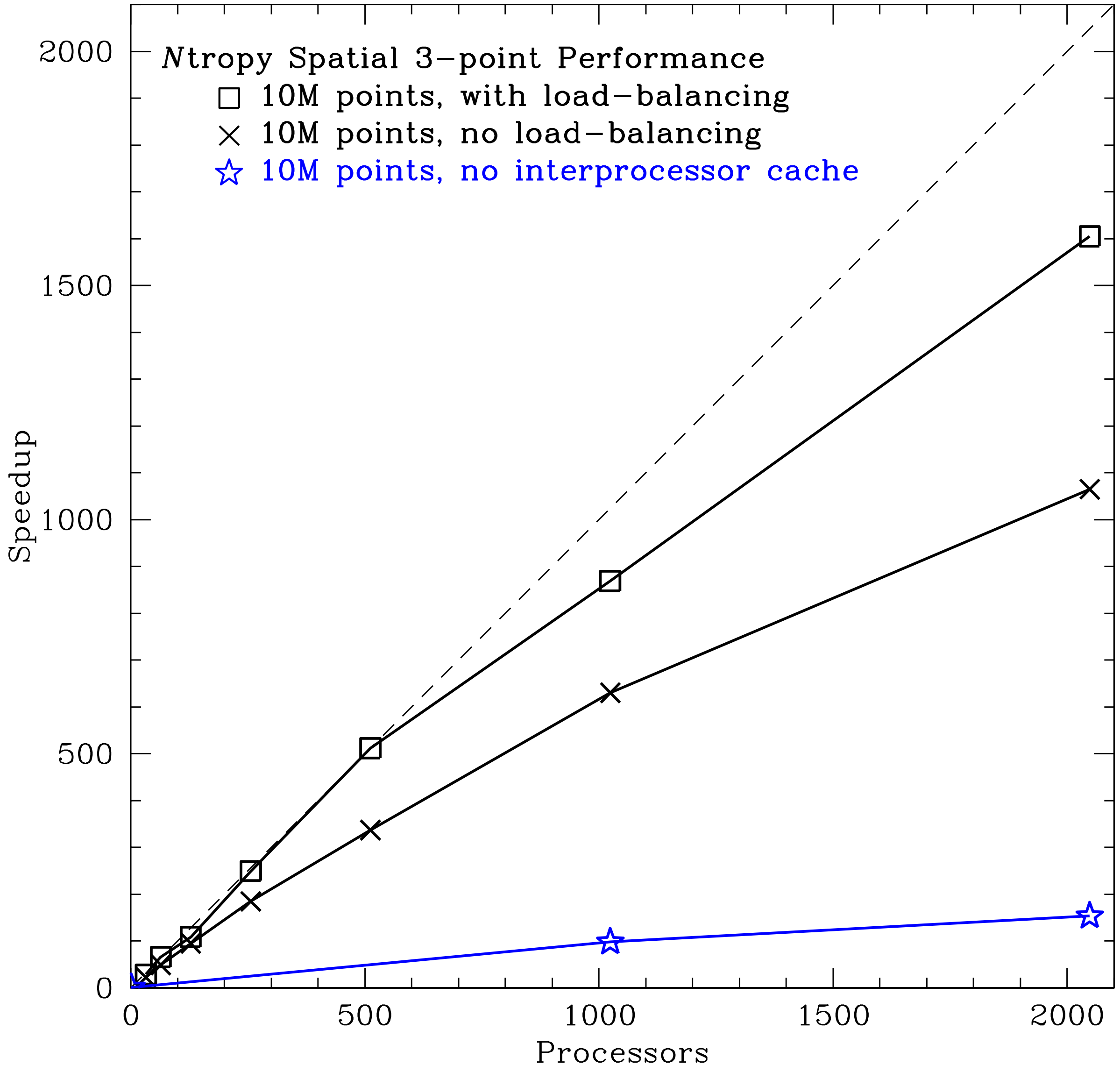


Figure 3.4 The effects on scaling of interprocessor data caching and dynamic load balancing. The open squares show scaling for data caching and dynamic load balancing, while crosses demonstrate the effects of turning off load balancing. The open stars illustrate the further consequences of disabling the interprocessor data cache. The dashed line shows ideal scaling where each CPU divides up the work perfectly. This scaling test is for a single spatial 3-point calculation on a fixed problem size of 10 million particles randomly distributed in the Sloan Digital Sky Survey volume. It was performed on the PSC Cray XT3. The results shown are reported by Gardner et al. (2007).

npt (Moore et al., 2001; Gray et al., 2004; Nichol et al., 2006) (both are based on kd-tree based searches). This highlights the impressive capability of the *N*tropy framework to attain maximal efficiency in serial as well as parallel scalability with a *single* implementation.

More complete details of *N*tropy and its performance are given in Gardner et al. (2007).

3.4 *N*TROPY: FRIENDS-OF-FRIENDS GROUPFINDER

We use the *N*tropy framework to implement a parallel group finder using the friends-of-friends (FoF) algorithm (Davis et al., 1985). First, groups of particles are constructed using efficient tree searches, rapidly finding nearest neighbors to link together into associated groups. Afterward, cross processor groups are reconnected using an iterative graph-based procedure (Shiloach and Vishkin, 1982). This implementation was greatly sped up using the enhanced caching system developed in *N*tropy. Its parallel performance showed marked improvements over other implementations (such as the FoF algorithm in non-public parallel `subfind`, see Springel et al. (2001)). The accuracy of this implementation was verified by detailed comparisons with the University of Washington FoF code (UW FoF) which is publicly available at `http://www-hpcc.astro.washington.edu/tools/fof.html`.

In parallel programs, especially on large massively parallel platforms, care must be taken to properly handle input-output (IO) of data. It is known that IO can be a major bottleneck for these computing clusters. We designed a binary group catalog (BGC) format for FoF output that collected the particles per group in a sensible format to be easily processed by serial codes offsite (e.g. for mock making). This output format uses all running processes to stream parallel output, and quickly implemented in parallel using the *N*tropy framework. For example, using 1000 processors a single FoF calculation (input data was 1000^3 particles) took less than 10 minutes when using the BGC output. The same calculation on the same computing cluster took almost an hour (over 50 minutes) when using an ascii output format equivalent to that of the UW FoF code – a factor of 5 difference!

4.0 GALAXY MEASUREMENTS IN THE SDSS

In this chapter, we present clustering measurements of the SDSS galaxy samples detailed in Chapter 2 (see Table 2.1). We discuss important details of the measurements in §4.1. We introduce measurements of the two-point correlation function (2PCF) in §4.2 to confirm clustering properties seen in other studies. We continue with §4.3, focusing on the shape or *configuration* dependence of the reduced three-point correlation function (3PCF). Between our three samples, we investigate the dependence on luminosity and color in both redshift and projected space. We present full covariance matrices of the measurements in §4.4 an integral part of using these measurements in quantitative analyses. Finally, we address how super structures, such as the Sloan Great Wall (SGW; Gott et al., 2005), affect these measurements in §4.5. We summarize the main results of the chapter in §4.6.

4.1 INTRODUCTION

We reviewed the basics of correlation functions in the introduction (Chapter 1), which includes the general description, relating predictions from gravitational collapse to galaxy distributions, as well as the means of estimating the correlation functions from observational galaxy samples. We use unbiased and minimal variance estimators (see Szapudi and Szalay, 1998) to correct for edge effects with random catalogs used to characterize the volume. We use random catalogs that have a density a factor of 5-10 greater than the data. We find this sufficient to keep the shot noise contribution of randoms below that of the data for all triplet counts, as well as small enough to be computational manageable.

The two-point correlation function (2PCF) is characterized by a separation, generally denoted as r. When we refer to redshift space measurements, we note this separation as s. Redshift space distances can be decomposed into the line-of-sight (π) and projected separation (r_p). For projected space measurements, we integrate along the line-of-sight (π direction) and keep the 2PCF characterized by a single dependent variable: r_p. We chose to integrate along the line-of-sight to $\pi_{max} = 20\ h^{-1}$Mpc. For more details, refer to the discussion in Chapter 1 and justification in Chapter 6. The 3PCF is a function with three dependent variables and we continue the same notation: s or r_p for redshift and projected space separations respectively, and r can generically refer to real, redshift or projected space separations.

4.2 TWO POINT CORRELATION FUNCTION

We first investigate the 2PCF of our samples. While we later focus our clustering measurements on the reduced 3PCF, $Q(r_1, r_2, \theta)$, it is instructive to look directly at the 2PCF since Q is the ratio of the connected 3PCF, $\zeta(r_1, r_2, \theta)$ divided by products of the 2PCF, $\xi(r)$, such that $Q \propto \zeta/\xi^2$. For the 2PCF, we chose equal bins in $\log r$.

For comparison with the observed 2PCF, we have included a *fiducial* power law model:

$$\xi(r) = \left(\frac{r}{r_o}\right)^{-\gamma} \ ; \ \xi_{fid} = \left(\frac{r}{5h^{-1}\mathrm{Mpc}}\right)^{-1.8} . \tag{4.1}$$

Given the precision of modern measurements, recent work has noted significant departures of the galaxy 2PCF from a power law (Zehavi et al., 2004) where the data is better described by the halo model (see Cooray and Sheth, 2002, for a review). Nevertheless, a power law provides a simple and convenient comparison which we use for illustration (to see the difference between power law and halo model, look at theoretical predictions in Figure 1.3). We plot the redshift space 2PCF, $\xi(s)$, in Figure 4.1 for our three galaxy samples. At larger scales, $\xi(s)$ runs almost parallel to the *fiducial* model for all galaxy samples. The galaxy 2PCF is notably higher than the *fiducial* model, especially for brighter galaxy samples. We interpret this as brighter galaxies having a larger *bias*, in general agreement with similar studies on SDSS samples (Zehavi et al., 2005). We discuss this concept in more detail in the next chapter (galaxy-mass bias, Chapter 5). We note the reduction in strength of $\xi(s)$ at separations below $\approx 5h^{-1}$Mpc in Figure 4.1. This is a result of non-linear collapse where the peculiar velocity cannot be disentangled from the radial distance obtained from the galaxy redshift. This redshift space distortion is significant, and in large dark matter halos such as galaxy clusters can be tens of h^{-1}Mpc in length. These so called *fingers-of-god* make galaxy pairs at close separations less likely and reduce the small scale correlation function.

We plot the projected two-point correlation function, $w_p(r_p)$, for our three galaxy samples in Figure 4.2. In contrast to $\xi(s)$, the $w_p(r_p)$ has no turn-over at small projected separations, and remains close to a power law for small projected separations. We expect the $w_p(r_p)$ to be less affected by redshift distortions, especially at the smaller projected scales – visibly confirmed in Figure 4.2. At large r_p, there is a reduction of power resulting from truncating the line-of-sight integration at a $\pi_{max} = 20\ h^{-1}$Mpc. Assuming a perfect power law in real space, i.e. the *fiducial model*, we show the resulting projection for $w_p(r_p)$ in Figure 4.2. The dotted line uses the same π_{max} as the data. The solid line is the full projection ($\pi_{max} = \infty$), which yields the expected property that projecting a power law results in another power law. We note that brighter galaxies exhibit stronger clustering when compared to fainter galaxies, which was also apparent in the redshift space $\xi(s)$ measurements. These observations agree with detailed measurements on SDSS galaxies (Zehavi et al., 2005).

4.3 THREE POINT CORRELATION FUNCTION

4.3.1 Equilateral 3PCF

The simplest analog to the 2PCF is the equilateral 3PCF (Q_{eq}), a slight reformulation of the reduced 3PCF we introduced in Chapter 1, specifically (1.8). As each side of the triangle formed by the triplet corresponds to the same scale, Q_{eq} can easily be characterized by a single separation, such that

$$Q_{eq}(r) = \frac{\zeta(r)}{3\,\xi(r)^2}\ . \tag{4.2}$$

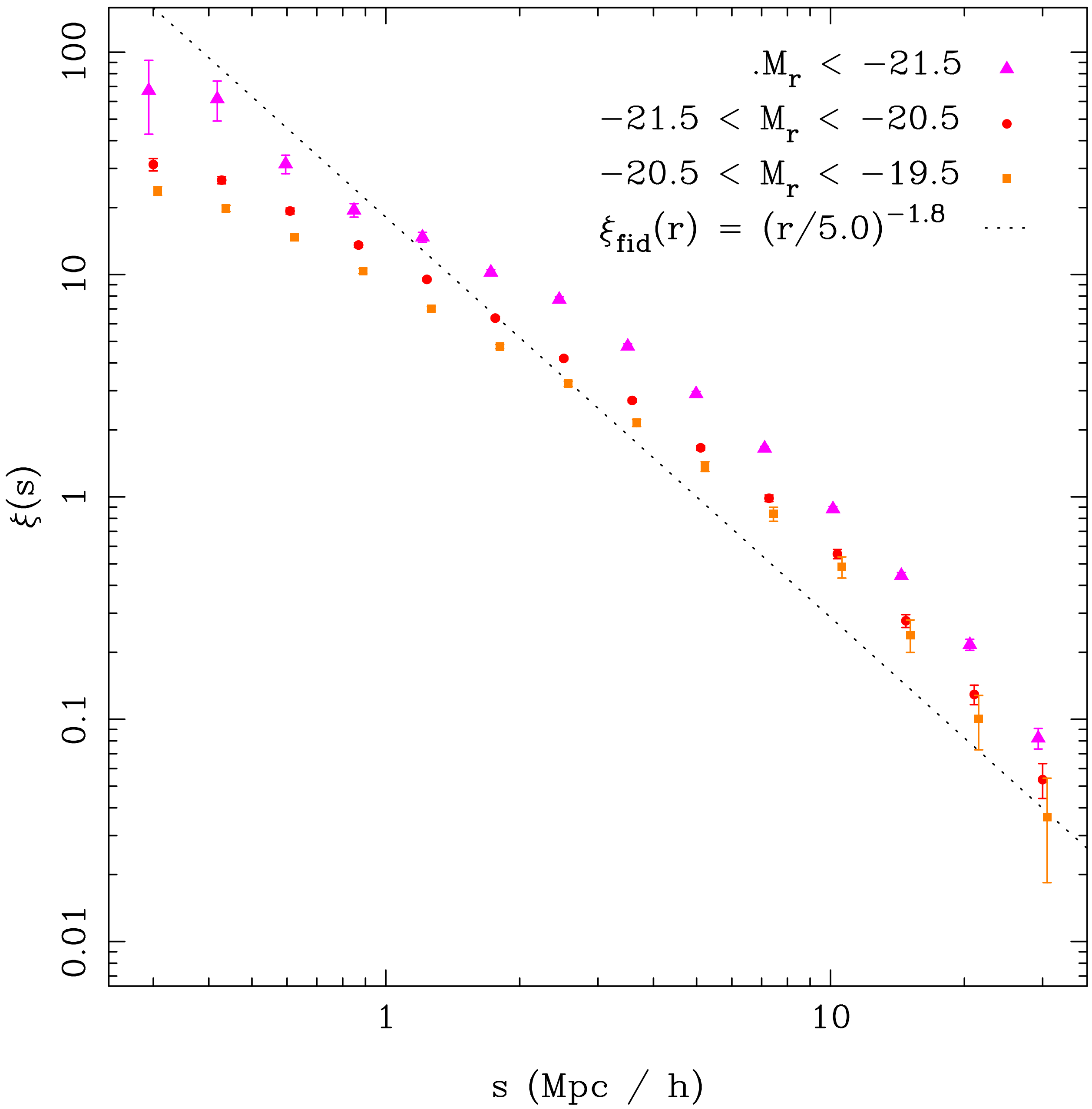


Figure 4.1 The redshift space 2PCF, $\xi(s)$, for the three SDSS samples. The small scales do not continue the power law shape because redshift distortions reduce the clustering power at small separations.

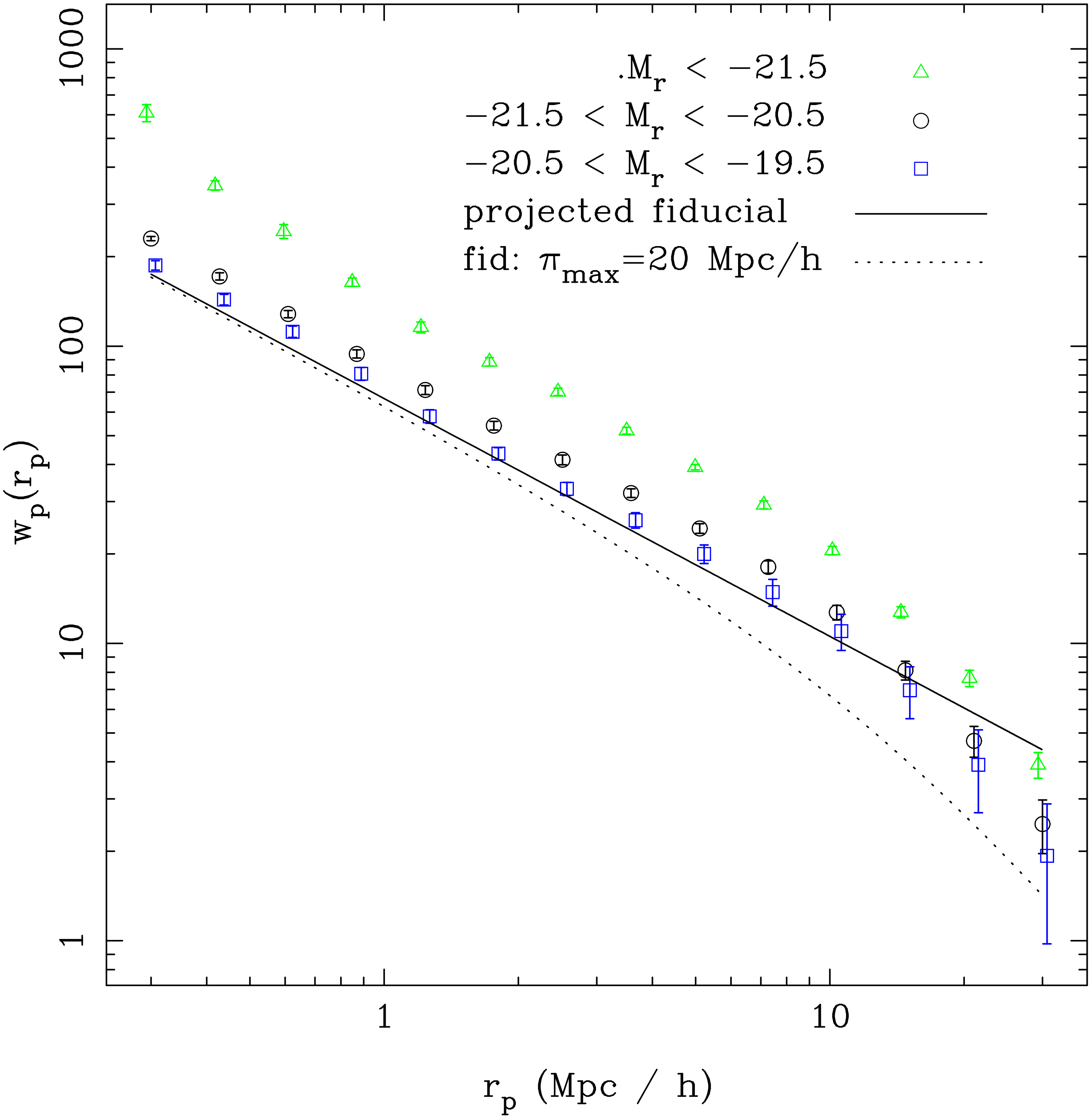


Figure 4.2 The projected 2PCF, $w_p(r_p)$, for the three SDSS samples. Brighter galaxy samples exhibit more clustering (e.g. higher amplitude) than fainter samples, in agreement with previous analyses of SDSS galaxies (Zehavi et al., 2005).

Q_{eq} is not sensitive to shape, but it provides a clear sense of scale dependence in the 3PCF. When $Q_{eq} \approx 1$, the number of triplets exactly corresponds to those expected from the 2PCF; when Q_{eq} is above or below 1, there are more or less triplets respectively. Like the 2PCF, we chose equal width bins in $\log r$.

We show Q_{eq} in both redshift and projected space for all three of our galaxy samples in Figure 4.3. In redshift space, Q_{eq} appears flat and never exceeds 1, showing very little difference in terms of luminosity. In the projected measurement, small scales reveal $Q_{eq} > 1$ until $r_p \approx 2 - 3\ h^{-1}$Mpc where it approximately reproduces the amplitude of the redshift space measurement. We expect larger scales to dip below one, as the number of triplets, or more specifically $\zeta(r)$, drops more rapidly than pairs in the 2PCF for the same scale. Intuitively, the number of equilateral triplets of a set side length quickly become more rare than the number of available pairs of the same separation length.

On small scales, redshift distortions destroy small scale triplets, which are recovered by the projected Q_{eq} in Figure 4.3. This interpretation is supported by a similar trend seen in a theoretical comparison of the real and redshift space Q_{eq} in Marín et al. (2008) (see their Figure 1). We notice a slight dependence on sample luminosity, as we saw in the 2PCF, but not nearly as pronounced (we discuss this point below in further detail). We notice that the fainter samples show increased errors at larger separations, suggesting the measurements are limited in signal by the finite volume of the galaxy sample. Since the faintest volume-limited sample also occupies the smallest volume, a deficiency of triplets will happen before the brighter sample which correspond to a larger volume. Finite volume effects have been shown to cause a rapid reduction in amplitude for higher order measurements of clustering (Scoccimarro, 2000).

4.3.2 Triplet Configurations

The full 3PCF is a function of three variables that characterize both the size and shape of triplets. A natural and unique description of a triplet is the length of each side of the triangle that connects the points: r_1, r_2 and r_3 where connectivity is assumed by $\vec{r}_3 = \vec{r}_1 + \vec{r}_2$. Please note our slight change to notation from Chapter 1, such that side lengths are numbered

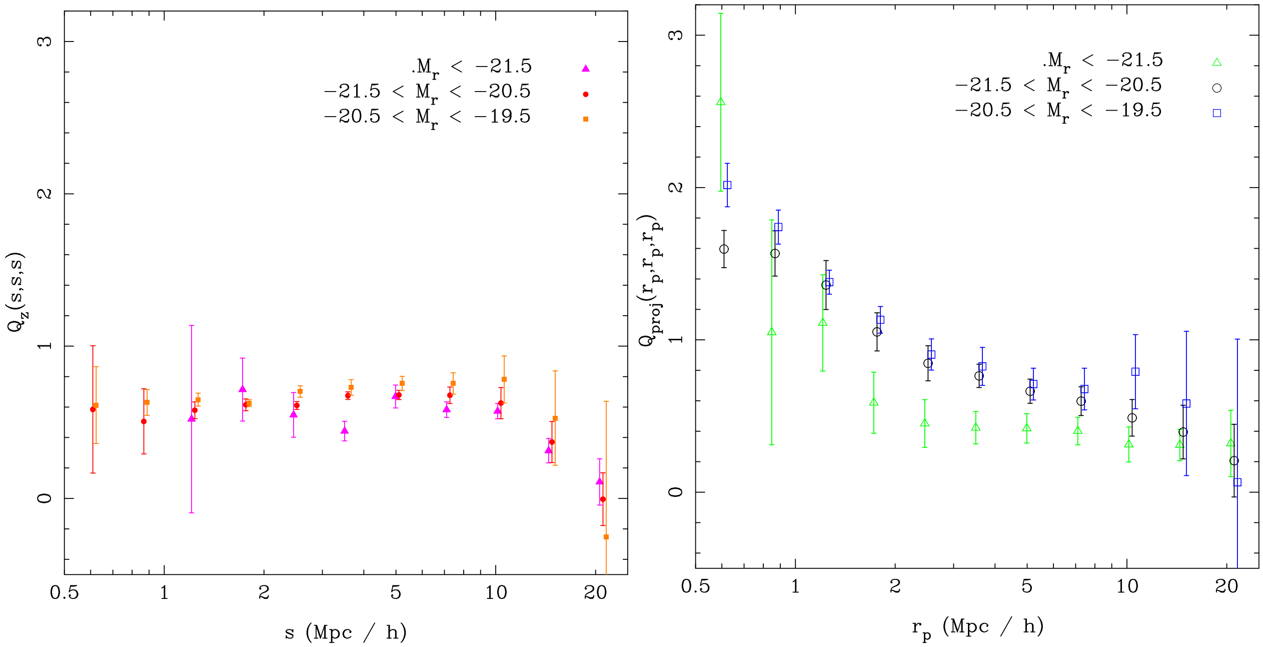


Figure 4.3 The equilateral 3PCF, $Q_{eq}(r)$, for our three SDSS galaxy samples, both in redshift space (left) and projected space (right). The equilateral 3PCF is characterized by a single scale. We chose a bin-size to be equal width in $\log r$, like we did for the 2PCF.

rather than points (e.g. $r_1 = r_{12}$, etc.).

Another parameterization employed by Peebles (1980) orders the sides such that $r_1 < r_2 < r_3$, and then defines the following:

$$s = r_1 \; ; \; u = \frac{r_2}{r_1} \; ; \; v = \frac{r_3 - r_2}{r_1} \; . \tag{4.3}$$

This helps to conceptually decouple triangle shapes and scales. The "s" is equal to the smallest side of the triangle and still retains units of length, describing the overall "scale" of the triangle. The other two variables are unit-less ratios describing shapes of triangles, where $u > 1$ and $0 < v < 1$ due to the explicit side ordering.

In perturbation theory, it is most common to see triangles described by two side lengths (r_1 and r_2) and the opening angle between them (θ) defined by the cosine rule:

$$\cos\theta = \frac{r_1^2 + r_2^2 - r_3^2}{2r_1 r_2} \; . \tag{4.4}$$

We find this last characterization the most natural and intuitive and will use it to describe our 3PCF measurements. When $\theta \approx 0$ or $\theta \approx \pi$, triangles are "collapsed" or "elongated" with two sides being very close to co-linear. As θ approaches $\pi/2$ the triplet forms a right triangle which we will refer to as a "perpendicular" configuration. The triangle shape, or "configuration" dependence, describes a function of θ where "*strong* configuration dependence" means a significant amplitude difference in the reduced 3PCF (Q) between collapsed and perpendicular configurations and "*weak* configuration dependence" shows little or no change of Q with θ. This terminology is identical to that used in Gaztañaga and Scoccimarro (2005).

While it might appear this description is both trivial and pedantic, we caution that these mappings are non-linear. The transformations between descriptions are completely equivalent for points describing triangles vertices. However, they are not exact transformations for small volumes around the points, as is the case with counts in bins of finite size used to calculate the correlation functions. We note that significant discrepancies in the measurement can arise between the parameterization used for the calculation and the one that is modeled. This will be discussed in more detail in Chapter 6.

We choose r_1, r_2, and θ to parametrize our 3PCF measurements. We feel it is a fair balance between a description decoupling the shape and size of triangles yet still simple enough to interpret the effects of binning. This parameterization is conceptual only, we measure the 3PCF using (r_1, r_2, r_3), converting to θ using (4.4). Even though we treat the reduced 3PCF as a function of three variables, $Q(r_1, r_2, \theta)$, we may often denote it $Q(\theta)$ or even just Q for simplicity.

4.3.3 Configuration Dependence in the 3PCF

We focus our investigation on the shape or *configuration* dependence of the reduced 3PCF, $Q(r_1, r_2, \theta)$, for SDSS DR6 galaxies at three scales, which we specify by the smallest side of the triangle (r_1). We chose r_2 such that the ratio of the first two sides stays fixed at $r_2/r_1 = 2$. We measure the opening angle between the sides, θ, regularly spaced between 0 and π. The resulting scale we probe varies from r_1, when $\theta = 0$ (i.e. "collapsed") to $(r_1 + r_2)$, when $\theta = \pi$ (i.e. "elongated"). Specifically we measure triplets on scales between $3 - 9$, $6 - 18$ and $9 - 27$ h^{-1}Mpc corresponding respectively to $r_1 = 3, 6$ and 9 h^{-1}Mpc.

The choice of binning is not as straight forward for the configuration dependence of the

3PCF. Since the bins are tightly packed to measure $Q(\theta)$, a choice of bin-width based on $\log r$ will not suffice as it is far too large. On the other side, a small bin-size will quickly become under-sampled, as triplets are characterized by three variables. We discuss these problems in detail in Chapter 6. We choose linearly spaced bins in θ, which define the midpoint r_3 as per the cosine rule in (4.4). We chose a bin-width as a fraction, f, of the measured scale, r, such that $\Delta_r = fr$ and a bin at r is measured between $(r - \frac{\Delta_r}{2}, r + \frac{\Delta_r}{2})$. This binning scheme was found to be useful in an independent study of the 3PCF by Marín et al. (2008). We compare measurements between three volume-limited galaxy samples which have very different number densities (see Table 2.1), and chose a single binning scheme to apply to all samples for a fair comparison. We require the smallest scale measurements ($r_1 = 3\ h^{-1}$Mpc) on the lowest density sample ($M_r < -21.5$) to be reasonably sampled, which set our fractional bin-width at 25% ($f = 0.25$). This choice produces relatively wide bins that can physically overlap, resulting in two observable consequences: (1) a slight smoothing of the 3PCF near $\theta \approx \pi$, and (2) an induced correlation between measured points. Qualitatively, this should not pose a problem since all comparisons use the same scheme (including N-body results). Quantitatively, we account for this in the covariance matrix.

We introduce our measurement of the reduced 3PCF (Q) on the brightest galaxy sample, i.e. $M_r < -21.5$, in Figure 4.4. The data points are measurements on the entire sample, and the uncertainties shown are the 1σ bounds from the variance in 30 jackknife samples. (We describe estimating the uncertainties in more detail in our discussion of the covariance matrix below.) The black line is a measurement using dark matter particles from the Hubble Volume (HV) simulation. Using N-body simulations enables reliable predictions well into the non-linear regime where accurate analytic models do not exist. In addition, we can include observational systematics by trimming HV particles to match the exact selection and volume of the observed galaxy sample. The HV measurement serves as a comparison between clustering of the observed galaxies and that expected from gravitational evolution of a ΛCDM mass field. Assuming the same cosmology, the difference between clustering of galaxies and the underlying mass commonly referred to as *galaxy-mass bias* which depends on galaxy properties. Remember, the nature of $Q(\theta)$ makes it insensitive to cosmology (discussed in Chapter 1).

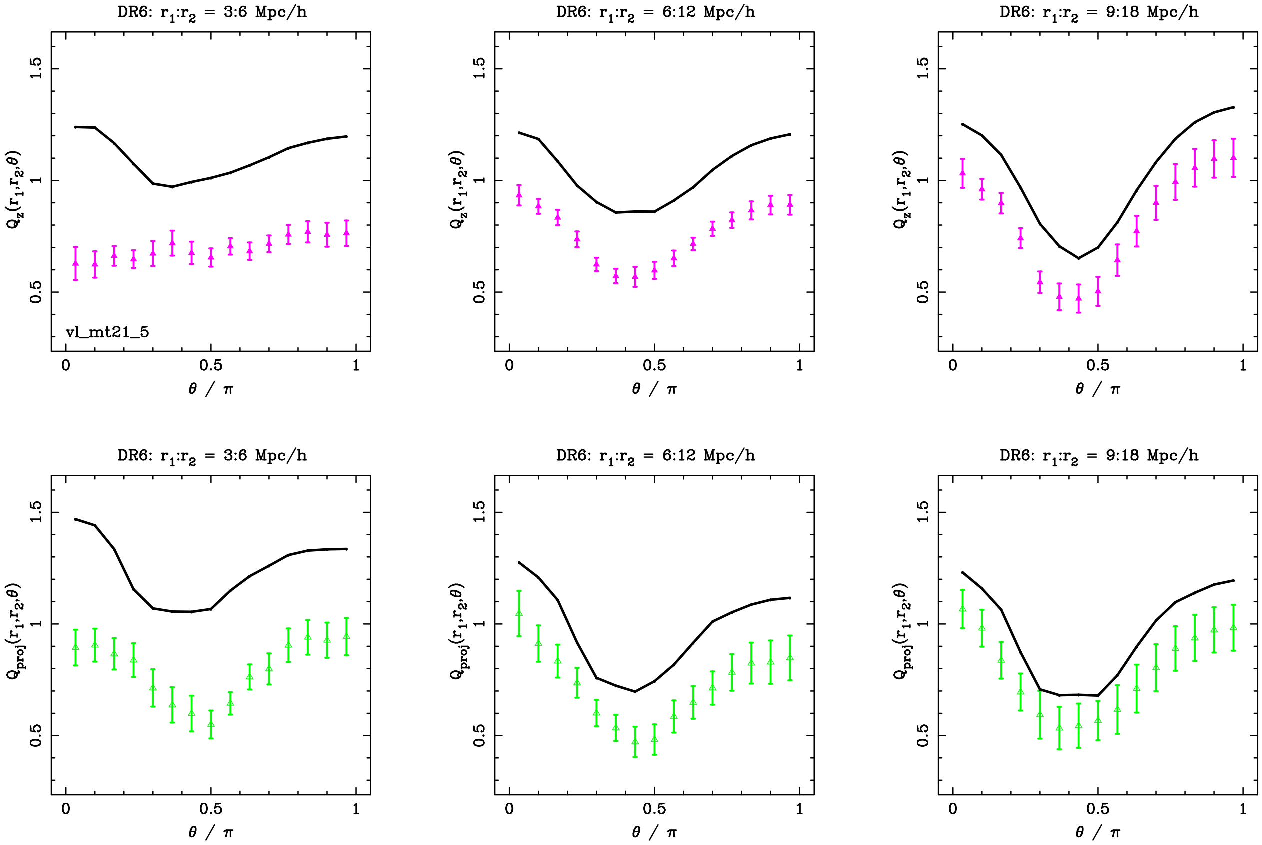


Figure 4.4 We show the configuration dependence of the reduced 3PCF of SDSS DR6 galaxies with $M_r < -21.5$. The top row consists of redshift space measurements, and the bottom row depicts projected measurements. The three columns correspond to different scales specified by the first side of the triangle (r_1) which represents the smallest scale measured. Uncertainties on the data measurements are 1σ dispersion from 30 jackknife samples. The black line denotes measurements on dark matter in the Hubble Volume simulation, matching the selection of the galaxies and includes redshift distortions.

There are a few things to note about Figure 4.4. It is clear that galaxies do not cluster exactly the same as the simulated mass field, depicted as the vertical offset between data points and the solid black line. However, there are significant commonalities. The overall shape is the same, and the predominant effect is an offset. There is a large volume of work showing that galaxies are known to be biased tracers of the mass field, which offers a natural explanation of the offset (Cooray and Sheth, 2002). We quantify this difference in the next chapter.

We notice a "V-shape" at $r_1 = 9\ h^{-1}$Mpc, present in both redshift and projected space measurements in Figure 4.4. We interpret this as a statistical signal of filamentary structure that isn't present at the smaller scales: an over-abundance ($Q > 1$) of elongated and collapsed triplets with an under-abundance ($Q < 1$) of perpendicular configurations. This characteristic V-shape is predicted from gravitational perturbation theory (see Gaztañaga and Scoccimarro, 2005, for a theoretical comparison with simulations). On smaller scales (left side of Figure 4.4) we see weaker configuration dependence in the reduced 3PCF (i.e. smaller change in Q with θ) than at larger scales (right side). We identify the effect of redshift distortions on the small scales in the redshift space reduced 3PCF in two ways: (1) almost no configuration dependence in Q_z and (2) $Q_z < 1$ for all configurations, showing a lack of triplets due to non-linear redshift distortions. The projected measurement recovers some of the configuration dependence, but still exhibits a lower Q than larger scales. Since we have chosen $\pi_{max} = 20h^{-1}$Mpc, redshift distortions can still play a role, albeit in a reduced capacity.

4.3.4 Color Dependence

We investigate the color dependence of the reduced 3PCF using two volume-limited samples. They are designed to have unit magnitude bins with limiting r-band magnitude above and below L_*. We divided the galaxies into "red" and "blue" sub-samples to probe the color dependence as described in Chapter 2.

The 3PCF for our LSTAR sample ($-21.5 < M_r < -20.5$) is shown in Figure 4.5. This sample is fainter than the one presented in Figure 4.4, and has approximately three times as many galaxies. In the 2PCF, we noticed that fainter galaxy samples exhibit lower power in clustering as in figures 4.1 and 4.2. Since the reduced 3PCF is normalized by the 2PCF, this is not as evident in Q. Returning to the reduced 3PCF of Figure 4.5, the volume-limited LSTAR sample has a luminosity closer to L_* (hence the name) and therefore has a higher density of galaxies (factor of 10 greater than the BRIGHT sample). We see trends of the reduced 3PCF that we noted in the brighter sample, but the statistical strength of this larger sample has increased the significance of the smaller scale measurements which appear to have less shot-noise. Although the difference is small at large scales, it appears that the blue sub-sample might show a stronger configuration dependence than the red sub-sample. If true, this suggests that blue galaxies more commonly populate the filamentary structures at these scales. In projected space, we see the opposite trend at smaller scales ($r_1 = 3\ h^{-1}$Mpc) where the blue sub-sample shows less configuration dependence than the red. Both Q_z and Q_{proj} show significant differences in the "red" and "blue" populations for the smaller triangle scales.

Next, we examine our faintest sample ($-20.5 < M_r < -19.5$) in Figure 4.6. We notice that the uncertainties are much larger than measurements on the brighter samples. This sample still contains over 76,000 galaxies, a factor of two greater than the brightest sample, so the increased errors are not due to shot noise. This fainter galaxy sample occupies a smaller volume, and a few large structures dominate the errors in the large scale measurements. We see configuration dependence at all scales, and the V-shape becomes more prevalent on larger scales – analogous to the behavior observed in the brighter samples. However, the measurements on red and blue galaxies in the FAINT sample show a greater difference than in the brighter LSTAR sample, especially at $r_1 = 3h^{-1}$Mpc in both the redshift and projected space reduced 3PCF.

4.3.5 Luminosity Dependence

To highlight the effect of luminosity on the 3PCF, we (re)plot the full measurements of all three samples in Figure 4.7. As we saw with the 2PCF, there is a clustering dependence on luminosity. However, the trend is seemingly reversed – brighter galaxies show *lower* values

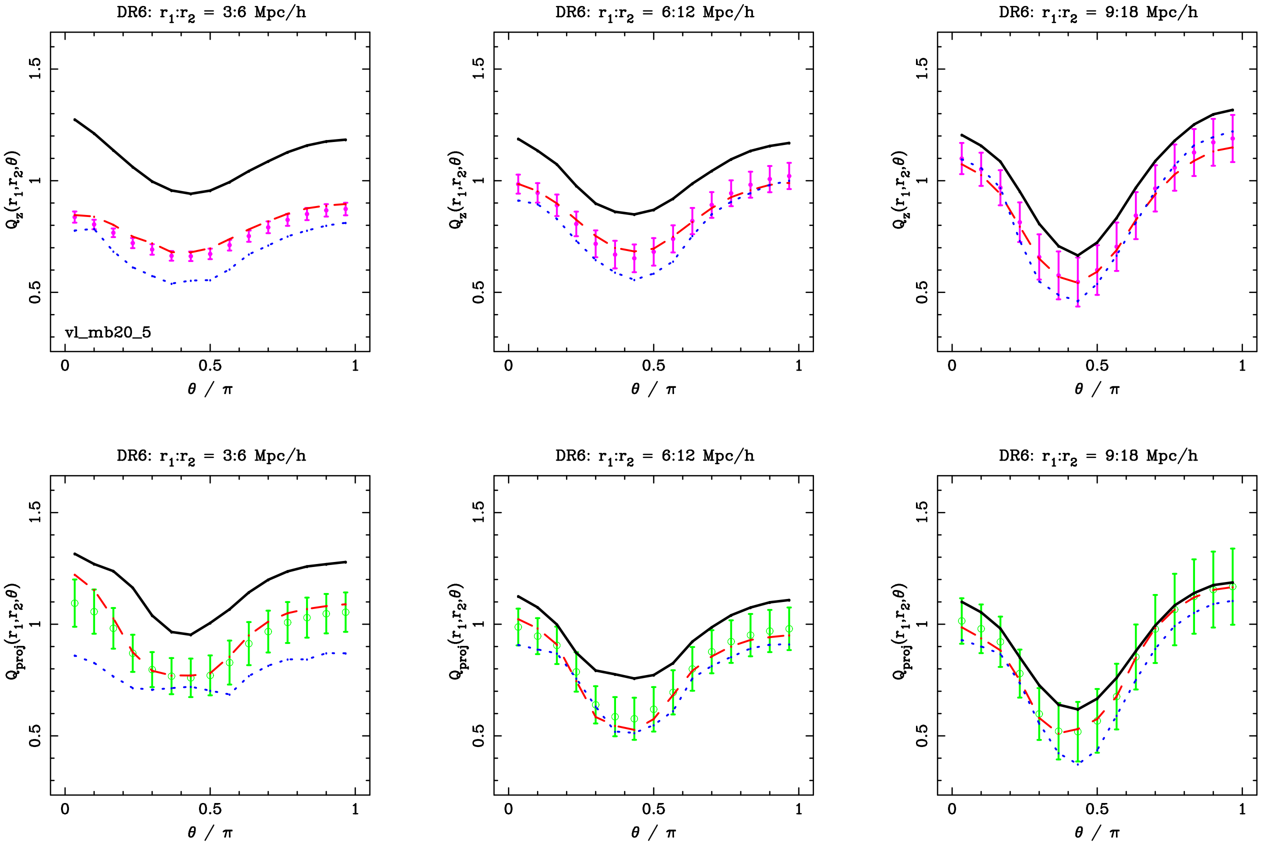


Figure 4.5 We show the configuration dependence of the reduced 3PCF, as in Figure 4.4, but for a fainter galaxy sample defined by $-21.5 < M_r < -20.5$. We divided this sample into "red" (red, dashed line) and "blue" (blue, dotted line) sub-samples based on SDSS $g - r$ color (see sample description for details). The top row consists of redshift space measurements, and the bottom row depicts projected measurements. The three columns correspond to different scales specified by the first side of the triangle (r_1) representing the smallest scale measured. The black line denotes measurements on dark matter in the Hubble Volume simulation, matching the selection of the galaxies and includes redshift distortions.

