Analysing Galaxy Clustering for Future Experiments Including the Dark Energy Survey

by

Kelly Nock

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of
Doctor of Philosophy
of the
University of Portsmouth

2010

Copyright

Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

To my mum,

I miss you everyday. This is for you. I love you.

Abstract

The use of Baryon Acoustic Oscillations (BAO) as a standard ruler in the 2-point galaxy clustering signal has proven to be an excellent probe of the cosmological expansion. With the abundance of good quality galaxy data predicted for future large sky surveys, the potential to conduct precision cosmology using clustering analyses is immense. Many of the next generation sky surveys, including the Dark Energy Survey (DES), the Panoramic Survey Telescope and Rapid Response System (PanStarrs), and the Large Synoptic Survey Telescope (LSST), will utilise photometric redshift estimation techniques, which will make it possible to probe wider and deeper regions of the Universe than spectroscopic redshift surveys in an equivalent amount of time. The use of photometric techniques to estimate galaxy redshifts however, induces errors on inferred radial distances. Consequently, the amplitude of the power spectrum and correlation function is reduced in the radial direction by this smoothing. In this regime, precise measurements of the BAO signal will be difficult. Because of this, there is an urgent need to obtain a better understanding of exactly how photometric redshift uncertainties affect 3D clustering analyses, and to investigate alternative clustering analysis techniques that may be used in future experiments.

In this thesis, I investigate the systematic effects arising in the projected correlation function due to redshift-space distortions, and introduce a new binning scheme to eradicate the problem. I also consider the level of systematic uncertainty induced in realistic measurements of the 3D correlation function from conflicting photometric redshift estimation techniques, and highlight a requirement for empirical test results to be incorporated into model predictions of the anisotropic correlation function for future surveys. Finally, I collate my results to make predictions about how BAO can be optimally used in future photometric redshift experiments like the DES.

Preface

The work of this thesis was carried out at the Institute of Cosmology and Gravitation, University of Portsmouth, United Kingdom.

The results of Chapters 3 and 5 were obtained via a collaboration with Dr Will Percival and Dr Ashley Ross, and have been published in a paper entitled *The effect of redshift-space distortions on projected 2-pt clustering measurements* (Nock et al., 2010).

Acknowledgements

Firstly, I would like to thank my supervisor Dr Will Percival for his patience, guidance and support during my time at the ICG. Working under your supervision has been both inspirational and enlightening, which has made for a very enjoyable PhD experience.

I would also like to thank everyone at the ICG, both past and present, for creating a unique, lively and fun place to work over the last few years. It has been an honour and a pleasure to have been part of such a superb research department, and it would not have been possible without support from the Science and Technology Facilities Council and the University of Portsmouth.

Heaps of love and gratitude goes to some very special people who have consistently tolerated my incessant ramblings and endless TV show quotations, constant need for karaoke and 80s movie marathons, and drunken swims in the solent: Gaby Caldera, "*Uncle*" Jim Cresswell, Fabio Silva, Chiara "*kiki*" Tonini, Cris Sabiu, Ben Hoyle, Dominic "*I ♡ CJ*" Galliano, Jo Lester, Jason Foster, Vicky Tunbridge and Dave Griffiths.

A big shout out to all of my musically inclined buddies who have endured my singing or guitar playing over the last few years: Lukas Hollenstein, Andrea Vimercati, Antonio "*Sugarpuss*" Cardoso, Chaz "*You thought you were the bomb*" Shapiro, Gustav Stromback, and Beppe di Risi. My dreams of being a rock star have been realised on many an occasion thanks to you guys.

Very special thanks go to my family who have supported me unconditionally throughout my PhD, especially my parents Pam and Fred.

And finally, I couldn't have done this without you, Jenny. Thank you. I love you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern cosmology is based fundamentally on Einstein's theory of General Relativity (GR) and the Copernican Principle, which states that the Universe is isotropic and homogeneous on large-scales. The current standard model successfully describes observational phenomena across an extremely wide range of times and scales. Over the past century, observations have been continually conducted and refined to provide us with cosmological constraints that are consistent with theoretical predictions and it is often claimed that we have now entered an era of precision cosmology. The past decade in-particular has seen one of the most exciting cosmological discoveries in recent years: the late-time acceleration of the expansion of the Universe (Riess et al., 1998; Perlmutter et al., 1999). Understanding the nature of this acceleration, or *dark energy*, has become one of the main challenges facing cosmologists today.

1.1 The Standard Model of Cosmology

In this chapter we give a general introduction to the concordance cosmological model, highlighting the theory and observations that have moulded it into its current form. We present some of the existing theoretical models proposed to explain the accelerated expansion of the Universe and discuss how clustering analysis techniques may be used to further understand the nature of dark energy.

1.1.1 General Relativity and Einstein's Equations

GR considers the relationship between geometry and matter-energy through Einstein's field equations

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.1}$$

where $G_{\mu\nu}$ is the Einstein tensor and describes geometry, $R_{\mu\nu}$ is the Ricci tensor which depends on the metric and its derivatives, R is the Ricci scalar, and $g_{\mu\nu}$ is the metric tensor. The matter distribution is characterized by the energy-momentum tensor $T_{\mu\nu}$ where $T_{00} = \rho$ is the energy density and $T_{ii} = P_i$ are the pressure components for $i = 1, 2, 3$. The indices $\mu = 0, 1, 2, 3$ correspond to the spacetime coordinates $x^\mu = (x^0, x^1, x^2, x^3) = (ct, x, y, z)$.

In order to describe how space-time and matter-energy interact in our Universe we need to define a metric with which we can work.

Friedmann-Lemaître-Robertson-Walker Metric

Friedmann and Lemaître were the first people to attempt to solve Einstein's equations in a framework that is consistent with the large-scale distribution of matter in the Universe. They achieved this task by considering a simple homogeneous and isotopic mass distribution, which is expected from the Copernican Principle. Observational evidence has since verified that this is a good choice of prior as the Universe exhibits both of these properties on large-scales (Smoot et al., 1991).

Under these assumptions the solutions to Einstein's field equations are fairly straightforward. Robertson and Walker independently proposed a general spherically symmetric metric of the form

$$ds^2 = -dt^2 - a(t)^2 \left[\frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + sin^2\theta \; d\phi^2) \right] \tag{1.2}$$

where t is a time coordinate, $a(t)$ is the scale factor which represents the expansion of the universe and is set to $a = 1$ today, and K is the curvature parameter taking the values 1, 0, or -1 depending on the geometry of the universe being closed, flat or open, respectively. This metric can now be used with Einstein's equations to describe how space-time is curved by the presence of matter-energy in the Universe, and vice-versa.

The Friedmann and Fluid Equations

To solve Einstein's equations for the FLRW metric we require knowledge of the distribution of matter-energy in the Universe. We get this information from the energy-momentum tensor $T_{\mu\nu}$. A perfect fluid approximation is completely defined by the rest frame energy density ρ and the isotropic rest frame pressure P. In this case the energy-momentum tensor takes the form,

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix} \tag{1.3}$$

Using Eq. 1.3 in Eq. 1.1 we can solve the time-time (G^0_0) and space-space (G^i_i) components to obtain the Friedmann and acceleration equations:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2}, \tag{1.4}$$

$$\left(\frac{\ddot{a}}{a}\right) = -\frac{4\pi G}{3}(\rho + 3P), \tag{1.5}$$

where ρ corresponds to the energy densities in the Universe with pressure P, and a is the scale factor (see Appendix A for a full derivation).

The evolution of the material in the Universe can be described by the fluid equation, which is obtained by substituting the acceleration equation (Eq. 1.5) into the time derivative of the Friedmann equation (Eq. 1.4) and rearranging such that

$$\dot{\rho} + 3\frac{\dot{a}}{a}(P + \rho) = 0. \tag{1.6}$$

By specifying an equation of state $w = P/\rho$, where w is assumed to be constant, we can predict the evolution of the material in the Universe. Solving the Friedmann equations for a flat Universe with $K = 0$ gives:

$$a(t) \propto (t - t_0)^{\frac{2}{3(1+w)}}, \tag{1.7}$$

$$\rho \propto a^{-3(1+w)}, \tag{1.8}$$

where t_0 is constant.

The Critical Density

In order to fully assess contributions of different forms of energy density in the Universe to the total expected we need to define the critical density ρ_c, which quantifies the density required to produce a flat Universe ($K = 0$). In the FLRW framework, we can use Eq. 1.4 to show that

$$\rho_c = \frac{3H^2}{8\pi G}, \tag{1.9}$$

where the Hubble function $H \equiv \dot{a}/a$. We can use this critical density to ascertain the abundances of different forms of energy density in the Universe today by saying that

$$\Omega_i(t) = \frac{\rho_i(t)}{\rho_c(t)}, \tag{1.10}$$

where the subscript *i* denotes the species under consideration. The Universe is comprised of numerous different components of energy density. Accordingly, it is best described by a multi-fluid model. We consider the main contributions to the total energy density of the Universe is the following section.

1.1.2 The Multi-Fluid Model of the Universe

The Universe is comprised from several components of energy density including: radiation, matter, dark energy and curvature, which need to be incorporated into the FLRW model. Assuming that the total density $\rho(t) = \sum_i \rho_i(t)$, where the subscript i denotes the species under consideration, we can now use Eq. 1.10 to decompose the total density parameter into a sum of its parts:

$$\Omega(t) = \Omega_r(t) + \Omega_m(t) + \Omega_k(t) + \Omega_\Lambda(t). \tag{1.11}$$

The density parameter for matter is further decomposed such that $\Omega_m(t) = \Omega_b(t) + \Omega_{CDM}(t)$. Subscript indices denote matter m, baryons b, cold dark matter CDM, radiation r, dark energy Λ and curvature K. The energy density of each species depends upon its equation of state w and can be found from the general solution

$$\Omega_i(t) = \frac{\Omega_i(t=0)}{a^{3(1+w_i)}}, \tag{1.12}$$

where a is the scale factor. The special case of $\Omega(t) = 1$ implies by definition that $K = 0$, and since K is a fixed constant it can be concluded that $\Omega(t) = 1$ at all times. In the following subsections we will describe these components of energy density in more detail, introducing the equation of state for each.