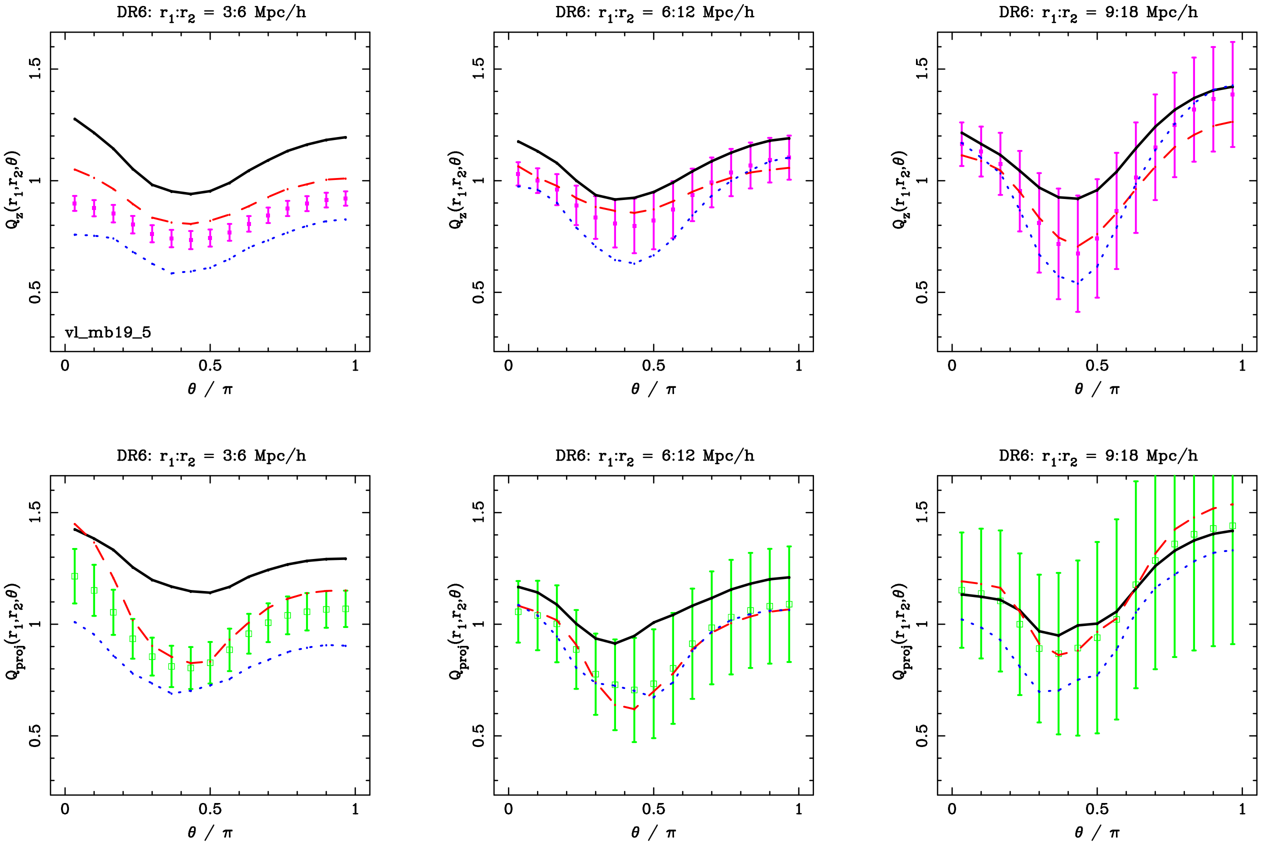


Figure 4.6 We show the configuration dependence of the reduced 3PCF, as in figures 4.4 and 4.5, for our faintest galaxy sample with $-20.5 < M_r < -19.5$. We divided this sample into "red" (red, dashed line) and "blue" (blue, dotted line) sub-samples based on SDSS $g-r$ color (see sample description for details). The top row consists of redshift space measurements, and the bottom row depicts projected measurements. The three columns correspond to different scales specified by the first side of the triangle (r_1) representing the smallest scale measured. The black line denotes measurements on dark matter in the Hubble Volume simulation, matching the selection of the galaxies and includes redshift distortions.

of Q than fainter galaxies. This trend is mirrored in the equilateral measurements (Q_{eq}, see Figure 4.3). Remember in the 2PCF, brighter galaxies had *higher* measured values of $\xi(s)$ and $w_p(r_p)$. We expected to see this trend with our characterization of bias (refer to the introduction in Chapter 1). If brighter galaxies exhibit stronger clustering, they will have a correspondingly higher linear bias b_1. Since $\xi \propto b_1^2$, it will show a *higher* value than the 3PCF since $Q \propto 1/b_1$. We can speculate that some of the measurements, specifically those at $r_1 = 6h^{-1}$Mpc, might suggest that fainter samples have a slightly stronger configuration dependence, but the effect is weak given the measured uncertainties. Physically, we expect bright red galaxies to predominantly live in galaxy groups or clusters, centered at the "knots" in LSS. Fainter galaxies will populate the field and filamentary structure more, thereby showing more configuration dependence in the 3PCF.

4.3.6 Redshift vs Projected Space

Let us compare directly redshift and projected space 3PCF measurements. The purpose of projecting the correlation function is to alleviate the effects of redshift distortions on clustering. We certainly expect redshift distortions to play a role on these scales, so we can ask if there are specific differences between the two measurements. Is the projection recovering additional information of the density field? Are there trade offs to doing this projection? Measurements on the 2PCF indicate the method works quite well (in this work as well as a huge number of studies in the literature, e.g. Zehavi et al. (2005)).

We overlay the redshift and projected space reduced 3PCF for each sample and scale in Figure 4.8. On small scales, we immediately see that Q_{proj} has both a higher amplitude and exhibits a stronger configuration dependence in comparison with Q_z. As we see in comparing the 2PCF (figures 4.1 and 4.2), the projected measurements are larger in amplitude. However, Q_{proj} is normalized by $w_p(r_p)$ which means we expect Q_{proj} to again be close to 1 in value. As a result, the increased amplitude of Q_{proj} is attributed to recovering triplets that were lost to Q_z from non-linear redshift distortions. We also consider the stronger configuration dependence of Q_{proj} reflective of the true galaxy distribution, as a simple projection on the sky should not induce this dependence. As the scale increases, Q_{proj} and Q_z generally

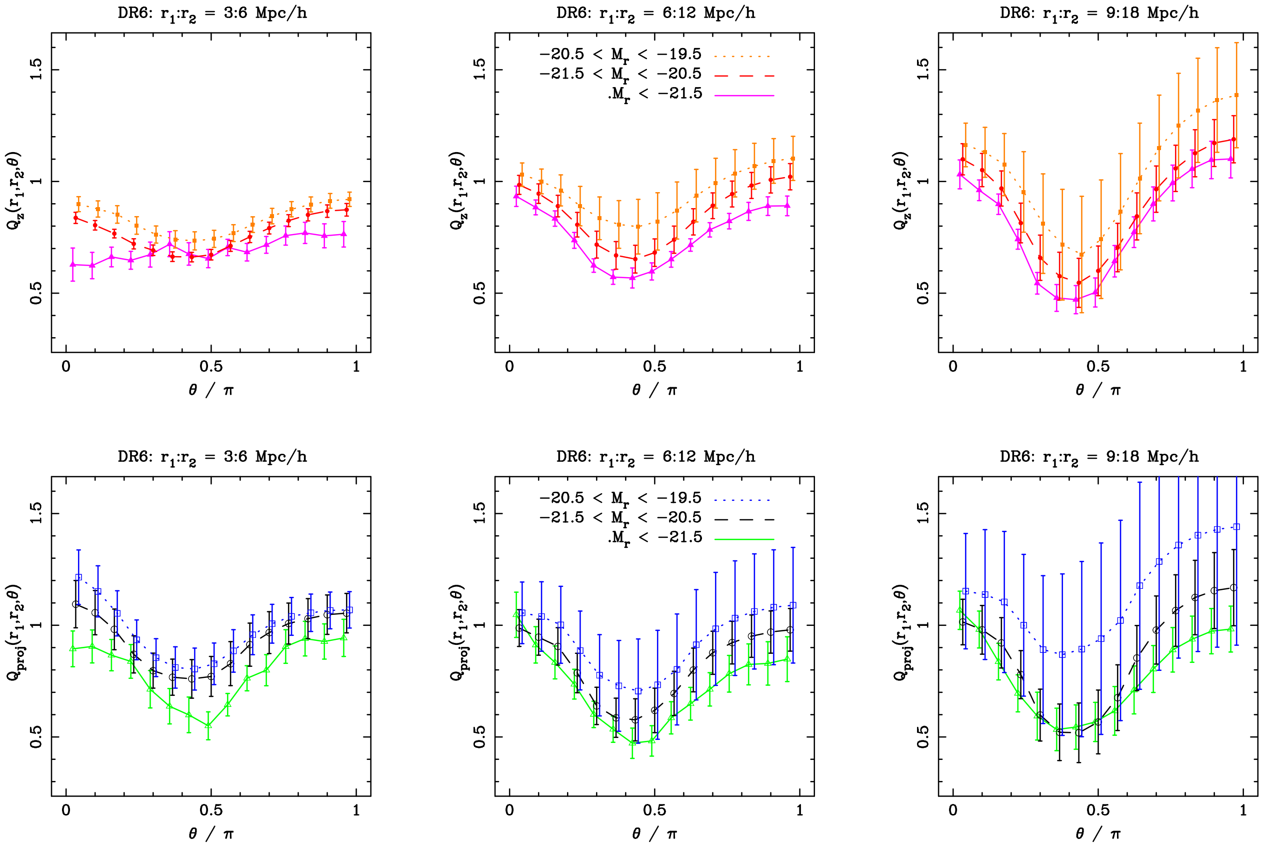


Figure 4.7 The reduced 3PCF on SDSS DR6 galaxies, comparing the three samples of different magnitude limits. Triangles correspond to $M_r < -21.5$; circles with $-21.5 < M_r < -20.5$ and squares with $-20.5 < M_r < -19.5$. The top row contains redshift space measurements, and the bottom row depicts the projected measurements. The three columns are different scales, specified by the first side of the triangle (r_1) representing the smallest scale measured.

converge within the uncertainties. We see the uncertainties are larger on Q_{proj} at all scales. The latter point is explained naturally by the projection: a given scale in r_p really represents a lower bound on the redshift space scale (s) where the upper bound is $s = \sqrt{r_p^2 + \pi_{max}^2}$. Since Q at any given value of r_p is sensitive to scales greater than the corresponding s, it will include the properties of larger scales. We previously noted that the uncertainty grows with scale, which is reflected in the larger uncertainties for Q_{proj}.

4.3.7 Discussion

We can clearly see a few common trends in the galaxy samples:

1. $Q(\theta)$ exhibits configuration dependence at all scales ($3 - 27h^{-1}$Mpc), and is generally not consistent with Q being simply a constant.
2. Larger scales ($9 - 27$ h^{-1}Mpc) exhibit a stronger configuration dependence, displaying the V-shape predicted by the ΛCDM model.
3. All galaxy samples show significantly different clustering than the mass estimates.
4. The discrepancy with the mass predictions is larger for brighter galaxies.
5. Uncertainties on the 3PCF grow for larger scales and with smaller sample volume.

Generally, our SDSS galaxy measurements appear consistent with predictions from ΛCDM realized by the Hubble Volume simulation. Galaxies are known to be biased tracers of mass, which naturally accounts the differences in the 3PCF between observed galaxies and the mass field. We will quantify the *galaxy-mass bias* for our 3PCF in the next chapter.

The size of uncertainties growing with scale is understood to represent the limited volume of the sample. Basically, the volume is insufficient to reliably sample the finite number of structures of a given size making the survey volume less representative of the universe. The fainter samples occupy a smaller effective volume and consequently larger scales exhibit a larger uncertainties.

The division of two volume-limited samples into sub-populations based on color led to these observations:

1. The $g - r$ color split isolates populations with different clustering properties.
2. Red sub-samples have a higher average Q value than blue.

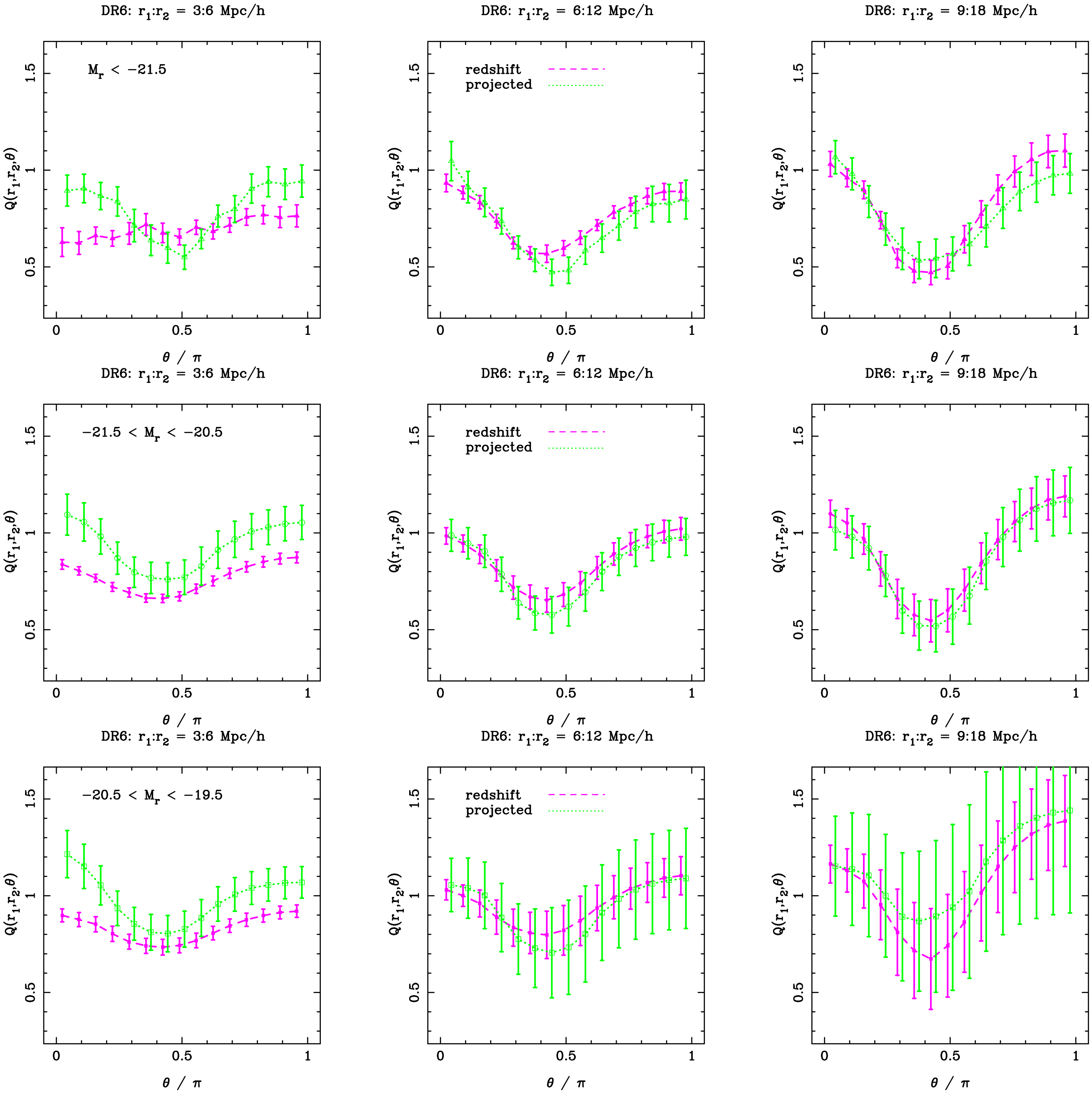


Figure 4.8 The reduced 3PCF on SDSS DR6 galaxies, comparing projected to redshift measurements. The filled symbols are redshift measurements, and the hollow symbols are projected. Triangles correspond to $M_r < -21.5$, circles with $-21.5 < M_r < -20.5$ and squares with $-20.5 < M_r < -19.5$. The three columns are different scales, specified by the first side of the triangle (r_1) representing the smallest scale measured.

3. The difference between red and blue sub-samples is greatest at small scales, which includes changes to the configuration dependence.

Overall, the luminosity dependence of the 3PCF differs from color dependence in these ways:

1. Luminosity dependence affects all scales at about the same level, and primarily changes the average amplitude of Q.
2. The effects of color produce a greater change in Q than luminosity at small scales, where galaxy color significantly affects the configuration dependence.

An analysis of the 2PCF of SDSS galaxies by Zehavi et al. (2005) conceptually explains the clustering difference of brighter galaxy samples, where brighter galaxies are "more biased" than fainter ones and hence there is a larger difference between bright galaxies in comparison with mass. We see a similar result in our measurements of the 3PCF. The stronger effect of color separation on smaller scales ($3-9h^{-1}$Mpc) agree with the conclusions by Blanton et al. (2005a) that galaxy color is strongly tied to local environment. We noticed that the configuration dependence of Q_{proj} was stronger for red sub-samples at smaller scales ($r_1 = 3\ h^{-1}$Mpc). Almost all dark matter halos are expected to have radii significantly smaller than the $3-9\ h^{-1}$Mpc this probes. However, it might suggest that regions outside large halos have an anisotropic distribution of red galaxies surrounding them, perhaps tracing infalling regions. This interpretation supports observations of the angular distribution of satellite galaxies by Azzaro et al. (2007), where they find an alignment of satellites along the major axis of their host. They find the strongest evidence when considering red satellites of red hosts which exactly represents our red sub-sample. On measurements of large scale triangles ($r_1 = 9\ h^{-1}$Mpc), we conclude that both red and blue populations exhibit similar configuration dependence given the large errors. However, we speculate that the blue galaxies occupy regions that show stronger shape dependence.

A detailed study of the reduced 3PCF in redshift space was conducted on a different redshift survey (Gaztañaga et al., 2005), the two-degree field galaxy redshift survey (2dFGRS; Colless et al., 2001). Their parameterization of triplets is very close to ours, and the analysis is in qualitative agreement with our measurements of Q_z. They also divided a volume-limited sample into "red" and "blue" sub-samples, noting that red galaxies typically have a larger value of Q_z below 12 h^{-1}Mpc. A more specific comparison is difficult, as their samples are based on a different survey with samples defined by a different pass-band and color criteria.

The 3PCF of SDSS galaxies was also studied in redshift space by Kayo et al. (2004). Their measurements that are sensitive to shape (see their Figure 13) show a configuration dependence that grows with scale. Their Q_z for $s = 2.5$ h^{-1}Mpc shows very little configuration dependence with $Q_z \leq 1$ for all θ which is comparable to our measurements at $s = 3.0$ h^{-1}Mpc, and confirm the effects of redshift distortions on these scales. In contrast to their conclusions, we conclude there is a significant luminosity dependence of the 3PCF which is apparent in the equilateral Q_{eq} (Figure 4.3) as well as the configuration dependent $Q(\theta)$ (Figure 4.7). The disagreement between our results and Kayo et al. (2004) is likely due to the statistical significance between the galaxy samples – since DR6 contains many more galaxies in a much larger volume the weak luminosity dependence is noticeable. We also note differences between the results with respect to color dependence. Kayo et al. (2004) see very little clustering difference between populations on our small scales (middle panels of their figure 13). In addition, their measurements show the blue population has larger value of Q than the red on larger scales (bottom panels their same figure), whereas our results show little difference (our right hand panels of figures 4.5 and 4.6). However, the latter effect is within 1σ given the uncertainties of both studies, and is therefore not significant. The differences might be accounted for by one of the many subtle differences of the analyses, which include (1) different SDSS sample, (2) different galaxy sample and color definitions, and (3) different parameterization of triangles and binning. Sorting out these systematic differences would require a detailed joint comparison which is not warranted by the low statistical significance of the discrepancy.

Finally, we compare redshift and projected space 3PCFs to find:

1. Q_{proj} successfully recovers configuration dependence of the 3PCF at small scales, which is lost in Q_z.

2. Q_{proj} has a higher amplitude at small scales than Q_z.

3. Q_z and Q_{proj} both converge at large scales.

Both Q_z and Q_{proj} were measured by Jing and Börner (2004). Although they analyze an entirely different galaxy data (2dFGRS), we expect trends between redshift and projected space measurements to be similar. However, they do not agree completely with our findings on comparable scales and triplet configuration. While they notice an increase in Q amplitude for their "full" sample (compare their figures 7 & 12 for $r = 3.25\ h^{-1}$Mpc with $u = 2$), it is not as much of a difference as we see. They also see no configuration dependence in Q_{proj}. We suspect their measurements artificially obscure these features due to their choice of binning and triangle parameterization, as described in Gaztañaga and Scoccimarro (2005). A separate analysis of 2dFGRS data by Gaztañaga et al. (2005), which uses a parameterization similar to what we employ, presents Q_z measurements that show more configuration dependence than Jing and Börner (2004), supporting our results on the discrepancy with Q_{proj}. We conclude that our measurements are the most reliable since we have carefully considered binning effects, as well as used a larger, more statistically significant galaxy sample.

4.4 COVARIANCE MATRICES

It is essential to take into account the correlation between measurements to properly utilize observational data for theoretical constraints. Neglecting this information can result in an unphysical bias or misrepresentation of significance.

We expect measurements of clustering to be correlated from even simple theoretical models of gravitational collapse. The fourier transform of the 2-pt correlation function, the power spectrum, is often used on large scales as we expect the power between modes to be independent in the linear regime. The structure of the covariance matrix, with non-zero off-diagonal terms, encodes this "independence" of modes for the 2-pt correlation function. Even in the linear regime, observational datasets have a correlation between modes imposed on them by the window function of the survey; it is not enough to just use the power spectrum. The covariance must be used for accurate data constraints, regardless of whether the analysis is done in configuration or fourier space.

Determining the covariance matrix, however, can be difficult. Theoretical predictions quickly become tedious, even for the simplest assumptions. For example, the covariance of the 2-pt correlation function depends on both the 4-pt and 3-pt functions. For the 3-pt, the dominant term is the 6-pt function (for a concise description see Szapudi, 2005). Without theory, we are left with empirical methods of estimating the covariance matrix. All empirical methods introduce sampling noise. Ideally, the noise can be minimized by using a very large number of independent realizations, which are difficult to obtain in practice. In any case, empirical methods estimate the covariance to within some finite accuracy.

To demonstrate why it is important to consider the uncertainty in the covariance, let us examine a simple χ^2 statistic. We construct a vector of data measurements ($\vec{d}$) with uncertainties (σ) that we want to compare to some model ($\vec{d}_{model}$), where we have measured the covariance in the correlation matrix ($\boldsymbol{\mathcal{C}}$), then

$$\chi^2 = \left(\frac{\vec{d} - \vec{d}_{model}}{\sigma} \right)^T \boldsymbol{\mathcal{C}}^{-1} \left(\frac{\vec{d} - \vec{d}_{model}}{\sigma} \right) . \tag{4.5}$$

We see that χ^2 depends on the inverse of the covariance matrix. If we think of the covariance matrix in terms of its eigenmodes (λ_i), we can see the problem. We assume a signal dominated estimate of the covariance, i.e. the observed structure reflects the true correlation. The largest variance will be contained by the most significant eigenmode (largest λ) with the noise in the least significant mode (smallest λ). Since we invert the covariance matrix, χ^2 is proportional to $\frac{1}{\lambda}$, and we see the dominant contribution is from eigenmodes with the smallest λ. This, however, can be corrected by trimming the unresolved "noisy" modes before calculating χ^2. The problem then becomes one of determining the appropriate threshold between "signal" and "noise" in the covariance matrix.

4.4.1 Estimating the covariance

We measure the correlation between measurements by empirically calculating the covariance matrix. Given a number of realizations, N, a fractional error on Q can be defined as

$$\Delta_i^k = \frac{Q_i^k - \bar{Q}_i}{\sigma_i} , \tag{4.6}$$

for each realization (k) and bin (i) given a mean value ($\bar{Q}_i$) and variance (σ_i^2) for each bin over all realizations. We use Q as a general placeholder for any measured statistic (2PCF, 3PCF, etc). We construct the normalized covariance matrix using the standard unbiased estimator:

$$\mathcal{C}_{ij} \equiv \frac{1}{N-1} \sum_{k=1}^{N} \Delta_i^k \Delta_j^k \; . \tag{4.7}$$

Equation 4.7 assumes that each realization is independent. In practice, a number of mock galaxy catalogs can be used, making this a tractable approach. Galaxy mock catalogs are commonly constructed in one of two ways: (1) placing artificial galaxies in high-resolution N-body simulations or (2) an ingenious amalgamation of halo model approximations with perturbation theory predictions via PTHalos (Scoccimarro and Sheth, 2002). Both methods generate a spatial distribution of galaxies that can be adjusted to match important aspects of observational samples (e.g. the SDSS footprint). However, running N-body simulations in the first place is computationally expensive. Even with the huge TeraFlop computing resources available today, running the appropriate simulations is a large task. They must have both the mass resolution to resolve halos that observed galaxies live in as well as cover a significant volume of that probed by current galaxy redshift surveys. PTHalos can achieve similar distributions with orders of magnitude less computational power at the cost of an approximate treatment. PTHalos has been shown to work well at large scales, but can be deficient where non-linear effects become important or for higher order clustering moments (R. Scoccimarro, private communication).

4.4.2 Jackknife Re-sampling

If mock catalogs appropriate to the galaxy sample are not available, a covariance matrix can be estimated from the data itself using a *leave one out cross validation* method, more commonly referred to in the community as jackknife re-sampling (Lupton et al., 2001). Briefly described, a small subset of the total volume is omitted and treated as a different realization. For twenty jackknife samples, the omitted region would be 1/20 of the total volume leaving each realization to consist of 19/20 of the full sample volume. The number of jackknife samples (which we still denote N) are not independent realizations, but the covariance matrix can be estimated by:

$$\mathcal{C}_{ij}^{(jack)} = \frac{(N-1)^2}{N}\mathcal{C}_{ij} = \frac{N-1}{N}\sum_{k=1}^{N}\Delta_i^k\Delta_j^k\ . \tag{4.8}$$

Jackknife re-sampling has been shown to be reliable on scales up to $30h^{-1}$Mpc for the 2PCF on spectroscopic galaxy samples when compared with independent mocks (Zehavi et al., 2002). It is not yet clear if it is as effective for the 3PCF.

We estimate our covariance matrices using the jackknife re-sampling method, as a sufficient number of mock catalogs were not available that matched our galaxy samples. We generate our jackknife samples using the SDSSpix pixelization scheme which can account for the irregular geometry of the SDSS data. The implementation, `jack_random_polygon`, has been used in studies of angular clustering (Scranton et al., 2002; Connolly et al., 2002). The jackknife regions are selected by angular span on the sky and maintain equal unmasked area. In the SDSS samples we use, survey depth does not vary over the sky which makes equal area consistent with equal volume. We choose 30 jackknife samples where the regions are ≈ 180 square degrees. We discuss this decision, and its validity with respect to some independent mock catalogs in Chapter 6.

4.4.3 Covariance of Galaxy Samples

We show the normalized covariance matrix of the 2PCF for our three galaxy samples in Figure 4.9. The BRIGHT sample on the right appears to have a covariance matrix that is undersampled, with the most significant contribution being noise. We can see a few trends in the two fainter samples. First, the projected $w_p(r_p)$ measurements appear to be much more correlated than the redshift space $\xi(s)$. Since w_p projects along the line of sight, measurements of scales between $(s = r_p)$ and $(s = \sqrt{r_p^2 + \pi^2})$ become mixed, inducing a correlation. Seeing this effect supports the validity our of covariance estimation. We also notice that the correlation appears to increase for the fainter galaxy sample.

The covariance matrices for the reduced 3PCF on the $M_r < -21.5$ sample are shown in Figure 4.10. On small scales $(r_1 = 3)$, the measurements of $Q(\theta)$ appear largely uncorrelated,

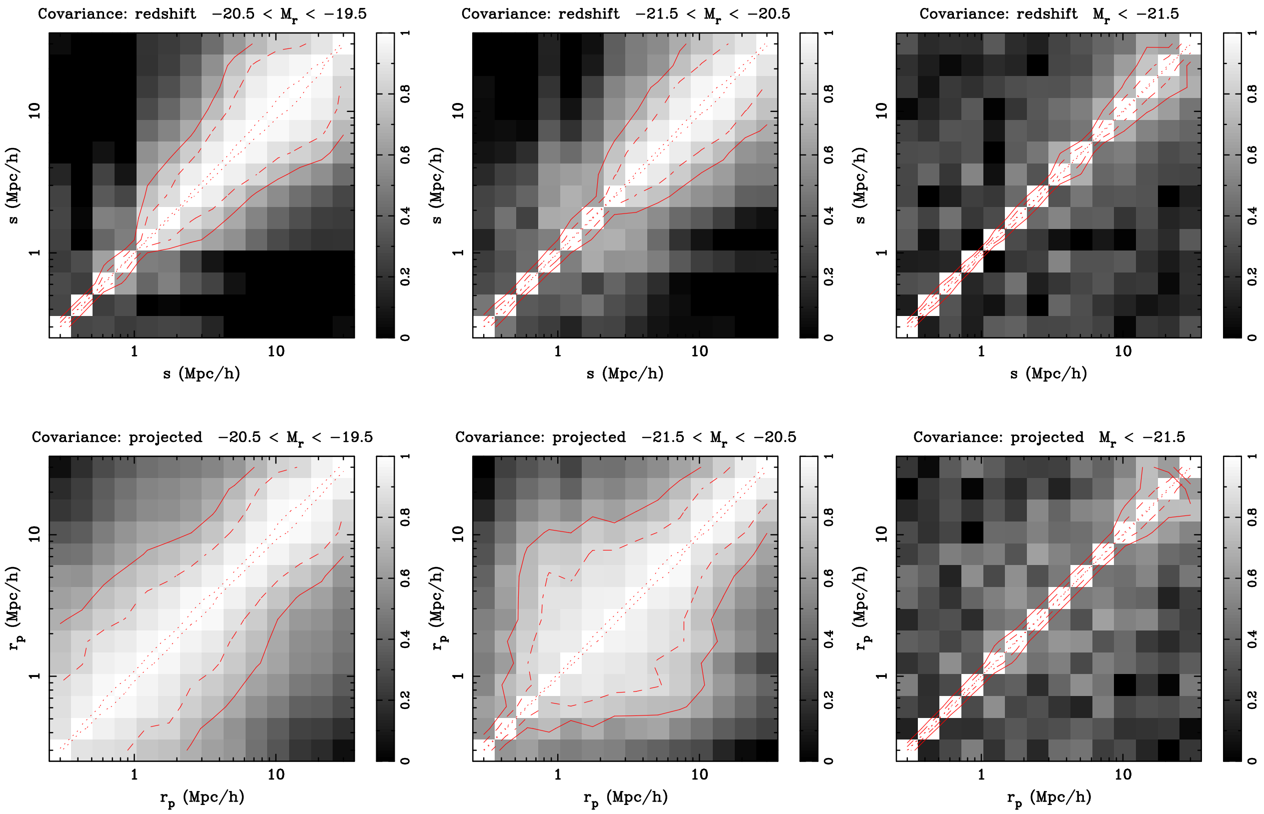


Figure 4.9 The normalized covariance matrix for the 2PCF, both in redshift (top row) and projected space (bottom row) for our three DR6 galaxy samples of different magnitudes. From left to right, the panels indicate FAINT to BRIGHT galaxy samples. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99.

except at large opening angle ($\theta \approx \pi$). Because this BRIGHT sample has a low number density (see Table 2.1), the small scale 3PCF has few counts resulting in covariance dominated by shot-noise which is uncorrelated. However, we can still see structure in non-zero elements that are off-diagonal. At all scales, the increased correlation at $\theta \approx \pi$ is due to physical overlap of our binning scheme (the same triplet can be counted more than once). Note the significant off-diagonal structure for the $r_1 = 6$ & $9\ h^{-1}$Mpc, which is comparable to that seen in theoretical studies (see figure 9 in Gaztañaga and Scoccimarro, 2005).

We now examine our LSTAR sample ($-21.5 < M_r < -20.5$) in Figure 4.11. The immediately noticeable difference is the much greater correlation at all scales. Since the number density is higher, we expect a better sampling with respect to $M_r < -21.5$ galaxies for the $r_1 = 3;\ h^{-1}$Mpc scale – which is evident. The projected 3PCF at $r_1 = 3\ h^{-1}$Mpc displays a much larger amount of correlation than in redshift space. Off-diagonal elements reveal structure at all scales. We note that almost all measurements display significant correlation (over 0.5, e.g. 50%). Clearly the use of a diagonal covariance matrix would be a poor assumption.

Figure 4.12 shows the covariance for $-20.5 < M_r < -19.5$ galaxies. At all scales the measured covariance indicate more correlation in the projected measurements as opposed to redshift space. The two larger scale projected covariances are almost singular, with all values above 0.85. They are so correlated that there is almost no structure visible, although a wobble in the 0.99 contour can be seen for the $r_1 = 9;\ h^{-1}$Mpc matrix.

4.4.4 Discussion

We have seen that the covariance of the 3PCF yields significant structure at all resolved scales and in all of our galaxy samples. While our choice of wide binning contributes to the covariance especially when θ approaches zero or π, we conclude the structure of the covariance is physical (we justify this claim in Chapter 5). Measuring the configuration dependence of the 3PCF requires closely packed measurements. The range of scales probed by a specific $Q(r_1, r_2, \theta)$ measurement is the change in scale of the third side and equal to $2r_1$. Empirically estimating the covariance enables both systematic correlations (i.e.

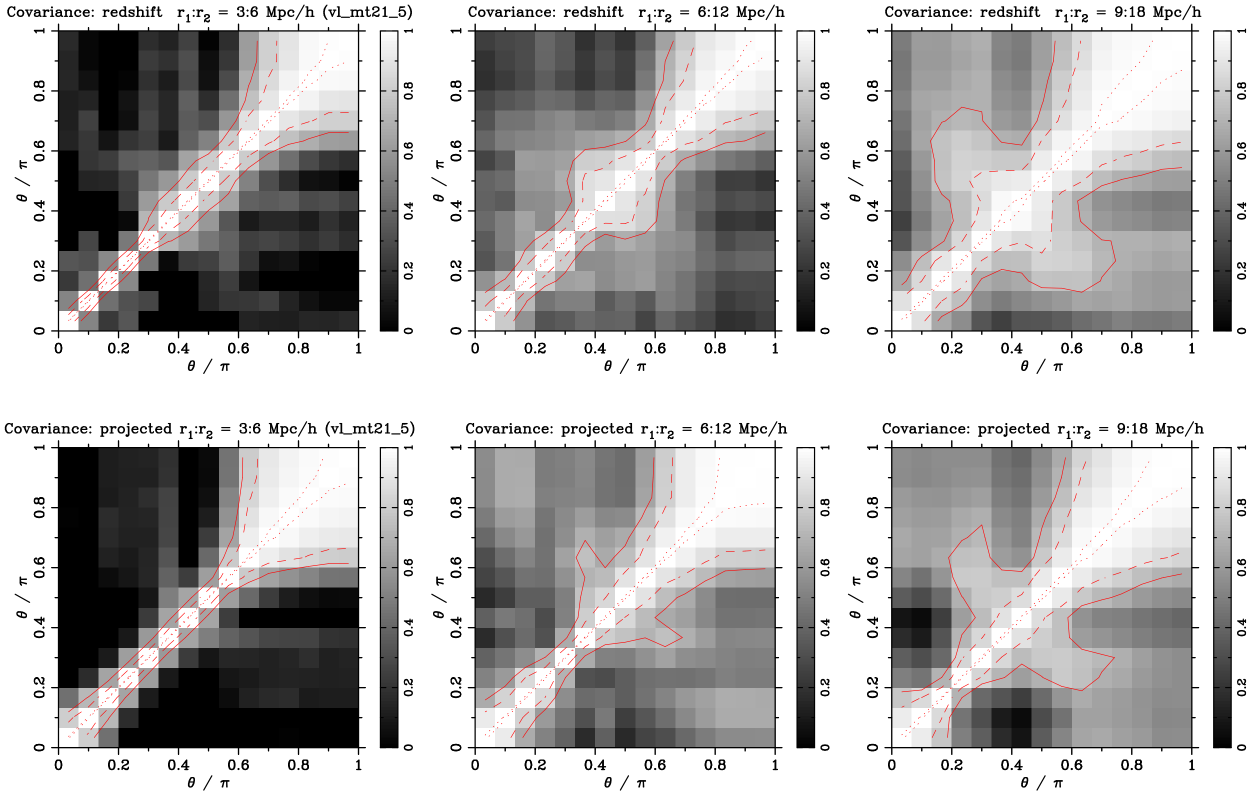
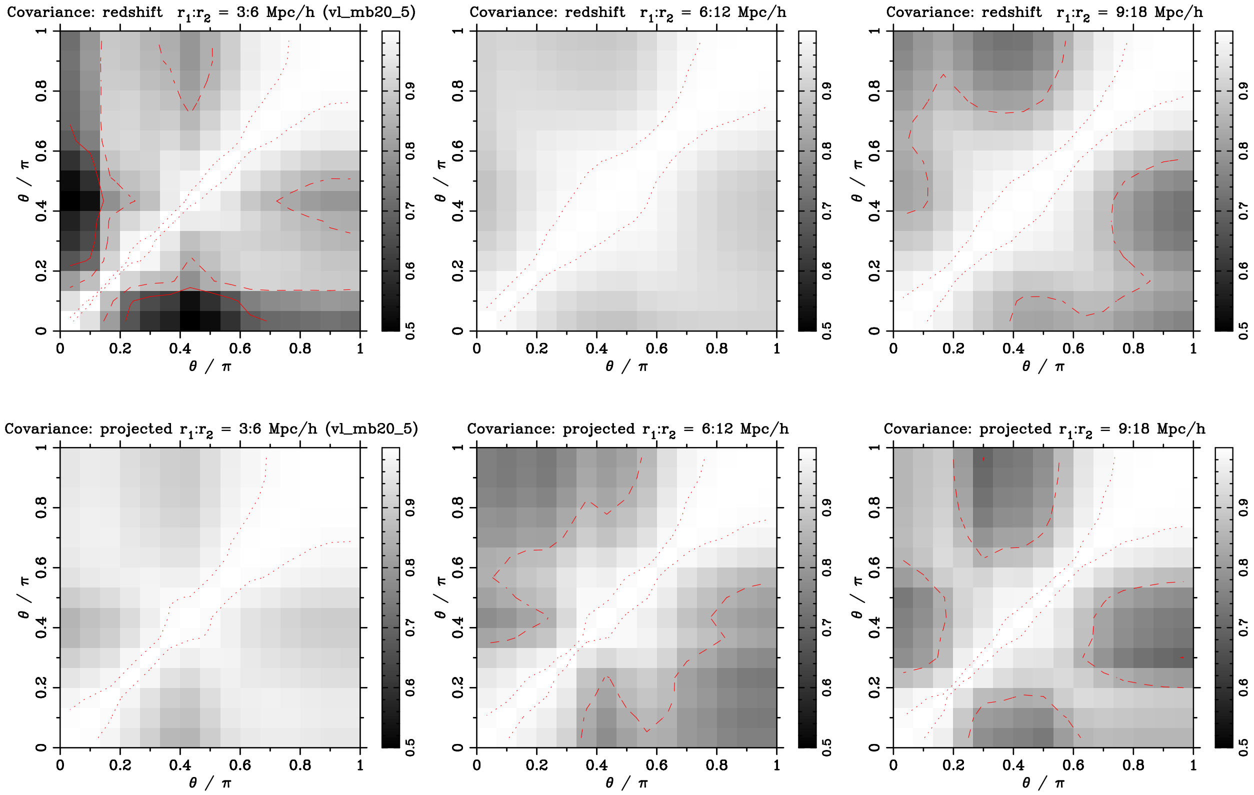


Figure 4.10 The normalized covariance matrix of the 3PCF for SDSS DR6 galaxies with $M_r < -21.5$. The matrix is normalized such that diagonal elements are set to unity, rather than the 1-sigma error. The panel order corresponds to that of the measurements presented in Figure 4.4. That is, the top three panels correspond to redshift space and the bottom three to projected space. Going from left to right, the scale of triangle increases as denoted by $r_1 : r_2$. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99.



Covariance: redshift r_1:r_2 = 3:6 Mpc/h (vl_mb20_5)

Covariance: redshift r_1:r_2 = 6:12 Mpc/h

Covariance: redshift r_1:r_2 = 9:18 Mpc/h

Covariance: projected r_1:r_2 = 3:6 Mpc/h (vl_mb20_5)

Covariance: projected r_1:r_2 = 6:12 Mpc/h

Covariance: projected r_1:r_2 = 9:18 Mpc/h

Figure 4.11 The normalized covariance matrix of the 3PCF for SDSS DR6 galaxies with $-21.5 < M_r < -20.5$. The matrix is normalized such that diagonal elements are set to unity, rather than the 1-sigma error. The panel order corresponds to that of the measurements presented in Figure 4.5. That is, the top three panels correspond to redshift space and the bottom three to projected space. Going from left to right, the scale of triangle increases as denoted by $r_1 : r_2$. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99. Please note that black now corresponds to a higher minimum value, specifically 0.5 as opposed to zero in the previous covariance matrices.

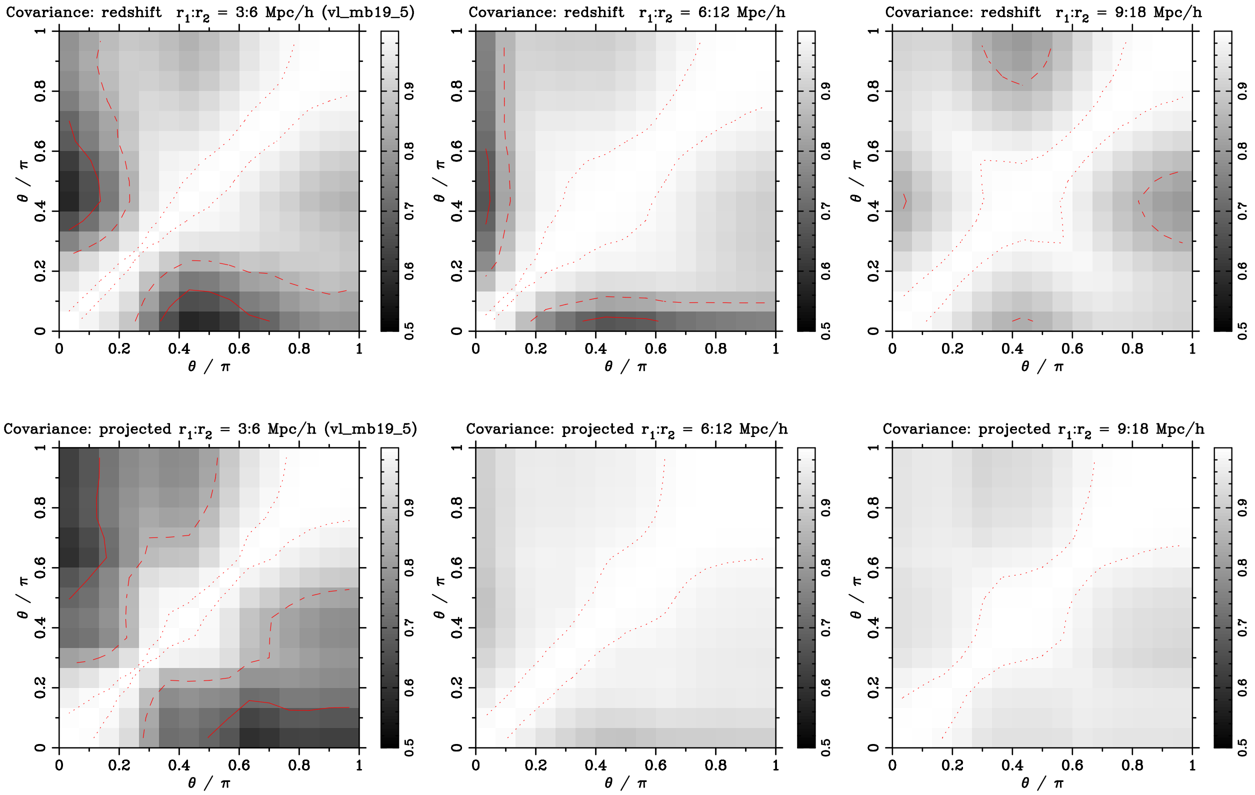


Figure 4.12 The normalized covariance matrix of the 3PCF for SDSS DR6 galaxies with $-20.5 < M_r < -19.5$. The matrix is normalized such that diagonal elements are set to unity, rather than the 1-sigma error. The panel order corresponds to that of the measurements presented in Figure 4.6. That is, the top three panels correspond to redshift space and the bottom three to projected space. Going from left to right, the scale of triangle increases as denoted by $r_1 : r_2$. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99. Please note that black now corresponds to a higher minimum value, specifically 0.5 as opposed to zero which is similar to Figure 4.11 and different from other covariance matrices.

overlapping bins) and physical covariance between 3PCF values to be taken into account for quantitative constraints. There is a limit, however. If we bin the correlation function too finely, such that pair and triplet counts become poorly sampled, the uncertainties will be dominated by Poisson noise and the covariance matrix will look diagonal. This would be a false representation of the correlation, even if the full covariance was used for constraints. We note that we approached this limit for our brightest sample ($M_r < -21.5$) even with our wide binning scheme (bin-width was $0.25r$, where r is the scale probed).

If we restrict ourselves to the quasi-linear regime where $\xi < 1$, represented in the 3PCF by the largest scale measured ($r_1 = 9\ h^{-1}$Mpc), we see significant structure to the covariance matrix. This type of structure, almost an "X" pattern, has been seen in simulations (e.g. Gaztañaga and Scoccimarro, 2005). We notice that the overall correlation increases as the galaxy sample becomes fainter for this scale. This luminosity dependence of covariance has been seen in the projected 2PCF (Zehavi et al., 2005). This work is the first time it has been resolved in observational measurements of the 3PCF, which highlights the importance of accurately determining the covariance. We identify the trend in both redshift and projected space. One new implication of our results, which is often overlooked (see Kayo et al. (2004), Nichol et al. (2006) and Zheng and Weinberg (2007)), is that care must be taken to properly estimate the covariance matrix specific to the galaxy sample being studied. The correlation matrix of one galaxy sample is not necessarily representative of another.

4.5 EFFECTS OF *SUPER* STRUCTURES

Large coherent structures or "super structures", such as the Sloan Great Wall (SGW; Gott et al., 2005), can have a dramatic effect on clustering measurements. Detailed analyses on SDSS galaxy samples have documented this in both the 2PCF (Zehavi et al., 2002, 2005) and redshift space 3PCF (Nichol et al., 2006). One advantage of using jackknife re-sampling methods for error analysis is that we probe the variation between different spatial regions essentially "for free". We investigate this variation using our 30 jackknife samples. We (re)investigate the effects of superstructures within the context of our clustering measurements for several reasons. First, our samples are based on a newer SDSS sample (DR6) which has both a larger volume than previous studies as well as new regions of the sky. Second, as we have repeatedly pointed out, measurements of the 3PCF are affected by a chosen binning scheme making it important to document the effects of structures given our parameterization. Finally, we want to understand the impact on detailed measurements of the projected 3PCF.

We identify 6 out of 30 regions of the sky which show large deviations in the reduced 3PCF. These "jackknife regions" characterize a jackknife sample by being *omitted* from a clustering measurement. The northern SDSS DR6 footprint in shown in Figure 4.13, with the entire sample displayed in gray and the six regions highlighted by color. Two regions encapsulate the majority of the SGW, specifically the red and magenta regions at a J2000 declination of zero. Overall, the jackknife regions appear contiguous and rectangular (neglecting the false geometry of the Aitoff sky projection). Please note the black region, however, which is split between two sides of the survey. The algorithm to define the jackknife regions must occasionally make such divisions.

Before we look at the 3PCF for these regions, let us briefly review the jackknife resampling method. We excise a jackknife region from the full sample and measure the clustering. This means that a measurement on a specific jackknife sample represents the clustering of the entire sample *omitting* the jackknife region. If the clustering on the jackknife sample deviates strongly from the average of all samples, it means that a specific jackknife region dominates the measurement for the entire sample. Without that region, the overall clustering would be significantly different. This is somewhat profound, since we do not expect such clustering differences given the volume of the SDSS survey.

We investigate the affects on the 3PCF in two of our three galaxy samples. Although the jackknife regions based on sky location are consistent across both, the redshift limits vary. This means the volume is not identical between them, although it does overlap. We use our LSTAR ($-21.5 < M_r < -20.5$) and FAINT ($-20.5 < M_r < -19.5$) samples; both include the SGW which has a mean redshift at $z \approx 0.08$.

To highlight the clustering deviations between samples, we plot the *normalized residual* of the 3PCF. We subtract the mean 3PCF of the 30 samples from each individual measurement,

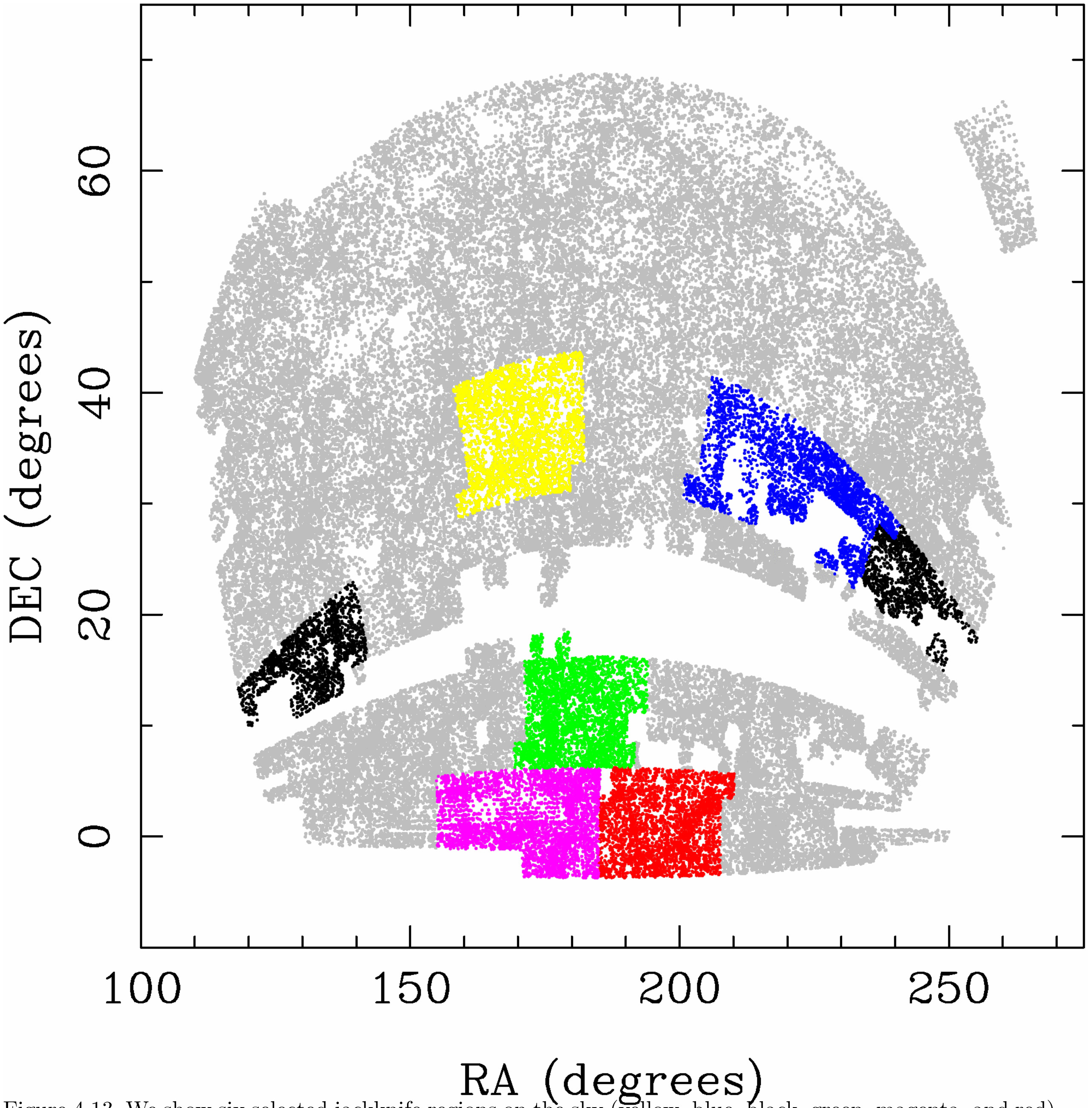


Figure 4.13 We show six selected jackknife regions on the sky (yellow, blue, black, green, magenta, and red) in comparison with the full galaxy sample (grey points) for the SDSS North footprint in J2000 equatorial coordinates. The Sloan Great Wall (Gott et al., 2005) is located at ≈ 0 in declination, and is included in two of the six selected jackknife regions (red and magenta).

and normalize this difference by the jackknife variance. As noted in (4.8) the prefactor for the jackknife variance is $(\frac{N-1}{N})$, and not the familiar $\frac{1}{N}$.

Figure 4.14 shows the *normalized residual* of the 3PCF for the LSTAR galaxy sample with $-21.5 < M_r < -20.5$. The left three panels are results for redshift space, Q_z, and the right correspond to projected space, Q_{proj}. The scale increases downward going in order of $r_1 = 3, 6$ and $9\ h^{-1}$Mpc. The gray lines represent results from the $30 - 6 = 24$ "ordinary" jackknife samples. The colored lines correspond to jackknife samples that omit the jackknife regions highlighted in Figure 4.13. We notice that most of the regions are close to the mean value (i.e. around zero). We also notice several samples deviate, which always happens in the negative direction. If a structure in the jackknife region, which is included in *all* other samples, boosts $Q(\theta)$ then a negative residual results. In the $-21.5 < M_r < -20.5$ sample, Q_z appears to be boosted at all scales by galaxies in the blue jackknife region. As scale increases, structure in the red region also seems to increase Q_z. Both regions also have an effect on Q_{proj}, with red clearly dominating at $r_1 = 9\ h^{-1}$Mpc and blue at the small scale $r_1 = 3\ h^{-1}$Mpc measurement. In Q_{proj}, we note the black region, which is a physical neighbor to the blue region, strongly affects the small scale 3PCF.

The effects of the red region, especially at large scales, is likely due to the SGW. This has been seen before (e.g. Nichol et al. 2006). The blue and neighboring black regions were first included in the DR5 release. At smaller scales, these clearly dominate the 3PCF measurement. While the jackknife regions are 1/30 of the total sky coverage, the jackknife region could possibly contain multiple structures that coincidentally occupy the same jackknife region. We do not believe this to be the case, however. We investigated the galaxies in the blue region, and found a single projected overdensity was responsible for the clustering difference in the region, which appears to have a median redshift of $z \approx 0.11$. This is a distinct physical structure from the SGW. This structure overlaps with the black region, producing an anomalous result for that volume. This suggests clustering deviations within a jackknife region arise from a single dominant "super" structure. We refer to them as "super" structures, since like the SGW, they are coherent overdensities but are too large to be gravitationally self-bound (in contrast to a galaxy cluster). Individually, these regions shift the 3PCF by less than 10% in the jackknife samples – but this analysis excludes cumulative

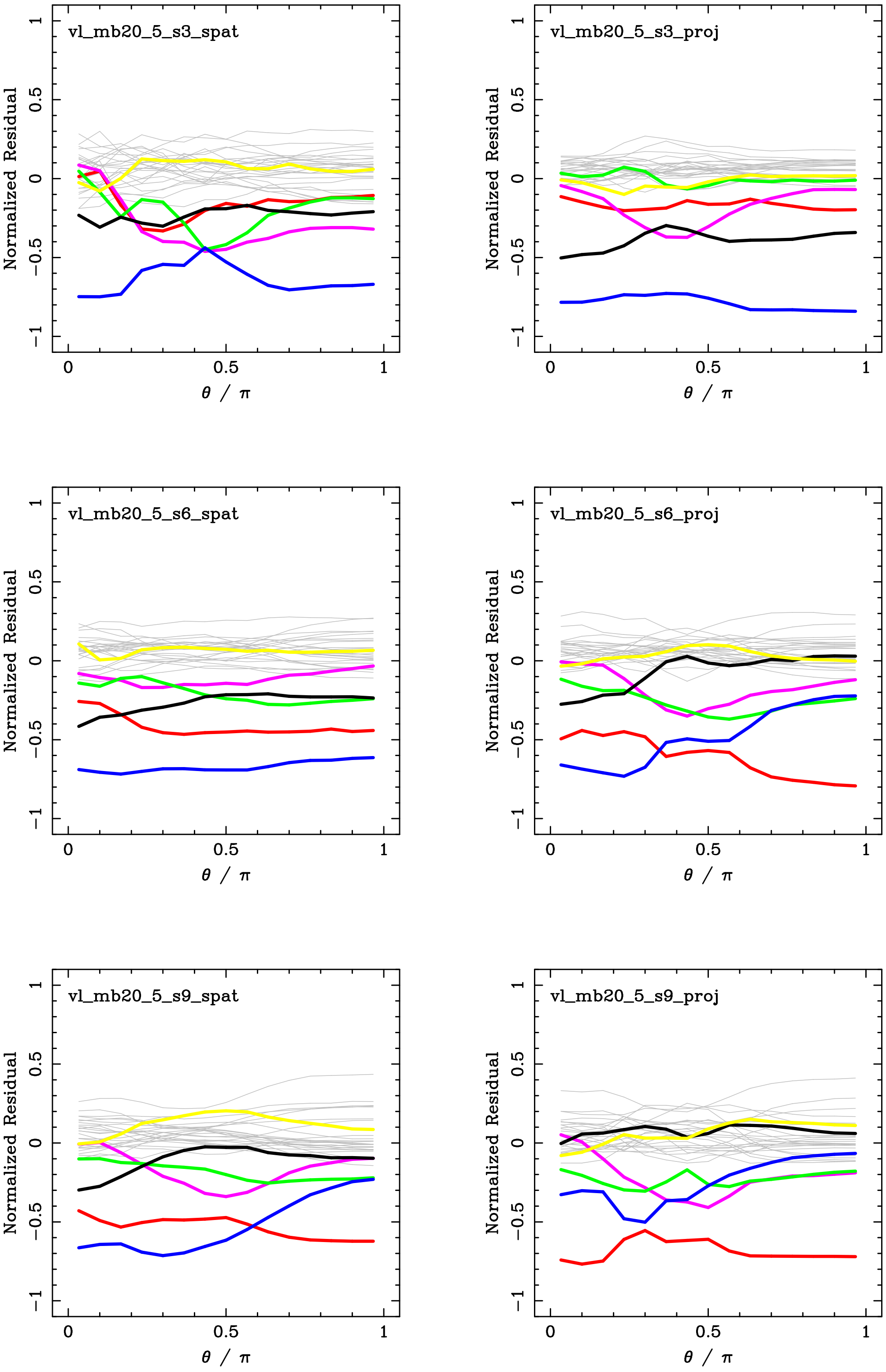


Figure 4.14 The normalized residuals for 30 jackknife samples of the reduced 3PCF in redshift and projected space for the DR6 galaxies with $-21.5 < M_r < -20.5$. A jackknife sample is defined by taking the full sample and excluding a specific jackknife region. Colors of the six selected jackknife samples correspond to excluding the matching color in Figure 4.13.

effects (for example, when the blue region is omitted part of the same "super" structure remains in the sample from the black region).

We look at the FAINT galaxy sample ($-20.5 < M_r < -19.5$) in Figure 4.15. Recall from the 3PCF measurement, specifically Figure 4.6, that the uncertainties were quite large, much more than the previous $-21.5 < M_r < -20.5$ sample. Looking at the residuals, we see the large scales are completely swamped from the clustering in the red region. Excluding this region changes the 3PCF by approximately 20%, primarily due to galaxies clustered in the SGW. The black and blue regions seem inconsequential for this sample. This makes sense as this fainter sample has a maximum redshift of $z = 0.086$ and has trimmed the structure at $z \approx 0.11$ which accounted for the significant 3PCF difference in the brighter $-21.5 < M_r < -20.5$ sample. We also see that a new region, denoted by yellow, dominates the differences at small scales.

Investigating the jackknife samples enabled us to view the clustering of galaxies in SDSS samples with an alternative perspective. A few regions, which each likely contain one or perhaps two "super" structures, dictate the clustering of the entire sample – even with the sizable volume probed by the SDSS galaxies. Different structures affect different scales. The standard proposed solution is to continue collecting bigger samples in the hope of a large enough volume to average over multiple rare "super" structures. This also highlights the difficulty of identifying trends with luminosity, as creating volume-limited samples results in galaxies that occupy different volumes, where a single structure might affect clustering of one sample and not the other.

We have used the reduced 3PCF to probe for differences, which is more sensitive to the effects of structures than lower order statistics (Nichol et al., 2006). We also notice marked differences in the 2PCF for the regions, in agreement with other studies (Zehavi et al., 2005). With the 3PCF, deviations seemed to be largely an amplitude offset as structures affected triplets of all configurations equally. It would be interesting to see if this continues to be the case for smaller scales that approach the size of large halos. The configuration dependence at the scales we measure are predominately due to filamentary structures, so it makes sense that one specific region would not dramatically enhance or erase this signal. Clearly, these super structures can affect quantitative descriptions of bias. How best to handle this in

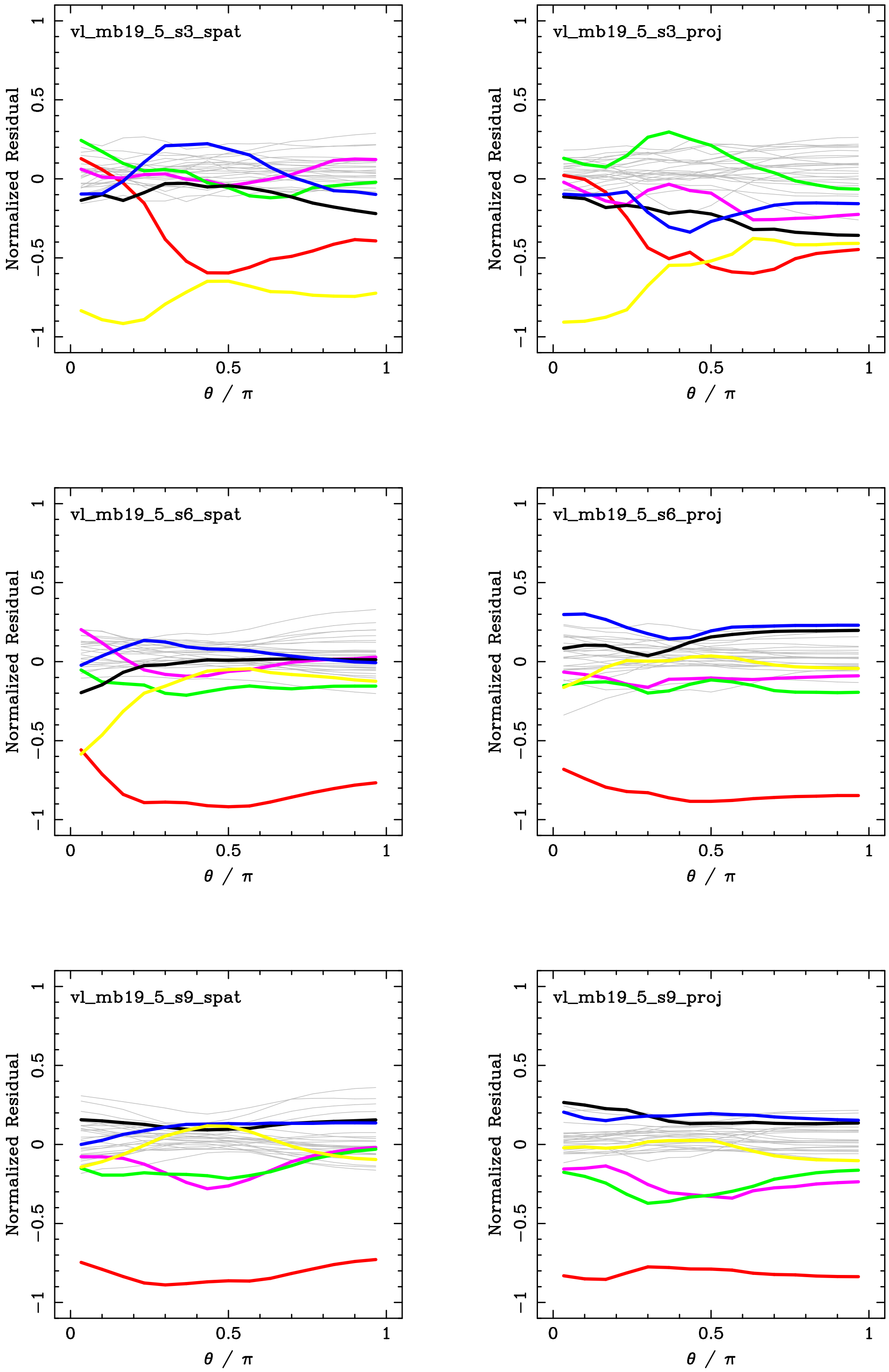


Figure 4.15 The analogous figure to 4.14, the normalized residuals for 30 jackknife samples of the reduced 3PCF in redshift and projected space for the DR6 galaxies with $-20.5 < M_r < -19.5$. A jackknife sample is defined by taking the full sample and excluding a specific jackknife region. Colors of the six selected jackknife samples correspond to excluding the galaxies correspondingly colored in Figure 4.13.

detailed analyses remains an open question.

4.6 SUMMARY

In this chapter, we presented our clustering measurements on SDSS galaxies. We investigated the 2PCF in §4.2 and found our galaxy samples exhibit similar trends to previous analysis of SDSS data (Zehavi et al., 2005). Specifically, we noted the 2PCF is approximately consistent with a power-law model and brighter galaxies result in stronger clustering at all measured scales ($0.3 - 30\ h^{-1}$Mpc).

We considered the 3PCF in §4.3. We found significant configuration dependence on intermediate to large scales ($3 - 27\ h^{-1}$Mpc), in general agreement with predictions from ΛCDM. This is in contrast to the hierarchical ansatz where the reduced 3PCF shows no dependence on triplet shape. Below 6 h^{-1}Mpc, the redshift space 3PCF showed reduced power and weak configuration dependence in comparison with projected measurements. Our results indicate that redshift distortions, and not galaxy bias, can make the 3PCF appear consistent with the hierarchical ansatz. Compared to the lower order 2PCF, the 3PCF exhibited a weaker dependence on luminosity with no significant dependence on scales above 9 h^{-1}Mpc. On scales less than 9 h^{-1}Mpc, the 3PCF showed a greater dependence on color than on luminosity.

We resolved the covariance matrices of our clustering measurements in §4.4, calculated by jackknife re-sampling using 30 samples. We found significant structure in the covariance with large off-diagonal elements depicting strong correlations. Our results demonstrate that an assumption of a diagonal covariance matrix is a poor choice and the correlations must be taken into account for any quantitative analysis. We discussed that the covariance matrix can be improperly resolved, such as when measurement bins are too small, which can incorrectly result in the assumption of a diagonal covariance matrix. We noted that the overall correlation generally increased for fainter galaxy samples, suggesting a luminosity dependence to the structure of the covariance. We conclude that care must be taken to properly estimate the covariance matrix specific to the galaxy sample being used for quantitative constraints – a fact which is often overlooked in recent work on the 3PCF.

Finally in §4.5, we demonstrated how large coherent structures, referred to as *super structures*, affect these clustering measurements. We use 30 independent regions on the sky, and show that 6 of the 30 produce anomalous deviations. These regions, which each contain one or perhaps two super structures, dictate the clustering of the entire sample – even with the sizable volume of the SDSS galaxy samples. Two of these regions are coincident with the huge structure known as the Sloan Great Wall (SGW; Gott et al., 2005), which has been shown to strongly affect clustering (Zehavi et al., 2005; Nichol et al., 2006). We demonstrated that how much a specific region affects clustering measurements depends on both galaxy sample and scale. We showed no one structure dominates all scales, but almost all measurements are anomalously affected by at least one region.

5.0 GALAXY-MASS BIAS

In this chapter, we quantify the clustering differences between observational galaxy samples and mass as realized by dark matter (DM) particles of N-body calculations in the Hubble Volume (HV) simulation. We introduce the context and motivation of this approach in §5.1. We review our method of analysis in §5.2, an eigenmode analysis that utilizes the covariance of measurements in the reduced 3PCF. In §5.3, we investigate the eigenvectors of the reduced 3PCF measurements. We utilize these to constrain the linear and quadratic terms of the local galaxy-mass bias model in §5.4, including the significance of quadratic bias in §5.4.1 and the implications for cosmology (σ_8) in §5.4.2. We evaluate our constrained bias parameters by investigating the relative bias between galaxy samples in §5.5. Finally in §5.6, we summarize the main results of the chapter.

5.1 INTRODUCTION

The galaxies measured in redshift surveys, such as the SDSS, are commonly understood to be imperfect tracers of the underlying mass. We want to constrain the galaxy-mass bias described by (1.9) using the full configuration dependence of the 3PCF in the quasi-linear and linear regime. At appropriately large scales, specifically when the Taylor expansion of overdensity holds, we interpret (1.12) with $B = b_1$ and $C = b_2/b_1$, where b_1 and b_2 represent the linear and quadratic bias parameters in (1.9). On smaller scales, such as the strongly non-linear regime, (1.12) can still be used as a fitting formula, however the interpretation of the B and C parameters as local bias parameters is no longer valid.

Empirically, $\xi < 1$ denotes the quasi-linear regime and we restrict our analysis to scales above 6 h^{-1}Mpc, corresponding to our two largest measurements with $r_1 = 6$ & 9 h^{-1}Mpc. We investigate galaxy-mass bias in two samples, the BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$) where the covariance is well determined.

We expect redshift distortions to affect this bias relation, which we partially neglect. In particular, we account for the effects of redshift distortions by applying a distortion distance to the dark matter particles based on their velocities for our mass measurement. However, this will not be completely sufficient as redshift distortions alter the bias relation (1.12), especially for Q_z. We expect that Q_{proj} will be predominantly unaffected and roughly equivalent to real space measurements for this parameterization (see e.g. Zheng, 2004).

While this bias prescription is simple, it remains an important empirical step for analyzing galaxy data. In the linear regime, characterizing bias helps us to understand the clustering of mass through the galaxy distribution – allowing observational constraints on cosmology. The bias itself, especially with variations due to galaxy properties such as luminosity and color, parameterizes statistical results from observations to help in understanding galaxies, their evolution and the environments in which they reside.

By using a simple prescription for galaxy-mass bias, we investigate effects of binning and covariance resolution in a quantitative analysis with a clear and simple model. An important part of our analysis is a baseline comparison of the projected 3PCF with respect to the more commonly studied redshift space measurements.

5.2 EIGENMODE ANALYSIS

We constrain bias parameters using the full information in the covariance matrix. We use a method that minimizes the effects of numerical noise in the estimated covariance for fitting parameters. We utilize an *eigenmode* analysis (Scoccimarro, 2000), which is the same method often referred to as a discrete Karhunen-Loeve transform or principal component analysis (PCA). We adopt the eigenmode terminology, which has been recently studied in the context of galaxy-mass bias (Gaztañaga and Scoccimarro, 2005).

The basic idea is to isolate the primary contributing eigenmodes of the reduced 3PCF based on the structure of the normalized covariance matrix. This allows us to trim unresolved modes and perform a fit in a basis which minimizes the non-Gaussianity of the residuals. To summarize, the covariance matrix can be cast in terms of a singular value decomposition (SVD),

$$\boldsymbol{\mathcal{C}} = \boldsymbol{U}\ \boldsymbol{\Sigma}\ \boldsymbol{V}^T \quad ; \quad \Sigma_{ij} = \lambda_i^2 \delta_{ij}\ . \tag{5.1}$$

where δ_{ij} is the Kronecker delta function making $\boldsymbol{\Sigma}$ a diagonal matrix containing the singular values, λ_i^2. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal rotations to diagonalize the covariance into $\boldsymbol{\Sigma}$ where $\boldsymbol{V}^T$ denotes the transpose of $\boldsymbol{V}$.

Applying the SVD to the covariance matrix yields a rotation into a basis where the eigenmodes are independent (i.e. the covariance matrix becomes diagonal). The rotation matrix from the covariance can be applied directly to our signal forming the *Q-eigenmodes*,

$$\widehat{Q}_i = \sum_j U_{ij} \frac{Q_j}{\sigma_j}\ . \tag{5.2}$$

The singular values provide a weight on the importance of each eigenvector. Specifically, a multiplicative factor of $1/\lambda_i^2$ gets applied when $\mathcal{C}$ gets inverted. With this in mind, a natural interpretation of a signal-to-noise (S/N) arises:

$$\left(\frac{S}{N}\right)_i = \left|\frac{\widehat{Q}_i}{\lambda_i}\right|\ . \tag{5.3}$$

We note that this S/N estimate is a *lower* bound on the true S/N because of the SVD. To remove noise and avoid numerical instabilities, we trim eigenmodes corresponding to low singular values. A suggestion by Gaztañaga and Scoccimarro (2005) is to keep eigenmodes corresponding to modes resolved better than the sampling error in the covariance matrix. Since our covariance matrices are normalized (i.e. the diagonal elements are set to one), the singular values are directly related to sampling error, and we require the so-called "dominant modes" (Gaztañaga and Scoccimarro, 2005) to satisfy:

$$\lambda_i^2 > \sqrt{2/N_m}\ . \tag{5.4}$$

This definition seems well motivated for independent realizations, such as mock catalogs. It is not yet clear if this choice is reliable for other estimates of the covariance, such as the jackknife re-sampling we use. In general, choosing which principal components, or modes, to keep is a non-trivial process.

The advantage to using this eigenmode analysis for fitting is threefold. First, the full covariance between measurements is correctly accounted for. Second, by performing the fit in the rotated basis of the eigenmodes, the residuals of the fit are more Gaussian and the degrees of freedom are properly addressed (e.g. 3 eigenmodes really only fits over 3 numbers). Finally, using only dominant modes removes artifacts and noise in the estimated covariances. For example, when using the full covariance but not trimming modes, noise can cause a fit to converge on incorrect values with artificially small errors (and falsely high S/N). This effect becomes worse as the covariance becomes more unresolved. Conversely, fitting over dominant eigenmodes helps to eliminate any problems with the quality of error estimation (Gaztañaga and Scoccimarro, 2005, figure 13), including properly dealing with singular covariance matrices.

5.3 EIGENVECTORS IN THE 3PCF COVARIANCE MATRIX

We investigate the structure of the normalized covariance matrices by examining the eigenvectors (EVs), or principal components, from their singular value decomposition. The EVs are contained in the $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices in (5.1) and (5.2). The first EV is associated with the largest singular value (SV), and accounts for the largest variance in the covariance matrix. The second is the next largest SV and so on. If the structure in the covariance matrix represents predominately "true" signal, the lowest ranked EVs encapsulate the noise. While their amplitude is meaningless without the corresponding SV, they still represent the variation between bins in an orthogonal basis such that the true measurement is a simple linear combination.

We show the top three EVs for the BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$) galaxy samples in figures 5.1 and 5.2 respectively. In all cases there appear to be consistent features in the eigenmodes. The first EV represents weighting all bins equally. Typically, the second EV highlights the difference between "perpendicular" and "co-linear" configurations. Finally, the third EV tracks a roughly monotonic change from small to large θ, possibly accounting for the scale difference of the continually increasing third side of the triangle. We point out that this structure evident in observational galaxy samples agrees well with theoretical predictions from simulations in Gaztañaga and Scoccimarro (2005). We point out these EVs are obtained by deconstructing the covariance matrix without using the measurement signal directly. We conclude that structure evident in the reduced 3PCF is clearly encoded in the error estimates which is not consistent with random noise.

If we look in detail at the eigenmodes, we note there is not always a clear separation between the second and third EVs. Since the full structure is a linear combination of all modes, the configuration dependence and scale variation effects can be blended. The SVs of the two effects are almost equivalent for our measurements, making their numerical distinction in the SVD somewhat arbitrary. This is not a concern, as it appears the linear combination of these two EVs is consistent with the above interpretation, even if they are mixed. Higher EVs show less coherent structure as they represent noisy modes in the covariance, as we would expect.

By examining the eigenmodes of the covariance matrices, we note signal consistent with our 3PCF measurements. This supports our assessment that we utilize a signal dominated estimate of the covariance matrix. We can therefore use a combination of the most significant eigenmodes in a quantitative comparison to galaxy-mass bias models. In the following section, we ensure that at least the first three eigenmodes, clearly identified with signal structure, are included in any fit.