Radiation

Contributions to the energy density of radiation come from all electromagnetic radiation and relativistic matter, such as neutrinos. The relationship between the radiation energy density and temperature can be obtained by integrating the Bose-Einstein distribution function for photons:

$$\rho_\gamma = \frac{\pi^2}{15} T^4 \propto a^{-4}. \tag{1.13}$$

Relating this to the solutions of the Friedmann equations (Eqns. 1.7, 1.8) and the general energy density (Eq. 1.12), we can see that in a radiation dominated Universe the expansion $a(t) \propto t^{1/2}$, pressure $P = \rho/3$ and $w = 1/3$, reducing the radiation energy density to

$$\Omega_r(t) = \frac{\Omega_r(t=0)}{a^4}. \tag{1.14}$$

The current contribution of this species to the total energy density, as measured from the Cosmic Microwave Background, is $\sim 10^{-4}$ times smaller than that of matter and dark energy. Consequently, it is often neglected.

Matter

The two main non-relativistic matter components of energy density in the Universe are comprised of luminous baryonic matter $\Omega_b(t)$ and non-luminous dark matter $\Omega_{DM}(t)$. The majority of the former is contained in stars, gas and dust, whereas the existence of the latter is only inferred from observations of its effect on the former.

The need for dark matter is physically motivated and was introduced to explain discrepancies between theoretical and measured shapes of rotation curves (Zwicky, 1937), but has never been detected directly. The nature of dark matter is dissimilar to that of luminous matter because it only interacts via gravity. Candidates for dark matter come from both theoretical particle physics and astronomy. Possible candidates include:

- Weakly Interacting Massive Particles (WIMPs): Massive particles that only interact through the weak nuclear force and gravity.
- Axions: Theoretical particles that couple with photons in the presence of magnetic fields.
- Massive Astrophysical Compact Halo Objects (MACHOs): Massive baryonic objects such as black holes or brown dwarf stars that emit little to no light.

The way in which large-scale structure in the Universe has formed suggests that dark matter must be non-relativistic. Accordingly, it is usually referred to as "cold" dark matter

(CDM). The non-relativistic nature of CDM suggest that it has close to zero pressure, which means that the equation of state $w = 0$. This means that, after radiation, the Universe will enter a matter dominated era where $a(t) \propto t^{2/3}$. The total matter energy density $\Omega_m(t) = \Omega_b(t) + \Omega_{CDM}(t)$ depends on the physical volume of the Universe such that

$$\Omega_m(t) = \frac{\Omega_m(t=0)}{a^3}. \tag{1.15}$$

Curvature

Observational evidence suggests that the curvature of the Universe is very close to zero (Komatsu et al., 2010). Setting $K = 0$ in the Friedmann equation (Eq. 1.4) and treating it as a perfect fluid with an equation of state $w = 0$, we find that the energy density of curvature is

$$\Omega_K(t) = \frac{1}{a^2}. \tag{1.16}$$

Dark Energy

The accelerated expansion of the Universe is driven by the energy density of an exotic substance dubbed dark energy, and is often assumed to be a cosmological constant Λ (see § 1.3.1). Including Λ in the fluid equation (Eq. 1.6) results in $P = -\rho$, which gives an equation of state $w = -1$. This means that the energy density as a function of time is simply

$$\Omega_\Lambda(t) = \Omega_\Lambda, \tag{1.17}$$

and that it is significantly smaller than $\Omega_m(t) \propto a^{-3}$ and $\Omega_r(t) \propto a^{-4}$ at early times, but comes to dominate at late-times as the other components are diluted by the increasing volume of the Universe.

1.2 Observational Cosmology

The precise mapping of the large-scale structure in the Universe is a prerequisite for the observational validation of the standard cosmological model. In this section we introduce the theory required in order to achieve this, and discuss some of the main observational contributions made to this effort.

1.2.1 Hubble's Law

Hubble's law was formulated in 1929 by Edwin Hubble when he observed that other galaxies are receding from us at a rate proportional to their distance from us (Hubble, 1929). The Hubble parameter $H(t)$ is the time-dependent constant of proportionality relating recession velocity v and (proper) distance d in an expanding Universe. For nearby

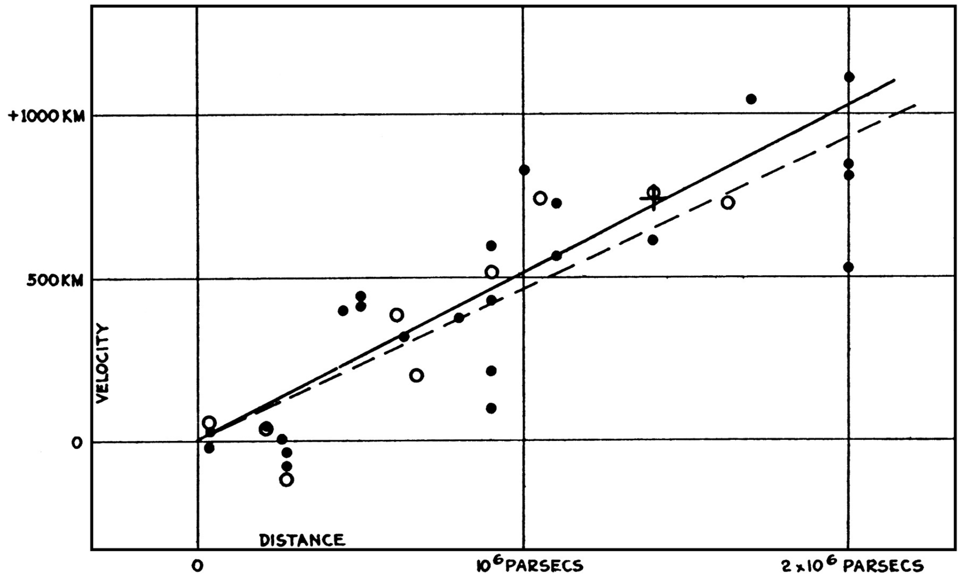

Figure 1.1: The original Hubble diagram showing the velocity-distance relation between extra-galactic nebulae. Filled circles represent individual galaxies, open circles represent stacked galaxies, and the cross represents 22 combined galaxies with inaccurate measurements. Credit: Hubble (1929)

galaxies at the present epoch

$$v = H_0 d. \tag{1.18}$$

H_0 has dimensions of $1/t$ but is generally written as

$$H_0 = 100\ h\ \mathrm{kms^{-1}Mpc^{-1}}, \tag{1.19}$$

where h is a dimensionless number that parameterizes the uncertainty on H. The reciprocal of h_0 is Hubble time,

$$t_H \equiv \frac{1}{H_0}, \tag{1.20}$$

and the Hubble distance is defined as

$$D_H \equiv \frac{c}{H_0} = 3000\ \ h^{-1}\,\mathrm{Mpc}, \tag{1.21}$$

where c is the speed of light. The current constraint on the Hubble parameter from the Wilkinson Microwave Anisotropy Probe (WMAP) 7 data is $H_0 = 71.0{\pm}2.5\ \mathrm{kms^{-1}Mpc^{-1}}$ (Larson et al., 2010).

1.2.2 Redshift

Hubble's Law implies that the wavelength of light emitted by receding galaxies will be stretched to a value proportional to the rate of expansion of the Universe. This *stretch* factor can be quantified by the redshift z of the galaxy and can be used to deduce its radial distance (see § 1.2.4). Redshift is defined as

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{\nu_e}{\nu_o}, \tag{1.22}$$

where λ_o and ν_o are the observed wavelength and frequency, and λ_e and ν_e are the emitted.

The cosmological redshift of an object z_{cos}, which is due solely to the Hubble flow, is different from its observed redshift z_{obs}. The difference between these two quantities can be attributed to the peculiar velocity of the object

$$v_{pec} = c\frac{(z_{obs} - z_{cos})}{(1 + z)}, \tag{1.23}$$

for $v_{pec} \ll c$ (see § 1.2.3).

For small v/c in the expanding Universe, the recession velocity of an object is linearly proportional to its distance (for small redshifts), such that

$$z_{cos} \approx \frac{v}{c} = \frac{d}{D_H}, \tag{1.24}$$

where D_H is the Hubble distance defined in Eq. 1.21. Cosmological redshift is directly related to the scale factor $a(t)$ via

$$1 + z_{cos} = \frac{a(t_o)}{a(t_e)}, \tag{1.25}$$

where $a(t_o)$ corresponds to the size of the Universe at the time of observation, and $a(t_e)$ is the size at the time the light was emitted.

1.2.3 Redshift-Space Distortions

The distribution of galaxies that we observe in sky surveys, where we measure radial distances from spectroscopic or photometric redshifts, is not a true 3D picture. We observe an apparent clustering pattern in *redshift-space*, which is systematically different from

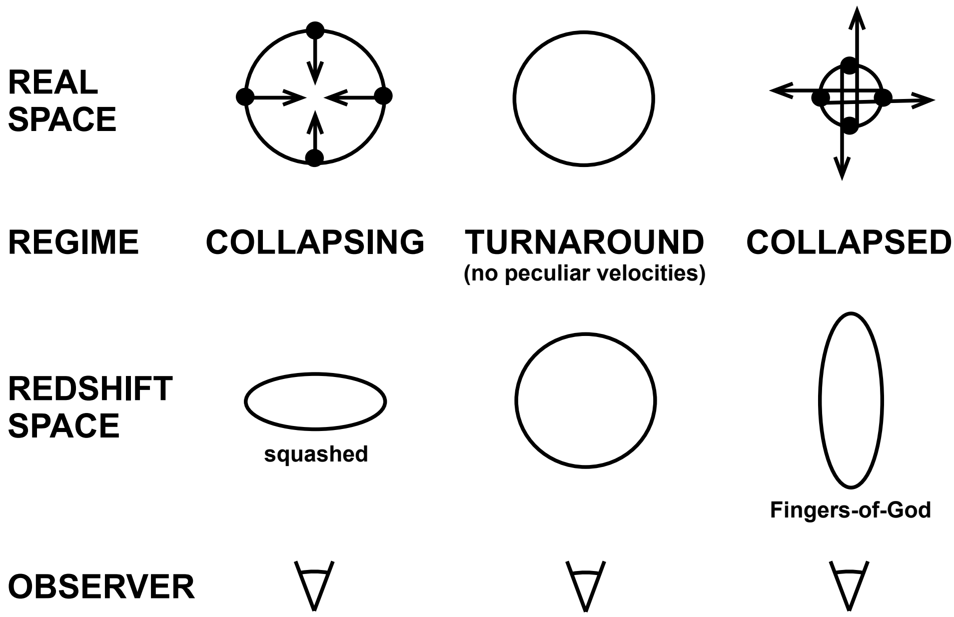

Figure 1.2: An illustration showing the effects of redshift-space distortions on the observed galaxy distribution. Dots represent galaxies falling into spherical overdensities with peculiar velocities shown by arrows.

the true distribution in *real-space*. Redshifts of galaxies are systematically altered from their Hubble flow values by peculiar velocities in two ways:

- **Fingers of God** The virialisation process that occurs for collapsing objects increases their velocity dispersion. The random internal velocities act to smear out the collapsed object along the line-of-sight producing linear structures pointing towards the observer known as Fingers-of-God (FoG). When we infer galaxy distances assuming that the total velocity relative to the observer comes from the Hubble expansion flow, the result is that we see a distorted density field. This is a small-scale effect.