5.4 GALAXY-MASS BIAS CONSTRAINTS

We constrain the galaxy-mass bias using a maximum likelihood analysis by calculating a simple χ^2 statistic where the likelihood $\ln \mathcal{L} \propto (-\chi^2/2)$ and

$$\begin{aligned} \chi^2 &= \vec{\Delta}^T \boldsymbol{\mathcal{C}}^{-1} \vec{\Delta} \,, \\ \Delta_i &= \frac{Q_i - Q_i^{(t)}}{\sigma_i} \,. \end{aligned} \tag{5.5}$$

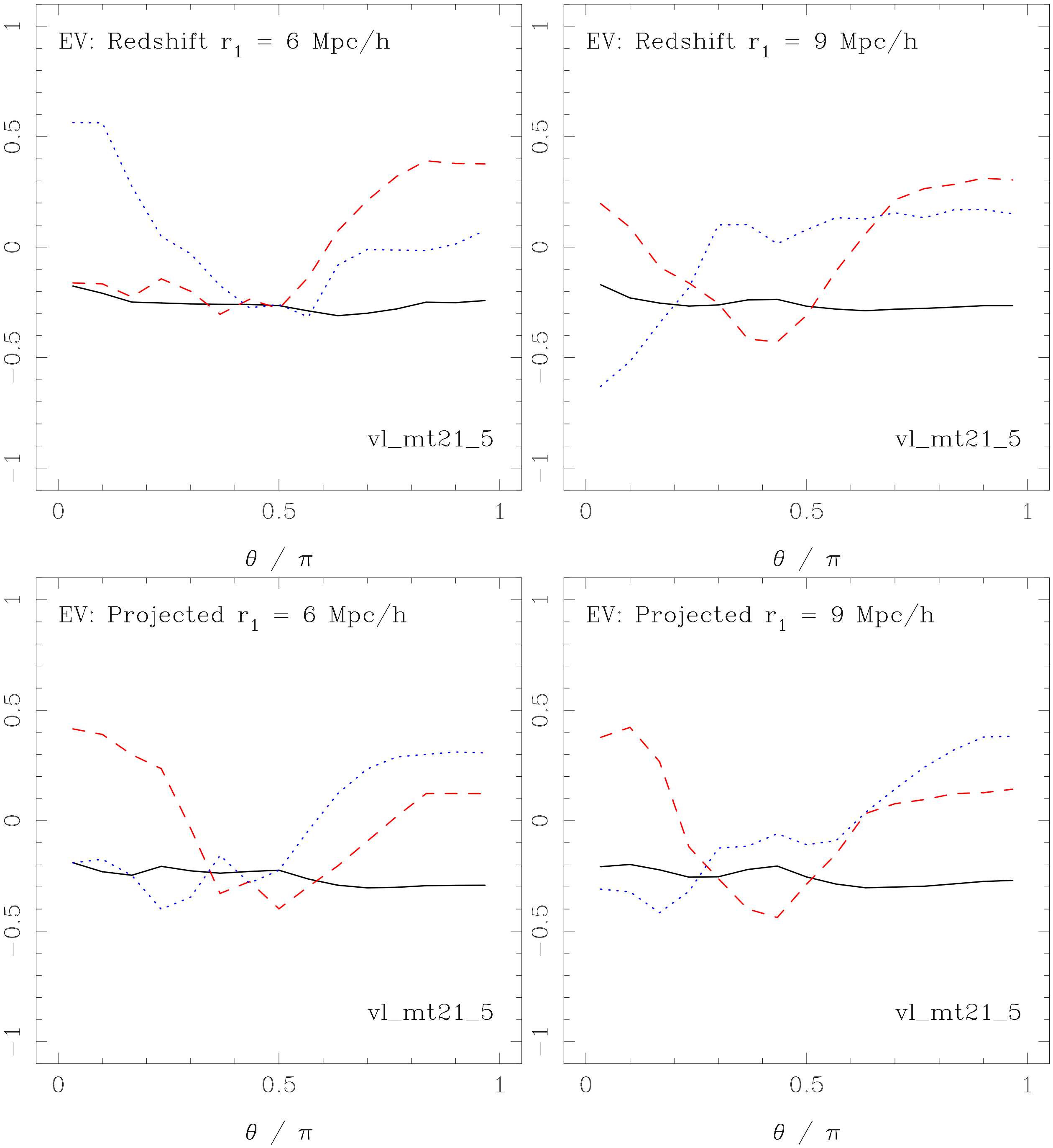


Figure 5.1 Top three eigenvectors (EVs) chosen from the normalized covariance matrix in the $M_r < -21.5$ galaxy sample. The sign of the EV is arbitrary. The first EV (solid black) shows equal weights for all bins. The second (dashed red) and third (dotted blue) EV display the configuration difference between perpendicular and co-linear triangles as well as the scale variation as the scale of the third side increases.

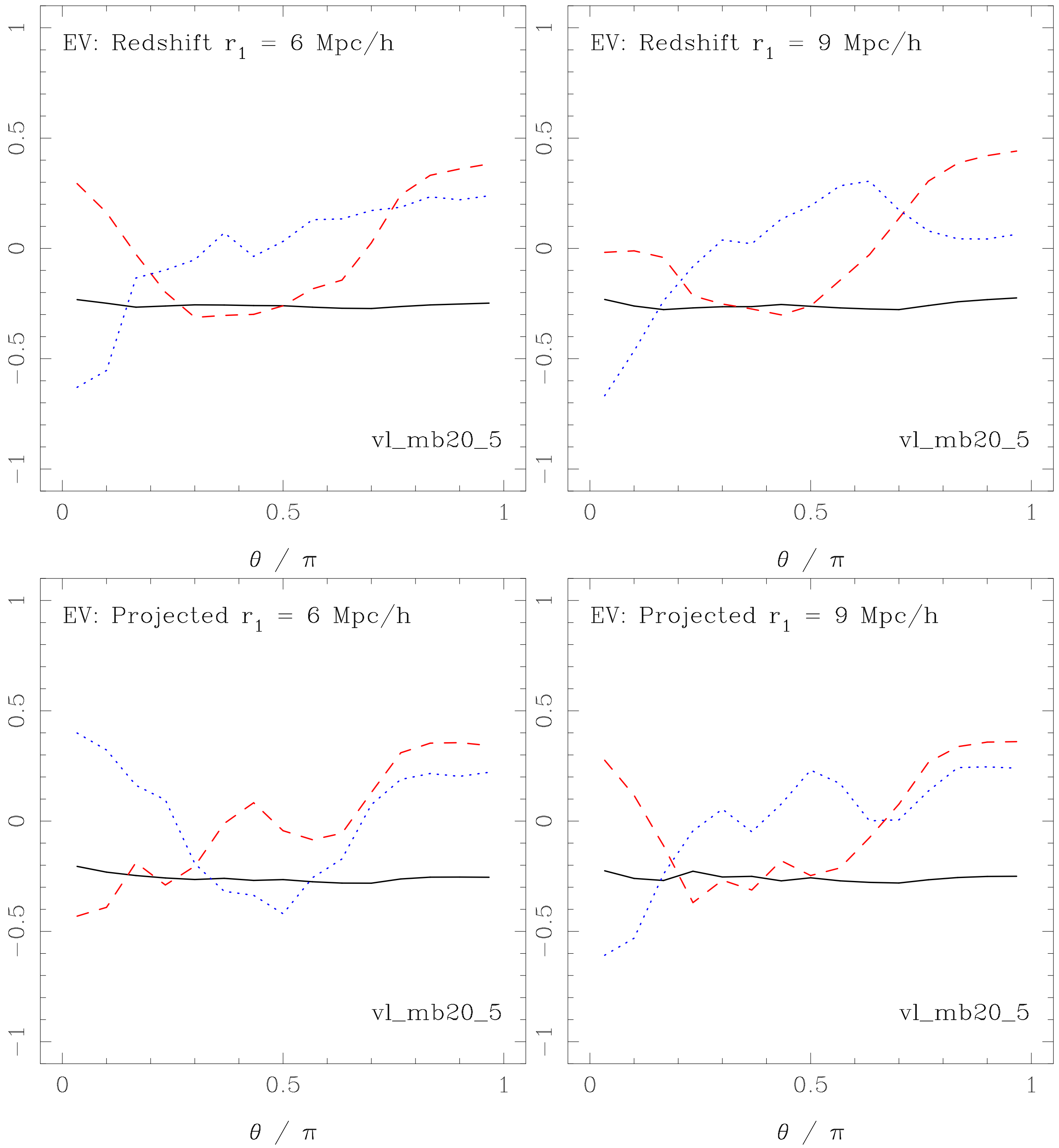


Figure 5.2 Like Figure 5.1 but for the $-21.5 < M_r < -20.5$ galaxy sample. The top three eigenvectors (EVs) chosen from the normalized covariance matrix. The sign of the EV is arbitrary. The first EV (solid black) shows equal weights for all bins. The second (dashed red) and third (dotted blue) EV display the configuration difference between perpendicular and co-linear triangles as well as the scale variation as the scale of the third side increases.

We determine the theoretical model, $Q^{(t)}$, by scaling the mass measurement from the HV simulation, Q_m, with bias parameters B and C as per (1.12). We evaluate $\mathcal{L}$ on a grid using the ranges: $B = 0.1 \ldots 3.0$ and $C = -1.5 \ldots 1.5$ with a resolution of 0.01. We tested for discrepancies using a factor of 10 finer spacing between grid elements with no significant differences to the fitted results.

We remove higher eigenmodes from the fits where the SV is at or below a value we expect for Poisson noise by applying the criteria in (5.4). We investigate the effects of using a different number of modes in the next chapter which discusses systematics (Chapter 6). For each sample, we perform six independent fits: a series of three different scales for measurements in of redshift and projected space. We use the full configuration dependence for triangles with $r_1 = 6$ and 9 h^{-1}Mpc, as well as a joint fit using both scales. For the joint fit, we estimate the full combined covariance matrix to correctly account for overlap and correlation and use the same eigenmode analysis. This changes the number of available modes from 15 in the individual fits to 30 modes for the combined joint fit.

We first examine the BRIGHT DR6 sample ($M_r < -21.5$), with the likelihood space of the six 2-parameter fits displayed in Figure 5.3. We include contours for Gaussian $1, 2$ and 3σ levels which identify regions of probabilities for $68.3, 95.5$ and 99.7%. We calculate these from the $\Delta\chi^2$ distribution for a 2-parameter fit (i.e. two degrees-of-freedom). We include two reference points for comparison, the *unbiased* result where ($B = 1$,$C = 0$) along with the a potential negative quadratic bias term accounting for the entire galaxy bias ($B = 1$,$C = -0.3$).

We note a few interesting characteristics in Figure 5.3. The degeneracy between B and C is clearly visible as the elongated diagonal contour. Larger values of B remain likely with larger values of C, consistent with our expectation of degeneracy by inspecting the bias relation in (1.12). The size of the errors are notably larger for projected space measurements, as well as lower values for the overall S/N. This results from the larger uncertainties in the projected measurements, as we noted in the previous chapter. Since the scale r_p represents a projection that incorporates larger scales (determined by the line-of-sight integration π_{max}), projected measurements are more sensitive to the dominant uncertainty from cosmic variance that increases with scale. In all cases, the *unbiased* ($B = 1$,$C = 0$) model is ruled out at greater than a 2σ level. To see the success of the fit "by eye", we plot the 3PCF for dark matter, galaxies and best fit scaled model for this sample in Figure 5.4. Both the "individual" fits and "combined" joint fit produce models that well match the data.

We also constrain galaxy-mass bias parameters for the LSTAR sample ($-21.5 < M_r < -20.5$). This sample spans a unit bin in magnitude, and consists only of galaxies fainter than the previous bright sample. The results of the fit with likelihood contours are shown in Figure 5.5. The uncertainties appear reduced in size – a striking difference with respect to the BRIGHT sample in Figure 5.3. In addition, the slope of the "line of degeneracy" between B and C has shifted. We reason this is in part due to the increased statistical significance of the larger sample, as both the measurements and covariance are better resolved. Due to the higher number density of galaxies, we re-measured the 3PCF using a finer binning scheme (fractional bin-width of $f = 0.1$ as opposed to $f = 0.25$). With the finer binning, we see a stronger configuration dependence, which will alter the degeneracy between B and C. We note that many of the best fit B values appear smaller, which we expect for a fainter sample. The same line of reasoning suggests that the "unbiased" model ($B = 1$,$C = 0$) should be more likely to fit. As before, we plot the respective best fit model in comparison with the dark matter and galaxy 3PCF in Figure 5.6. There is a smaller difference between HV (mass) and galaxy measurements, as this sample is fainter. We notice some noise of the HV measurement for Q_{proj}, making the model not quite as smooth. We note that by eye, Q_z on larger scales indicates a slight bias for the "combined fit", with the model undershooting the data and 1σ uncertainties. Significant off-diagonal structure in the covariance matrix can produce a fit where "chi-by-eye" doesn't work. Since the $r_1 = 6\ h^{-1}$Mpc triangle in Q_z has much smaller errors, it drives the fit making the $r_1 = 9\ h^{-1}$Mpc seem the poorest fit.

The results of the two parameter galaxy-mass bias fits for the BRIGHT and LSTAR samples are summarized in Table 5.1. The magnitudes represented in the BRIGHT sample ($M_r < -21.5$) are at least a magnitude larger than $L_* \sim -20.4$ (Blanton et al., 2003b). We typically consider L_* galaxies to have a linear bias where $B \sim 1$, and we might expect this brighter sample to have a larger B value. The constraints from projected measurements appear to follow this logic. In agreement, the best fit values on Q_{proj} in fainter LSTAR sample ($-21.5 < M_r < -20.5$) are lower with $B \sim 1$. Redshift space measurements, Q_z,

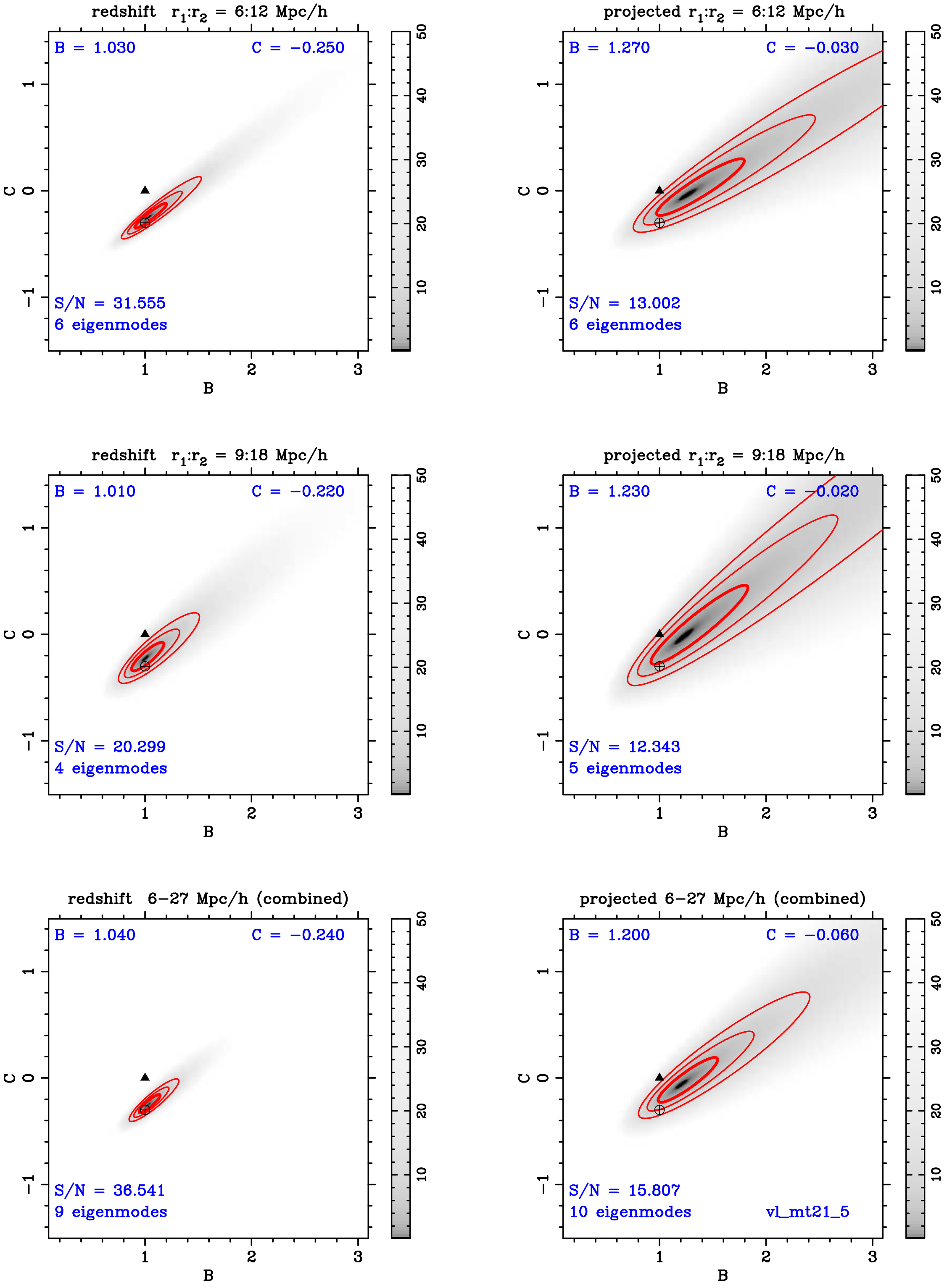


Figure 5.3 Constraints on the galaxy-mass bias parameters using the $M_r < -21.5$ galaxy sample and the HV simulation. The left column corresponds to analysis in redshift space (Q_z) with the right column showing projected space (Q_{proj}). The top two panels of each column represent individual fits to the triangles of $r_1 = 6$ and 9 h^{-1}Mpc as indicated. The bottom panel is a joint fit between both scales. There are two points of comparison marked: the unbiased result where ($B = 1.0; C = 0$) and ($B = 1.0; C = -0.3$). The contours denote the $1, 2$ and 3σ levels from the $\Delta\chi^2$ distribution of two parameters.

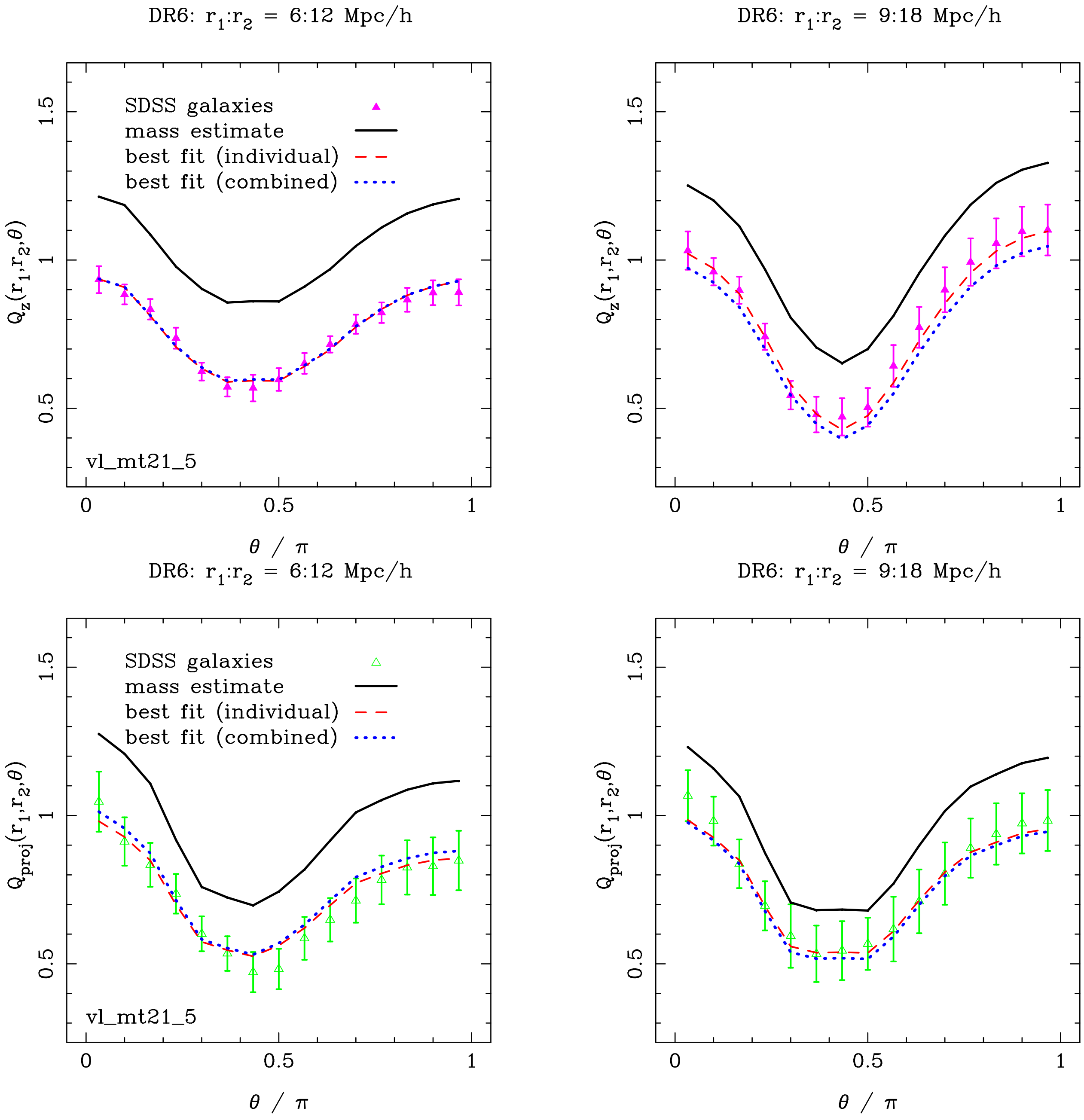


Figure 5.4 The reduced 3PCF for the $M_r < -21.5$ sample showing the mass scaled to the "best fit" galaxy-mass bias parameters. The top two panels correspond to redshift space, and the bottom two to projected space. From left to right, the scale of the triangle increases as noted. The red (dashed) line represents an individual fit only to that triangle scale, and the blue (dotted) line shows a joint fit between both scales.

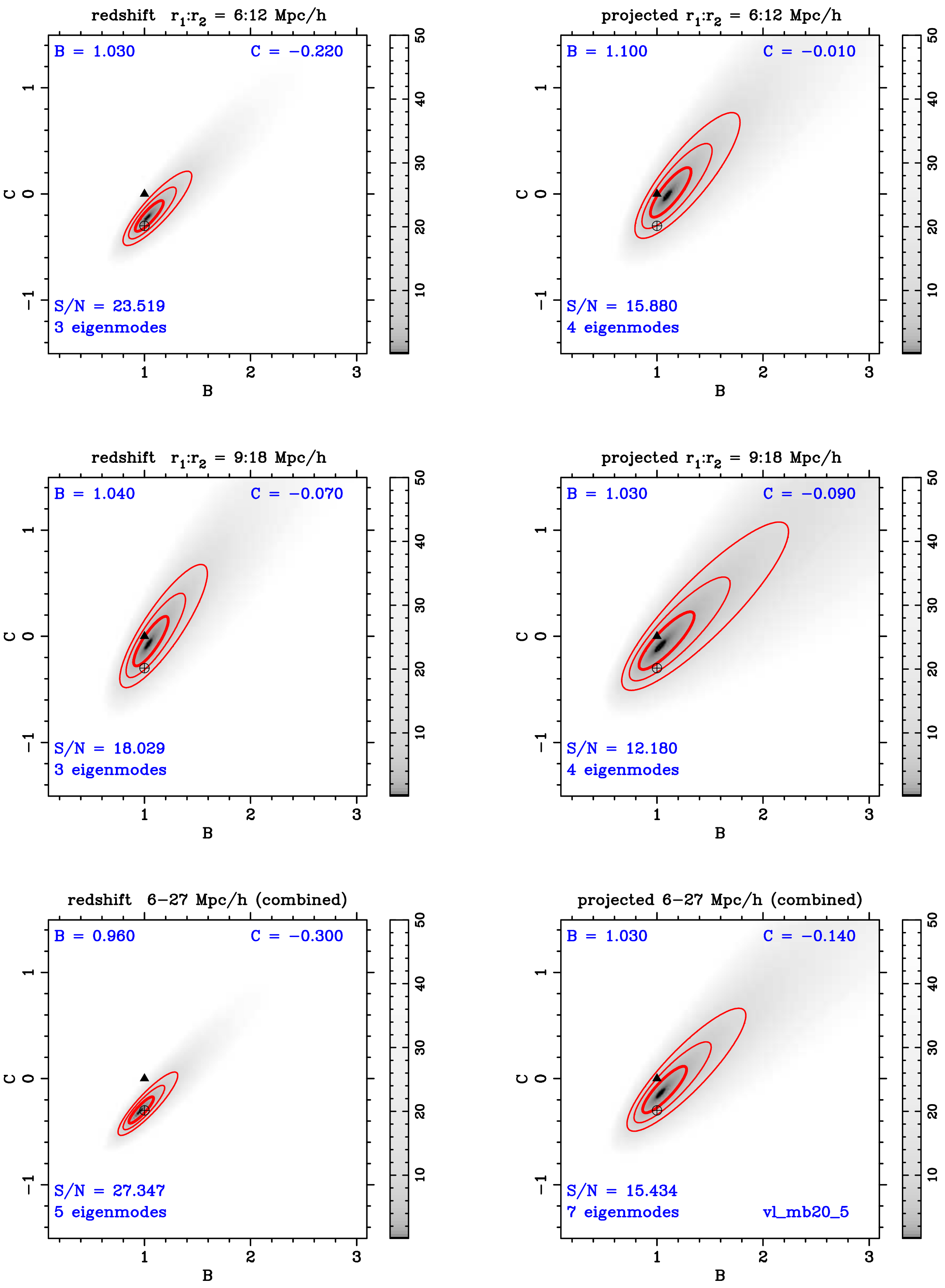


Figure 5.5 Analogous to Figure 5.3, but for $-21.5 < M_r < -20.5$ galaxies. Constraints on the galaxy-mass bias parameters with respect to the HV simulation. The left column corresponds to analysis in redshift space (Q_z) with the right being projected space (Q_{proj}). The top two panels of each column represent individual fits to the triangles of $r_1 = 6$ and 9 h^{-1}Mpc as indicated. The bottom panel is a joint fit between both scales. There are two points of comparison marked: the unbiased result where ($B = 1.0; C = 0$) and ($B = 1.0; C = -0.3$). The contours denote the $1, 2$ and 3σ levels from the $\Delta\chi^2$ distribution of two parameters.

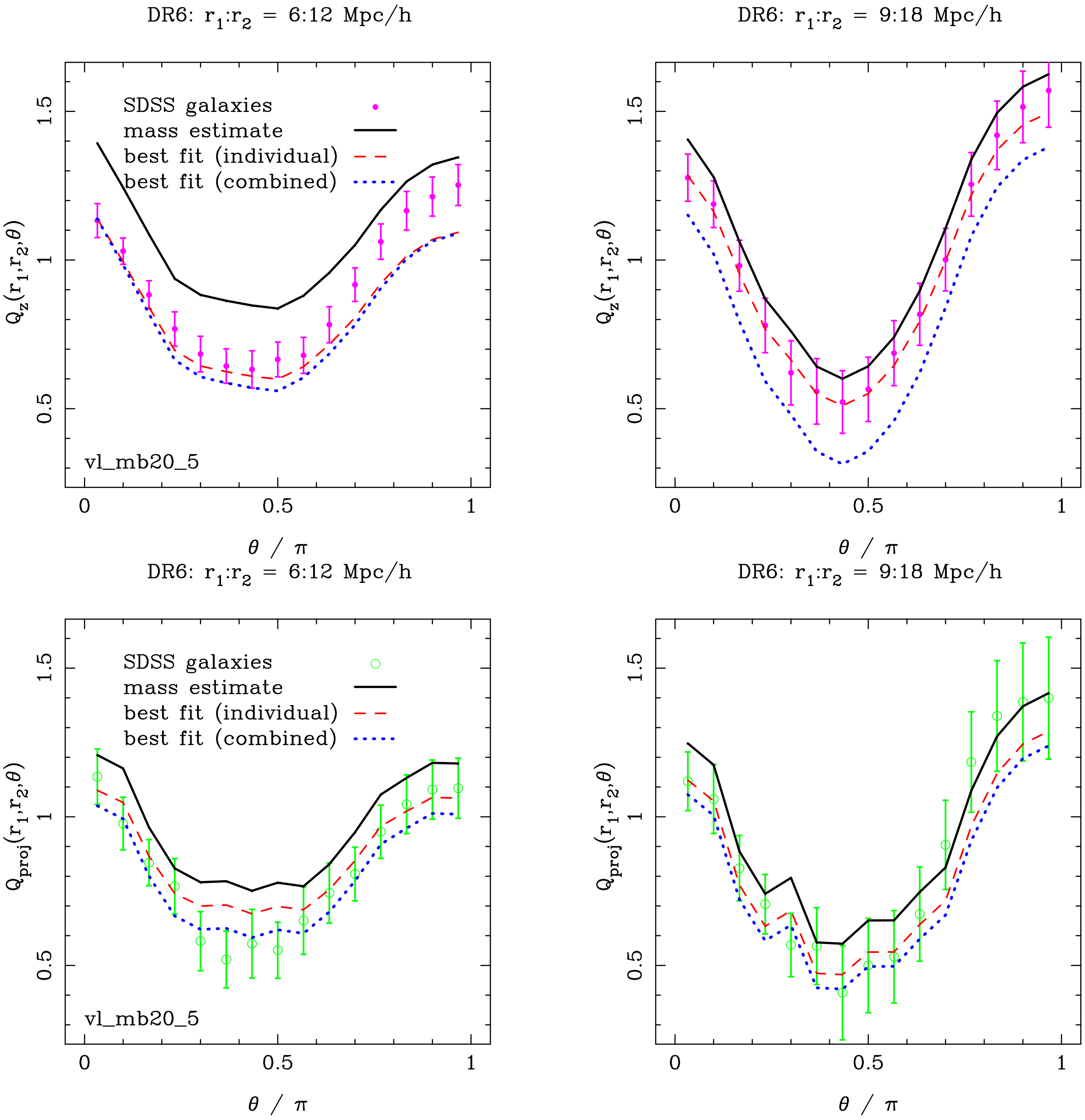


Figure 5.6 Like Figure 5.4 but for the $-21.5 < M_r < -20.5$ sample. The reduced 3PCF showing the mass scaled to the "best fit" galaxy-mass bias parameters. The top two panels correspond to redshift space, and the bottom two to projected space. From left to right, the scale of the triangle increases as noted. The red (dashed) line represents an individual fit only to that triangle scale, and the blue (dotted) line shows a joint fit between both scales.

appear consistent with $B \sim 1$ for all fits, but at the same time values of C are lower, which is the degeneracy of the B and C parameters. The reduced χ^2_ν show an acceptable fit in almost all cases; the exceptions are the two Q_z fits using the $r_1 = 6\ h^{-1}$Mpc triangles for the LSTAR sample. Consequently, the joint fit is the one that looks poor in Figure 5.6. The $\Delta\chi^2$ in Table 5.1 displays the likelihood an unbiased model is from the best fit parameters. We find an unbiased model is ruled out for the BRIGHT galaxy sample at greater than 4.8σ in redshift space and 2.6σ in projected. We cannot conclude the same for the LSTAR sample, which is largely consistent with an unbiased model. We generally consider BRIGHT galaxies to be "more" biased (Zehavi et al., 2005), and the LSTAR sample, being a magnitude bin around L_*, is likely in accordance with expected bias values.

5.4.1 Non-zero Quadratic Bias?

With our two parameter likelihood space, we can ask about the statistical significance of a non-zero quadratic bias term (b_2 which is encapsulated in $C = b_2/b_1$). We use the same configuration dependence of the 3PCF, and the covariance – but restrict the two parameter fit such that $C = 0$. We evaluate the best fit B, the quality of the fit (via the reduced χ^2_ν) as well as the $\Delta\chi^2$ for the best two-parameter fit, which we present in Table 5.2. For the BRIGHT sample, we notice the B values are equivalent across both Q_{proj} and Q_z for the same scales on the same sample. Since we removed the degeneracy (as C is zero), this behavior makes sense and in agreement with the measurements. We also note that the typical B values are larger for the BRIGHT sample ($M_r < -21.5$), and lower for the fainter LSTAR sample ($-21.5 < M_r < -20.5$). For Q_{proj} our constraints find little statistical significance for a non-zero quadratic bias term, the likelihood difference is small and a linear bias term is sufficient to quantify the bias for both the BRIGHT and LSTAR samples. Overall, measurements in redshift space (Q_z) more strongly suggest that $C \neq 0$, especially when using the smaller scale triangles ($r_1 = 6\ h^{-1}$Mpc).

Galaxy-Mass Bias Parameters from SDSS

Sample	Scales (h^{-1}Mpc)	B	C	χ^2_ν	D.o.F.	unbiased $\Delta\chi^2$
BRIGHT-z	6-18	$1.03^{+0.11}_{-0.08}$	$-0.25^{+0.08}_{-0.06}$	1.48	6-2	118.86 (10.7σ)
BRIGHT-proj	6-18	$1.27^{+0.30}_{-0.21}$	$-0.03^{+0.19}_{-0.14}$	0.78	6-2	16.43 (3.6σ)
BRIGHT-z	6-27	$1.04^{+0.06}_{-0.06}$	$-0.24^{+0.05}_{-0.05}$	0.83	9-2	132.54 (11.3σ)
BRIGHT-proj	6-27	$1.20^{+0.21}_{-0.14}$	$-0.06^{+0.15}_{-0.11}$	0.45	10-2	18.70 (3.9σ)
BRIGHT-z	9-27	$1.01^{+0.10}_{-0.09}$	$-0.22^{+0.09}_{-0.08}$	0.60	4-2	26.90 (4.8σ)
BRIGHT-proj	9-27	$1.23^{+0.34}_{-0.22}$	$-0.02^{+0.27}_{-0.18}$	0.34	5-2	9.44 (2.6σ)
LSTAR-z	6-18	$1.03^{+0.09}_{-0.07}$	$-0.22^{+0.10}_{-0.08}$	13.47	3-2	28.07 (4.9σ)
LSTAR-proj	6-18	$1.10^{+0.13}_{-0.11}$	$-0.01^{+0.16}_{-0.14}$	0.85	4-2	3.00 (1.2σ)
LSTAR-z	6-27	$0.96^{+0.08}_{-0.07}$	$-0.30^{+0.08}_{-0.08}$	3.22	5-2	45.85 (6.5σ)
LSTAR-proj	6-27	$1.03^{+0.15}_{-0.11}$	$-0.14^{+0.16}_{-0.12}$	1.07	7-2	5.86 (1.9σ)
LSTAR-z	9-27	$1.04^{+0.11}_{-0.09}$	$-0.07^{+0.16}_{-0.14}$	0.07	3-2	2.37 (1.0σ)
LSTAR-proj	9-27	$1.03^{+0.19}_{-0.13}$	$-0.09^{+0.19}_{-0.15}$	1.75	4-2	1.93 (0.9σ)

Table 5.1 The two-parameter best fits for galaxy-mass bias, using (1.12) with the configuration dependence in the reduced 3PCF from SDSS DR6 galaxy samples in comparison with dark matter clustering from the Hubble volume simulation. Fits are performed separately on two galaxy samples BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$) using measurements in redshift space (denoted with "z") as well as projected space ("proj"). The second column lists the range of scales used for the respective fit. The errors are marginalized 1σ bounds calculated by the range within $\Delta\chi^2 \leq 1$ from the best fit value. The quality of the best fit value is stated with the reduced chi-square $\chi^2_\nu = \chi^2/\text{D.o.F.}$ The degrees of freedom (D.o.F.) correspond to the number of eigenmodes used minus the number of parameters (2). The last column lists the $\Delta\chi^2$ value to quantify the likelihood of an "unbiased" model matching the data, i.e. ($B = 1;C = 0$), with a likelihood expressed in the number of σ from the standard Gaussian assumption for the $\Delta\chi^2$ distribution.

Galaxy-Mass Bias without Quadratic Term

Sample	Scales (h^{-1}Mpc)	B	χ^2_ν	D.o.F.	$\Delta\chi^2$ from best fit
BRIGHT-z	6-18	$1.34^{+0.04}_{-0.04}$	2.42	6-1	6.16 (2.0σ)
BRIGHT-proj	6-18	$1.31^{+0.11}_{-0.09}$	0.63	6-1	0.04 (0.0σ)
BRIGHT-z	6-27	$1.30^{+0.03}_{-0.03}$	2.23	9-1	12.01 (3.0σ)
BRIGHT-proj	6-27	$1.27^{+0.09}_{-0.07}$	0.42	10-1	0.16 (0.1σ)
BRIGHT-z	9-27	$1.24^{+0.06}_{-0.06}$	1.85	4-1	4.35 (1.6σ)
BRIGHT-proj	9-27	$1.25^{+0.11}_{-0.09}$	0.26	5-1	0.01 (0.0σ)
LSTAR-z	6-18	$1.21^{+0.05}_{-0.04}$	8.68	3-1	3.90 (1.5σ)
LSTAR-proj	6-18	$1.11^{+0.07}_{-0.06}$	0.57	4-1	0.01 (0.0σ)
LSTAR-z	6-27	$1.23^{+0.04}_{-0.04}$	4.59	5-1	8.70 (2.5σ)
LSTAR-proj	6-27	$1.15^{+0.07}_{-0.07}$	1.02	7-1	0.76 (0.4σ)
LSTAR-z	9-27	$1.08^{+0.06}_{-0.05}$	0.14	3-1	0.21 (0.1σ)
LSTAR-proj	9-27	$1.11^{+0.09}_{-0.08}$	1.25	4-1	0.25 (0.1σ)

Table 5.2 The single-parameter best fits for galaxy-mass bias using (1.12) where we constrain $C = 0$. We fit the configuration dependence of the reduced 3PCF from SDSS DR6 galaxy samples in comparison with dark matter clustering from the Hubble volume simulation. Fits are performed separately on two galaxy samples BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$) using measurements in redshift space (denoted with "z") as well as projected space ("proj"). The second column lists the range of scales used for the respective fit. The errors are marginalized 1σ bounds calculated by the range within $\Delta\chi^2 \leq 1$ from the best fit value. The quality of the best fit value is stated with the reduced chi-square $\chi^2_\nu = \chi^2/\text{D.o.F.}$ The degrees of freedom (D.o.F.) correspond to the number of eigenmodes used minus the number of parameters (in this case, just one). The last column lists the $\Delta\chi^2$ value to quantify the difference in likelihood of this model with $C = 0$ compared with the best fit of a two-parameter fit (i.e. Table 5.1.

5.4.2 Implications for Cosmology: σ_8

Better understanding galaxy-mass bias, or at the least accurately parameterizing it, allows us to "calibrate" out the effects of galaxies and infer properties of the underlying mass distribution to constrain cosmology. A cosmological parameter of interest is the variance of mass in spheres of a certain radius, linearly extrapolated from a very early epoch until today. We commonly use this variance to normalize the amplitude of the matter power spectrum, $P(k)$, the Fourier transform of the 2PCF discussed in the Chapter 1. A somewhat arbitrary but historical choice is to define this for spheres of 8 h^{-1}Mpc in radius, making the definition of this variance:

$$\sigma_8^2 = 4\pi \int_0^\infty W^2(k, R = 8\ h^{-1}\text{Mpc}) P_{lin}(k) \frac{k^2 dk}{(2\pi)^3}\ , \tag{5.6}$$

where $W(k, R)$ is a top-hat window function in Fourier space for mode k and smoothing radius R and $P_{lin}(k)$ is the linear power spectrum.

In terms of our fitting formula on the 3PCF in (1.12), we expand the bias relation for the 2PCF to highlight its dependence on σ_8

$$\xi_g(r) = B^2 \left(\frac{\sigma_8}{0.9}\right)^2 \xi_m(r)\ . \tag{5.7}$$

Formally, the mass 2PCF already encodes a value of σ_8. As ξ_m scales linearly with a change in the square of σ_8, we include an explicit scaling factor to account for a difference in σ_8 between the underlying mass of the observed galaxy distribution and that assumed in our estimate of mass clustering from N-body results. In our case, we use dark matter from the HV simulation where $\sigma_8 = 0.9$, explaining the denominator on the right hand side of (5.7). We can see that an incorrect assumption of σ_8 in the estimate of mass will directly translate into a different value of the best B describing galaxies. Even if we use the above relation in (5.7), the two parameters are completely degenerate when solely considering the 2PCF.

Using the information in higher order correlations, specifically the reduced 3PCF (Q), enables us to break this degeneracy between B and σ_8. The "ratio statistic" of the reduced 3PCF is predominantly insensitive to cosmology, including the specific value of σ_8 (Bernardeau et al., 2002) and the fitting formula for 3PCF bias in (1.12) remains unchanged.

Implied values of σ_8			
Sample	Scales (h^{-1}Mpc)	B	σ_8
BRIGHT-z	9-27	$1.24^{+0.06}_{-0.06}$	0.96-1.13
BRIGHT-proj	9-27	$1.25^{+0.11}_{-0.09}$	1.02-1.12
LSTAR-z	9-27	$1.08^{+0.06}_{-0.05}$	0.88-0.97
LSTAR-proj	9-27	$1.11^{+0.09}_{-0.08}$	0.83-0.97

Table 5.3 We use galaxy-mass bias constraints from the configuration dependence of the 3PCF, $Q(\theta)$, with measurements of the 2PCF to estimate the implied values of σ_8, as per (5.7). We use the largest triangle configurations for our two samples, and the 1-parameters constraints on B. The range of σ_8 does not represent formal uncertainties, rather we calculate values from the range of uncertainties stated in B, neglecting additional errors from the 2PCF. For reference, WMAP-5 (with SN and BAO) suggest $\sigma_8 = 0.82$ (Komatsu et al., 2009).

Formally, this is only true to leading order, as loop corrections to Q will add cosmological dependence; we neglect these higher order corrections for our analysis.

Constraining B from $Q(\theta)$ yields a value to use in (5.7) resulting in an estimate of σ_8 from measurements of the 2PCF. Ideally, we could construct a three-parameter fit to jointly constrain B, C and σ_8 (e.g. Pan and Szapudi, 2005, on 2dFGRS data). Or as an further extension, we could jointly fit over several samples, since they each have the same underlying σ_8. However, this additional complexity is beyond the scope of this analysis as our current uncertainties would yield poor constraints on σ_8. However, we can crudely estimate the value of σ_8 implied by best fit bias parameters. We restrict this estimate to the largest scale triangles ($r_1 = 9$ h^{-1}Mpc) to ensure we approach the linear regime (i.e. the scales we are most confident with using the local bias model). We present these projections in Table 5.3.

We estimate the values of σ_8 to be between 0.83 and 1.13 based on the BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$) galaxy samples. The values we obtain are contingent on a specific model of mass clustering, where we have chosen to use N-body simulations (specifically the HV), and redshift distortions (which we include through velocity information to distort particle positions in the HV simulation). For comparison, constraints of σ_8 from a joint analysis of the cosmic microwave background (CMB), supernova data (SN) and baryon acoustic oscillations (BAO) find $\sigma_8 = 0.82$ (Komatsu et al., 2009). Our lower values are in good agreement with this value. The high end values are significantly above, but our results are in reasonable agreement with an analysis of a related statistic, the monopole moment of the 3PCF, where they find best fit σ_8 values between 0.9 and 1.07 (see Table 3 in Pan and Szapudi, 2005) using 2dFGRS galaxies (Colless et al., 2001). Please note, although the value of σ_8 is comparable with 2dFGRS, the specific values of bias are not since the galaxy sample is different. If we underestimate the value of linear bias, effectively B here, (5.7) shows that the implied value of σ_8 will be overestimated. This might help explain the larger values of our estimates in comparison to WMAP analyses. Our projections for σ_8 use clustering measurements between $9 - 27\ h^{-1}$Mpc and exploit the configuration dependence of $Q(\theta)$, which is significantly different that the monopole measurement (many more scales, no configuration dependence) and WMAP results (huge amount of combined data).

5.4.3 Discussion

We investigated the constraints of galaxy-mass bias parameters in the local bias model, using the configuration dependence of the 3PCF in redshift and projected space. We conclude that galaxies are *biased* tracers of mass in both samples, with brighter galaxies corresponding to increased *bias*. These results are consistent with detailed analysis of SDSS galaxies from the 2PCF (Zehavi et al., 2005) where they quantify how bias increases clustering for brighter galaxy samples. Our results indicate that a linear bias model yields reasonable approximations to the observations, in agreement with (Hikage et al., 2005). However, non-linear bias produces slightly better agreement. We notice a strong correlation between linear and quadratic bias, as expected from inspection of (1.12), and consistent with measurements of SDSS galaxies using the bispectrum (Nishimichi et al., 2007). We find that our redshift space measurements predict significantly negative quadratic bias with a linear bias near one. This was seen in a very similar analysis conducted on 2dFGRS galaxies (Gaztañaga et al., 2005). Interestingly, projected measurements suggest a larger linear bias with near zero quadratic bias for the same samples – but this is not confirmed given the large uncertainties. We obtain reasonable projections for σ_8 with these samples that are comparable to higher order constraints on the 2dFGRS (Pan and Szapudi, 2005). We find increased linear bias values are preferred to match σ_8 to predictions from joint analysis of WMAP-5, SN and BAO measurements (Komatsu et al., 2009).

The local bias model is a simple way to characterize galaxy-mass bias. Alternative descriptions exist based on the halo model (reviewed in Cooray and Sheth, 2002), which form phenomenological models with a wider range of parameters. Two well used formulations include the halo occupation distribution (HOD; Berlind and Weinberg, 2002) and the conditional luminosity function (CLF; Yang et al., 2003; van den Bosch et al., 2003). Analyses using these models have characterized SDSS data, e.g. see Tinker et al. (2005); Zheng et al. (2007) for recent HOD results and Cacciato et al. (2009) for the CLF, but both predominantly use two-point statistics. There are formulations for the 3PCF, however, their accuracy is not as well determined when compared with data (see Takada and Jain, 2003; Wang et al., 2004; Fosalba et al., 2005). A significant advantage of the HOD method is the ability to use well determined measurements of the small scales for constraints (non-linear regime in gravitational perturbation theory). Understanding the projected 3PCF Q_{proj}, a major component of this work, provides a critical link to obtain reliable measurements at these smaller scales from observational galaxy samples. Constraining local bias parameters allows an investigation in a regime where redshift and projected measurements can be compared, and the implications for bias and cosmology are better studied for higher order moments.

5.5 RELATIVE BIAS

The *relative* bias characterizes the relative clustering strength between different galaxy samples – an alternative to the "absolute" galaxy-mass bias of the previous section. Relative bias is insensitive to σ_8 and does not require assumptions to determine mass clustering. We can use the relative bias to check consistency with linear and quadratic bias parameters obtained above in section 5.4. For the 2PCF, the relative bias is simply:

$$b_{rel}^{(2)} = \sqrt{\frac{\xi_{BRIGHT}}{\xi_{LSTAR}}}\,, \tag{5.8}$$

where ξ can refer to redshift or projected space measurements.

We show $b_{rel}^{(2)}$ from the 2PCF in Figure 5.7, using the linear bias parameters obtained from the best two-parameter fit (i.e. Table 5.1). Both projected and redshift space measurements agree and produce a flat relative bias, even down to non-linear scales below a few h^{-1}Mpc. Two obvious discrepancies arise when comparing observational data to "best fit" values. First, neither redshift nor projected space appear to match data. Earlier we noted a substantial degeneracy between the linear and quadratic bias terms. The quadratic bias term is accounting for more of the clustering bias when we constrain with $Q(\theta)$, which isn't noticeable in the 2PCF. This suggests we underpredict values of linear bias, either just for the BRIGHT sample or in unequal portions for both. Second, there is a significant difference between these two estimates given the same galaxy samples, although the projected measurement appears closer to agreement.

It might appear that our best fit bias parameters do not describe the relative bias of the data – but let us consider the relative bias of the reduced 3PCF. Since Q is proportional to $1/B$, the relation becomes

$$b_{rel}^{(3)} = \frac{Q_{LSTAR}}{Q_{BRIGHT}} \; . \tag{5.9}$$

Figure 5.8 presents the relative bias of our DR6 galaxies for the equilateral 3PCF (Q_{eq}). An obvious difference with respect to the 2PCF is the much larger uncertainties. We can see the predicted $b_{rel}^{(3)}$ from the galaxy-mass bias constraints look much better. They agree with the observational data, and show a much smaller discrepancy between redshift and projected space. The quadratic bias term (C) properly accounts for the clustering difference that was missing in the relative bias of the 2PCF. We note, the relative bias for 3PCF uses equilateral triangles, and not the specific $Q(\theta)$ measurements used to fit the galaxy-mass bias.

We conclude that the brighter galaxy sample is a *more biased* realization from the relative bias of both the 2PCF and 3PCF, consistent with other analyses (Zehavi et al., 2002, 2005). Relative bias provides a consistency check on the "absolute" galaxy-mass bias parameters we constrain, suggesting a combination of linear and quadratic bias terms are consistent with observations. However, looking at the 2PCF alone suggests that our two parameter bias model fits underpredict the value of linear bias necessary to explain the observations. Finally, we see a hint that constraints from projected measurements appear to be less affected – although we caution that this trend has weak statistical significance given the larger

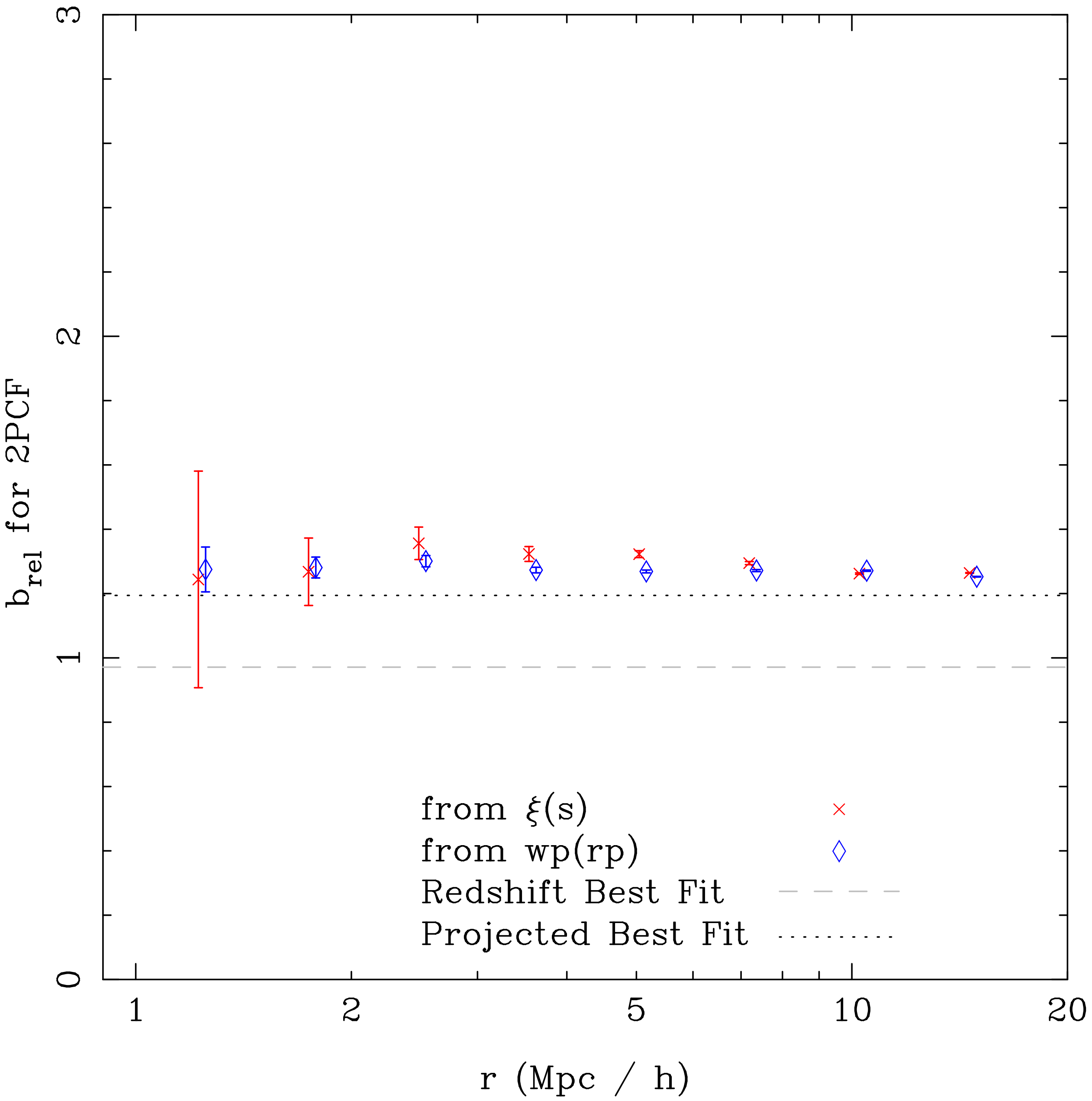


Figure 5.7 The relative bias $b_{rel}^{(2)} = \sqrt{\xi_{BRIGHT}/\xi_{LSTAR}}$ using measurements in redshift (red 'x' symbols) and projected space (blue diamonds). The uncertainties are propagated from the 1σ values from the 2PCF. The dotted and dashed line display results from our the linear bias terms in our two-parameter fit in Table 5.1 at the largest scales ($9 - 27h^{-1}$Mpc).

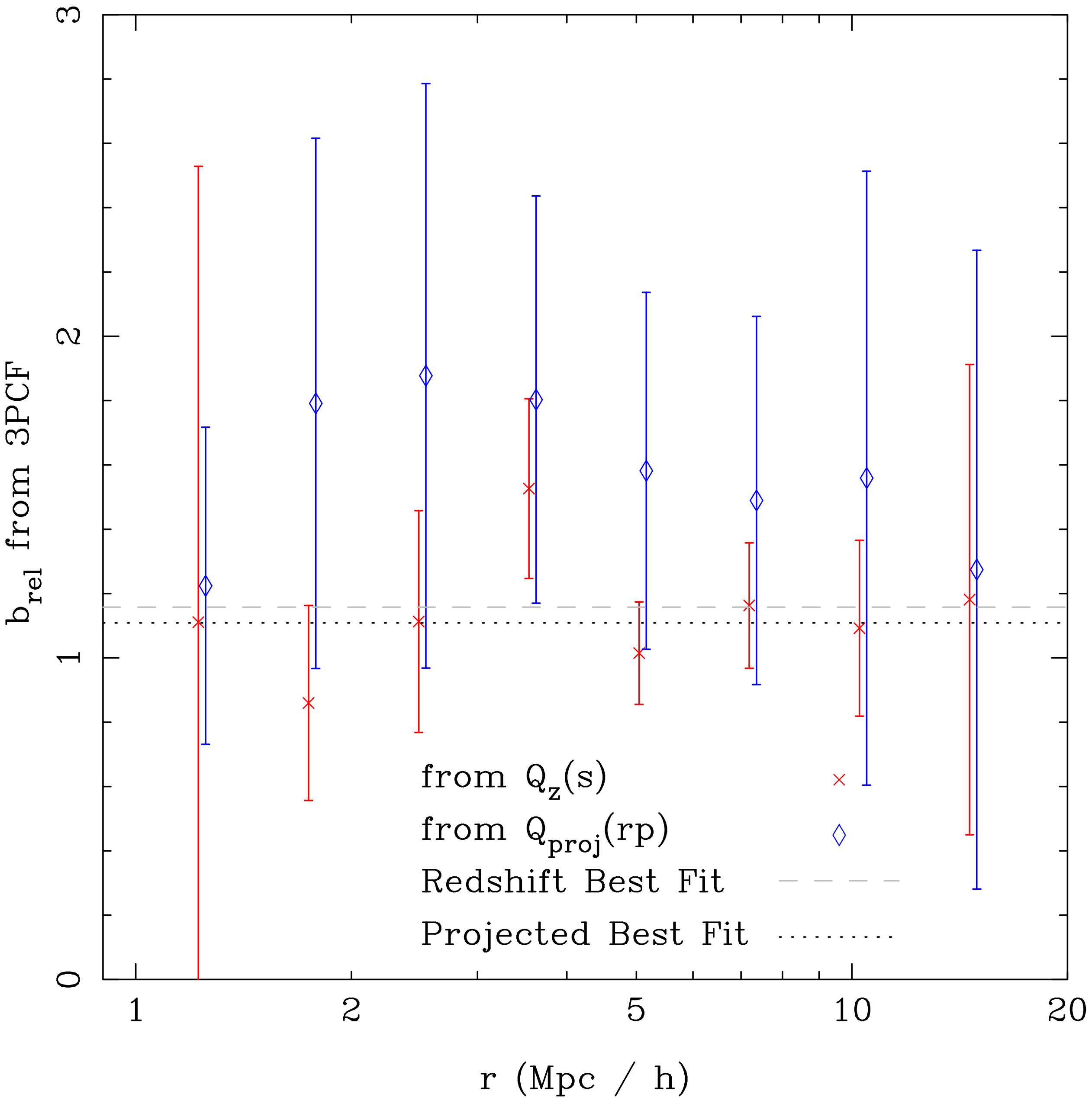


Figure 5.8 The relative bias $b_{rel}^{(3)} = Q_{LSTAR}/Q_{BRIGHT}$ using measurements of equilateral 3PCF in redshift (red 'x' symbols) and projected space (blue diamonds). The uncertainties are propagated from the 1σ values from the 3PCF. The dotted and dashed line display the best fit bias estimated from both linear and quadratic terms from the two-parameter fit in Table 5.1 at the largest scales ($9 - 27h^{-1}$Mpc).

uncertainties in projected space.

5.6 SUMMARY

In this chapter, we investigate the galaxy-mass bias model to quantify the clustering difference between galaxy samples and mass predictions from the Hubble Volume (HV) simulation. In §5.3, we investigated the eigenvectors of the full covariance matrix of the reduced 3PCF in two galaxy samples. We found the first three dominant modes show coherent structure consistent with variations seen in the $Q(\theta)$ measurements, supporting our claim that the covariance is signal dominated.

We performed a two parameter maximum likelihood analysis in §5.4 to constrain the linear (B) and quadratic (C) galaxy-mass bias as described by equation (1.12) for two galaxy samples: BRIGHT ($M_r < -21.5$) and LSTAR ($-21.5 < M_r < -20.5$). We resolve the strong degeneracy between B and C, which helps to explain the weak luminosity dependence of the reduced 3PCF. Best fit bias parameters show the brighter galaxy sample (BRIGHT) is a more biased realization than the fainter LSTAR sample. Conservatively using the largest scales we measure ($9 - 27h^{-1}$Mpc), we find the BRIGHT sample is a significantly biased description of mass clustering from dark matter in HV simulations at the 4.5σ level in redshift space and 2.5σ in projected space. Further, in section §5.4.1, our analysis indicated that a linear bias model, i.e. $C = 0$, produced a reasonable approximation to the observations. However, non-linear bias (non-zero C) produced a slightly better agreement. Given the degenerate axis of B and C, negative values of the quadratic bias C yield lower values of B. In §5.4.2, we use the bias parameters constrained from the reduced 3PCF in conjunction with the 2PCF to show the implications for the cosmological parameter σ_8. We find increased linear bias values are preferred to match σ_8 to predictions from joint analysis of WMAP-5, SN and BAO measurements (Komatsu et al., 2009).

We use the relative bias of galaxy samples in §5.5, which are independent of the mass predictions from the HV simulation and cosmological parameters, to evaluate our best fit bias parameters. The reduced 3PCF showed consistency with the best fit values. However, the relative bias of the 2PCF suggested the ratio of linear bias parameters should be larger to match observations (see Figure 5.7). As we noted above, a less significant non-zero bias term (C) would increase the linear bias parameter B in the BRIGHT sample. Taken together, the relative bias and σ_8 predictions implied increased values of B are preferred, which is more consistent with a linear bias model. While a quadratic bias produced a more likely fit, we can not significantly exclude the linear model for the triplet configurations we investigated with the reduced 3PCF.

6.0 SYSTEMATICS

In this chapter, we investigate the importance and sensitivity of our clustering measurements and analysis methods to potential systematics. These are necessary to understand an account for to trust the results presented in previous chapters. Specifically, we discuss the effects of sky completeness from the observational galaxy samples in §6.1, where we show a 12%-20% deviation in the reduced 3PCF if not correctly accounted for. In §6.2, we address the subtle effects of binning on measurements of the reduced 3PCF, the associated covariance, and the implications of binning for fitting galaxy-mass bias parameters. We cover the effectiveness of projected clustering measurements in minimizing redshift distortions in §6.3. Finally in §6.4, we evaluate the quality of our error estimation by comparing jackknife resampling methods to those obtained from mock galaxy catalogs. Overall, we address the importance of properly addressing these effects when doing quantitative constraints from clustering measurements.

6.1 EFFECTS OF SKY COMPLETENESS

Our measurements must take into account the angular sky completeness of the survey. We restrict our analysis to volume-limited galaxy samples, and as such, the radial selection function typically used in flux-limited samples plays no role (it is defined to be unity for all galaxy samples). The sky completeness can vary due to factors such as missing plates (specific regions in the sky), bad spectra and fiber collisions (discussed previously in Chapter 2). The completeness is well characterized using sectors, and calculated by comparing the number of targeted galaxies with the corresponding number of spectra obtained (Blanton et al., 2005b). We correct the estimated correlation functions by calculating a weight to apply to galaxies using the inverse completeness. For the large scales, this weighting *corrects* the clustering strength for regions where you know galaxies exist, and is standard practice for measurements (Zehavi et al., 2002, 2005). In addition, we conservatively restrict the analysis to regions of high completeness with average and median values above 96%.

Sky completeness is sometimes overlooked for large scale clustering analyses, as it is commonly understood to be a small contribution. While this appears true for measurements on the 2PCF in comparable datasets (Zehavi et al., 2002, 2005), there is little reason to believe this will remain so for higher orders. We investigate the effects of neglecting sky completeness on the reduced 3PCF in Figure 6.1. We notice a systematic offset that increases with scale in both redshift and projected measurements. On larger scales ($r_1 = 9\ h^{-1}$Mpc), we find about a 12% deviation in redshift space and almost 20% difference for projected space. We estimate the uncertainties in these measurements from 30 jackknife samples at approximately 12% and 17% respectively, and we conclude that accounting for sky completeness must be included for accurate 3PCF measurements. The systematic offset we find suggests neglecting sky completeness can translate into false best fit values for Galaxy-Mass bias parameters.

6.2 EFFECTS OF BINNING

The choice of binning scheme affects measurements and covariance matrices of the configuration dependence in the 3PCF. To be specific, we refer to the choice of bin-size as well as their spacing by "binning scheme". For the 2PCF, the most common scheme is to use log spaced bins in r to account for the dynamic range of scales and power law dependence of the 2PCF. In this scheme, bins at larger r correspond to larger bin-widths. When we measure $Q(\theta)$, the angular bins are closely packed over a much smaller scale range making a choice of log spaced bins impractical. If the bin-size is too large, we notice an attenuation or averaging out of the configuration dependence. On the other hand, if the bin-size is too small, we do not resolve the measurement nor the covariance. Both the sample size and number density will impact the efficiency of a binning scheme. Even if we find an acceptable choice for one sample, it might show different effects on another sample. This effect remains

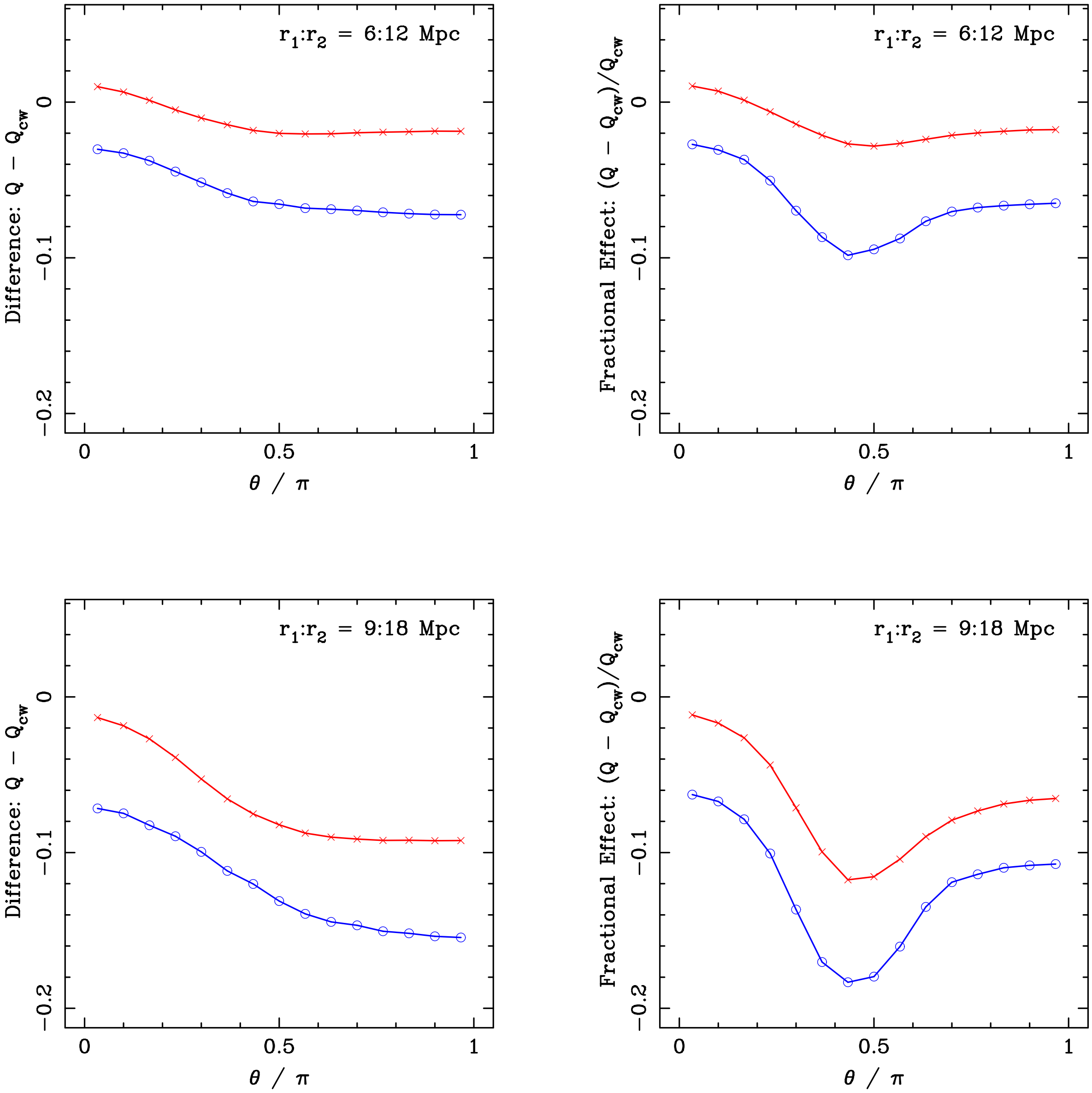


Figure 6.1 The effects of correcting for sky completeness in the reduced 3PCF. The Q_{cw} denotes the completeness weighted (corrected) quantity. The effect for both redshift space (red, x-marks) and projected space (blue, open circles) measurements are described by the absolute difference (left) and the fractional effect (right). Not accounting for sky completeness introduces a systematic offset that varies by scale, which can be seen in the left panel. The projected measurements appear more sensitive to sky completeness than redshift space ones.

more dramatic for the 3PCF due to the larger parameter space of potential configurations.

Naively, for a given data set of size N, we expect the number of available triplets (i.e. N^3) to be much larger than pairs (N^2). However, the configuration dependence of the 3PCF is a function of three dependent variables. Any specific configuration represents a tiny slice through the available data, making the 3PCF much harder to determine. For a rough idea, we can compare the number of triplets to the number of pairs for a specific 3PCF configuration. For a choice of binning scheme with *no* overlap for the $r_1 = 9h^{-1}$Mpc triangles, we find that approximately 1 in every 1000 pairs contribute to the 2PCF for the denominator of $Q_{proj}(\theta)$. For the 3PCF (the numerator), only 1 in $67,000,000$ contribute. Even if we consider that the 3PCF scales as the square of the 2PCF (i.e. $\zeta \propto \xi^2$), we notice the triplet count in the 3PCF remains smaller by a factor of ~ 67.

The impact of bin-size has only recently been addressed in the literature (see Gaztañaga and Scoccimarro, 2005). We believe this remained unresolved for two main reasons (1) the availability of large datasets to statistically determine finely binned higher order moments, and (2) the computational complexity of calculating them. For a small data sample, it may not be possible to measure the 3PCF with small enough bins. This is not the case with our SDSS samples. Most methods of estimation solve the computational complexity by performing counts after pre-gridding the data, thereby imposing a bin-size effect. While this can help mask the effect, it does not hide it entirely. Gaztañaga and Scoccimarro (2005) comment on the effect of bin-size using a pre-gridding technique by using a sufficiently fine grid. Our estimation of correlation functions uses an efficient counting algorithm to yield exact counts, so we do not need to pre-grid our data. At large scales, measurements using a sufficiently fine pre-gridding technique and those with an exact count converge. At small scales, pre-gridding becomes less effective at reducing the computational expense and may become prohibitively expensive.

Gaztañaga and Scoccimarro (2005) suggest a good test of bin-size is to measure the small scale configuration dependence of the 3PCF in redshift space. Redshift distortions produce elongated finger-of-god (FoG) structures, which should significantly amplify the signal of collapsed triangles at $\theta = 0$ and $\theta = \pi$. This characteristic "U-shape" will not be present if the bin-size is too large, since the FoG structures get effectively averaged over. For the

3PCF, we have to consider the bin-size in three parameters, i.e. each side of the triplet. In figure 6.2, we show three choices of bin-size in θ when two sides of the triangle (r_1, r_2) are tightly constrained ($\pm 0.1 h^{-1}$Mpc). Even with large bins in θ, all measurements equivalently show configuration dependence, demonstrating the importance of bin-size for r_1 and r_2. We also show a measurement with "large bins", constructed to match those used in Nichol et al. (2006) where the first side of the triangle (r_1) varies by $\pm 0.5 h^{-1}$Mpc. The larger bin-size completely masks the configuration dependence expected from the redshift distortions.

In Figure 6.3 we extend the comparison of bin-width to the three specific scales that we measure in redshift and projected space. We utilize our *fiducial* binning scheme, which consists of 15 linear spaced bins in θ with the bin-width chosen to be a fraction (denoted with f) of the scale of the bin midpoint. To be clear, the bin-size for r_1 and r_2 also change appropriately with f. We show results for $f = 0.1, 0.2$ and 0.3 with 1σ Poisson uncertainties calculated from bin counts. The larger bin-width smooths the configuration dependence at all scales, with a dramatic effect on the $r_1 = 9\ h^{-1}$Mpc triangles. This occurs in both redshift and projected space measurements of $Q(\theta)$. Physically, larger bins allow a greater range of configurations to be represented in each bin.

We keep the number of bins fixed (at 15) but vary the bin-width which results in an increased overlap of configurations, and hence impose a larger correlation in the covariance between neighboring bins. We characterize the overlap by plotting the measured scale of the third side of the triangle (r_3) versus the corresponding opening angle (θ) in Figure 6.4. First off, we note the non-linear mapping between r_3 and θ (see the cosine rule in (4.4)), which explains why the co-linear triangle configurations have increased correlation between neighboring bins. Basically, the physical overlap is simply larger for equally spaced θ bins when θ corresponds to ~ 0 or $\sim \pi$. We show the corresponding covariance matrices with these two configurations for $Q_z(\theta)$ and $Q_{proj}(\theta)$ in Figure 6.5. The significant configuration overlap in $f = 0.25$ results in a larger correlation, as we expected. However, we also see increased correlation in non-overlapping bins (see the $\theta \sim 0$ with $\theta \sim \pi$ bins; the top left and bottom right corners).

We identified differences from bin-size in both measurements of the 3PCF and its covariance. The general rule of thumb is that smaller bins, assuming they are well resolved,

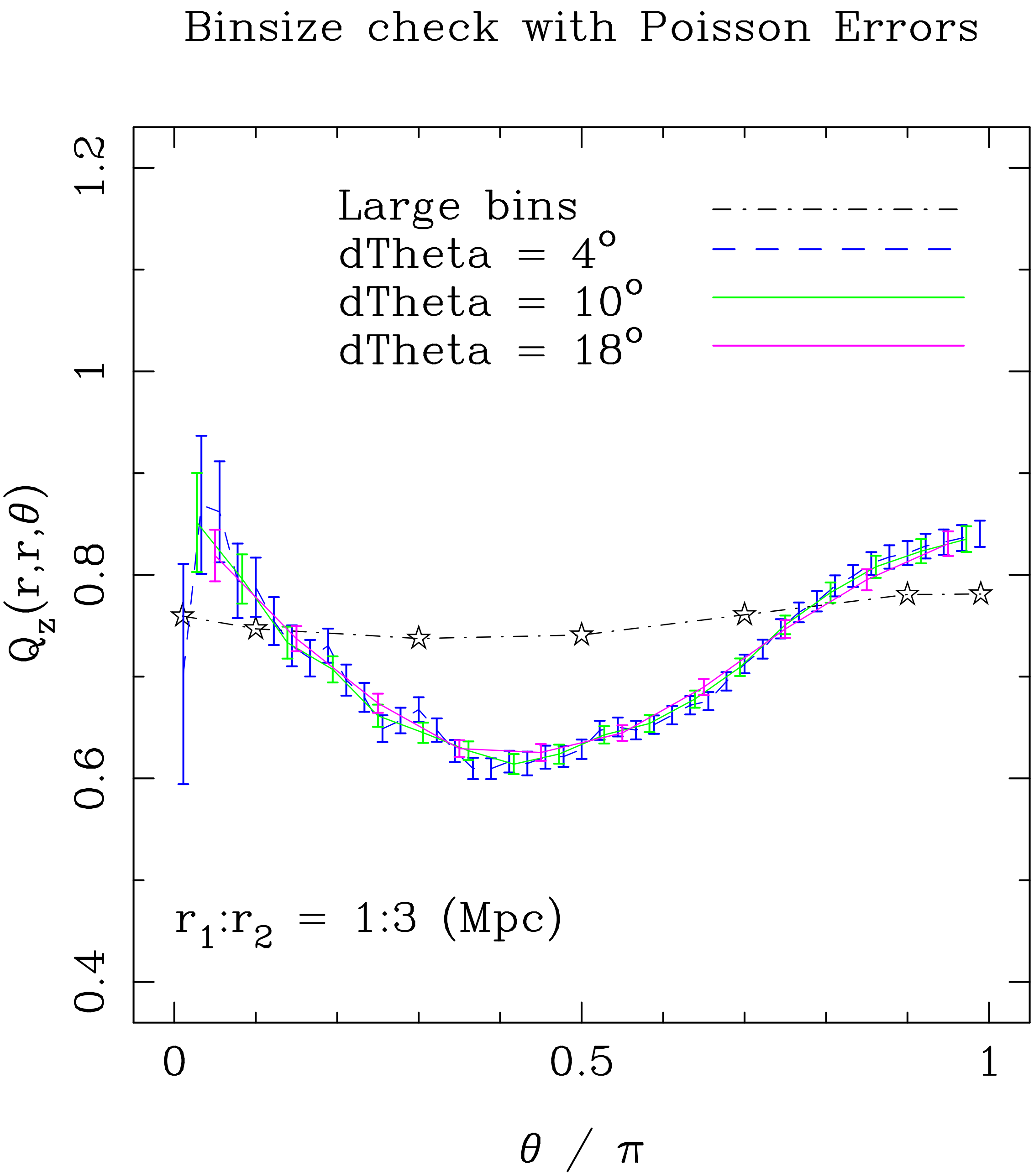


Figure 6.2 We show the reduced 3PCF on a subset of SDSS data. At small scales, we expect the redshift space distortions to cause a "U-shape" signal due to elongated fingers-of-god. This plot illustrates that small binning ($\pm 0.1h^{-1}$Mpc) in the two sides (r_1 and r_2) resolves the "U-shape" even with large θ bins. For comparison, we show "Large Bins" defined to correspond to the binning scheme of Nichol et al. (2006) (i.e. $\pm 0.5h^{-1}$Mpc). The error bars show the combined Poisson errors of the respective counts and signify the statistical significance of the bin-sizes.

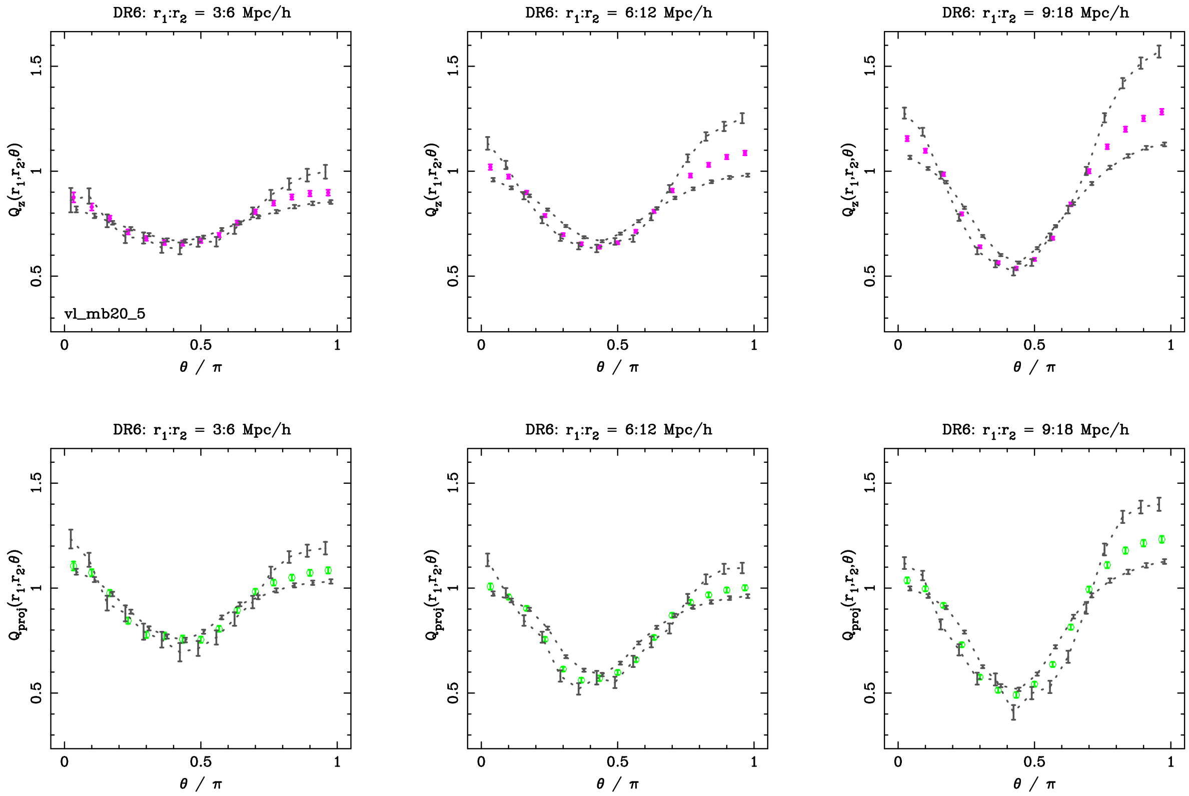


Figure 6.3 In our measurements on DR6 data (Chapter 4), we adopt a fiducial binning scheme using linear spaced bins in θ with a bin-size a set fraction of the midpoint. For the three triangle scales of interest, we compare this scheme using three fractional bin-widths on Q_z (top row) and Q_{proj} (bottom row): 10, 20 and 30%. The error bars represent Poisson errors on the counts from each bin. Larger bin-width measurements show smaller uncertainties and less configuration dependence. We use the LSTAR galaxy sample from DR6 selected to have $-21.5 < M_r < -20.5$.

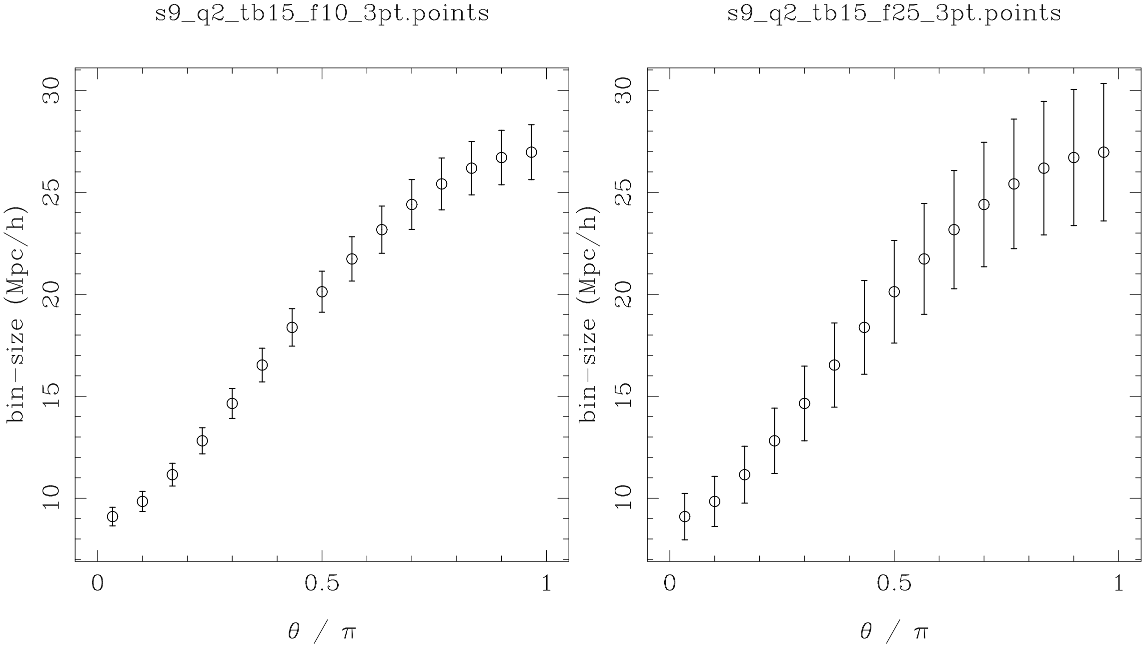


Figure 6.4 For our fiducial binning scheme we show the correspondence between opening angle θ and the respective scale with the chosen bin-width for 10% (left) and 25% (right) bin-widths. At both small and large θ, we note a significant amount of overlap between neighboring bins.

produce more accurate results. This basically is the so called "bias-variance trade-off", larger averaging produce lower variance estimates at the cost of a higher statistical bias. Smaller bin-size suggests smaller bias and therefore better accuracy, but high shot noise. But how do we quantify "well resolved" enough, especially with respect to the covariance? Furthermore, we use an eigenmode analysis which should help mitigate the effects of systematics since it removes unresolved modes in the covariance structure.