- **Kaiser effect** The growth of structure in the Universe occurs when objects fall in towards overdense regions via gravity. The infall velocity adds to the redshift, making the distance estimates using the Hubble flow incorrect. This means that clusters of galaxies are squashed in the radial direction, causing an increase in the measured power. This is a large-scale effect and enhances the appearance of walls and filaments (Kaiser, 1987; Hamilton, 1997).

This is illustrated in Fig. 1.2. The effect of these distortions on the correlation function can be described by considering a pair of galaxies with redshifts corresponding to velocities $\mathbf{v_1}$ and $\mathbf{v_2}$. The separation in redshift-space is given by,

$$\mathbf{s} = \mathbf{v_1} - \mathbf{v_2}, \tag{1.26}$$

and an observers line-of-sight is described by,

$$\mathbf{l} = \frac{1}{2}(\mathbf{v_1} + \mathbf{v_2}). \tag{1.27}$$

The separations parallel and perpendicular to the line-of-sight can now be described by

$$\pi = \frac{\mathbf{s} \cdot \mathbf{l}}{|\mathbf{l}|}, \tag{1.28}$$

and,

$$\sigma = \sqrt{\mathbf{s} \cdot \mathbf{s} - \pi^2}, \tag{1.29}$$

respectively. Fig. 1.3 shows how redshift-space distortions can affect the recovered clustering signal.

The magnitude of the effect of these distortions on the recovered galaxy power spectrum

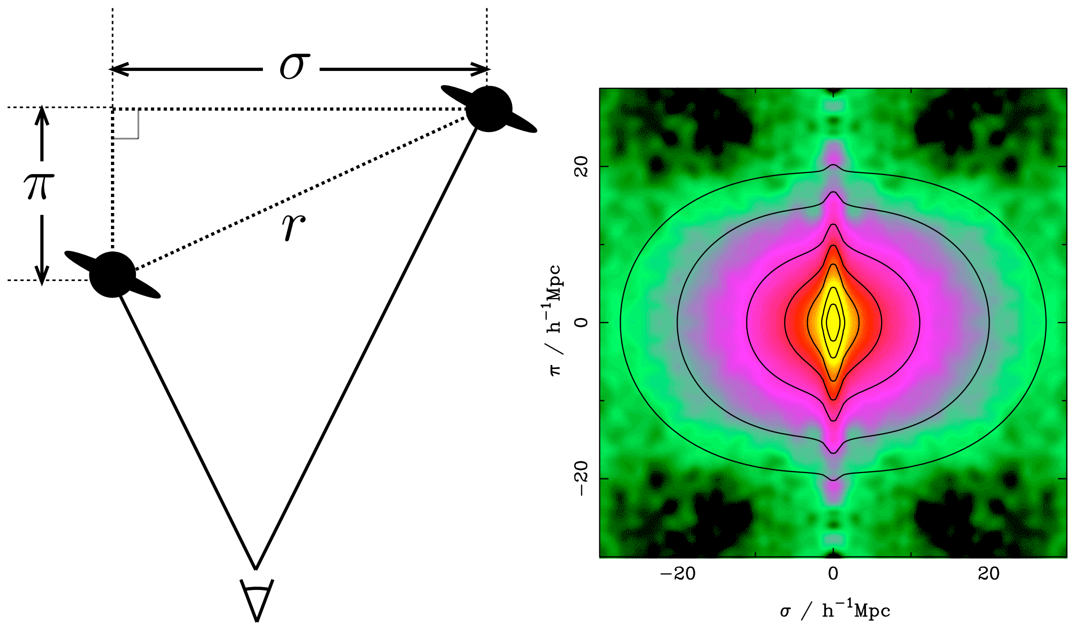

Figure 1.3: Left: Schematic showing the radial, π, and tangential, σ, components of separation for a pair of galaxies with 3D separation, r. Right: The redshift-space correlation function, $\xi(\sigma, \pi)$, for the 2-degree Field Galaxy Redshift Survey (2dFGRS) as calculated by Peacock et al. (2001). The effect of redshift-space distortions are apparent here on small-scales in the form of Fingers-of-God elongations, and on large-scales in the form of Kaiser flattening.

depends on the cosine of the angle, $\mu = \cos\theta$, between the line-of-sight, $\mathbf{l}$, and the separation, $\mathbf{s}$. In the linear regime we can correct for these effects with a model of the form

$$\frac{P_s}{P_r} = \frac{(1 + \beta\mu^2)^2}{(1 + k^2\mu^2\sigma_v^2)}, \tag{1.30}$$

where P_r and P_s are the spherically averaged real- and redshift-space power spectra respectively, $\beta = \Omega_m^{0.6}/b$, b is the galaxy bias, and σ_v is the pairwise velocity dispersion. The numerator in Eq. 1.30 corrects for the Kaiser effect (Kaiser, 1987) whilst the denominator corrects for FoG (Davis & Peebles, 1982; Mo et al., 1993; Fisher et al., 1994b). We will discuss the effects of peculiar velocities on clustering analyses in the following chapters.

1.2.4 Distance Measures in Cosmology

Our definitions of distance in static Euclidean geometry require modifications to account for the expansion of the Universe. We can start by considering the radial comoving distance D_C between two events, which remains constant over time if they are both moving according to the Hubble flow. It can be found by setting $ds^2 = 0$ in the FLRW metric (Eq. 1.2) and is given by

$$D_C = D_H \int_{z1}^{z2} \frac{dz}{E(z)}. \tag{1.31}$$

D_H is the Hubble distance and the term $E(z)$ comes from Eq. 1.4 such that

$$\begin{aligned} E(z) &= \frac{H(z)}{H_0} \\ &= \frac{\sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}}{H_0}, \end{aligned} \tag{1.32}$$

where $\Omega_m = 8\pi G\rho/3H_0^2$, $\Omega_k = K/(a_0 H_0)^2$, and $\Omega_\Lambda = \Lambda/3H_0^2$ (see § 1.3.1 for an explanation of the origins of Λ).

The proper distance D_P is defined by the distance travelled by a photon between two events. It can be found by setting the time-time component of Eq. 1.2 to $dD_P^2 = dt^2$ and is given by

$$D_P = D_H \int_{z1}^{z2} \frac{dz}{(1+z)E(z)}. \tag{1.33}$$

The time-of-flight of a photon crossing a redshift interval dz divided by the scale factor at that time is proportional to $dz/E(z)$. Given that the speed of light is constant, we can see that the radial comoving distance is just the proper distance divided by the scale factor.

In the case where we have two events at the same redshift that are separated on the sky by some angle $\delta\theta$, their comoving distance is given by $D_M\delta\theta$, where D_M is the transverse comoving distance or proper motion distance. D_M is related to D_C such that

$$D_{\mathrm{M}} = \begin{cases} D_{\mathrm{H}} \frac{1}{\sqrt{\Omega_k}} \sinh\left[\sqrt{\Omega_k}\, D_{\mathrm{C}}/D_{\mathrm{H}}\right] & \text{for } \Omega_k > 0 \\ D_{\mathrm{C}} & \text{for } \Omega_k = 0 \\ D_{\mathrm{H}} \frac{1}{\sqrt{|\Omega_k|}} \sin\left[\sqrt{|\Omega_k|}\, D_{\mathrm{C}}/D_{\mathrm{H}}\right] & \text{for } \Omega_k < 0 \end{cases} \tag{1.34}$$

where $\sinh$ and $\sin$ account for the curvature of space.

The angular diameter distance of an object D_A is defined by the ratio of its physical transverse size to its angular size and is related to the transverse comoving distance by

$$D_A = \frac{D_M}{1+z}. \tag{1.35}$$

Angular diameter distance does not increase indefinitely as $z \to \infty$, instead it turns over at $z \sim 1$. Consequently, more distant objects appear larger in angular size.

The luminosity distance D_L to an object is defined as

$$D_L \equiv \sqrt{\frac{L}{4\pi S}}, \tag{1.36}$$

where L and S are the bolometric luminosity and flux, respectively. It is related to the transverse comoving distance via

$$D_L = (1+z)D_M. \tag{1.37}$$

The surface brightness of a receeding object is reduced by a factor $(1+z)^{-4}$, and the angular area goes down as D_A^{-2}. The luminosity distance is therefore related to the angular diameter distance via

$$D_L = (1+z)^2 D_A. \tag{1.38}$$

Additionally, number densities of non-evolving objects that are moving with the Hubble flow and are constant with redshift occupy a comoving volume V_C. The comoving volume element in solid angle $d\Omega$ and redshift interval dz is given by

$$dV_C = D_H \frac{(1+z)^2 D_A^2}{E(z)} d\Omega\, dz, \tag{1.39}$$

where D_A is the angular diameter distance at redshift z. The total all-sky comoving volume out to a redshift z is

$$V_C = \begin{cases} \left(\frac{4\pi D_H^3}{2\Omega_k}\right)\left[\frac{D_M}{D_H}\sqrt{1+\Omega_k\frac{D_M^2}{D_H^2}} - \frac{1}{\sqrt{|\Omega_k|}}\mathrm{arcsinh}\left(\sqrt{|\Omega_k|}\frac{D_M}{D_H}\right)\right] & \text{for } \Omega_k > 0 \\ \frac{4\pi}{3}D_M^3 & \text{for } \Omega_k = 0 \\ \left(\frac{4\pi D_H^3}{2\Omega_k}\right)\left[\frac{D_M}{D_H}\sqrt{1+\Omega_k\frac{D_M^2}{D_H^2}} - \frac{1}{\sqrt{|\Omega_k|}}\mathrm{arcsin}\left(\sqrt{|\Omega_k|}\frac{D_M}{D_H}\right)\right] & \text{for } \Omega_k < 0 \end{cases} \tag{1.40}$$

where D_H^3 is often referred to as the Hubble volume.