As an example of the effects of bin-size for galaxy-mass bias constraints, we re-analyze the large galaxy sample ($-21.5 < M_r < -20.5$) using the two fractional bin-widths discussed above. First, we ignore the structure of the covariance matrix and show constraints using all the bins assuming perfect independence in Figure 6.6. While unphysical, this illustration allows us to probe the effect of shape differences in the 3PCF measurements without considering the resolution of the covariance matrix. Since larger bins smooth the configuration dependence, we expect a larger degeneracy between B and C which is apparent. We see the best fit values (symbols) stay within the respective 1σ contours, but just barely. Remember, this uses all modes and the exact same input data, suggesting that binning can result in a 1σ systematic bias.

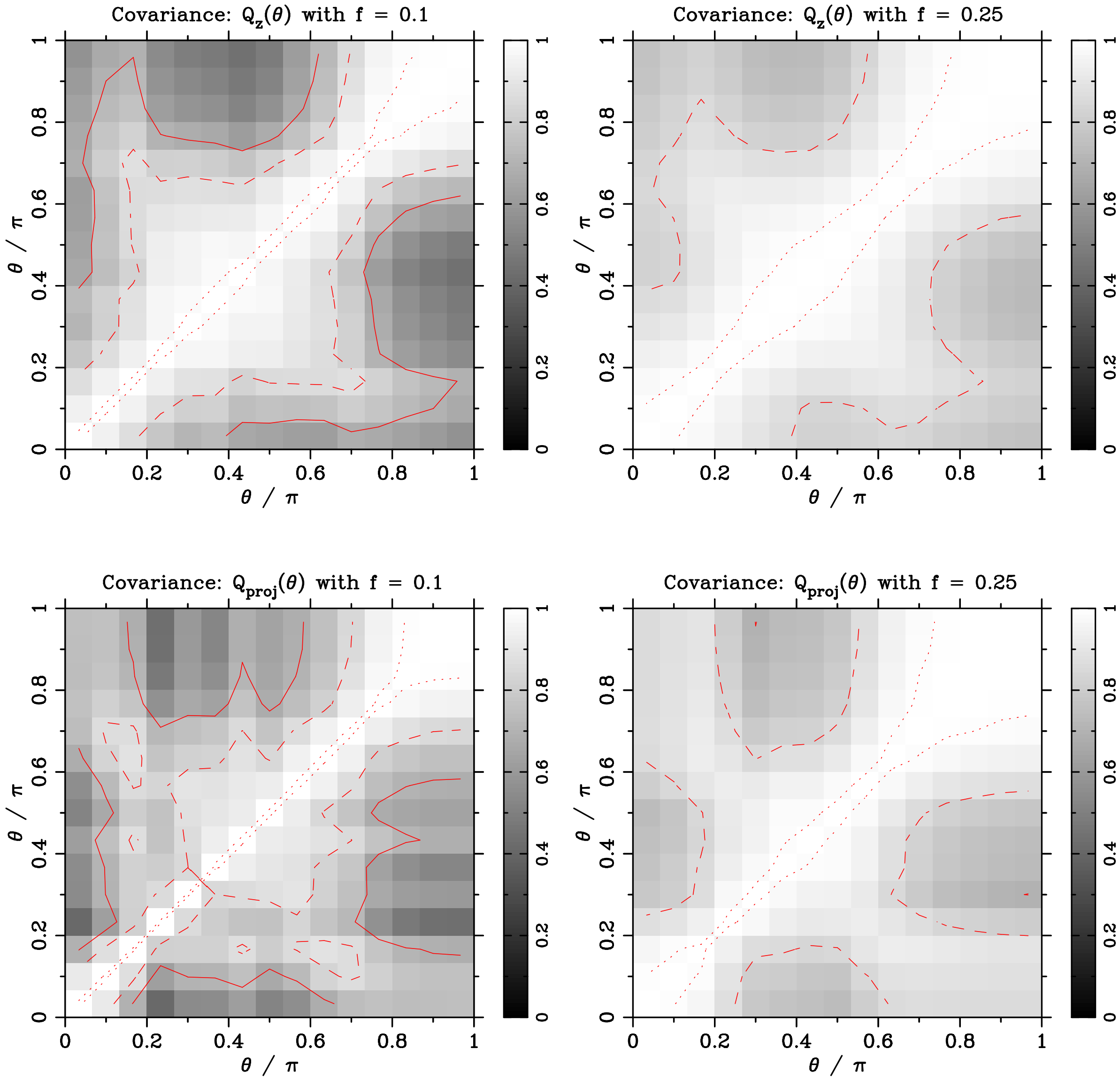


Figure 6.5 For our fiducial binning scheme, we show the covariance matrices for $Q_z(\theta)$ (top row) and $Q_{proj}(\theta)$ (bottom row) for the largest triangles with $r_1 = 9h^{-1}$Mpc. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99. We use the LSTAR galaxy sample from DR6 selected to have $-21.5 < M_r < -20.5$.

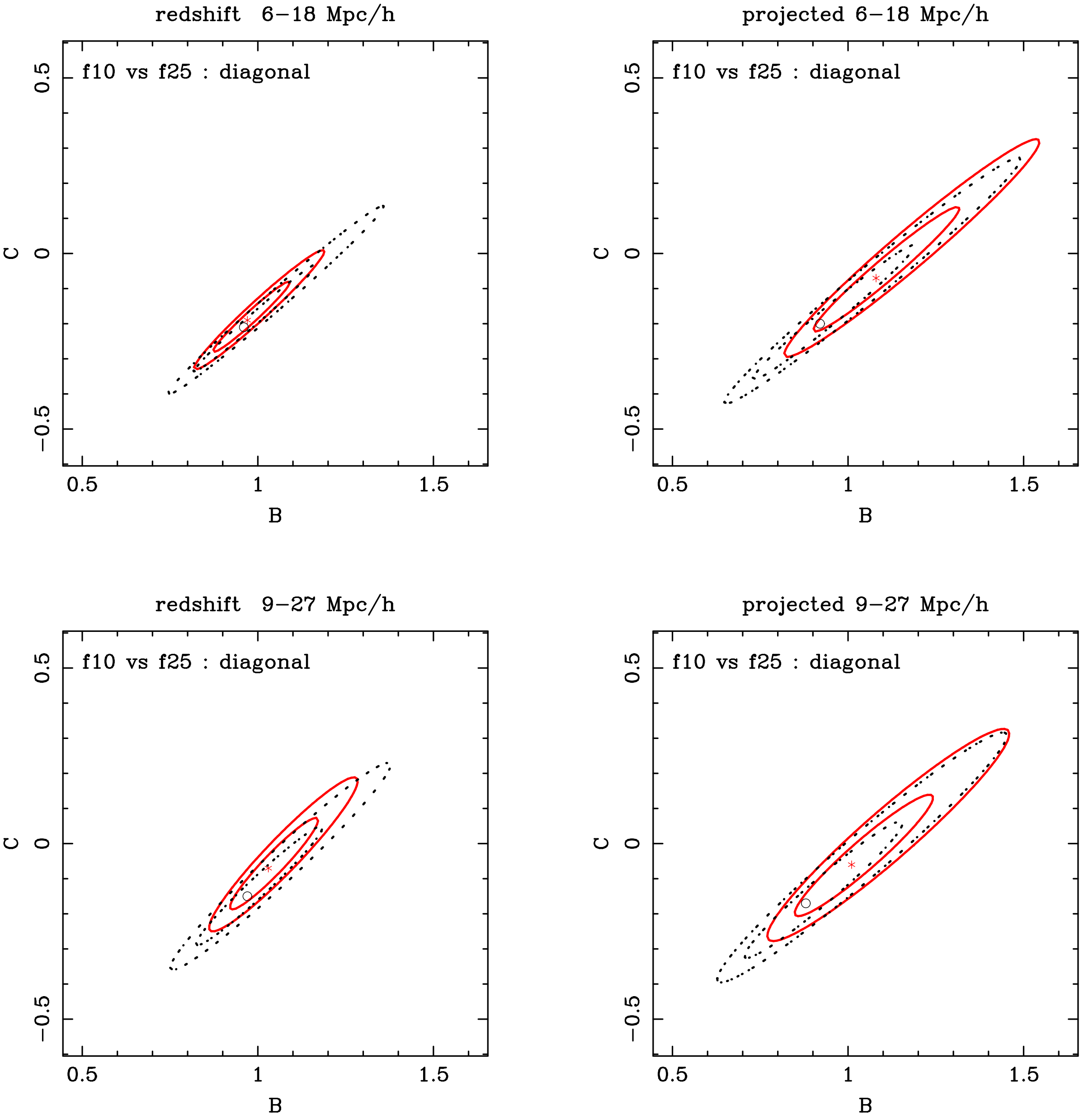


Figure 6.6 Analogous to the Galaxy-Mass bias constraints in Chapter 5, we show the constraints on B and C using the same data for our fiducial binning scheme with fractional bin-size of $f = 0.1$ (solid red contours) and larger $f = 0.25$ (block dotted). We neglect overlap and covariance, using the full 15 bins and assume independent diagonal errors. The contours correspond to the 1σ and 2σ confidence levels from the $\Delta\chi^2$ surface. We use the LSTAR galaxy sample from DR6 selected to have $-21.5 < M_r < -20.5$.

For Figure 6.7, we consider the full covariance as well as improvements obtained by using the eigenmode analysis in the galaxy-mass bias constraints. First, we notice the error contours appear less stretched, in accord with our expectations of using a non-diagonal covariance matrix. In most cases (excepting Q_z for $r_1 = 6h^{-1}$Mpc), the area of the contours appear of equal size or even decreased for the larger $f = 0.25$ measurements in contrast to the diagonal case. This makes sense, as the lower variance measurements of $f = 0.25$ appear better resolved as long as there are enough remaining modes to constrain two parameters. The best fit values appear discrepant, especially at the lower scales ($6 - 18\ h^{-1}$Mpc) where they disagree at more than a 1σ significance. While this causes some concern, it is not as drastic as the diagonal case. As the eigenmode analysis trims modes, it is throwing out information and the same input data ends up with a different statistical representation. In light of this, a 1σ difference becomes a statistical difference of analysis rather than a significant systematic effect.

In summary, we find lower galaxy-mass bias parameters with larger bin-widths, a potential artificial bias on the galaxy-mass parameters due to over-smoothing. Since we gain very little additional constraining power with the $f = 0.25$ bin-width, we argue the $f = 0.10$ bin-width represents the more conservative choice. This is the binning chosen for the analysis on this sample in Chapter 5. Although the $f = 0.10$ scheme represents smaller bins, they are still quite large and adequately resolve structure in the covariance.

6.3 ADDRESSING PROJECTED CORRELATION FUNCTIONS

The idea behind the projected correlation functions is to minimize the effects of redshift distortions on clustering measurements. We might ask about the effectiveness of the projection – is it achieving what we intend? High resolution N-body simulations, such as the Hubble Volume (HV) simulation, provide a suitable means to investigate this question. Since we want to apply our results to observational galaxy surveys, we construct realizations of dark matter (DM) that have the same footprint and volume of an observation galaxy sample. This should incorporate any edge effects or issues of finite volume that is also present in galaxy

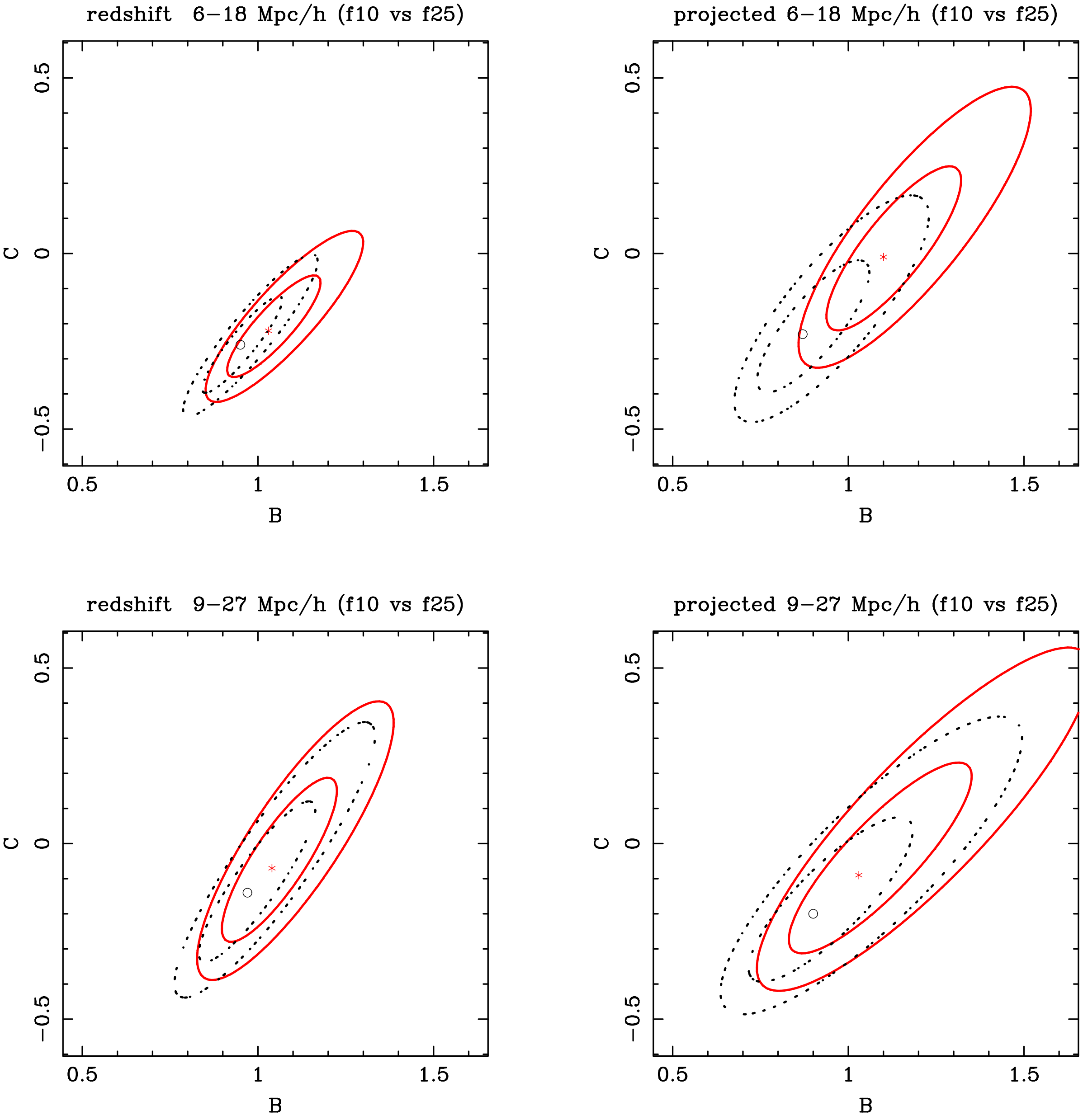


Figure 6.7 Similar to the above in Figure 6.6, we show the constraints on B and C using the same data for our fiducial binning scheme with fractional bin-size of $f = 0.1$ (solid red contours) and larger $f = 0.25$ (black dotted). However, we utilize the full covariance and only fit the dominant modes in an eigenmode analysis. The contours correspond to the 1σ and 2σ confidence levels from the $\Delta\chi^2$ surface. We use the LSTAR galaxy sample from DR6 selected to have $-21.5 < M_r < -20.5$.

data. We use line-of-sight peculiar velocities to create a redshift space distribution of DM. In a few cases, we compare this to a similar distribution without any redshift distortions, which we call *real* space. Two important decisions need to be made when using projected measurements: (1) how large of $\Delta\pi$ bins do we use to integrate over and (2) what is the maximum line-of-sight distance for the integration (π_{max}).

The first question stated above is relatively easy to address. We want to choose bins of $\Delta\pi$ that remain small enough to prevent a smoothing bias. In Figure 6.8, we show the anisotropy introduced by redshift distortions on the 2PCF, the $r_p - \pi$ diagram, within our HV test sample. Using these data, we calculate the projected 2PCF by integrating to $\pi_{max} = 40h^{-1}$Mpc with different values of $\Delta\pi$. We compare each with the fiducial measurement using $\Delta\pi = 2h^{-1}$Mpc, and show the fractional difference in Figure 6.9. We notice a very small effect and the largest deviation appears at small projected separation where $\xi(r_p, \pi)$ changes rapidly (see Figure 6.8). The systematic effect remains below the 2% level even with bins as wide as $\Delta\pi = 20h^{-1}$Mpc.

The second question to address remains more subtle: what is an appropriate π_{max}? Formally, it is preferred that $\pi_{max} = \infty$, which preserves the nice property that a power law spatial correlation function ξ produces a power law projected 2PCF w_p. Realistically, galaxy samples represent finite volumes and the correlation function can only be well estimated to a certain maximum distance. We want to keep our π_{max} well below this limit. On the other side, we want π_{max} to (re-)capture clustering strength lost due to redshift distortions. We compute $w_p(r_p)$ using two same volume realizations of DM, one in real space (no distortions) and one in redshift space, which we compare in Figure 6.10. We expand π_{max} to range between 20 and $80h^{-1}$Mpc, and chose a more standard log based binning in r_p to better match previous work (e.g. Zehavi et al., 2005). We see several interesting features. First, we note both real and redshift space $w_p(r_p)$ are roughly power laws, with the strongest power corresponding to the largest π_{max}. We expected this, as the projected 2PCF has units of distance, a larger integration equates to a higher functional value. We notice a much smaller difference between the smallest and largest π_{max} in redshift space. We understand this as a result of the large scale infall (Kaiser, 1987), as clustering strength is compressed to smaller line-of-sight separation which we see in Figure 6.8. Looking at the fractional

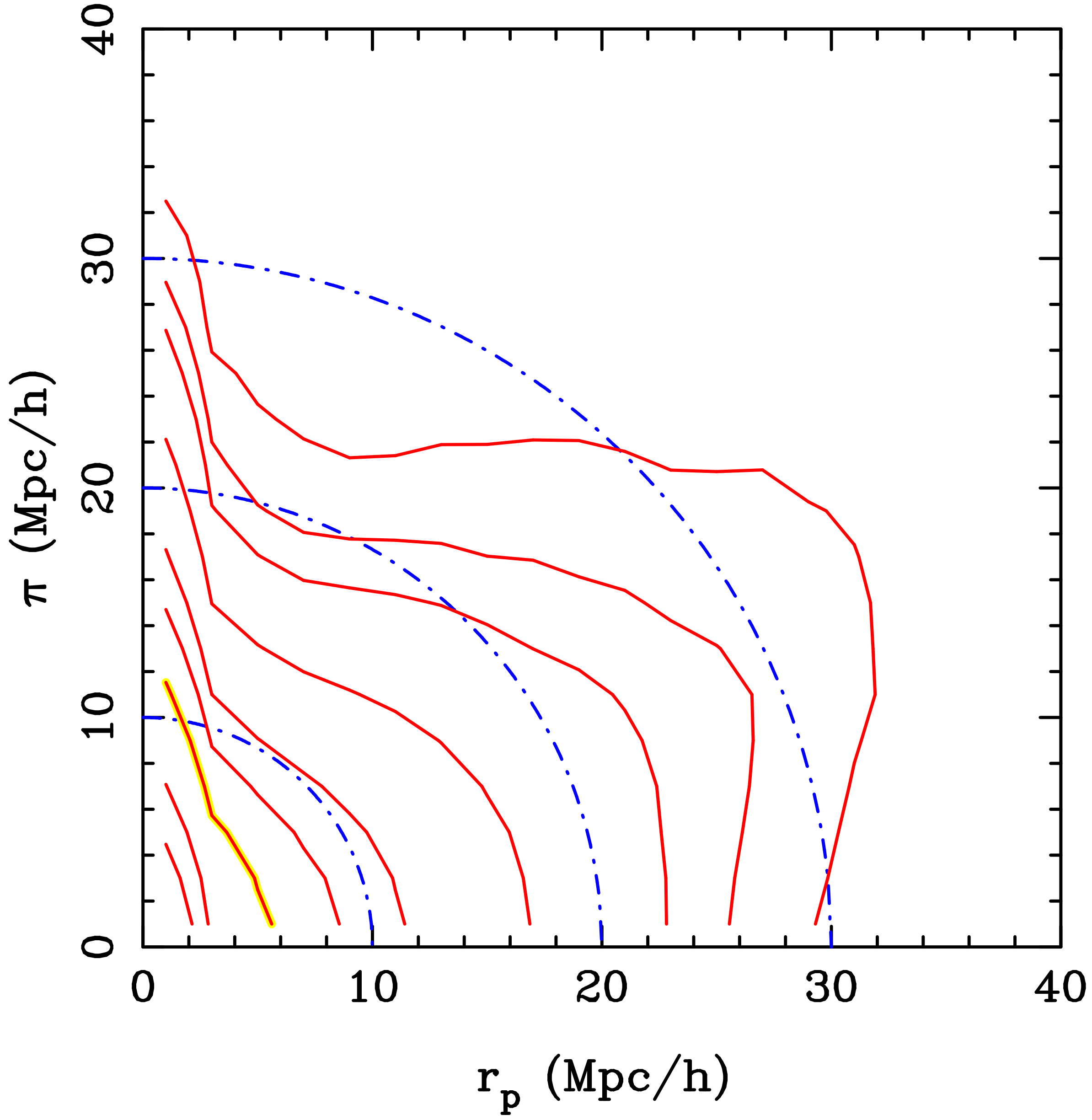


Figure 6.8 The $r_p - \pi$ diagram from the 2PCF on dark matter particles from the Hubble Volume simulation, using a grid of $2h^{-1}$Mpc in both r_p and π. The HV particles are trimmed to match the same volume and footprint of SDSS data, with velocities used to construct radial redshift distortions. The red (solid) lines are contours of a specific value of the CF, $\xi(r_p, \pi)$, with the yellow highlighted one corresponding to $\xi(r_p, \pi) = 1$. The blue (dot-dashed) semi-circles show a perfect isotropic correlation for comparison.

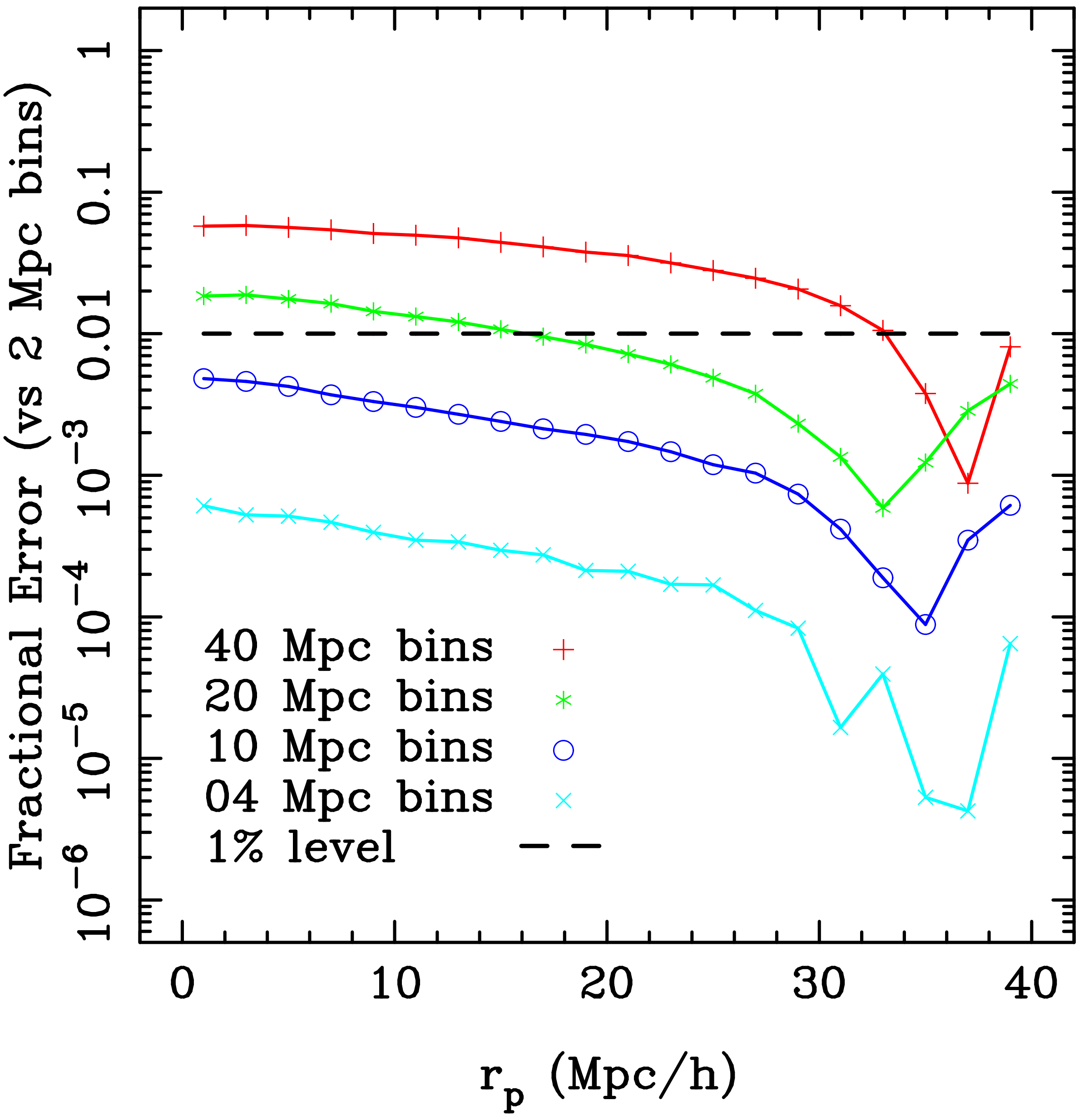


Figure 6.9 The accuracy of the $w_p(r_p)$ integration using different sized π bins, which we show as a function of projected scale (r_p). We obtain the "truth" value by comparing to a fiducial $\Delta\pi = 2h^{-1}$Mpc, and our $\pi_{max} = 40h^{-1}$Mpc. We show the 1% level of accuracy with the dashed line.

difference between real and redshift space measurements, we find a higher π_{max} does what we expect in reducing the difference between real and redshift space $w_p(r_p)$ measurements. However, we also see that different values of π_{max}, dramatically change the 2PCF from the "ideal case", inducing a systematic difference which can range from right below the 10% level at $r_p = 10; \pi_{max} = 80h^{-1}$Mpc to 80% at the largest scales and smallest π_{max}. Finally, we conclude that even at $\pi_{max} = 80h^{-1}$Mpc, the effects of redshift distortions cannot be completely negated by the projected 2PCF.

We extend the analysis to look at the reduced 3PCF, using the two largest triangle configurations ($r_1 = 6$ & $9h^{-1}$Mpc). Since $Q(\theta)$ is normalized by the respective 2PCF, we expect the same amplitude between spatial and projected measurements. We show the real and redshift space HV distributions where we measure $Q(\theta)$ for the spatial 3PCF, and projected with $\pi_{max} = 20, 30$ and $40h^{-1}$Mpc in Figure 6.11. We see little discrepancy in redshift space, with the only significant difference at the largest scales for elongated triangles. Real space, however, shows a significant configuration dependence for the spatial 3PCF that appears lost in redshift space. On these large scales, both redshift distortions as well as projecting the correlation function appear to dramatically reduce the configuration dependence. We caution that this effect is only a suggestion by this analysis, as the errors of these measurements (which we do not display) remain large. We add redshift distortions with a simplified model, and we downsample the HV dark matter realizations for computational efficiency. A more detailed analysis is required to appropriately interpret the effects of redshift distortions on these measurements – beyond the scope of our analysis.

Projected space measurements for the largest triangles in Figure 6.11 seem to decrease in amplitude with increased π_{max}. The effect is most pronounced at large θ (larger scale for the third triangle size). Since triplet counts decrease more rapidly than pairs in the 2PCF, we might be seeing some limited volume effects (for effects on bispectrum see Scoccimarro, 2000). That is, the range of triplets available in the survey is not sufficient to measure these scales. Remember, a higher π_{max} means increased sensitivity to larger scales with the same r_p. We noticed this trend of reduced amplitude for larger scales in the equilateral 3PCF when $r_p > 10h^{-1}$Mpc (see Figure 4.3). Interestingly, the discrepancy between different π_{max} values in Q_{proj} remains more pronounced in real space than in redshift space, a result we noted in

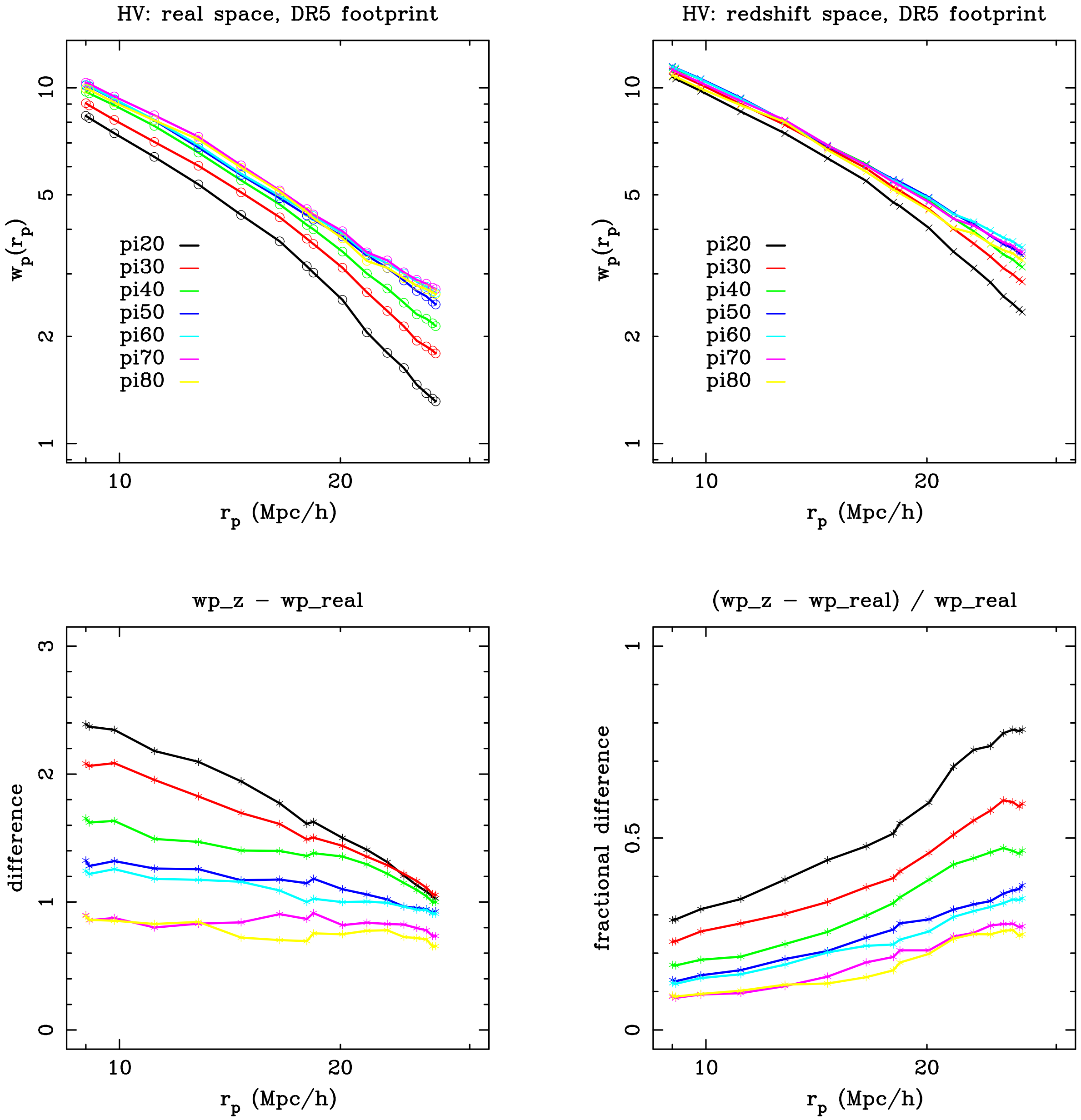


Figure 6.10 We show the projected 2PCF, $w_p(r_p)$, for dark matter particles in the Hubble Volume simulation both in real space (top left panel) and redshift space (top right panel). We show the absolute difference due to redshift distortions (bottom left panel) and the fractional difference (bottom right panel). The different lines correspond to different π_{max} integrations, as denoted. We estimate $w_p(r_p)$ by integrating in bins of $\Delta\pi = 5h^{-1}$Mpc.

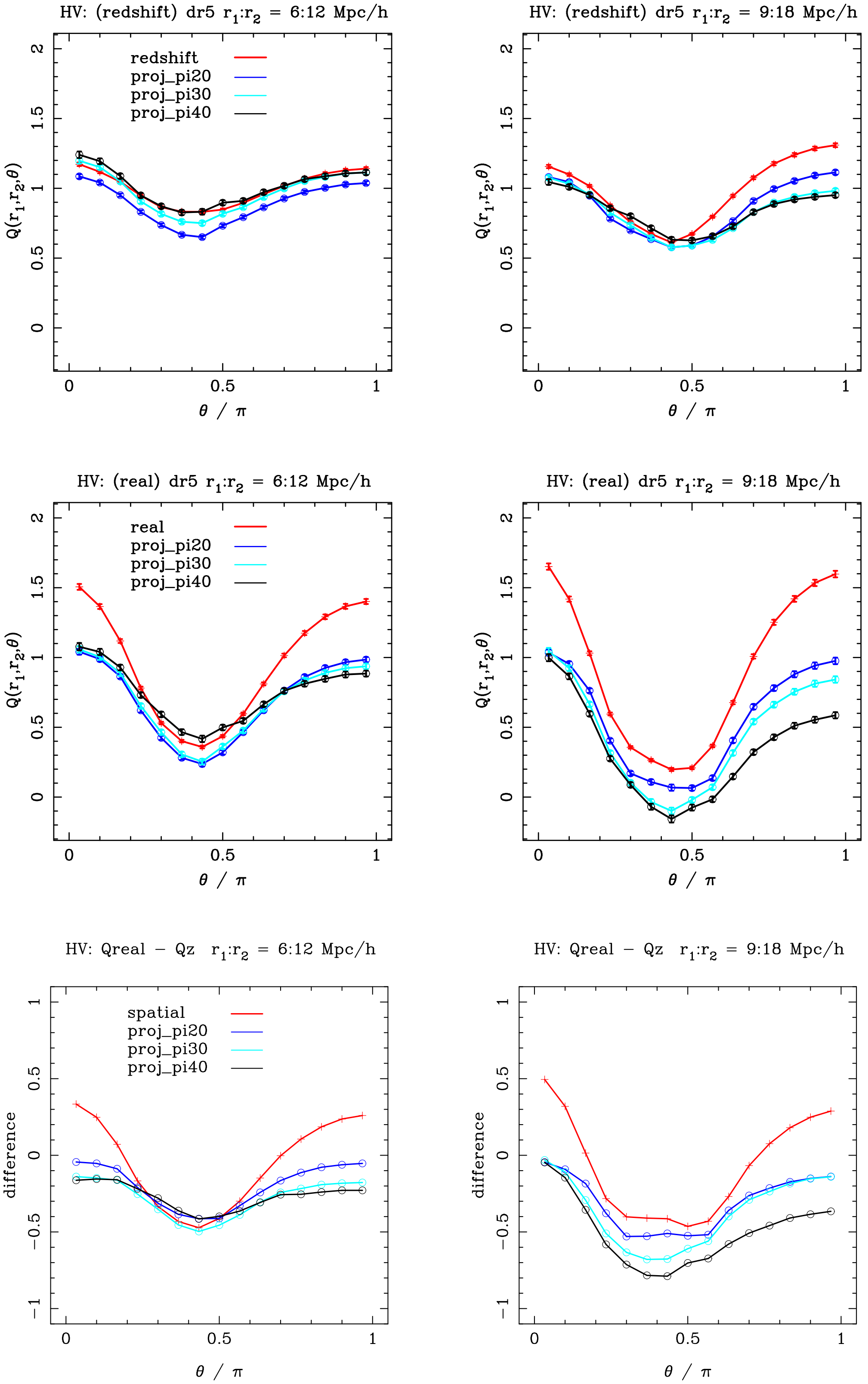


Figure 6.11 We plot reduced 3PCF, $Q(\theta)$, between real and redshift space at the two triangle scales we use for galaxy-mass bias constraints, $r_1 = 6$ & $9h^{-1}$Mpc. We compare three different values of π_{max}, namely $20, 30$ and $40h^{-1}$Mpc with the *spatial* 3PCF in redshift space. We use the dark matter distribution derived from the Hubble Volume simulations, and match to the SDSS DR5 footprint and corresponding volume.

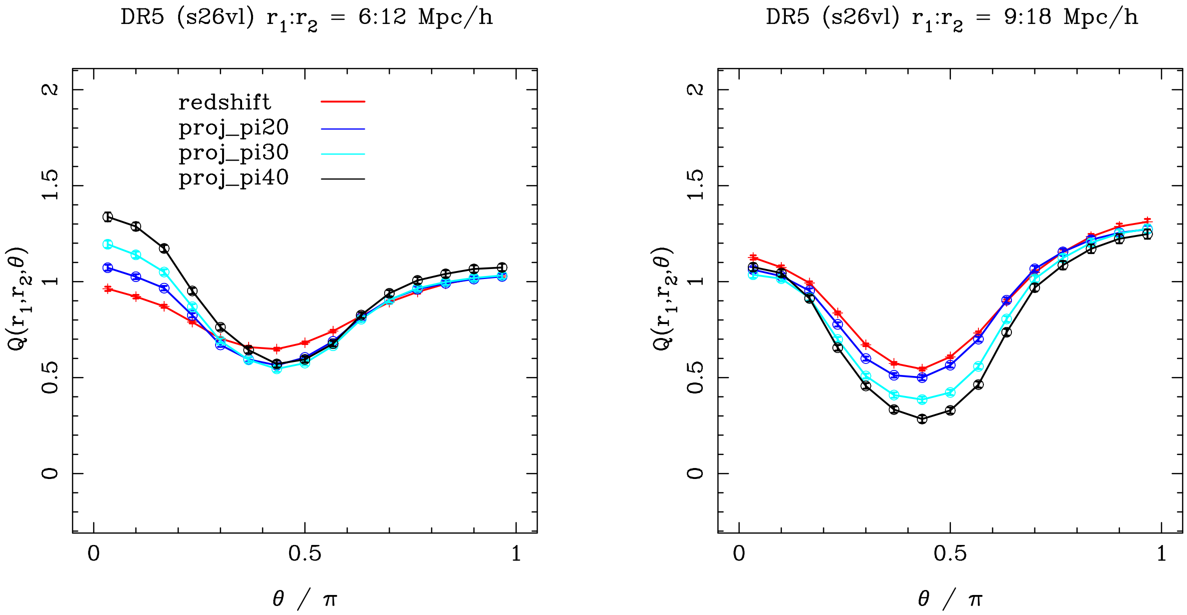


Figure 6.12 We plot reduced 3PCF, $Q(\theta)$, between real and redshift space at the two triangle scales we use for galaxy-mass bias constraints, $r_1 = 6$ & $9h^{-1}$Mpc. We compare three different values of π_{max}, namely $20, 30$ and $40h^{-1}$Mpc with the *spatial* 3PCF in redshift space. We plot the effects on a galaxy sample from SDSS DR5.

the 2PCF (Figure 6.10). Looking at the smaller triangles in redshift space for the HV (top left panel of Figure 6.11), we expected to see a larger difference in Q_{proj} between the π_{max} values. Specifically, we had hoped to notice *some* change in the configuration dependence.

We consider the same π_{max} test measurements using observational galaxy data in Figure 6.12. Right away, we see different effects than we noticed in the HV for dark matter. For observational galaxies, it appears that Q_{proj} recovers some configuration dependence with larger π_{max}. The $r_1 = 9h^{-1}$Mpc triangles show decreased amplitude perpendicular configurations and we note the "splitting" of collapsed triangles ($\theta \sim 0$) in the smaller triangles. In both cases, a decreased π_{max} yields a $Q(\theta)$ that approaches the redshift space measurement, in agreement with our expectations.