1.2.5 Galaxy Bias

Galaxies do not trace the distribution of matter in the Universe directly. There is a difference between the spatial distribution of luminous objects and dark matter that is known as galaxy *bias*. The existence of galaxy bias is supported by the fact that galaxies of different types have different clustering strengths (Dressler, 1980). The biasing of density peaks in a Gaussian random field is well known (Bardeen et al., 1986), and provides a theoretical framework for the origin of galaxy density biasing. In the linear regime, it is often assumed that the galaxy overdensity δ_g is related to the mass overdensity δ_m via the relation

$$\delta_g(\mathbf{x}) = b\delta_m(\mathbf{x}), \tag{1.41}$$

where b is the biasing parameter and is independent of scale. In this regime, the galaxy-galaxy and mass-mass correlation functions and power spectrums are related via the equations:

$$\xi_{gg}(r) = b^2\xi_{mm}(r), \tag{1.42}$$

$$P_{gg}(k) = b^2 P_{mm}(k). \tag{1.43}$$

However, this simplistic model is not reasonably physically motivated. By definition, galaxy overdensities with $\delta_g < -1$ are forbidden, so if we have a bias $b > 1$, for example, the model will break down in deep voids. Even in the simplistic case of constant comoving number density, the linear biasing relation is not preserved during the growth of fluctuations. Consequently, a nonlinear biasing model where b varies as a function of δ_m is required.

The Halo Model

The *halo model* provides a theoretical framework with which we can model the gravitational nonlinearities in the galaxy density field. The origins of the halo model can be attributed to Neyman et al. (1953), who proposed that the distribution of galaxies in the Universe can be considered simply as an amalgamation of clusters of various sizes. Consequently, a model may be formulated that depends on the distributions of galaxies and clusters of galaxies, and the clustering of the clusters themselves.

Halos are identified as peaks in the initial density field (Kaiser, 1984), where the size of the halos are related to the height of the peaks. The density around a high peak (large halo) is shallower than that around a low peak (small halo) (Bardeen et al., 1986), since larger halos are less centrally concentrated than smaller ones. This trend can be modelled by the NRW density profile (Navarro et al., 1997)

$$\rho(r|m) = \frac{\rho_c \delta_c}{(r/r_s)(1 + r/r_s)^2}, \tag{1.44}$$

where ρ_c is the critical density for collapse, δ_c is the characteristic overdensity, c is the halo concentration, and $r_s = r_{200}/c$ is a characteristic radius, with r_{200} defined as the radius at which the mean density enclosed is equal to $200\rho_c$. The characteristic overdensity is related to the concentration via the equation

$$\delta_c = \frac{200}{3} \frac{c^3}{\ln(1 + c) - \frac{c}{1+c}}. \tag{1.45}$$

See Cooray & Sheth (2002) for a recent review of halo models.

We can combine the halo model with the relation between the halos and luminous objects to better understand galaxy biasing. This is done by considering a halo occupation distribution (HOD), which simply relates the number of galaxies in a halo to the halo mass. The average number of galaxies residing in a halo of mass M can be described by the halo occupation number $N_g(M)$. The form of the HOD varies across the literature, but here we consider a simple parametric form:

$$N_g(M) = \begin{cases} (M/M_1)^\alpha & \text{for } M > M_{min} \\ 0 & \text{for } M < M_{min} \end{cases} \tag{1.46}$$

where M_{min} represents the minimum mass of halos that host the population of galaxies, M_1 is a normalisation parameter that represents the critical mass above which halos typically host more than one galaxy, and α is the power-law index of the mass dependence of the efficiency of galaxy formation. Using this relation, the number density of the corresponding galaxy population is given by

$$n_g = \int_{M_{min}}^{\infty} dM\ n_{halo}(M) N_g(M), \tag{1.47}$$

where $n_{halo}(M)$ is the halo mass function.

In the context of the halo model, the galaxy correlation function consists of two contributions from the 1-halo term and the 2-halo term:

$$\xi_g^{tot}(r) = \xi_g^{1h}(r) + \xi_g^{2h}(r). \tag{1.48}$$

The 1-halo term contains galaxy pairs that are located within the same halo, whilst the 2-halo term contains galaxy pairs that span two different halos. We can model the 1-halo term by specifying the distribution of galaxies in individual halos. For convenience, we introduce a normalised halo density profile $u_M(r) = \rho(r)/M$, where M is the total halo mass and ρ is the density profile. The normalised number density distribution of satellite galaxies can be written $u_s(r) = n_s(r)/N_s(M)$ so that

$$\int_{V_{200}} d^3x\ u_{M,s}(r) = 1, \tag{1.49}$$

where V_{200} is the volume of the sphere defined by the virial radius r_{200}. It is generally assumed that the central galaxy is located at the halo centre and that the satellite galaxies follow a normalised number density distribution given by $u_s(r)$. For simplicity, we can make the assumption that $u_s(r) = u_M(r)$, ie. that the number density of satellite galaxies is the same as that of the dark matter particles within the halos.

We can now write the 1-halo term as

$$\xi_g^{1h}(r) = \frac{2}{\bar{n}_g^2} \int_0^{\infty} dM\ n_{halo}(M) \langle N_{pair} \rangle f(r), \tag{1.50}$$

where $\langle N_{pair} \rangle$ represents the average number of galaxy pairs within a halo of mass M, $f(r) 4\pi r^2 \Delta r$ is the fraction of pairs with separation in the range $r \pm \Delta r/2$, and $f(r)$ is the galaxy pair distribution function within a halo of mass M:

$$f_M(r) = 4\pi \int_0^{r_{200}} ds\ u_M(s) s^2 \int_0^{\pi} d\theta\ u_M(s'(\theta)) \sin\theta, \tag{1.51}$$

where $s'(\theta) = (s^2 + r^2 - 2sr\cos\theta)^{1/2}$, and r is the pair separation. The assumed spherical symmetry of the halo density profile means that the 1-halo term can be obtained via a 3-dimensional integration over the halo mass M, the radial distance s within the halo and the angle θ of the galaxy-galaxy pair relative to the halo centre.

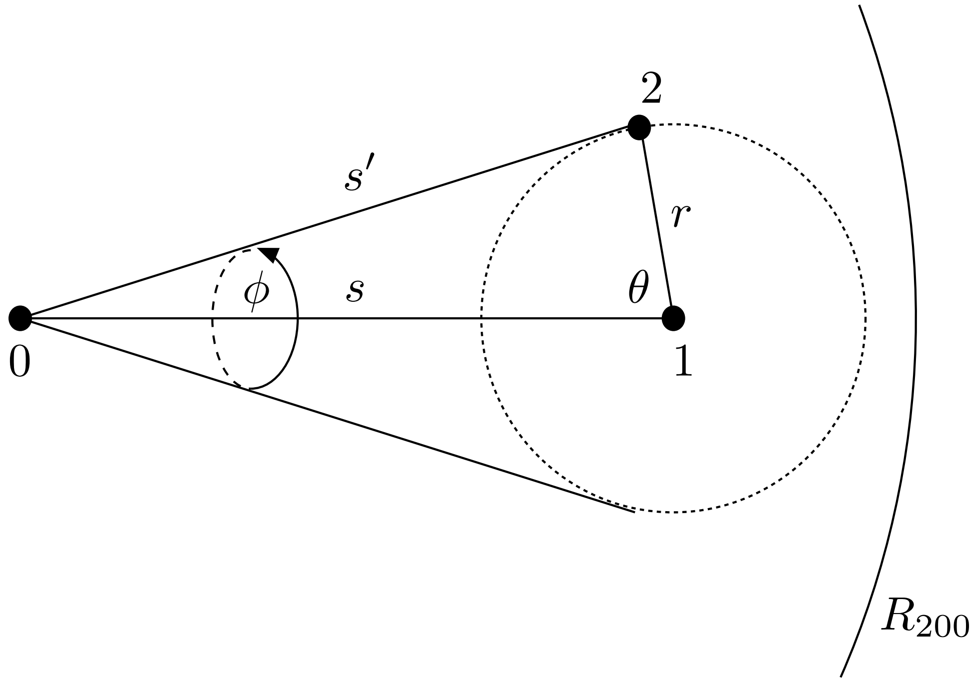

Figure 1.4: Schematic representation of the integration over the normalised halo density profile at two points, as defined in Eq. 1.51. The centre of the halo is placed at the origin. The position of the first integral over the radial distance within the halo is denoted at point 1. The position of the second integral is denoted at point 2.

The mean number of pairs as a function of separation, $\langle N_{pair}\rangle f(r)$, can be divided into contributions from central-satellite pairs and satellite-satellite pairs,

$$\langle N_{pair}\rangle f(r) = \langle N_{cs}\rangle u_s(r) + \langle N_{ss}\rangle f_s(r), \tag{1.52}$$

where $f_s(r)$ follows from Eq. 1.51 upon substituting u_s for u_M. The mean number of central-satellite pairs is

$$\langle N_{cs}\rangle = \langle N_c N_s\rangle, \tag{1.53}$$

and the mean number of satellite-satellite pairs is

$$\langle N_{ss}\rangle = \frac{\langle N_s(N_s-1)\rangle}{2}. \tag{1.54}$$

The 2-halo term can be written as

$$\xi_g^{2h}(r) = \frac{1}{\bar{n}_g^2} \int_0^\infty dM_1\, n_{halo}(M_1) N(M_1) b(M_1) \int_0^\infty dM_2\, n_{halo}(M_2) N(M_2) b(M_2) \xi_{dm}^{lin}(r), \tag{1.55}$$

where $\xi_{dm}^{lin}(r)$ is the linear dark matter correlation function and $b(M_1)$ is the halo bias factor.

1.2.6 Dark Energy

Fitting models with observations offers a route to accurately constraining the energy-momentum tensor and thus better understanding the exact nature of dark energy. Various cosmological probes have the power to constrain various cosmological parameters. However, there is no single observational technique that does it all. As such it has become customary to combine results from cosmological probes, thus helping to break degeneracies, reduce systematic errors, and better constrain the cosmological model. In this section, we will discuss some of the main observational contributions made to this effort, and highlight the requirement for a dark energy component in the current cosmological model.