We highlight the differences in $Q(\theta)$ between HV dark matter in redshift space and observational galaxies in Figure 6.13. Again, we see that lower π_{max} trends toward the redshift space $Q_z(\theta)$. We notice a clear difference between DM and galaxy measurements which are affected differently by π_{max}. Galaxy-mass bias presents an obvious explanation for this effect as galaxies can cluster differently than dark matter. Redshift space distortions

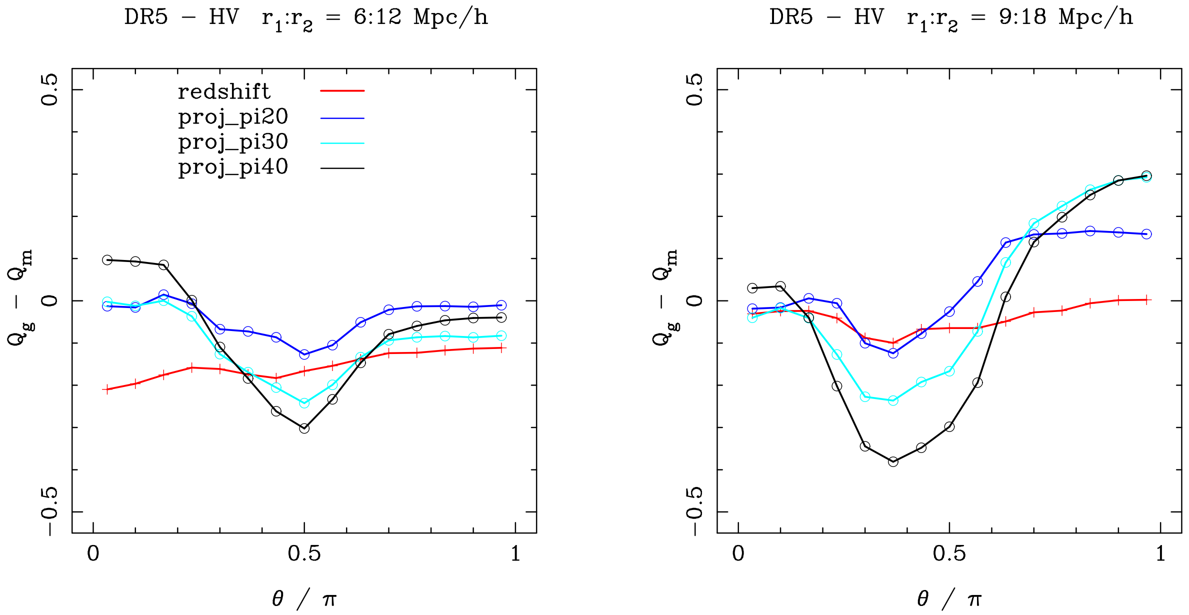


Figure 6.13 We plot the difference of the reduced 3PCF $Q(\theta)$ in redshift space between HV dark matter and observational galaxies in an SDSS DR5 sample. We show the two triangle scales we use for galaxy-mass bias constraints, $r_1 = 6$ & $9h^{-1}$Mpc. We compare Q_{proj} with three different values of π_{max}, namely $20, 30$ and $40h^{-1}$Mpc along with the *spatial* 3PCF in redshift space.

couple the velocity field with the density field, as the line-of-sight velocity is misinterpreted as a change in distance. This highlights any slight differences in velocities of the DM in the HV simulation from observational galaxy velocities, either from physical effects, generally referred to as *velocity bias* or simply inaccurate representation of distortions in the simulation. The latter can occur due to a large number of various assumptions, such as differences in cosmology, light-cone creation or even our postprocessing to add distortion distances. In addition, these measurements use a single bin in $\Delta\pi$ to reduce the computational cost. As we saw in the 2PCF (Figure 6.9), this approximation can introduce errors, especially for $\pi_{max} > 20h^{-1}$Mpc.

Overall, we conclude that the projected correlation function reduces the impact of redshift distortions on measurements of clustering. However, the projected statistic does not completely remove effects of the distortions for any of values of π_{max}. Our results here suggest a larger π_{max} might help to further minimize redshift distortions but a more thorough investigation is required to disentangle the systematics. Since the computational efficiency of estimating the 3PCF decreases dramatically with increased π_{max}, we chose to use

$\pi_{max} = 20h^{-1}$Mpc for the measurements presented in Chapter 4.

6.4 QUALITY OF ERROR ESTIMATION

We rely heavily on the structure of the error covariance matrix for constraints on galaxy-mass bias. It remains unclear how well we resolve the covariance with our jackknife re-sampling estimation. Higher orders add complexity and increased sensitivity to systematics, even with a "ratio statistic" such as the reduced 3PCF where the error sensitivity is canceled to first order. We must investigate the error resolution of jackknife resampling on the 3PCF, as tests on the 2PCF in angular correlation functions (Scranton et al., 2002) or redshift space SDSS (Zehavi et al., 2002) can not be assumed to be sufficient.

Another method to estimate errors consists of using a series of independent realizations of artificial galaxies, ideally created to match observational limitations such as the volume and geometry to the SDSS galaxy samples. Unfortunately, mocks that matched the characteristics of all the SDSS data samples were not readily available. However, we obtained 49 independent galaxy mock catalogs based on independent N-body simulations that have appropriate resolution to match the BRIGHT SDSS sample ($M_r < -21.5$). We use these independent mocks to estimate errors and compare with those obtained from jackknife resampling. A comparison of both methods of should provide some idea how effective jackknife re-sampling is for resolving the errors on the 3PCF. The BRIGHT galaxy sample has the lowest number density with the least number of galaxies over the largest volume. To help protect against undersampled measurements due to low bin counts, we restrict the comparison to the configuration dependence of the larger triangles ($r_1 = 9h^{-1}$Mpc sides).

We estimate the covariance matrix for the BRIGHT SDSS sample using different numbers of jackknife samples, specifically 15, 30, 49 and 105 jackknife regions. These numbers are not chosen at random. Since we measure 15 bins for $Q(\theta)$, we require at least 15 jackknife samples to prevent a singular covariance matrix. From there, we use twice this number (30) and then use the same number as the number of mocks (49). Finally, the 105 corresponds to the number of unique elements in the symmetric covariance matrix: $15(15-1)/2 = 105$.

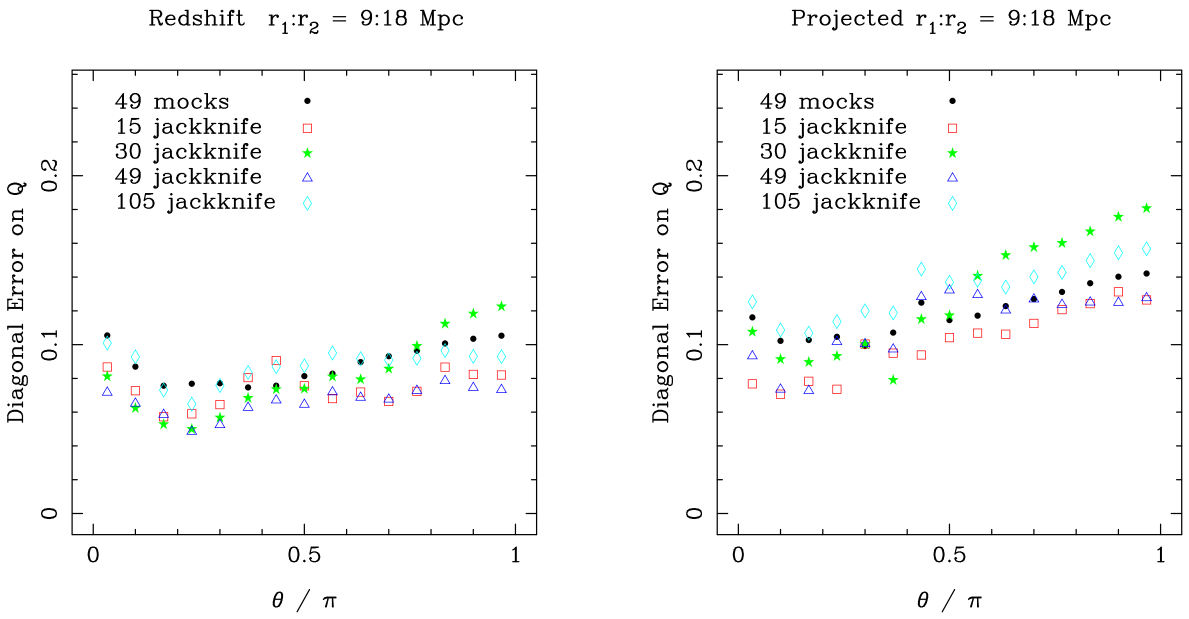


Figure 6.14 We compare the absolute (diagonal) errors between the different methods of estimation for the BRIGHT galaxy sample from SDSS DR5.

We caution that as the number of jackknife samples increase, the respective volume of each jackknife region decreases.

First, we investigate how the absolute (diagonal) errors compare. Since we use normalized covariance matrices, differences in the absolute errors might not be noticeable in the covariance structure. The 1σ absolute errors are shown in figure 6.14. We see little difference between any of the methods with the uncertainty typically ranging between 0.1 and 0.15.

We estimate the normalized covariance in both redshift and projected space for the reduced 3PCF, as depicted in Figure 6.15. We see that that jackknife re-sampling underestimates the correlation, but the general structure looks comparable. The jackknife estimate for the projected measurements appear noisy. This is not overly surprising given our method of jackknife creation. The jackknife regions are wedges chosen by equal area on the sky and use the full depth of the survey, which correlate most directly to the projected statistic. We clearly see that more samples produce a smoother, and more correlated, covariance matrix. However, not even the 105 jackknife sample estimate reproduces the correlation in 49 mocks. To estimate the covariance, we subtract the mean measurement from each realization. One important metric to consider in the reliability of the resulting covariance is the distribution of error residuals, which we also show in Figure 6.15. If the residuals are Gaussian, the covariance matrix accounts for all the necessary structure. We notice in several of the jackknife re-samplings a skew to the residuals, with a tail extending to lower values. As we discussed in Chapter 4, this is a result of cosmic variance in jackknife samples. A few rare structures affect the 3PCF, and when excluded in the occasional jackknife region the $Q(\theta)$ drops. The mock estimate shows a slight skew in the positive direction from the same effect. In mocks, when a rare structure exists in the probed volume the 3PCF rises producing a rare high measurement.

The eigenmode analysis we utilize relies on signal being the dominant contribution to the structure of the covariance matrix (as opposed to noise). Noise is commonly expected to be an independent or "diagonal" contribution. We look at the eigenvectors of the covariance matrix, also commonly referred to as principle components, to provide insight into the structure. By using the singular value decomposition (SVD), the eigenvectors are ordered by largest to least amount of variance explained in the covariance matrix. The first three eigenvectors are shown in Figure 6.16 for redshift and projected space. Similar structure appears in each of them. We interpret this as follows: the first eigenvector represents the general measurement, with all eigenmodes equally weighted. The second eigenvector shows the difference between "collapsed" and "perpendicular" configurations. Finally, the third eigenvector represents a scale dependence as the third side of the triplet ranges between $9h^{-1}$Mpc at $\theta \sim 0$ to $27h^{-1}$Mpc at $\theta \sim \pi$. In some of the estimates, the shapes of the second and third EVs appear either combined or transposed. Since the full measurement is a linear combination of all EVs, this lack of separation makes sense. In these cases, the statistical significance the two EVs remain similar. This interpretation of the structure follows the analysis by Gaztañaga and Scoccimarro (2005) for N-body simulations. The less significant eigenvectors (which we do not show) appear random, with the lowest being contributions from noise or numerical instabilities. We identify this by looking at the singular values (SVs) for each of the eigenmodes as shown in Figure 6.17. The SV can be understood as an "importance weighting" of each eigenmode with a log scale showing a rapid drop of significance. The first three eigenvectors cumulatively account for over 99.9% of the variance of the covariance matrix. This strongly supports our assertion that meaningful signal, and not noise, dominates the structure of the covariance matrix.

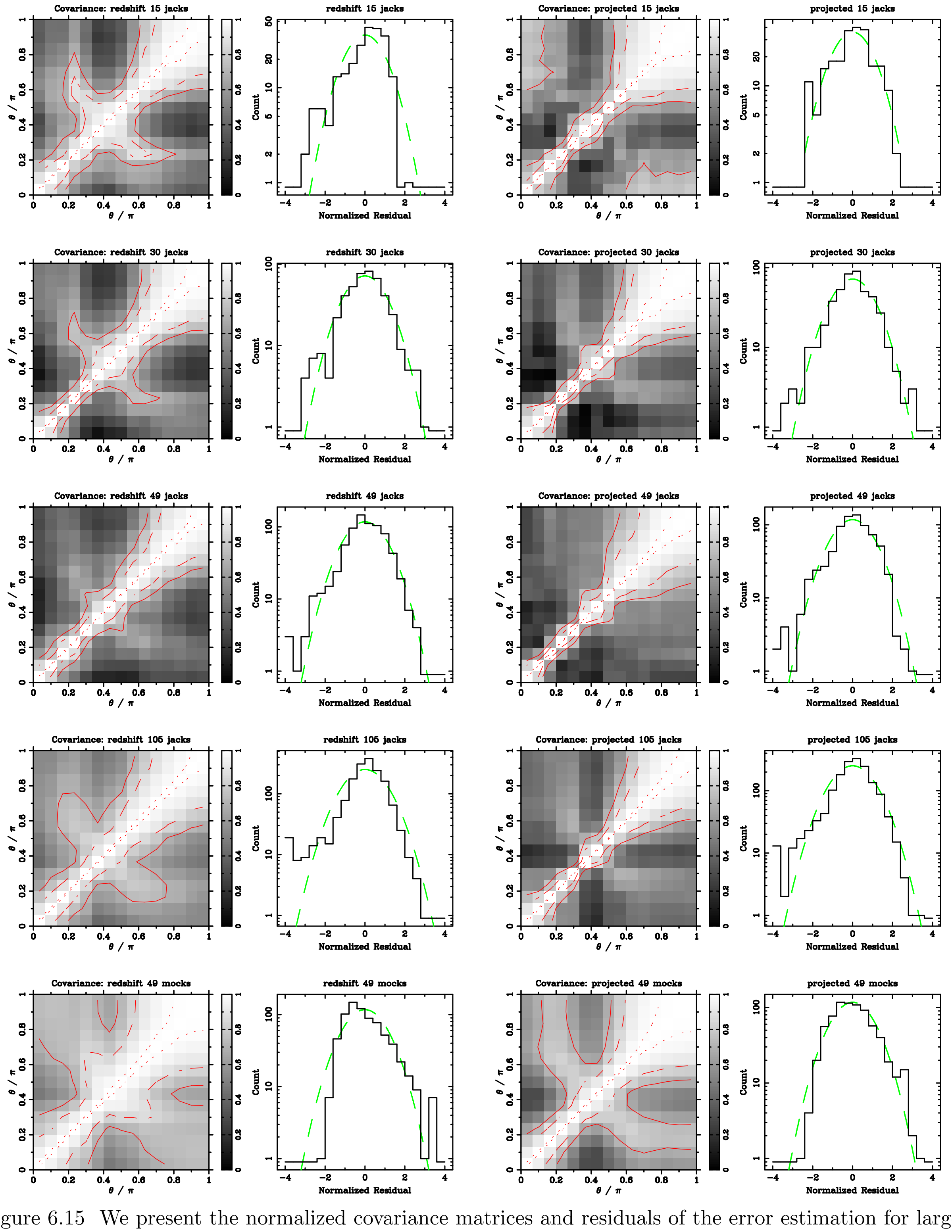


Figure 6.15 We present the normalized covariance matrices and residuals of the error estimation for large triangles ($9 - 27h^{-1}$Mpc). The left and right columns pertain to redshift and projected space respectively. We estimate errors using 15,30, 49 and 105 jackknife regions and compare with 49 independent mocks. The solid, dashed and dotted contours correspond respectively to values of 0.70, 0.85 and 0.99.

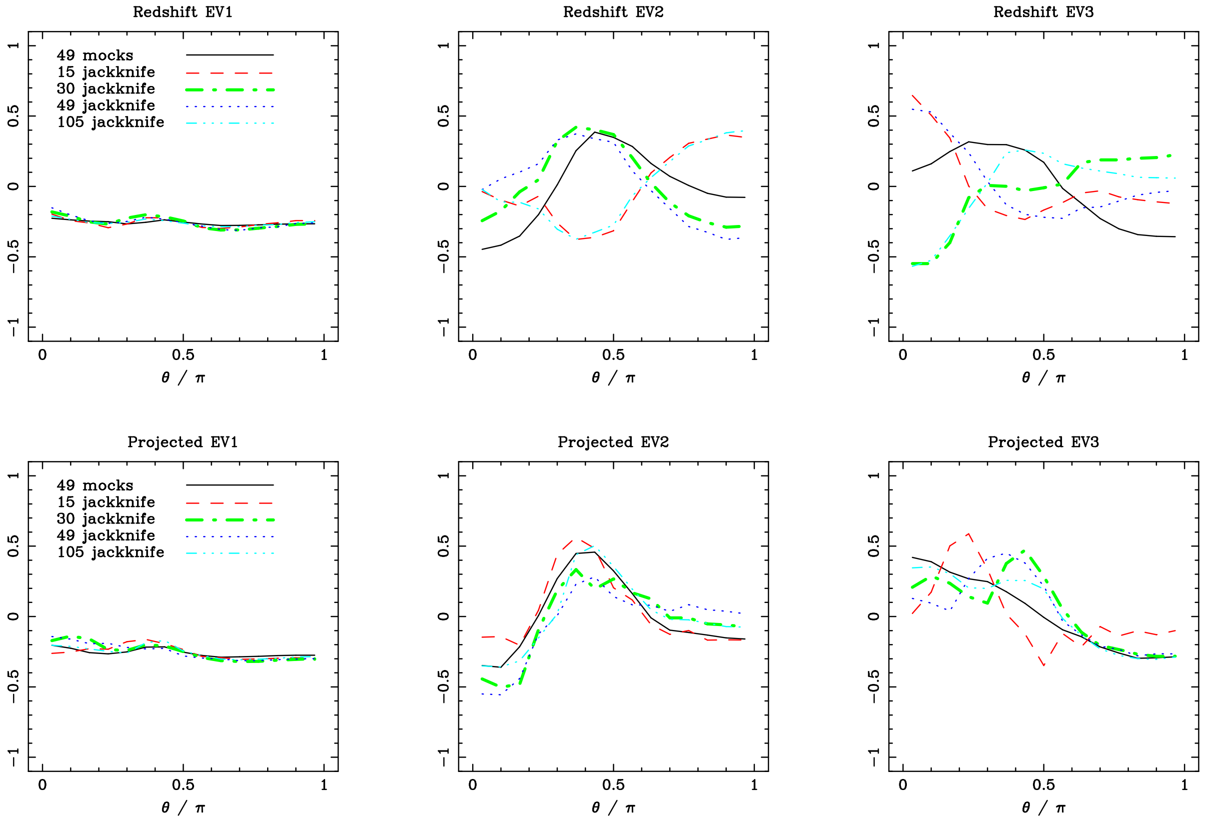


Figure 6.16 Top three eigenvectors (EVs) chosen from the normalized covariance matrix for the different error estimates for $Q_z(\theta)$ and $Q_{proj}(\theta)$. The sign of the EV is arbitrary. The first EV (left panels) shows equal weights between all bins. The second (middle) and third (right) EV display the configuration difference between perpendicular and co-linear triangles as well as the scale variation as the scale of the third side increases.

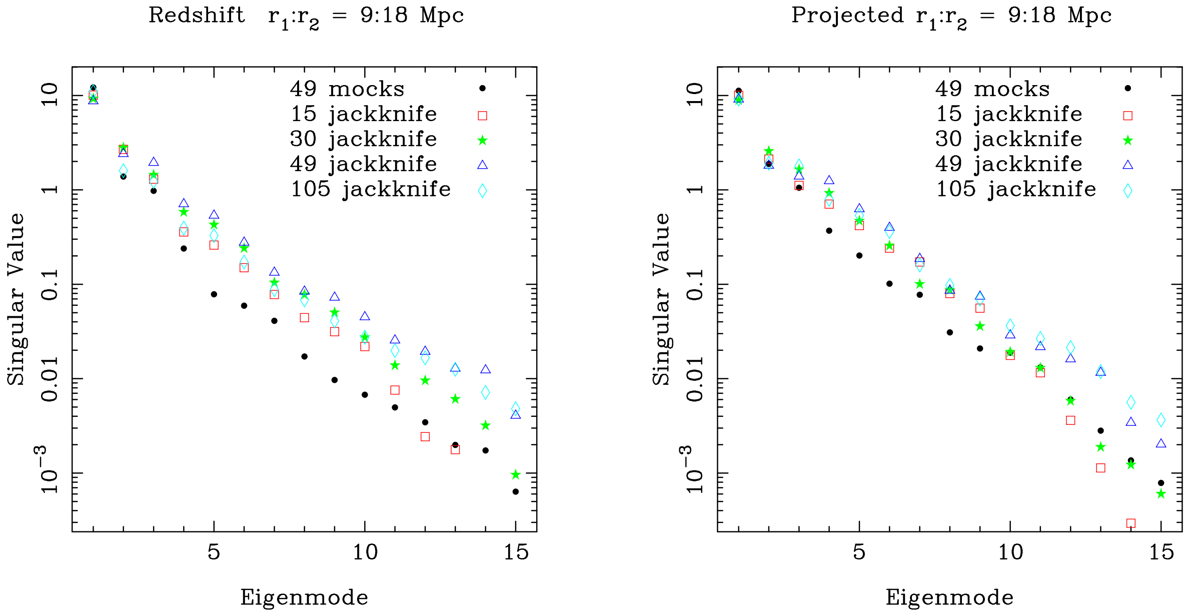


Figure 6.17 We show the singular values obtained from the SVD of the normalized covariance matrix for each of our error estimates. Larger values of the SV correspond to more statistically significant eigenmodes in the structure of the covariance.

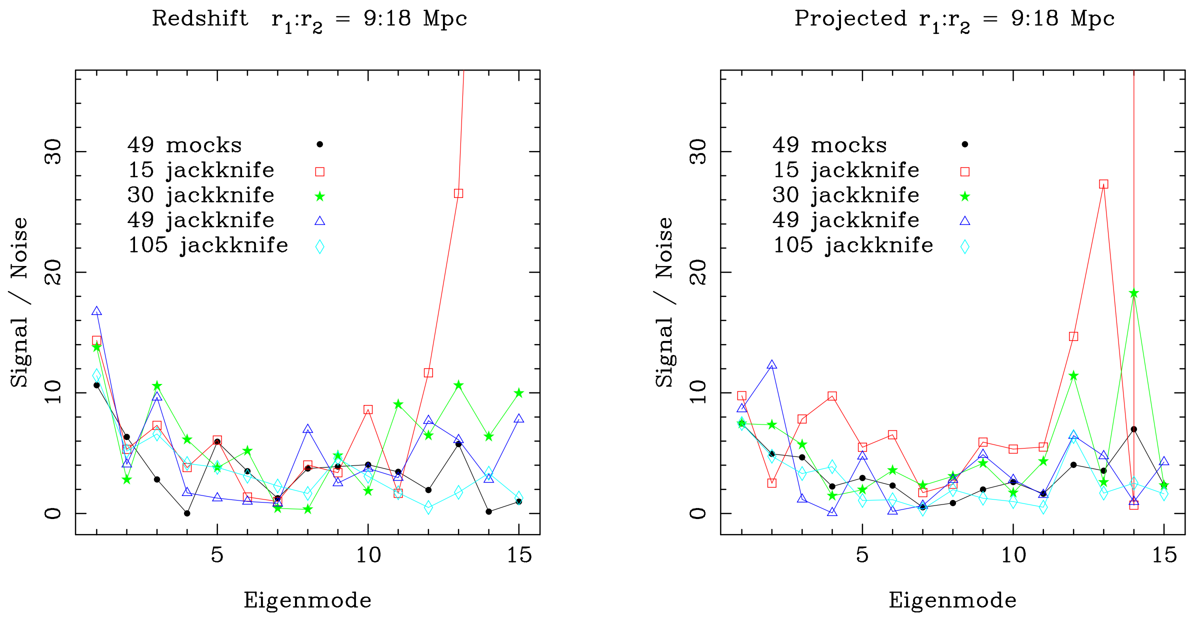


Figure 6.18 We show the signal-to-noise ratio for each eigenmode, ordered in terms of importance. The total signal-to-noise of a measurement is calculated by adding each individual eigenmode in quadrature.

The signal-to-noise (S/N) ratio of each eigenmode can be computed, which we describe in Chapter 5, and show in Figure 6.18. The mocks in both redshift and projected space depict a slow decline in S/N over the first few eigenmodes, supportive or our interpretation of relative significance. This trend is not as clear in the jackknife estimates for redshift space, although it appears consistent in projected space. We see the first half of the modes appear resolved, with well behaved S/N. For the lower significant eigenmodes, the noisier error estimates using fewest jackknife samples (especially the 15 jackknife sample one) result in unrealistically high S/N ratios. By including the non-negligible values of these noise dominated modes, the total S/N would increase dramatically and artificially. Using all modes of the covariance matrix in these cases would be a mistake. To make the point clearer, we show the cumulative S/N ratio in Figure 6.19 which highlights the necessity of trimming unresolved modes.

We can compare the *subspace* that a set of eigenmodes probe between two error estimates. This uses the same formalism discussed in Yip et al. (2004, see section 4), where we obtain a fractional "match" between a series of eigenvectors. The *subspace* comparison is based on the dot product. For example, two orthogonal unit vectors would have a vector subspace of 0 whereas two identical unit vectors would be 1. This is easily extended to the orthogonal

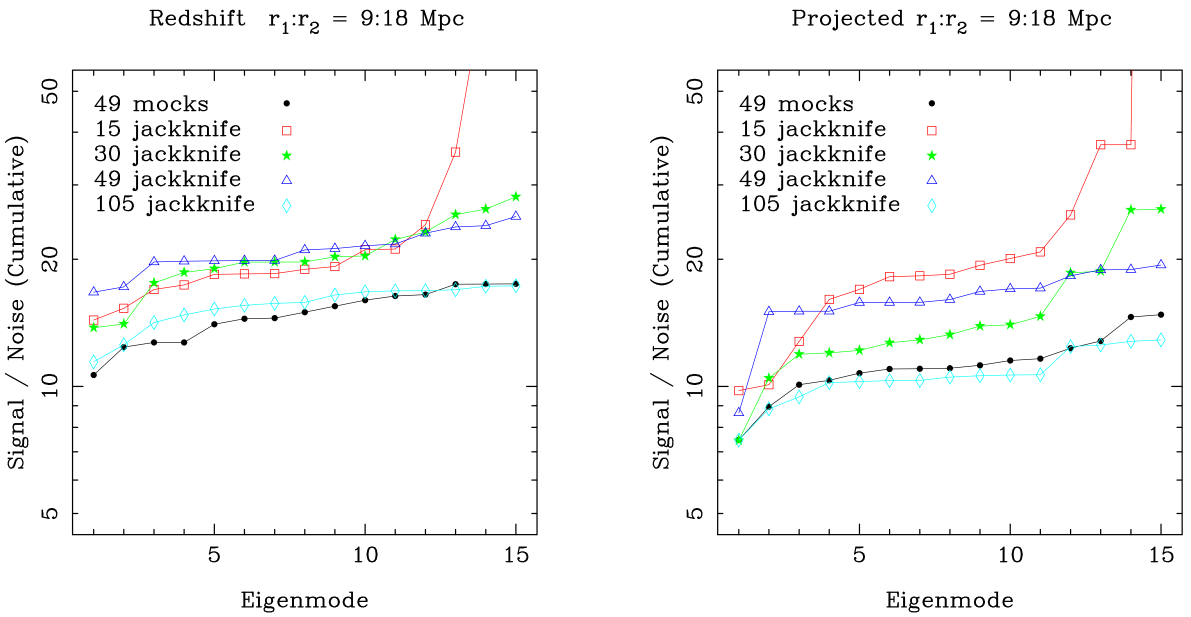


Figure 6.19 We show the cumulative signal-to-noise ratio for each eigenmode, ordered in terms of importance. The cumulative total is calculated by summing in quadrature the more significant modes.

eigenvectors of the covariance matrices. We use the covariance of the 49 mocks as "truth", and test the fractional compatibility of the jackknife estimates for covariance in $Q_z(\theta)$ and $Q_{proj}(\theta)$, which we plot in Figure 6.20. Please note, when all the modes are considered, the *subspace* becomes the full space and the comparison yields unity. We notice the projected measurements never appear more discrepant than 75%. Once past the first few modes, this remains true in redshift space as well. With the exception of the 15 jackknife sample estimate, the 3 eigenmode mark appears 90% compatible or better in all cases. This quantifies our argument of the top three EVs in Figure 6.16, where the second and third eigenvectors appear different (predominantly in redshift space), but their linear combination remains very consistent with each other. This comparison only takes into account the compatibility of the *direction* of each eigenvector, and not their relative strengths.

We can use the subspace comparison to investigate differences between redshift and projected space covariance matrices, instead of how compatible jackknife estimates appear with respect We show the fractional comparison of between them in Figure 6.21. We first note that the mocks estimates appear compatible almost across the board, where they compare at ~ 0.85 or better. We also see that the sum of three eigenmodes remain an optimum number for a compatible subspace with a fractional comparison above 0.9, once again excepting the

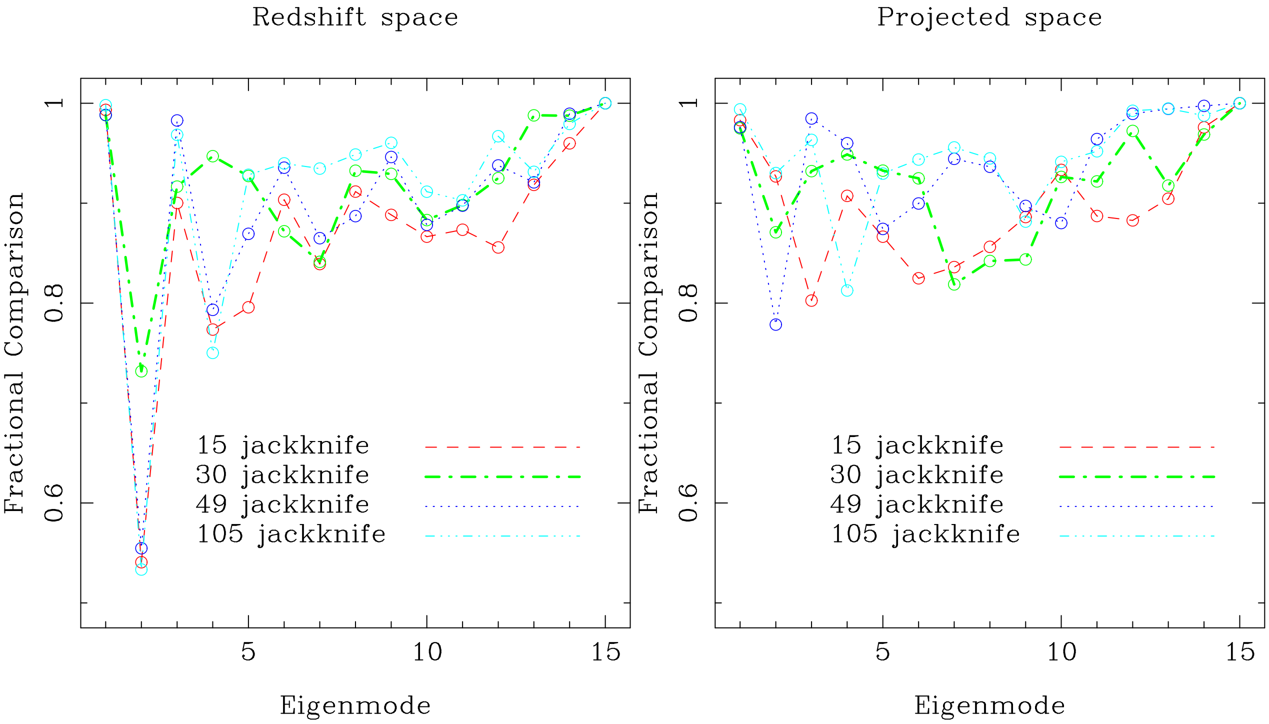


Figure 6.20 We compute the compatibility of the *subspace* contained in a series of eigenvectors such that the y-axis can be interpreted as the fractional ”match” between the spaces the eigenvectors probe. A value of 1.0 means no mismatch in the space they probe, and 0.0 means no overlap (i.e. orthogonal). The left panel uses the eigenvectors of the redshift space covariance matrix determined by the 49 mocks as the reference value. The right plot does the same, but uses measurements in projected space. The comparison is cumulative (eigenmode 3 means the sum of the 1st 3 modes).

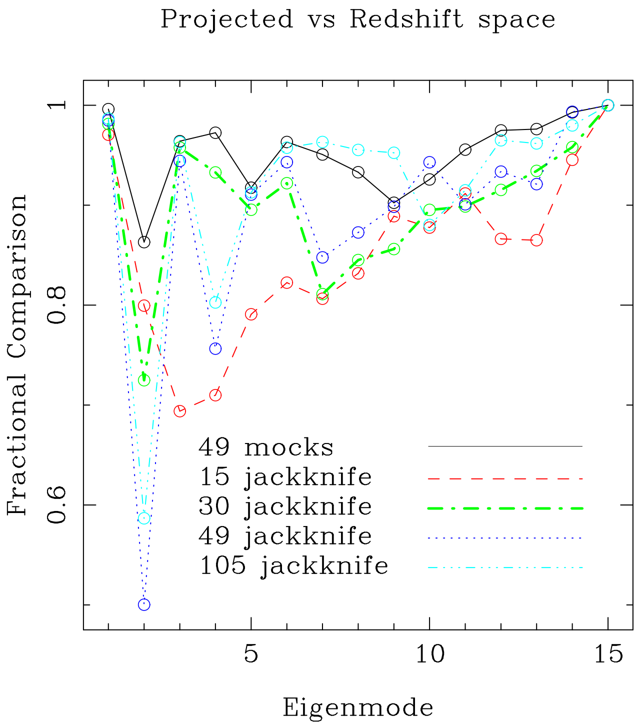


Figure 6.21 We use the subspace comparison of eigenvectors to estimate the difference in space probed between similar numbers of eigenmodes in redshift and projected covariance matrices for each of the error estimates.

15 jackknife sample that is less than 0.7.

We compared several properties of covariance matrices estimated from jackknife re-sampling and independent galaxy mock catalogs. While we noted some concerning discrepancies, we found these typically affected only the least significant eigenmodes. We found many similarities, including physical descriptions for the first three eigenmodes which account an overwhelming majority of the variance. We established the need to trim noisy, unresolved modes from the covariance. When trimmed, and the eigenmode analysis is properly utilized, we noted few significant differences except in the case of 15 jackknife samples. We conclude that our use of 30 jackknife samples should be sufficient for our constraints. We caution that the resolution of errors and the choice of binning scheme relate in a non-trivial manner. In fact, we chose “large” bins (fiducial scheme with $f = 0.25$) to ensure a smooth, signal dominant structure in the covariance matrix. Overall, this error comparison comparison supports our claim that accurate results can still be obtained even with less-than-optimal error estimation such as jackknife re-sampling.

7.0 CONCLUSIONS

We presented a detailed analysis of the shape or *configuration* dependence of the three-point correlation function (3PCF) using SDSS galaxy samples. In this chapter, we summarize the main scientific results of this work. Measurements of the reduced 3PCF, primarily discussed in Chapter 4, are summarized in §7.1, where we investigated intermediate to large scales ($3 - 27h^{-1}$Mpc) encompassing the non-linear, quasi-linear and linear regime by using three volume-limited galaxy samples characterized by luminosity and color. In §7.2, we summarize our fits to a non-linear galaxy-mass bias model. We review the effects of systematics on these analyses in §7.3. We summarize the new insight we obtained from the projected measurements in §7.4. We discuss future extensions in §7.5. This work constitutes the most detailed analysis to date of the configuration dependence in the reduced 3PCF for the SDSS Main galaxy sample.

7.1 SDSS CLUSTERING MEASUREMENTS

In Chapter 4, we employed our novel, parallel n-point calculator to investigate clustering in three SDSS DR6 galaxy samples, both in redshift and projected space. We found our measurements show consistency with ΛCDM predictions at all scales in the 2PCF, equilateral 3PCF and configuration dependence of the 3PCF. We agreed favorably even when including non-linear effects of the ΛCDM model as predicted by a high resolution N-body simulation (Hubble Volume; Colberg et al., 2000; Evrard et al., 2002). We found that larger scales show increased configuration dependence, the "V-shape" signature of filamentary structures.

We found significant configuration dependence on all scales in the reduced 3PCF, $Q(\theta)$, in disagreement with the *hierarchical ansatz*. We noted weaker configuration dependence for the smaller scale triangles ($3-9h^{-1}$Mpc), especially in the redshift space $Q_z(\theta)$, which might explain why other work using less detailed measurements (e.g. Kayo et al. 2004; Nichol et al. 2006) find little or no configuration dependence at these (and smaller) scales. In agreement with Kayo et al. (2004), we showed that the redshift space equilateral 3PCF, Qeq, agrees with hierarchical scaling (flat for $1-10h^{-1}$Mpc with a value $Q_{eq} \approx 0.8$). However, our results in projected space showed Q_{eq} varying with scale below $5h^{-1}$Mpc, strongly suggesting that the effects of redshift distortions make Q_{eq} *appear* to follow the hierarchical scaling.

In Chapter 4, we presented measurements for three samples composed of galaxies with differing luminosity. Our measurements for the 2PCF showed strong agreement with Zehavi et al. (2005), with stronger clustering for brighter galaxy samples. Results on the 3PCF showed much weaker luminosity dependence, statistically significant only at the smaller scales ($\sim 3-10h^{-1}$Mpc). The slight changes in clustering due to luminosity appeared to stay constant over scale but the errors grew with scale thereby reducing the significance. These findings agree with previous analyses of 2dFGRS galaxies (Jing and Börner, 2004; Wang et al., 2004), but are in contrast to previous analyses of SDSS galaxies (Kayo et al., 2004). Given the weak luminosity dependence, we attribute the discrepancy to subtle differences in measurements as described in Gaztañaga and Scoccimarro (2005), as well as the increased significance of our measurements using a much larger galaxy sample based on DR6.

We found that the 3PCF depends on galaxy color and luminosity in different ways. On small scales ($3-9h^{-1}$Mpc), we found a stronger color dependence between red and blue color samples in agreement with predictions by Takada and Jain (2003). Qualitatively, this result agrees with statements in Blanton et al. (2005a) that galaxy color is strongly tied to local environment. On small scales, the red population showed stronger configuration dependence than the blue galaxies, which was most evident in $Q_{proj}(\theta)$. This result suggests an anisotropic distribution of galaxies outside large halos, perhaps tracing the infalling region. This interpretation is consistent with results from observations of angular distributions of satellite galaxies (Azzaro et al., 2007), which find that red satellites correlate well with the major axis of red hosts. On large scales ($9-27h^{-1}$Mpc), we showed no statistically significant difference between "red" and "blue" subsamples given the size of the errors.

We showed that the covariance matrix of the 3PCF displays significant structure and correlation in agreement with theory (Gaztañaga and Scoccimarro, 2005). We conclude from this that the assumption of "diagonal" errors is not valid, and the full covariance must be used for any detailed constraints. We pointed out that insufficiently resolved errors (bins that are too small) will *not* properly represent the covariance, as Poisson errors will be independent. We showed that projected measurements have increased correlation between bins, as we expect from a projection. We also conclude that the structure of the covariance matrix varies with luminosity, emphasizing that care must be taken to properly estimate covariance for individual measurements.

In agreement with Nichol et al. (2006), we found "super structures" (SS) strongly affect measurements of the 3PCF. On large scales, the Sloan Great Wall has a dramatic effect. On smaller scales, other SS can dominate measurements and significantly change the amplitude of the 3PCF as well as bias estimates of the covariance (when using jackknife re-sampling methods). This argues that even the impressive volume of the SDSS Main galaxy sample might not be big enough to obtain "true" measurements of the galaxy field, and are biased by a few SS at all scales we investigated ($3 - 27h^{-1}$Mpc).

7.2 GALAXY-MASS BIAS

The configuration dependence of the 3PCF probes both the linear and quadratic bias parameters, without being degenerate with cosmology as is the case with the 2PCF (specifically linear bias and σ_8). In Chapter 5, we resolved the strong correlation between the linear and quadratic galaxy-mass bias terms, which can account for the weak luminosity dependence seen in the 3PCF, often reported as insignificant in previous studies such as Kayo et al. (2004).

In Chapter 5, we concluded that galaxies remain a *biased* tracer of the mass field with a stronger bias associated with greater luminosity, in qualitative agreement with expectations and other analyses (Zehavi et al., 2005; Gaztañaga et al., 2005). For our BRIGHT ($M_r < -21.5$) galaxy sample, we rule out galaxies being an unbiased tracer of the mass at greater than 4.5σ for $Q_z(\theta)$ and better than 2.5σ for $Q_{proj}(\theta)$. For a fainter sample around L_* ($-21.5 < M_r < -20.5$), we find it is biased at about the 1σ level. While not significant in itself, we expect $\sim L_*$ galaxies to have a linear bias around unity (Zehavi et al., 2005). To obtain these results, we used the full two parameter likelihood space for the linear and quadratic terms, which included their correlation (see Table 5.1).

We constrained values of both the linear and quadratic bias, but could not rule out a zero quadratic bias at greater than the 1.6σ level in the BRIGHT sample with $Q_z(\theta)$ and not at all for Q_{proj} when we restricted ourselves to scales above 9 h^{-1}Mpc. The LSTAR sample ($-21.5 < M_r < -20.5$) is almost completely consistent with zero quadratic bias (see Table 5.2 for full details). Our bias values appeared consistent with a bispectrum analysis (Fourier transform of the 3PCF) on SDSS samples (Nishimichi et al., 2007, figure 8), but they do not report their errors. Our constraints are less precise with respect to an analysis on 2dFGRS (Gaztañaga et al., 2005), where they fit $Q_z(\theta)$ using a very similar analysis. The galaxy samples they analyze are selected using a different bandpass, so direct comparison is difficult. As our results show, different samples will result in different bias values where fainter galaxies are consistent with a zero quadratic bias term.

We presented implications for cosmology in Chapter 5 by calculating implied values of σ_8 based on the best fit linear bias term and measurements of the 2PCF. We restricted this prediction to only use the largest scales ($9-27$ h^{-1}Mpc) where the bias model is most robust. We showed values of σ_8 range between 0.83 and 1.13, in reasonable agreement with a joint bias and σ_8 fit from 2dFGRS (Pan and Szapudi, 2005), where they find $\sigma_8 = 0.9 - 1.07$. This is an impressive match, as we used far less data for our result (only the configuration dependence at one scale where $r_1 = 9h^{-1}$Mpc). We conclude our implied values are consistent as they correspond to a detailed analysis that includes WMAP5 results and find $\sigma_8 = 0.82$ (Komatsu et al., 2009).

Finally, we investigated relative bias and find our predictions for the 2PCF are in good agreement with other lower order analyses with $b/b_* \approx 1.3$ (Zehavi et al., 2005, figure 11). Our relative bias results suggested that joint two-parameter bias fits *underpredict* the value of linear bias when compared to our 2PCF data. We find this is more of a problem in redshift space at smaller scales ($r_1 = 6h^{-1}$Mpc) than projected. This is due to the strong correlation between the two bias terms, and we pointed out that the simple prescription of bias we used might be breaking down at the small scales in redshift space. When we considered the best fit χ^2 values of the fits, we concluded the inclusion of the smaller triangles ($r_1 = 6h^{-1}$Mpc) in redshift space produced poor fits to the simple model. Projected space might reduce the effects of redshift distortions, making the bias model more accurate or the larger errors might obscure the accuracy of the model. Overall, the relative bias of the 3PCF showed good consistency with the best fit bias results and provided a good consistency check on the framework of our quantitative fits.

7.3 METHODS AND SYSTEMATICS

Throughout Chapter 6, we investigated several systematics that can affect the 3PCF, being careful to represent observational limitations (such as galaxy selection). We showed that not accounting for sky completeness significantly alters the amplitude of the 3PCF at the 10%-20% level. This occurred even though we carefully analyzed a large clean galaxy sample, and can potentially bias derived constraints. We showed that a choice of binning scheme can mask or alter the expected signal in the 3PCF and its covariance, in agreement with Gaztañaga and Scoccimarro (2005). We found that a subtle tradeoff exists between using bins large enough to resolve the 3PCF signal and errors, and oversmoothing that can increase degeneracy between fitted parameters and produce singular covariance matrices. We conclude the problems of oversmoothing can be largely mitigated by a sufficient analysis technique such as an eigenmode analysis.

We investigated the reliability of cross validation error estimation (jackknife re-sampling) with respect to independent mock catalogs based on 49 independent N-body simulations. We concluded that although diagonal errors are well approximated, none of our jackknife estimates (ranging between 15 and 105 jackknife regions) could exactly reproduce the correlation represented in the mock based covariance matrix. We highlighted the danger of using too few jackknife samples, as well as using the full covariance matrix that contained noisy modes. These effects can result in false best fit values with severe overestimates of the signal-to-noise. However, we found an eigenmode analysis can account for unresolved differences if care is taken to isolate the main eigenmodes. We concluded that with sufficient analysis techniques, both methods agree and should not bias our results.

7.4 PROJECTED MEASUREMENTS TO MITIGATE REDSHIFT DISTORTIONS

In Chapters 4 and 5, we investigated the reduced 3PCF in redshift and projected space. We investigated dark matter particles in the Hubble Volume simulation where we added redshift distortions from particle velocities. In Chapter 6, we concluded that projected measurements can mitigate the impact of redshift distortions but could not entirely negate the effects.

In measurements of SDSS galaxy samples presented in Chapter 4, projected space recovers power at small scales that is lost due to redshift distortions, both in the equilateral 3PCF and configuration dependence in $Q(\theta)$. At scales approaching the linear regime (greater than $9h^{-1}$Mpc), projected measurements appeared consistent with redshift space, but with larger uncertainties since the line-of-sight projection mixes scales. We saw in Chapter 5 that this results in lower signal-to-noise and less optimal constraints obtained from Q_{proj}, in contrast to theoretical predictions (Zheng, 2004). We caution that this result might be due to limited volume effects that more severely affect the 3PCF in the finite size of SDSS Main galaxy samples. A larger galaxy sample volume might allow a larger projection to be robust. For example, SDSS-III intends to produce such a sample of luminous red galaxies (LRGs) in the BOSS survey. We noted modest improvements and simplicity in modeling the quasi-linear regime with Q_{proj}, but the weak statistical improvement makes this no more than a suggestion. We conclude that projected space 3PCF measurements show the best promise at smaller scales where it can recover shape dependence otherwise obscured due to non-linear effects of redshift distortions.

7.5 FUTURE WORK

7.5.1 Galaxy Evolution: Higher Order Constraints on the Halo Model

A natural extension of this work is to further understand the luminosity and color dependence in the galaxy 3PCF, especially on small to medium scales ($r < 5h^{-1}$Mpc). The correlation functions encode many properties of galaxy formation and evolution; small scale measurements probe the local environment of galaxies, as well as having substantial statistical power in galaxy surveys. The halo model (reviewed in Cooray and Sheth, 2002) and halo occupation distributions (HOD; Berlind and Weinberg, 2002) provide more physically motivated parameterizations of galaxy-mass bias. In contrast to the galaxy-mass bias parameters, which require large scales approximating the linear regime, the halo model can be best constrained at small scales. HOD parameters effectively encode vital statistical information of observations, providing an accessible means to compare with models and better understand galaxy formation, including non-linear processes such as merging and feedback.

Measurements of the projected 2PCF on SDSS galaxies (Zehavi et al., 2005) have been used as a basis for detailed HOD constraints (Tinker et al., 2005; Zheng et al., 2007). Measurements of the 3PCF can break degeneracies in parameters unresolvable with the 2PCF alone. For example, the 3PCF can help disentangle two degenerate parameters in the 2PCF: the slope of the average number of galaxies per halo mass and the minimum mass of a halo to contain a single galaxy (Kulkarni et al., 2007). The use of projected correlation functions is crucial for reliable measurements in the highly non-linear regime. A thorough investigation with the projected 3PCF could give much stronger constraints on these parameters.

We also have shape information encoded in the 3PCF. Although often assumed to be spherical, dark matter halos (mass overdensities where we believe galaxies reside) might be elongated or triaxial in nature. Triaxial profiles have been incorporated into a halo model description, but the impact on the 2PCF remains minimal. Recent work on the bispectrum (Fourier analog of the 3PCF) suggests that a distinct triaxial signature exists (Smith et al., 2006). This information can help to further refine formulations of the halo model for the 3PCF (Takada and Jain, 2003; Wang et al., 2004; Fosalba et al., 2005). We expect galaxies to trace the halo shape, and a small scale measurement of galaxies might yield a statistical measure of halo triaxiality. This is an optimal use of the projected 3PCF, and a novel application to observations.

7.5.2 Understanding Errors: Making Mock Universes

The nature of the errors needs to be well understood to maximize constraints of the data. Higher order moments are known to be more sensitive to systematics and a more detailed understanding of the practical uncertainties in the 3PCF would be helpful. Recently, some of these complications have been addressed for the redshift space 3PCF (Gaztañaga and Scoccimarro, 2005) and the methods applied in analysis of the final release of the 2dF galaxy survey Gaztañaga et al. (2005), along with our analysis on the SDSS. However, many complications still exist that require careful analysis for specific galaxy samples, such as the luminosity dependence of the covariance (see Chapter 4).

The best way to resolve errors requires a large number of realistic mock catalogs generated from expensive N-body simulations. Ultimately, a series of independently evolved N-body simulations with the same cosmology would provide a good basis for realistic mock galaxy catalogs. We can place artificial galaxies corresponding to the actual statistical properties of SDSS galaxy samples creating an ensemble to adequately address error resolution and aid constraints on galaxy evolution and cosmology.

Producing these realistic galaxy mock catalogs from 200 high resolution N-body simulations has been the first major goal of the LasDamas (LArge Suite of DArk MAtter Simulation) collaboration, which consists of the following primary investigators: Andreas Berlind (Vanderbilt University), Roman Scoccimarro (New York University), Risa Wechsler (Stanford) and Frank van den Bosch (Yale). We have already released publicly available galaxy mock catalogs corresponding to five SDSS galaxy sample (three for Main samples and two for LRGs) generated from over 120 independent high-resolution N-body simulations. We use these galaxy mocks to resolve the covariance matrix of samples, and a platform to investigate additional methods for error estimation recently introduced into cosmology (Pope and Szapudi, 2008). In addition to the public catalogs, we can generate less realistic galaxy mocks to test effects of observational systematics (e.g. switch off redshift distortions and compare). An empirical approach is necessary to test the non-linear and subtle contributions in a robust way.

BIBLIOGRAPHY

K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, J. C. Barentine, B. A. Bassett, A. C. Becker, T. C. Beers, E. F. Bell, V. Belokurov, A. A. Berlind, E. F. Berman, M. Bernardi, S. J. Bickerton, D. Bizyaev, J. P. Blakeslee, M. R. Blanton, J. J. Bochanski, W. N. Boroski, H. J. Brewington, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, L. N. Carey, S. Carliles, M. A. Carr, F. J. Castander, D. Cinabro, A. J. Connolly, I. Csabai, C. E. Cunha, P. C. Czarapata, J. R. A. Davenport, E. de Haas, B. Dilday, M. Doi, D. J. Eisenstein, M. L. Evans, N. W. Evans, X. Fan, S. D. Friedman, J. A. Frieman, M. Fukugita, B. T. Gänsicke, E. Gates, B. Gillespie, G. Gilmore, B. Gonzalez, C. F. Gonzalez, E. K. Grebel, J. E. Gunn, Z. Györy, P. B. Hall, P. Harding, F. H. Harris, M. Harvanek, S. L. Hawley, J. J. E. Hayes, T. M. Heckman, J. S. Hendry, G. S. Hennessy, R. B. Hindsley, J. Hoblitt, C. J. Hogan, D. W. Hogg, J. A. Holtzman, J. B. Hyde, S.-i. Ichikawa, T. Ichikawa, M. Im, Ž. Ivezić, S. Jester, L. Jiang, J. A. Johnson, A. M. Jorgensen, M. Jurić, S. M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, K. Konishi, R. G. Kron, J. Krzesinski, N. Kuropatkin, H. Lampeitl, S. Lebedeva, M. G. Lee, Y. S. Lee, R. F. Leger, S. Lépine, N. Li, M. Lima, H. Lin, D. C. Long, C. P. Loomis, J. Loveday, R. H. Lupton, E. Magnier, O. Malanushenko, V. Malanushenko, R. Mandelbaum, B. Margon, J. P. Marriner, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, H. L. Morrison, F. Mullally, J. A. Munn, T. Murphy, T. Nash, A. Nebot, E. H. Neilsen, H. J. Newberg, P. R. Newman, R. C. Nichol, T. Nicinski, M. Nieto-Santisteban, A. Nitta, S. Okamura, D. J. Oravetz, J. P. Ostriker, R. Owen, N. Padmanabhan, K. Pan, C. Park, G. Pauls, J. Peoples, W. J. Percival, J. R. Pier, A. C. Pope, D. Pourbaix, P. A. Price, N. Purger, T. Quinn, M. J. Raddick, P. R. Fiorentin, G. T. Richards, M. W. Richmond, A. G. Riess, H.-W. Rix, C. M. Rockosi, M. Sako, D. J. Schlegel, D. P. Schneider, R.-D. Scholz, M. R. Schreiber, A. D. Schwope, U. Seljak, B. Sesar, E. Sheldon, K. Shimasaku, V. C. Sibley, A. E. Simmons, T. Sivarani, J. A. Smith, M. C. Smith, V. Smolčić, S. A. Snedden, A. Stebbins, M. Steinmetz, C. Stoughton, M. A. Strauss, M. Subba Rao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tanaka, M. Tegmark, L. F. A. Teodoro, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, Y. Wadadekar, S. Watters, D. H. Weinberg, A. A. West, S. D. M. White, B. C. Wilhite, A. C. Wonders, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, S. Zibetti, and D. B. Zucker. The Seventh Data Release of the Sloan Digital Sky Survey. ApJS, 182:543–558, June 2009. doi: 10.1088/0067-0049/182/2/543.

J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, T. C. Beers, V. Belokurov, A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, D. M. Bramich, H. J. Brewington, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, L. N. Carey, S. Carliles, M. A. Carr, F. J. Castander, A. J. Connolly, R. J. Cool, C. E. Cunha, I. Csabai, J. J. Dalcanton, M. Doi, D. J. Eisenstein, M. L. Evans, N. W. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, G. Gilmore, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, E. de Haas, P. B. Hall, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, G. S. Hennessy, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. A. Holtzman, S.-i. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Jester, D. E. Johnston, A. M. Jorgensen, M. Jurić, G. Kauffmann, S. M. Kent, S. J. Kleinman, G. R. Knapp, A. Y. Kniazev, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, B. C. Lee, R. F. Leger, M. Lima, H. Lin, D. C. Long, J. Loveday, R. H. Lupton, R. Mandelbaum, B. Margon, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, J. A. Munn, R. Nakajima, T. Nash, E. H. Neilsen, Jr., H. J. Newberg, R. C. Nichol, M. Nieto-Santisteban, A. Nitta, H. Oyaizu, S. Okamura, J. P. Ostriker, N. Padmanabhan, C. Park, J. J. Peoples, J. R. Pier, A. C. Pope, D. Pourbaix, T. R. Quinn, M. J. Raddick, P. Re Fiorentin, G. T. Richards, M. W. Richmond, H.-W. Rix, C. M. Rockosi, D. J. Schlegel, D. P. Schneider, R. Scranton, U. Seljak, E. Sheldon, K. Shimasaku, N. M. Silvestri, J. A. Smith, V. Smolčić, S. A. Snedden, A. Stebbins, C. Stoughton, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, D. H. Weinberg, A. A. West, S. D. M. White, B. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, S. Zibetti, and D. B. Zucker. The Fifth Data Release of the Sloan Digital Sky Survey. ApJS, 172:634–644, October 2007. doi: 10.1086/518864.

J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, B. A. Bassett, A. C. Becker, T. C. Beers, E. F. Bell, A. A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, S. Carliles, M. A. Carr, F. J. Castander, D. Cinabro, R. J. Cool, K. R. Covey, I. Csabai, C. E. Cunha, J. R. A. Davenport, B. Dilday, M. Doi, D. J. Eisenstein, M. L. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. T. Gänsicke, E. Gates, B. Gillespie, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, V. K. Gurbani, P. B. Hall, P. Harding, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. B. Hyde, S.-i. Ichikawa, Ž. Ivezić, S. Jester, J. A. Johnson, A. M. Jorgensen, M. Jurić, S. M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, S. Lebedeva, Y. S. Lee, R. F. Leger, S. Lépine, M. Lima, H. Lin, D. C. Long, C. P. Loomis, J. Loveday, R. H. Lupton, O. Malanushenko, V. Malanushenko, R. Mandelbaum, B. Margon, J. P. Marriner, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, H. L. Morrison, J. A. Munn, R. Nakajima, E. H. Neilsen, Jr., H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nieto-Santisteban, A. Nitta, S. Okamura, R. Owen, H. Oyaizu, N. Padmanabhan, K. Pan, C. Park, J. J. Peoples,

J. R. Pier, A. C. Pope, N. Purger, M. J. Raddick, P. Re Fiorentin, G. T. Richards, M. W. Richmond, A. G. Riess, H.-W. Rix, C. M. Rockosi, M. Sako, D. J. Schlegel, D. P. Schneider, M. R. Schreiber, A. D. Schwope, U. Seljak, B. Sesar, E. Sheldon, K. Shimasaku, T. Sivarani, J. A. Smith, S. A. Snedden, M. Steinmetz, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, Y. Wadadekar, D. H. Weinberg, A. A. West, S. D. M. White, B. C. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, and D. B. Zucker. The Sixth Data Release of the Sloan Digital Sky Survey. ApJS, 175:297–313, April 2008. doi: 10.1086/524984.

M. Azzaro, S. G. Patiri, F. Prada, and A. R. Zentner. Angular distribution of satellite galaxies from the Sloan Digital Sky Survey Data Release 4. MNRAS, 376:L43–L47, March 2007. doi: 10.1111/j.1745-3933.2007.00282.x.

I. K. Baldry, K. Glazebrook, J. Brinkmann, Ž. Ivezić, R. H. Lupton, R. C. Nichol, and A. S. Szalay. Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. ApJ, 600: 681–694, January 2004. doi: 10.1086/380092.

A. A. Berlind and D. H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ, 575:587–616, August 2002. doi: 10.1086/341469.

F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro. Large-scale structure of the Universe and cosmological perturbation theory. Phys. Rep., 367:1–3, September 2002. doi: 10.1016/S0370-1573(02)00135-7.

M. R. Blanton, D. W. Hogg, N. A. Bahcall, J. Brinkmann, M. Britton, A. J. Connolly, I. Csabai, M. Fukugita, J. Loveday, A. Meiksin, J. A. Munn, R. C. Nichol, S. Okamura, T. Quinn, D. P. Schneider, K. Shimasaku, M. A. Strauss, M. Tegmark, M. S. Vogeley, and D. H. Weinberg. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. ApJ, 592:819–838, August 2003a. doi: 10.1086/375776.

M. R. Blanton, H. Lin, R. H. Lupton, F. M. Maley, N. Young, I. Zehavi, and J. Loveday. An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey “Tiling” Algorithm. AJ, 125:2276–2286, April 2003b. doi: 10.1086/344761.

M. R. Blanton, D. Eisenstein, D. W. Hogg, D. J. Schlegel, and J. Brinkmann. Relationship between Environment and the Broadband Optical Properties of Galaxies in the Sloan Digital Sky Survey. ApJ, 629:143–157, August 2005a. doi: 10.1086/422897.

M. R. Blanton, D. J. Schlegel, M. A. Strauss, J. Brinkmann, D. Finkbeiner, M. Fukugita, J. E. Gunn, D. W. Hogg, Ž. Ivezić, G. R. Knapp, R. H. Lupton, J. A. Munn, D. P. Schneider, M. Tegmark, and I. Zehavi. New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. AJ, 129:2562–2578, June 2005b. doi: 10.1086/429803.

M. Cacciato, F. C. van den Bosch, S. More, R. Li, H. J. Mo, and X. Yang. Galaxy clustering and galaxy-galaxy lensing: a promising union to constrain cosmological parameters. MNRAS, 394:929–946, April 2009. doi: 10.1111/j.1365-2966.2008.14362.x.

J. M. Colberg, S. D. M. White, N. Yoshida, T. J. MacFarland, A. Jenkins, C. S. Frenk, F. R. Pearce, A. E. Evrard, H. M. P. Couchman, G. Efstathiou, J. A. Peacock, P. A. Thomas, and The Virgo Consortium. Clustering of galaxy clusters in cold dark matter universes. MNRAS, 319:209–214, November 2000. doi: 10.1046/j.1365-8711.2000.03832.x.

M. Colless, G. Dalton, S. Maddox, W. Sutherland, P. Norberg, S. Cole, J. Bland-Hawthorn, T. Bridges, R. Cannon, C. Collins, W. Couch, N. Cross, K. Deeley, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, D. Madgwick, J. A. Peacock, B. A. Peterson, I. Price, M. Seaborne, and K. Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. MNRAS, 328: 1039–1063, December 2001. doi: 10.1046/j.1365-8711.2001.04902.x.

A. J. Connolly, R. Scranton, D. Johnston, S. Dodelson, D. J. Eisenstein, J. A. Frieman, J. E. Gunn, L. Hui, B. Jain, S. Kent, J. Loveday, R. C. Nichol, L. O'Connell, M. Postman, R. Scoccimarro, R. K. Sheth, A. Stebbins, M. A. Strauss, A. S. Szalay, I. Szapudi, M. Tegmark, M. S. Vogeley, I. Zehavi, J. Annis, N. Bahcall, J. Brinkmann, I. Csabai, M. Doi, M. Fukugita, G. S. Hennessy, R. Hindsley, T. Ichikawa, Ž. Ivezić, R. S. J. Kim, G. R. Knapp, P. Kunszt, D. Q. Lamb, B. C. Lee, R. H. Lupton, T. A. McKay, J. Munn, J. Peoples, J. Pier, C. Rockosi, D. Schlegel, C. Stoughton, D. L. Tucker, B. Yanny, and D. G. York. The Angular Correlation Function of Galaxies from Early Sloan Digital Sky Survey Data. ApJ, 579:42–47, November 2002. doi: 10.1086/342787.

A. Cooray and R. Sheth. Halo models of large scale structure. Phys. Rep., 372:1–129, December 2002. doi: 10.1016/S0370-1573(02)00276-4.

M. Crocce, S. Pueblas, and R. Scoccimarro. Transients from initial conditions in cosmological simulations. MNRAS, 373:369–381, November 2006. doi: 10.1111/j.1365-2966.2006.11040.x.

M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. ApJ, 292:371–394, May 1985. doi: 10.1086/163168.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. pages 137–150, December 2004. URL http://www.usenix.org/events/osdi04/tech/dean.html.

A. E. Evrard, T. J. MacFarland, H. M. P. Couchman, J. M. Colberg, N. Yoshida, S. D. M. White, A. Jenkins, C. S. Frenk, F. R. Pearce, J. A. Peacock, and P. A. Thomas. Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations. ApJ, 573:7–36, July 2002. doi: 10.1086/340551.

P. Fosalba, J. Pan, and I. Szapudi. Cosmological Three-Point Function: Testing the Halo Model against Simulations. ApJ, 632:29–48, October 2005. doi: 10.1086/432906.

J. N. Fry and E. Gaztanaga. Biasing and hierarchical statistics in large-scale structure. ApJ, 413:447–452, August 1993. doi: 10.1086/173015.

J. P. Gardner, A. Connolly, and C. McBride. A Framework for Analyzing Massive Astrophysical Datasets on a Distributed Grid. In R. A. Shaw, F. Hill, & D. J. Bell, editor, *Astronomical Data Analysis Software and Systems XVI*, volume 376 of *Astronomical Society of the Pacific Conference Series*, pages 69–+, October 2007.

E. Gaztañaga and R. Scoccimarro. The three-point function in large-scale structure: redshift distortions and galaxy bias. MNRAS, 361:824–836, August 2005. doi: 10.1111/j.1365-2966.2005.09234.x.

E. Gaztañaga, P. Norberg, C. M. Baugh, and D. J. Croton. Statistical analysis of galaxy surveys - II. The three-point galaxy correlation function measured from the 2dFGRS. MNRAS, 364:620–634, December 2005. doi: 10.1111/j.1365-2966.2005.09583.x.

J. R. I. Gott, M. Jurić, D. Schlegel, F. Hoyle, M. Vogeley, M. Tegmark, N. Bahcall, and J. Brinkmann. A Map of the Universe. ApJ, 624:463–484, May 2005. doi: 10.1086/428890.

A. G. Gray, A. W. Moore, R. C. Nichol, A. J. Connolly, C. Genovese, and L. Wasserman. Multi-Tree Methods for Statistics on Very Large Datasets in Astronomy. In F. Ochsenbein, M. G. Allen, & D. Egret, editor, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, pages 249–+, July 2004.

J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. de Haas, Ž. Ivezić, G. Knapp, R. Lupton, G. Pauls, R. Simcoe, R. Hirsch, D. Sanford, S. Wang, D. York, F. Harris, J. Annis, L. Bartozek, W. Boroski, J. Bakken, M. Haldeman, S. Kent, S. Holm, D. Holmgren, D. Petravick, A. Prosapio, R. Rechenmacher, M. Doi, M. Fukugita, K. Shimasaku, N. Okada, C. Hull, W. Siegmund, E. Mannery, M. Blouke, D. Heidtman, D. Schneider, R. Lucinio, and J. Brinkman. The Sloan Digital Sky Survey Photometric Camera. AJ, 116:3040–3081, December 1998. doi: 10.1086/300645.

A. H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. Phys. Rev. D, 23:347–356, January 1981. doi: 10.1103/PhysRevD.23.347.

A. J. S. Hamilton. Toward Better Ways to Measure the Galaxy Correlation Function. ApJ, 417:19–+, November 1993. doi: 10.1086/173288.

A. J. S. Hamilton. Linear Redshift Distortions: a Review. In D. Hamilton, editor, *The Evolving Universe*, volume 231 of *Astrophysics and Space Science Library*, pages 185–+, 1998.

A. J. S. Hamilton and M. Tegmark. A scheme to deal accurately and efficiently with complex angular masks in galaxy surveys. MNRAS, 349:115–128, March 2004. doi: 10.1111/j.1365-2966.2004.07490.x.

C. Hikage, T. Matsubara, Y. Suto, C. Park, A. S. Szalay, and J. Brinkmann. Fourier Phase Analysis of SDSS Galaxies. PASJ, 57:709–718, October 2005.

D. W. Hogg. Distance measures in cosmology. *ArXiv Astrophysics e-prints*, May 1999.

D. W. Hogg, I. K. Baldry, M. R. Blanton, and D. J. Eisenstein. The K correction. *ArXiv Astrophysics e-prints*, October 2002.

E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15:168–173, March 1929. doi: 10.1073/pnas.15.3.168.

Y. P. Jing and G. Börner. The Three-Point Correlation Function of Galaxies Determined from the Two-Degree Field Galaxy Redshift Survey. ApJ, 607:140–163, May 2004. doi: 10.1086/383343.

N. Kaiser. Clustering in real space and in redshift space. MNRAS, 227:1–21, July 1987.

I. Kayo, Y. Suto, R. C. Nichol, J. Pan, I. Szapudi, A. J. Connolly, J. Gardner, B. Jain, G. Kulkarni, T. Matsubara, R. Sheth, A. S. Szalay, and J. Brinkmann. Three-Point Correlation Functions of SDSS Galaxies in Redshift Space: Morphology, Color, and Luminosity Dependence. PASJ, 56:415–423, June 2004.

M. Kerscher, I. Szapudi, and A. S. Szalay. A Comparison of Estimators for the Two-Point Correlation Function. ApJ, 535:L13–L16, May 2000. doi: 10.1086/312702.

E. Komatsu, J. Dunkley, M. R. Nolta, C. L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. Limon, L. Page, D. N. Spergel, M. Halpern, R. S. Hill, A. Kogut, S. S. Meyer, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright. Five-Year Wilkinson Microwave Anisotropy Probe Observations: Cosmological Interpretation. ApJS, 180:330–376, February 2009. doi: 10.1088/0067-0049/180/2/330.

G. V. Kulkarni, R. C. Nichol, R. K. Sheth, H.-J. Seo, D. J. Eisenstein, and A. Gray. The three-point correlation function of luminous red galaxies in the Sloan Digital Sky Survey. MNRAS, 378:1196–1206, July 2007. doi: 10.1111/j.1365-2966.2007.11872.x.

S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. ApJ, 412:64–71, July 1993. doi: 10.1086/172900.

A. R. Liddle and D. H. Lyth. *Cosmological Inflation and Large-Scale Structure.* April 2000.

J. Loveday. Evolution of the galaxy luminosity function at z ¡ 0.3. MNRAS, 347:601–606, January 2004. doi: 10.1111/j.1365-2966.2004.07230.x.

R. Lupton, J. E. Gunn, Z. Ivezić, G. R. Knapp, and S. Kent. The SDSS Imaging Pipelines. In F. R. Harnden, Jr., F. A. Primini, and H. E. Payne, editors, *Astronomical Data Analysis Software and Systems X*, volume 238 of *Astronomical Society of the Pacific Conference Series*, pages 269–+, 2001.

F. A. Marín, R. H. Wechsler, J. A. Frieman, and R. C. Nichol. Modeling the Galaxy Three-Point Correlation Function. ApJ, 672:849–860, January 2008. doi: 10.1086/523628.

A. W. Moore, A. J. Connolly, C. Genovese, A. Gray, L. Grone, N. I. Kanidoris, R. C. Nichol, J. Schneider, A. S. Szalay, I. Szapudi, and L. Wasserman. Fast Algorithms and Efficient Statistics: N-Point Correlation Functions. In A. J. Banday, S. Zaroubi, & M. Bartelmann, editor, *Mining the Sky*, pages 71–+, 2001. doi: 10.1007/10849171_5.

R. C. Nichol, R. K. Sheth, Y. Suto, A. J. Gray, I. Kayo, R. H. Wechsler, F. Marin, G. Kulkarni, M. Blanton, A. J. Connolly, J. P. Gardner, B. Jain, C. J. Miller, A. W. Moore, A. Pope, J. Pun, D. Schneider, J. Schneider, A. Szalay, I. Szapudi, I. Zehavi, N. A. Bahcall, I. Csabai, and J. Brinkmann. The effect of large-scale structure on the SDSS galaxy three-point correlation function. MNRAS, 368:1507–1514, June 2006. doi: 10.1111/j.1365-2966.2006.10239.x.

T. Nishimichi, I. Kayo, C. Hikage, K. Yahata, A. Taruya, Y. P. Jing, R. K. Sheth, and Y. Suto. Bispectrum and Nonlinear Biasing of Galaxies: Perturbation Analysis, Numerical Simulation, and SDSS Galaxy Clustering. PASJ, 59:93–106, February 2007.

J. B. Oke and A. Sandage. Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies. ApJ, 154:21–+, October 1968. doi: 10.1086/149737.

J. Pan and I. Szapudi. The monopole moment of the three-point correlation function of the two-degree Field Galaxy Redshift Survey. MNRAS, 362:1363–1370, October 2005. doi: 10.1111/j.1365-2966.2005.09407.x.

P. J. E. Peebles. *The large-scale structure of the universe*. 1980.

A. C. Pope and I. Szapudi. Shrinkage estimation of the power spectrum covariance matrix. MNRAS, 389:766–774, September 2008. doi: 10.1111/j.1365-2966.2008.13561.x.

P. Schechter. An analytic expression for the luminosity function for galaxies. ApJ, 203: 297–306, January 1976. doi: 10.1086/154079.

R. Scoccimarro. The Bispectrum: From Theory to Observations. ApJ, 544:597–615, December 2000. doi: 10.1086/317248.

R. Scoccimarro. Transients from initial conditions: a perturbative analysis. MNRAS, 299: 1097–1118, October 1998. doi: 10.1046/j.1365-8711.1998.01845.x.

R. Scoccimarro and R. K. Sheth. PTHALOS: a fast method for generating mock galaxy distributions. MNRAS, 329:629–640, January 2002. doi: 10.1046/j.1365-8711.2002.04999.x.

R. Scranton, D. Johnston, S. Dodelson, J. A. Frieman, A. Connolly, D. J. Eisenstein, J. E. Gunn, L. Hui, B. Jain, S. Kent, J. Loveday, V. Narayanan, R. C. Nichol, L. O'Connell, R. Scoccimarro, R. K. Sheth, A. Stebbins, M. A. Strauss, A. S. Szalay, I. Szapudi, M. Tegmark, M. Vogeley, I. Zehavi, J. Annis, N. A. Bahcall, J. Brinkman, I. Csabai, R. Hindsley, Z. Ivezic, R. S. J. Kim, G. R. Knapp, D. Q. Lamb, B. C. Lee, R. H. Lupton, T. McKay, J. Munn, J. Peoples, J. Pier, G. T. Richards, C. Rockosi, D. Schlegel, D. P. Schneider, C. Stoughton, D. L. Tucker, B. Yanny, and D. G. York. Analysis of Systematic Effects and Statistical Uncertainties in Angular Clustering of Galaxies from Early Sloan Digital Sky Survey Data. ApJ, 579:48–75, November 2002. doi: 10.1086/342786.

Y. Shiloach and U. Vishkin. An O(log n) parallel connectivity algorithm. *Journal of Algorithms*, 3:57–67, February 1982.

R. E. Smith, P. I. R. Watts, and R. K. Sheth. The impact of halo shapes on the bispectrum in cosmology. MNRAS, 365:214–230, January 2006. doi: 10.1111/j.1365-2966.2005.09707.x.

V. Springel. The cosmological simulation code GADGET-2. MNRAS, 364:1105–1134, December 2005. doi: 10.1111/j.1365-2966.2005.09655.x.

V. Springel, S. D. M. White, G. Tormen, and G. Kauffmann. Populating a cluster of galaxies - I. Results at [formmu2]z=0. MNRAS, 328:726–750, December 2001. doi: 10.1046/j.1365-8711.2001.04912.x.

J. G. Stadel. *Cosmological N-body simulations and their analysis.* PhD thesis, AA(UNIVERSITY OF WASHINGTON), 2001.

M. A. Strauss, D. H. Weinberg, R. H. Lupton, V. K. Narayanan, J. Annis, M. Bernardi, M. Blanton, S. Burles, A. J. Connolly, J. Dalcanton, M. Doi, D. Eisenstein, J. A. Frieman, M. Fukugita, J. E. Gunn, Ž. Ivezić, S. Kent, R. S. J. Kim, G. R. Knapp, R. G. Kron, J. A. Munn, H. J. Newberg, R. C. Nichol, S. Okamura, T. R. Quinn, M. W. Richmond, D. J. Schlegel, K. Shimasaku, M. SubbaRao, A. S. Szalay, D. Vanden Berk, M. S. Vogeley, B. Yanny, N. Yasuda, D. G. York, and I. Zehavi. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. AJ, 124:1810–1824, September 2002. doi: 10.1086/342343.

I. Szapudi. Introduction to Higher Order Spatial Statistics in Cosmology. *ArXiv Astrophysics e-prints*, May 2005.

S. Szapudi and A. S. Szalay. A New Class of Estimators for the N-Point Correlations. ApJ, 494:L41+, February 1998. doi: 10.1086/311146.

M. Takada and B. Jain. The three-point correlation function in cosmology. MNRAS, 340: 580–608, April 2003. doi: 10.1046/j.1365-8711.2003.06321.x.

J. L. Tinker, D. H. Weinberg, Z. Zheng, and I. Zehavi. On the Mass-to-Light Ratio of Large-Scale Structure. ApJ, 631:41–58, September 2005. doi: 10.1086/432084.

F. C. van den Bosch, X. Yang, and H. J. Mo. Linking early- and late-type galaxies to their dark matter haloes. MNRAS, 340:771–792, April 2003. doi: 10.1046/j.1365-8711.2003.06335.x.

Y. Wang, X. Yang, H. J. Mo, F. C. van den Bosch, and Y. Chu. The three-point correlation function of galaxies: comparing halo occupation models with observations. MNRAS, 353: 287–300, September 2004. doi: 10.1111/j.1365-2966.2004.08141.x.

X. Yang, H. J. Mo, and F. C. van den Bosch. Constraining galaxy formation and cosmology with the conditional luminosity function of galaxies. MNRAS, 339:1057–1080, March 2003. doi: 10.1046/j.1365-8711.2003.06254.x.

C. W. Yip, A. J. Connolly, A. S. Szalay, T. Budavári, M. SubbaRao, J. A. Frieman, R. C. Nichol, A. M. Hopkins, D. G. York, S. Okamura, J. Brinkmann, I. Csabai, A. R. Thakar, M. Fukugita, and Ž. Ivezić. Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey. AJ, 128:585–609, August 2004. doi: 10.1086/422429.

D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C.-h. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, S.-i. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, and N. Yasuda. The Sloan Digital Sky Survey: Technical Summary. AJ, 120:1579–1587, September 2000. doi: 10.1086/301513.

I. Zehavi, M. R. Blanton, J. A. Frieman, D. H. Weinberg, H. J. Mo, M. A. Strauss, S. F. Anderson, J. Annis, N. A. Bahcall, M. Bernardi, J. W. Briggs, J. Brinkmann, S. Burles, L. Carey, F. J. Castander, A. J. Connolly, I. Csabai, J. J. Dalcanton, S. Dodelson, M. Doi, D. Eisenstein, M. L. Evans, D. P. Finkbeiner, S. Friedman, M. Fukugita, J. E. Gunn,

G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, G. R. Knapp, R. Kron, P. Kunszt, D. Q. Lamb, R. F. Leger, D. C. Long, J. Loveday, R. H. Lupton, T. McKay, A. Meiksin, A. Merrelli, J. A. Munn, V. Narayanan, M. Newcomb, R. C. Nichol, R. Owen, J. Peoples, A. Pope, C. M. Rockosi, D. Schlegel, D. P. Schneider, R. Scoccimarro, R. K. Sheth, W. Siegmund, S. Smee, Y. Snir, A. Stebbins, C. Stoughton, M. SubbaRao, A. S. Szalay, I. Szapudi, M. Tegmark, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, B. Yanny, and D. G. York. Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data. ApJ, 571:172–190, May 2002. doi: 10.1086/339893.

I. Zehavi, D. H. Weinberg, Z. Zheng, A. A. Berlind, J. A. Frieman, R. Scoccimarro, R. K. Sheth, M. R. Blanton, M. Tegmark, H. J. Mo, N. A. Bahcall, J. Brinkmann, S. Burles, I. Csabai, M. Fukugita, J. E. Gunn, D. Q. Lamb, J. Loveday, R. H. Lupton, A. Meiksin, J. A. Munn, R. C. Nichol, D. Schlegel, D. P. Schneider, M. SubbaRao, A. S. Szalay, A. Uomoto, and D. G. York. On Departures from a Power Law in the Galaxy Correlation Function. ApJ, 608:16–24, June 2004. doi: 10.1086/386535.

I. Zehavi, Z. Zheng, D. H. Weinberg, J. A. Frieman, A. A. Berlind, M. R. Blanton, R. Scoccimarro, R. K. Sheth, M. A. Strauss, I. Kayo, Y. Suto, M. Fukugita, O. Nakamura, N. A. Bahcall, J. Brinkmann, J. E. Gunn, G. S. Hennessy, Ž. Ivezić, G. R. Knapp, J. Loveday, A. Meiksin, D. J. Schlegel, D. P. Schneider, I. Szapudi, M. Tegmark, M. S. Vogeley, and D. G. York. The Luminosity and Color Dependence of the Galaxy Correlation Function. ApJ, 630:1–27, September 2005. doi: 10.1086/431891.

Z. Zheng. Projected Three-Point Correlation Functions and Galaxy Bias. ApJ, 614:527–532, October 2004. doi: 10.1086/423838.

Z. Zheng and D. H. Weinberg. Breaking the Degeneracies between Cosmology and Galaxy Bias. ApJ, 659:1–28, April 2007. doi: 10.1086/512151.

Z. Zheng, A. L. Coil, and I. Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. ApJ, 667:760–779, October 2007. doi: 10.1086/521074.