Supernovae 1a

Type Ia Supernovae (SNe Ia) are thought to be accreting white dwarf stars that undergo a violent explosion when their stellar mass reaches the Chandrasakher limit ($\sim 1.38 M_\odot$). White dwarfs that explode via the accretion mechanism have a uniform mass and so produce a consistent peak luminosity, thus making them ideal candidates for use as standard candles. After luminosity curve corrections, their typical observed absolute magnitude is $M = -19.3$, which is $\sim 5 \times 10^6$ times brighter than the sun. This is bright enough for them to outshine their host galaxies, allowing them to be detected at high redshifts. Their observed magnitude m can be used to reconstruct their total flux and therefore their luminosity distance as a function of redshift, which is sensitive to cosmological parameters. This is often expressed in terms of distance modulus, which is defined as

$$\mu = m - M = 5 \log D_L + 25. \tag{1.56}$$

Observations made by two teams in 1998 (Riess et al., 1998; Perlmutter et al., 1999) revealed that distant SNe appear fainter than they should given a purely matter dominated expansion, suggesting that the Universe has entered an epoch of accelerated expansion. In order to explain these observations, we are required to introduce a late-time acceleration into our cosmological model in the form of a cosmological constant Λ (see § 1.1.2).

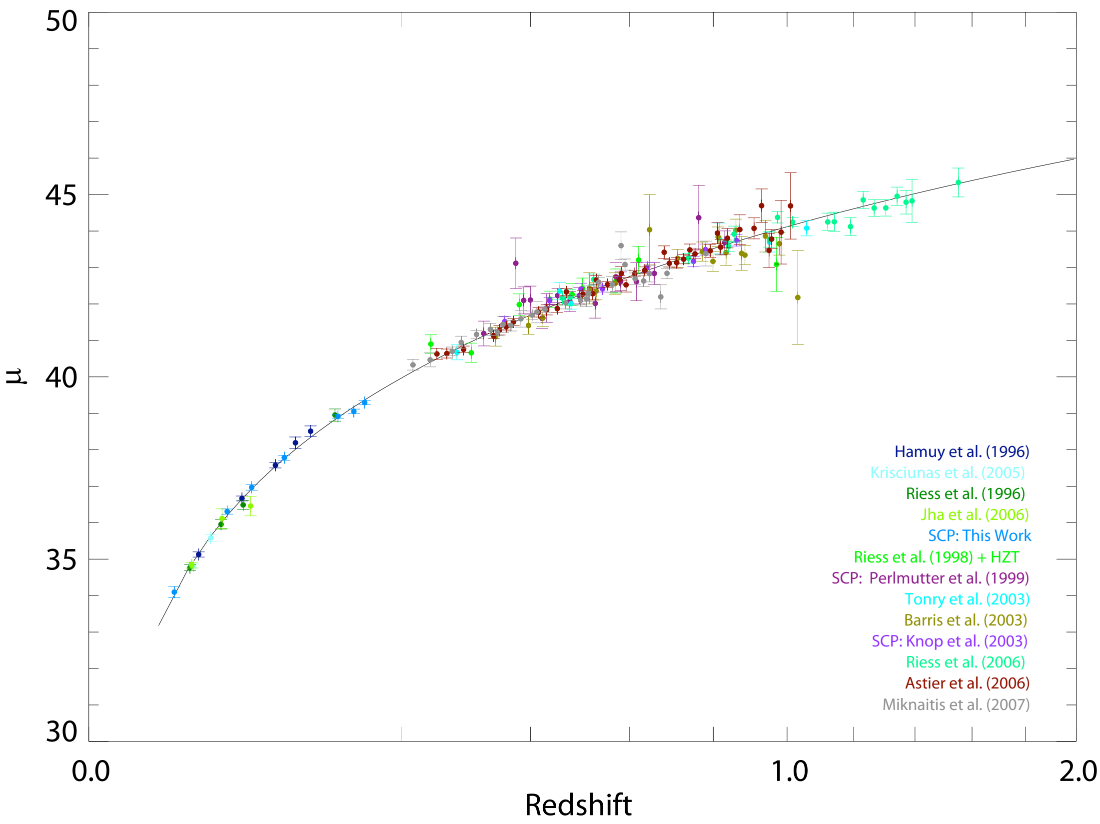

Figure 1.5: Current status of measurements of the Hubble diagram of Type Ia Supernovae. SNe distance modulus, $\mu = |m - M|$, is plotted as a function of SNe redshift. The line is the best fit model ($\Omega_m = 0.29$, $\Omega_\Lambda = 0.71$). Credit: Kowalski et al. (2008).

Fig. 1.5 shows a recent Hubble diagram of SNe compiled by Kowalski et al. (2008) for various different projects.

Cosmic Microwave Background

The early Universe consisted of a hot ionised plasma where photons and baryons were strongly coupled. As the Universe expanded and cooled, the photons of light became less energetic preventing them from ionising forming atoms. At a temperature of $\sim 3000\mathrm{K}$ neutral hydrogen was able to form, thus allowing the photons to decouple and free-stream to us. These relic photons can be detected today in the form of Cosmic Microwave Background (CMB) radiation, detectable as microwaves with a temperature of $\sim 2.7\mathrm{K}$. The CMB has undergone very little interaction with matter since decoupling and therefore provides us with a detailed picture of the state of the Universe when it was just $\sim 380,000$ years old.

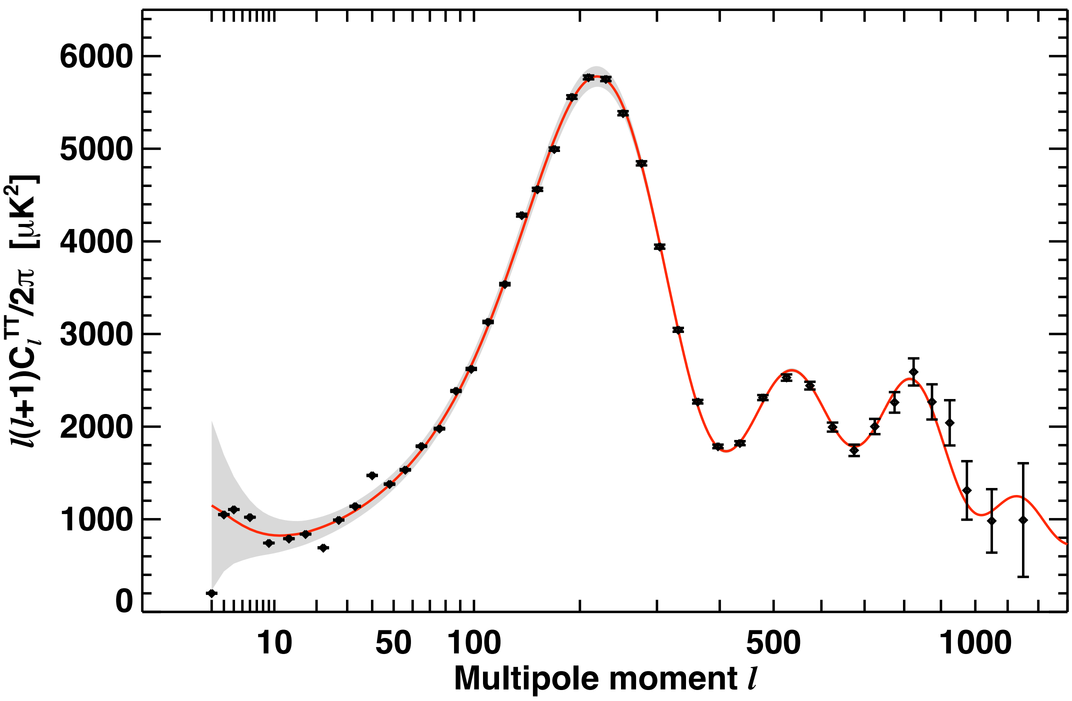

Figure 1.6: The CMB temperature power spectrum, $l(l+1)C_l/\pi$ for WMAP7 data plotted as a function of multipole, l, where $l = \pi/\theta$. The solid line shows the best-fitting ΛCDM model to the data. Credit: (Larson et al., 2010).

The CMB radiation is not perfectly isotropic. Instead, it has small temperature fluctuations as a function of angular position. These temperature fluctuations represent primordial density perturbations that are seeds for galaxy formation and present detailed, observable features that depend on the cosmological parameters. For example, the location of the first peak in the CMB temperature power spectrum corresponds to the size of the horizon at last-scattering. We can measure the distance to the surface of last-scattering from the redshift of the CMB, allowing us to probe the geometry of the Universe. This method measures the curvature energy density of the Universe, with subsequent peaks probing combinations of $\Omega_r(t)$, $\Omega_b(t)$, and $\Omega_m(t)$. Fig. 1.6 shows a current measurement of the CMB temperature power spectrum, calculated from WMAP7 data (Larson et al., 2010).

Galaxy Clustering

Perturbations created in the early Universe are imprinted in the distribution of matter at low redshifts (Silk 1968, Peebles & Yu 1970, Sunyaev & Zeldovich 1970, Bond &

Efstathiou 1984, Holtzman 1989). These perturbations encode characteristic scales according to the constituents of energy density present. Two main features that arise include Baryon Acoustic Oscillations (BAO) and a turnover.

Baryon Acoustic Oscillations BAO are evident in the distribution of matter in the Universe as a result of sound waves propagating in the baryon-photon plasma after inflation. Prior to recombination and decoupling, the Universe was filled with an extremely hot plasma in which baryons and photons were tightly coupled due to Thomson scattering. The competing forces between radiation pressure and gravity at this time were responsible for creating oscillations within the plasma. For a single spherical overdensity, a shell of baryonic material is driven away from the perturbation by a sound wave with a speed

$$c_s = c/\sqrt{3(1+R)}, \tag{1.57}$$

where

$$R \equiv 3\rho_b/4\rho_\gamma \propto \Omega_b/(1+z) \tag{1.58}$$

(Eisenstein et al., 2007), to a radius

$$\frac{r_S(z_*)}{h^{-1}\,\mathrm{Mpc}} \equiv \frac{1}{100\Omega_m^{1/2}} \int_0^{a_*} \frac{c_S}{(a+a_{eq})^{1/2}} da \tag{1.59}$$

$$\tag{1.60}$$

$$\sim 100\,h^{-1}\,\mathrm{Mpc}, \tag{1.61}$$

corresponding to the comoving sound horizon at recombination (Hu & Sugiyama, 1995). Here, the expansion factor $a \equiv 1/(1+z)$, and a_*, a_{eq} are the values at recombination and matter-radiation equality, respectively. When the photons and gas decouple, the photons are free to propagate throughout the Universe and form what we detect today as the CMB. We are left with a spherical shell of baryons surrounding a central concentration of dark matter with a small increase in density at a location corresponding to the sound horizon at the end of the Compton drag epoch. This increase in density can be detected as BAO in the power spectrum in Fourier space - in the same way that the transform of a top-hat function yields a sinc function - and a peak in the correlation function at $r_S(z_*)$ in real space. Fig. 1.7, taken from Eisenstein et al. (2007), shows the evolution of the radial mass profile as a function of comoving radius of an initially pointlike overdensity located at the origin.

Turnover Arising in the radiation dominated epoch, the growth of perturbations is initially linked to the Jeans scale; fluctuations smaller than this scale do not grow due

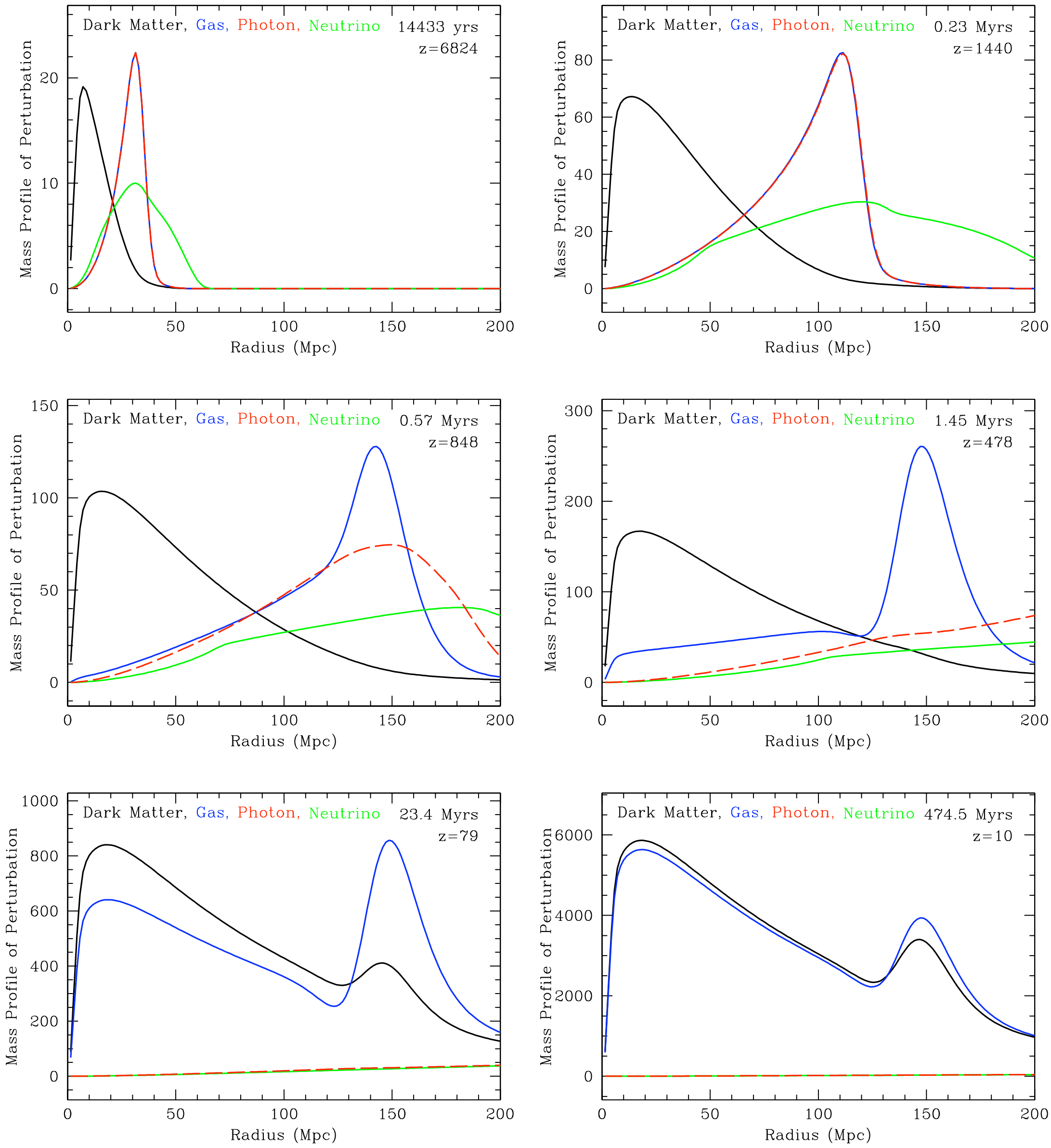

Figure 1.7: Snapshots of the evolution of the radial mass profile as a function of comoving radius of an initially pointlike overdensity located at the origin. Evolution is shown for the dark matter (black), baryon (blue), photon (red), and neutrino (green) perturbations, from early times ($z = 6824 \sim 14433$ yrs), to long after decoupling ($z = 10 \sim 474.5$ Myrs). Initially, the photons and baryons travel outwards like a pulse (top-left). Approaching recombination, the drag of the baryons and relativistic species on the dark matter is visible (top-right). This occurs because the dark matter only interacts gravitationally, so its perturbations lags behind the tightly coupled plasma. At recombination, the photons start to leak away from the baryonic perturbation (middle-left). By the time recombination is complete, the photons have streamed away entirely leaving us with a cold dark matter (CDM) perturbation toward the centre and a baryonic perturbation in a shell (middle-right). Gravitational instability now takes over, and new baryons and dark matter are attracted to the overdensities (bottom-left). At late times, the baryonic fraction of the perturbation is near the cosmic value, because all of the new material was at the cosmic mean (bottom-right). Credit: Eisenstein et al. (2007).

to pressure support from internal random velocities, whilst fluctuations larger than this scale are free to grow through gravity. In the radiation dominated era, the baryon-photon plasma makes the main contribution to the density and perturbations within this plasma are stabilised by high radiation pressure. Consequently, in a Universe containing just dark matter and radiation the Jeans scale grows to the size of the horizon until matter-radiation equality, after which it reduces to zero when the matter dominates. We therefore see an imprint of the horizon scale at matter-radiation equality in the fluctuation distribution, which marks a turnover in the growth rate of fluctuations.

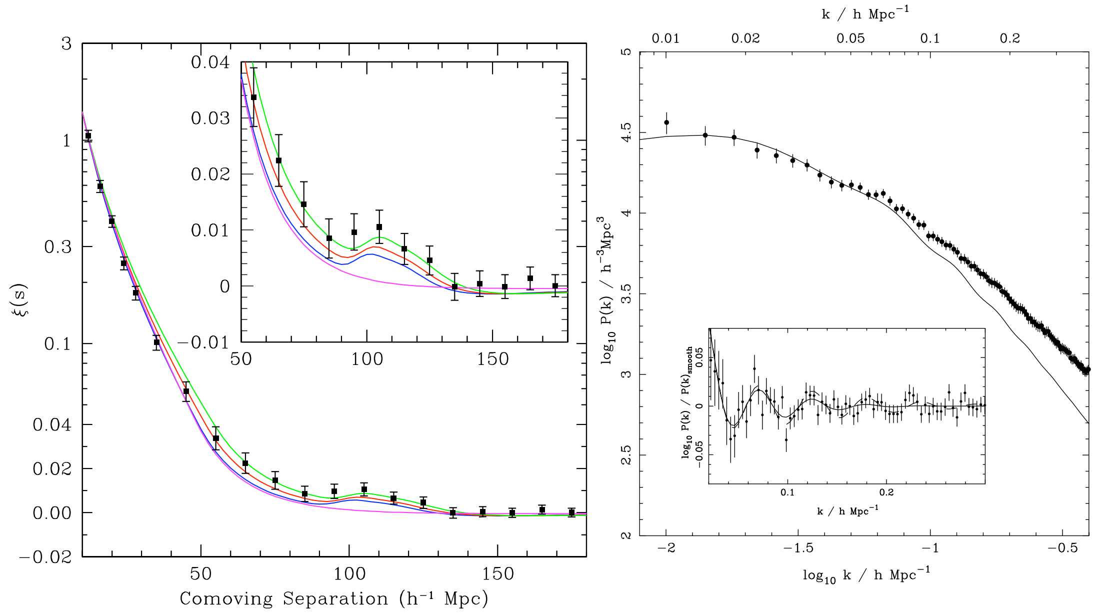

Figure 1.8: Left: The large-scale redshift-space correlation function of the SDSS LRG sample. The inset shows an expanded view with a linear vertical axis. The models are $\Omega_m h^2 = 0.12$ (top, green), 0.13 (red), and 0.14 (bottom with peak, blue), all with $\Omega_b h^2 = 0.024$ and $n = 0.98$ and with a mild non-linear prescription folded in. The magenta line shows a pure CDM model ($\Omega_m h^2 = 0.105$), which lacks the acoustic peak. The bump at $100\,h^{-1}\,\mathrm{Mpc}$ scale is statistically significant. Credit: Eisenstein et al. (2005).
Right: The redshift-space power spectrum recovered from the combined SDSS DR5 main galaxy and LRG sample, optimally weighted for both density changes and luminosity dependent bias (solid circles with 1-σ errors). A flat Λ cosmological distance model was assumed with $\Omega_M = 0.24$. For comparison, a model power spectrum calculated using the fitting formulae of Eisenstein & Hu (1998) for the best fit parameters calculated by fitting the WMAP 3-year temperature and polarisation data, $h = 0.73$, $\Omega_M = 0.24$, $n_s = 0.96$ and $\Omega_b/\Omega_M = 0.174$ (Spergel et al., 2007) is plotted as a solid line. Credit: Percival et al. (2007b).

The acoustic signature was convincingly detected for the first time in 2005 by two independent teams. Eisenstein et al. (2005) used 46,748 Luminous Red Galaxies (LRG) from the Sloan Digital Sky Survey (SDSS) at a median redshift $z = 0.35$ to calculate the redshift-space galaxy 2-point correlation function, as shown in Fig. 1.8. The BAO bump is clearly visible at $\sim 100\,h^{-1}\,\mathrm{Mpc}$. Cole et al. (2005) confirmed this discovery by finding evidence of BAO in the estimated power spectrum coming from the 2dF Galaxy Redshift Survey (2dFGRS) final dataset. It has subsequently been proven that the use of the BAO as a standard ruler is a superb probe of the acceleration history of the Universe (Percival et al. 2007a (see Fig. 1.8), Percival et al. 2007b, Huetsi 2006).

Weak Gravitational Lensing

Weak gravitational lensing, or cosmic shear, describes the shear and magnification of images of high redshift sources, due to the presence of intervening matter. Fluctuations in gravitational potential along the line-of-sight produce distortions in images of distant galaxies at the $\sim 1\%$ level (Munshi et al., 2008). These distortions can be used to obtain an understanding of the statistical properties of the density field and therefore the geometry of the Universe. Weak lensing (WL) surveys are complimentary to galaxy surveys

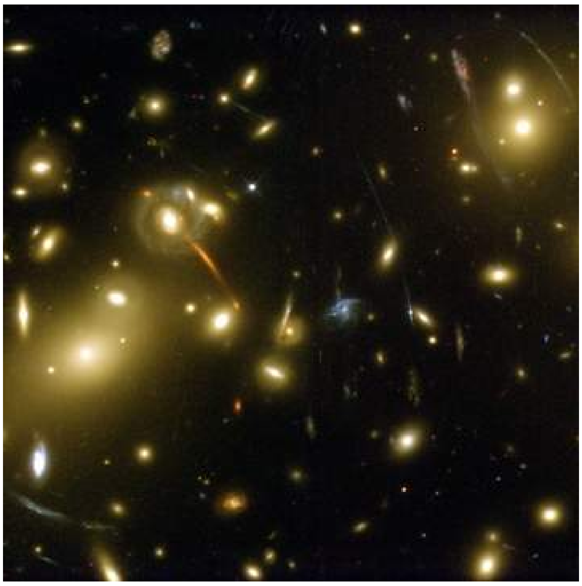

Figure 1.9: An example of gravitational lensing in the Abell 2218 cluster. Credit: A. Frutcher and the ERO team `http://apod.nasa.gov/apod/ap011007.html`.

and CMB observations, since they probe the non-linear matter power spectrum at modest redshifts. Currently, CMB and LSS observations alone provide accurate constraints on cosmological parameters, with WL merely confirming results. Consequently, the emphasis for weak lensing has shifted to understanding the nature of dark matter and dark energy using powerful 3D statistical analyses in future experiments such as Pan-STARRS and the DES.

1.2.7 Concordance Cosmology

A century of extensive observational research combined with theoretical ideas has provided us with a concordance cosmological model of the Universe: ΛCDM. This current model has a spatially flat geometry and a total energy density $\Omega = 1$, which is comprised from:

- Electromagnetic radiation and neutrinos ($< 1\%$ of the total energy density).
- Luminous baryonic matter in the form of stars, galaxies, gas and dust ($\sim 4\%$ of the total energy density).
- Non-baryonic dark matter describing the "missing" mass that is required to explain the measured shapes of galaxy rotation curves ($\sim 22\%$ of the total energy density).
- Dark energy describing the accelerated expansion of the Universe at late-times ($\sim 74\%$ of the total energy density).

In Table 1.1 we reproduce the WMAP7 ΛCDM concordance model constraints for several cosmological parameters that can be used to test the geometry of the Universe. These parameters include: Ω_b the energy density of luminous baryonic matter, Ω_m the total energy density of matter comprised from luminous baryonic matter and dark matter, Ω_Λ the energy density of dark energy describing the late-time accelerated expansion of the Universe, H_0 the Hubble constant, and σ_8 which measures the amplitude of the linear power spectrum on the scale of $8\,h^{-1}\,\mathrm{Mpc}$. In each case, results from the WMAP year 7 data-set have been combined with results from various other probes including:

- BAO - Baryonic Acoustic Oscillations (Percival et al. 2010).
- H_0 - Hubble constant (Riess et al. 2009).
- SNSALT - Type Ia supernovae from the extended SDSS dataset (Kessler et al. 2009), processed with SALT.

There is a clear reduction in errors on all of the measured parameters when probes have been combined.

The state of the concordance cosmological model raises the very important question: *what is dark energy?* We are living in a Universe that is governed largely by the dark energy component of energy density and so understanding the nature of it has become one of the main challenges facing cosmologists today. In the following section, we describe some of the main models of dark energy that are currently being considered.

Probe	Measured Cosmological Parameters				
	Ω_m	Ω_b	Ω_Λ	H_0	σ_8
WMAP	0.266 ± 0.029	0.0449 ± 0.0028	0.734 ± 0.029	71.0 ± 2.5	0.801 ± 0.030
WMAP+BAO	0.280 ± 0.018	0.0462 ± 0.0018	0.720 ± 0.018	69.8 ± 1.5	0.812 ± 0.024
WMAP+H_0	0.254 ± 0.022	0.0438 ± 0.0022	0.746 ± 0.022	$72.1^{+2.2}_{-2.0}$	0.792 ± 0.029
WMAP+ BAO+ H_0	$0.272^{+0.016}_{-0.015}$	0.0456 ± 0.0016	$0.728^{+0.015}_{-0.016}$	$70.4^{+1.3}_{-1.4}$	0.809 ± 0.024
WMAP+ BAO+ SNSALT	0.278 ± 0.015	0.0461 ± 0.0015	0.722 ± 0.015	69.9 ± 1.3	0.811 ± 0.023

Table 1.1: Current constraints for geometry sensitive cosmological parameters calculated for WMAP7 data combined with various other probes for a ΛCDM Universe.

1.3 Models of Dark Energy

In this section we review a selection of the current models of dark energy, following the reviews of Blanchard (2010) and Copeland et al. (2006).

1.3.1 ΛCDM

Einstein first introduced a *cosmological constant* term Λ into his field equations in 1917 to achieve a static Universe. He later removed it when Hubble observed that the Universe is expanding (see § 1.2.1), since it was no longer required. In recent years however, Λ has been re-introduced into the field equations to accommodate the accelerated expansion. It is by far the simplest model that can be constructed to explain current observations and is obtained by adding a cosmological constant term Λ to the Einstein de-Sitter matter model. The standard Einstein-Hilbert action for gravity with this additional Λ term included is given by

$$S = \frac{1}{16\pi G} \int d^4x \sqrt{-g}(R - 2\Lambda). \tag{1.62}$$

In a FLRW background, the modified form of Einstein's equations are:

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + \Lambda g_{\mu\nu}, \tag{1.63}$$

with the Friedmann and acceleration equations given by:

$$H^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3}, \tag{1.64}$$

$$\left(\frac{\ddot{a}}{a}\right) = -\frac{4\pi G}{3}(\rho + 3P) + \frac{\Lambda}{3}. \tag{1.65}$$

In this model, after the matter dominated era, we will enter a dark energy dominated era, where the scale factor will underdo exponential expansion such that $a(t) \propto e^{Ht}$, with $H = \sqrt{\Lambda/3}$.

We have seen that the ΛCDM model describes large-scale observations of the Universe extremely well. However, there are still various theoretical and astrophysical problems on small-scales that need to be addressed, including:

1. Fine tuning problem - A comparison of the theoretical vacuum density with that obtained from cosmological observations results in $\rho_v^{theory}/\rho_v^{observed} \sim 10^{120}$. Relating this to the equivalent masses, we see a huge discrepancy of ~ 30 orders of magnitude, which is probably the worst prediction in physics to date. Assuming that it is possible to cancel out contributions by summing over opposite signs, it is highly unlikely that we would get a full cancellation. It is even more unlikely that the resulting value would represent the small cosmological constant that we observe today. This is known as the *fine tuning problem*.

2. Coincidence problem - We are living in a cosmologically short epoch in which matter and vacuum energy densities are of the same order of magnitude, but *why now*?

3. Satellite problem - High resolution simulations of galactic sized dark matter halos reveal a plethora of substructures, which resemble scaled down versions of cluster mass dark halos (Moore et al., 1999a). Since clusters contain many galaxies, it makes sense to expect galactic halos to also contain numerous satellite galaxies in the cold dark matter model. However, the number of satellite galaxies detected to date in the local group is over an order of magnitude fewer than would be expected from simple predictions based on the mass and number of dark matter substructures.

4. Cuspy core problem - The rotation curves of low surface brightness galaxies appear to be less sharply peaked than predicted by CDM models, implying that the CDM halos are too cuspy (Moore et al., 1999b).

5. Downsizing problem - In the present day Universe more massive galaxies form new stars less efficiently than lower mass galaxies. It remains controversial however, as to whether this was also true when the Universe was younger. Some results suggest that the trend was weaker in the past, and that the star-forming efficiencies of the more massive and the less massive galaxies evolved differently. Consequently, as the Universe got older the contribution from lower mass galaxies became more important, an effect known as 'downsizing' (Thomas et al., 2010).

See Shanks (2005); Perivolaropoulos (2008); Baugh (2006) for more in-depth reviews of the problems with the ΛCDM model. These problems mean that the presence of a cosmological constant in the theory is largely unwanted. Other solutions to the acceleration problem are described briefly in the next sections.

1.3.2 Scalar Fields: Quintessence

The cosmological constant corresponds to a fluid with a constant equation of state $w = -1$. Observations that constrain the value of w today to be close to that of the cosmological constant say relatively little about the time evolution of w. We can broaden our horizons and consider a situation where the equation of state of dark energy changes with time, such as in inflationary cosmology. Scalar fields naturally arise in particle physics and can act as candidates for dark energy. A wide variety of scalar-field dark energy models have been proposed. In this section we consider quintessence.

Quintessence represents a scalar field that is coupled with gravity. Given a particular potential, quintessence can describe the late-time acceleration of the expansion of the Universe. The action for quintessence is given by

$$S = \int d^4x\sqrt{-g}\left[-\frac{1}{2}(\nabla\phi)^2 - V(\phi)\right], \tag{1.66}$$

where $(\nabla\phi)^2 = g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ and $V(\phi)$ is the potential of the field. In a flat FLRW Universe this action varies as

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} = 0 \tag{1.67}$$

with respect to ϕ. The stress-energy tensor has a form identical to that of an ideal fluid with pressure and density given by:

$$P = = \frac{1}{2}\dot{\phi}^2 + V(\phi), \tag{1.68}$$

$$\rho = = \frac{1}{2}\dot{\phi}^2 - V(\phi). \tag{1.69}$$

The equation of state parameter for the field ϕ is given by

$$w_\phi = \frac{P}{\rho} = \frac{\dot{\phi}^2 + 2V(\phi)}{\dot{\phi}^2 - 2V(\phi)}, \tag{1.70}$$

which suggests that $-1 \leq w \leq 1$. If the time evolution is slow $w \simeq -1$, and the field behaves like a slowly varying vacuum energy.

There is no reason to favour one form of the potential with respect to others. For example, the original scenario was proposed by Ratra & Peebles (1988) with a potential of the form $V(\phi) = \frac{M^{4+\alpha}}{\phi^\alpha}$, whereas Caldwell & Linder (2005) propose a quintessence theory consisting of two classes of *thawing* and *freezing* models, depending on whether the field accelerates or decelerates with time. Another radical approach involves a modification of the kinetic term, as in k-essence models (Armendariz-Picon et al., 2001).

Dynamical models can generally deliver some answers to the question of the nature of dark energy, but do not yet have the capability to provide a complete solution. Some classes of models have the capability to solve the coincidence problem, but the smallness of the cosmological constant, or minimum of the potential in this case, means that the fine tuning problem remains.

1.3.3 Modified Gravity: $f(R)$

Until now, we have only considered models that contain an additional term in the stress-energy tensor which have the properties required to account for the late-time acceleration of the Universe. An interesting alternative approach is to modify the geometrical part of the Einstein equation. The $f(R)$ modified gravity theories work by replacing the standard Einstein-Hilbert action by an arbitrary function of the Ricci scalar R. In Riemannian geometry, R describes the simplest curvature invariant of a Riemannian manifold. It assigns a single real number to each point on the manifold which is determined by the intrinsic geometry of the manifold near that point. In particular, R represents the amount by which the volume of a geodesic ball in a curved Riemannian manifold deviates from that of the standard ball in Euclidean space. In this section we follow the review of Lobo (2008).

The general action for a modified gravity field is given by

$$S = \frac{1}{16\pi G} \int d^4x \sqrt{-g} \left[f(R) + \mathcal{L}_m \right], \tag{1.71}$$

where $\mathcal{L}_m$ is the matter Lagrangian density. Variation of this action with respect to the metric $g^{\mu\nu}$ yields the field equation

$$FR_{\mu\nu} - \frac{1}{2}fg_{\mu\nu} - \nabla_\mu \nabla_\nu F + g_{\mu\nu}\Box F = 8\pi G T^m_{\mu\nu}, \tag{1.72}$$

where $F \equiv df/dR$ and $T^m_{\mu\nu}$ is the matter stress-energy tensor. This equation may be written as

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T^{eff}_{\mu\nu}, \tag{1.73}$$

where the new term $T^{eff}_{\mu\nu} = T^c_{\mu\nu} + T^m_{\mu\nu}/F$ describes the effective stress-energy tensor. All of the effects of the modification of gravity are now contained within the curvature part of the tensor $T^c_{\mu\nu}$.

For a FLRW metric, the generalised Friedmann and acceleration equations have an identical form to Eqns. 1.4 and 1.5 with $P_{eff} = P_m + P_c$ and $\rho_{eff} + \rho_m + \rho_c$. The pressure and density of the curvature field are given by:

$$P_c = \frac{1}{8\pi GF}\left(2\frac{\dot{a}}{a}\dot{R}f' + \ddot{R}F' + \dot{R}^2F'' - \frac{1}{2}[f - RF]\right), \tag{1.74}$$

$$\rho_c = \frac{1}{8\pi GF}\left(\frac{1}{2}[f - Rf] - 3\frac{\dot{a}}{a}\dot{R}F'\right). \tag{1.75}$$

If there is no matter $P_m = \rho_m = 0$ and the effective equation of state can be defined as $w_{eff} \equiv P_{eff}/\rho_{eff} = P_c/\rho_c$. Given the right choice of $f(R)$, these models can reproduce the late-time acceleration of the Universe.

$f(R)$ models however, are not without problems. For example, Chiba (2003) demonstrated that it is possible to derive the $f(R)$ equations of motions from a scalar-tensor theory by introducing a transformation $\{R, f\} \to \{\phi, V\}$, which has Lagrangian density $\mathcal{L}_{BD} = \phi R - V(\phi)$, ie. a Brans-Dicke parameter $\omega_{BD} = 0$. It has been shown by Bertotti et al. (2003) however, that this contradicts current solar system constraints. Many other $f(R)$ models also have a scale factor that evolves as $a \propto t^{1/2}$ in the matter dominated era, which is in contradiction with observations where $a \propto t^{2/3}$ (Amendola et al., 2007).

A different approach to modifying gravity includes extending the theory to account for extra dimensions. Superstring and supergravity theories possess astonishing properties in higher dimensional space and have therefore gained a lot of attention from theorists. Modern versions have been developed in the context of string theory (Maartens 2004, Koyama 2008) and are known as braneworlds or brane cosmology.

Braneworlds work by confining matter and radiation components to the brane, whilst gravity is allowed to move freely within the bulk. Extra dimensions are inaccessible from the brane, so the behaviour of the standard model of particle physics is left unchanged, whilst gravity can act in very different ways. The vacuum energy in the brane provides a tension term σ, with another vacuum energy present in the bulk Λ_B. The general action now contains terms involving an equivalent of the Ricci scalar, which corresponds to gravity in the higher dimension.

The Dvali-Gabadadze-Porrati (DGP) model (Dvali et al., 2000), where the Randall-Sundrum models (Randall & Sundrum 1999b, Randall & Sundrum 1999a) are modified at low energies, contain extra dimensions that are infinite (Deffayet, 2001). The effective action contains explicitly a 4D Einstein-Hilbert term on the brane in addition to the 5D term on the bulk. This introduces a scale r_c and two distinct regimes appear: on scales smaller than r_c gravity results from the 4D term and classical GR is recovered; on scales larger than r_c gravity is "leaking" and the expansion is eventually accelerated.

1.4 Thesis Summary

Understanding the nature of the accelerated expansion of the Universe is one of the main challenges facing cosmologists today. As shown in the previous section, there are a plethora of dark energy models available that are capable of describing the late-time acceleration, given the right conditions. The only way to tell these theories apart is by testing the evolution of the dark energy equation of state. By fitting observations of the accelerated expansion to our dark energy models we can hope to achieve *cosmic concordance*, ie. we can determine which model best describes the current state of the Universe. One of the key observational methods that will be used to help meet this challenge involves using Baryon Acoustic Oscillations (BAO) in the 2-point galaxy clustering signal as a standard ruler to make precise measurements of cosmological expansion.

Some of the next generation of sky surveys, including the Dark Energy Survey (DES `www.darkenergysurvey.org`), the Panoramic Survey Telescope and Rapid Response System (PanStarrs `pan-starrs.ifa.hawaii.edu`), and the Large Synoptic Survey Telescope (LSST `www.lsst.org`), will use photometric techniques to estimate galaxy redshifts, rather than more precise estimates from spectroscopic emission lines. The larger uncertainties on galaxy redshifts induce errors on inferred distances in the radial direction. The amplitude of the power spectrum and correlation function is reduced in the radial direction by this smoothing, removing information. Consequently, the main aim of this thesis is to investigate how alternative clustering measurement techniques may be used to extract information on dark energy for future photometric redshift surveys.

The layout of this thesis is as follows:

- Chapter 2 - We introduce the theory and methodology required to conduct accurate clustering analyses. We address the issue of which correlation function estimator to use, and explore algorithm optimisation techniques for large data-sets.
- Chapter 3 - We investigate systematic effects arising in the projected correlation function as a result of the coherent movement of galaxies between real- and redshift-space, and introduce a new binning scheme that alleviates the problem.
- Chapter 4 - We quantify the level of systematic error induced on a typical 3D clustering analysis by conflicting photometric redshift estimates.
- Chapter 5 - We predict how the systematic effects we have explored in the previous chapters will affect typical clustering analyses for future experiments, such as the Dark Energy Survey.

Chapter 2

Clustering Measurement Techniques

Galaxy clustering analysis techniques have been vastly refined to deal with increasingly larger and more sophisticated data-sets since their introduction (Neyman & Scott, 1952; Neyman et al., 1953; Peebles, 1973). In this chapter we introduce the statistical and methodological tools required to successfully quantify the clustering of galaxies in the Universe.

2.1 Introduction to $\xi(r)$ and $P(k)$

2.1.1 The 2-Point Correlation Function $\xi(r)$

The 2-point correlation function (2PCF) is a powerful statistic and has been used extensively in cosmology to quantify the clustering of galaxies (see Percival 2007 for a detailed review). In order to describe clustering analysis techniques in more detail let us first consider the dimensionless overdensity

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle\rho\rangle}{\langle\rho\rangle} \tag{2.1}$$

where $\rho(\mathbf{x})$ is the observed density at a location $\mathbf{x}$, and the expected mean density of the Universe $\langle\rho\rangle$ is invariant under translation due to statistical homogeneity. The autocorrelation function of the overdensity field can now be defined as

$$\begin{aligned} \xi(\mathbf{x}, \mathbf{x}+\mathbf{r}) &= \langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle & (2.2) \\ &= \xi(r) & (2.3) \end{aligned}$$

with Eq. 2.3 arising because of statistical isotropy and homogeneity and where r is the scale of interest. The angled brackets in Eq. 2.2 denote a spatial average.

The correlation function statistic can be understood by considering two small regions δV_1 and δV_2 separated by a distance r. The joint probability that galaxies are assigned to the elements δV_1 and δV_2 is given by

$$\delta^2 P_2 = \bar{n}^2[1 + \xi(r)]\delta V_1 \delta V_2 \tag{2.4}$$

where $\bar{n}$ is the mean number of galaxies per unit volume. $\xi(r)$ measures the excess clustering of galaxies at separation r and can be quantified as follows:

- $\xi(r) = 0$: the number of pairs present is a product of the expected number in volumes δV_1 and δV_2 and we recover a random Poisson distribution of galaxies.
- $\xi(r) > 0$: galaxies are strongly clustered.
- $\xi(r) < 0$: galaxies are anti-clustered.

On large-scales $\xi(r)$ is well described by a power-law function that scales with distance

$$\xi(r) = \left(\frac{r_0}{r}\right)^{\gamma} \tag{2.5}$$

where r_0 is a characteristic length scale. On small-scales however, this model is over-simplified and does not describe non-linear structure growth (eg. Zehavi et al. 2004).

2.1.2 The Power Spectrum $P(k)$

It is common in cosmology to build up a general field of fluctuations via the superposition of plane waves in Fourier space, allowing the fluctuations to evolve independently whilst in the linear regime (Peacock, 1999). If we construct a volume $V_U = L^3$ where L is much greater than the maximum scale at which there is significant structure due to perturbations, we can consider V_U as a fair sample of the Universe. In this case, it is possible to make a realisation of the Universe if we divide it up into cells of size V_U and assume periodic boundary conditions. Using these assumptions we can express the overdensity $\delta(\mathbf{x})$ that we defined in Eq. 2.1 as a Fourier series:

$$\delta(\mathbf{x}) = \sum_{\mathbf{k}} \delta_{\mathbf{k}} \exp(i\mathbf{k} \cdot \mathbf{x}) = \sum_{\mathbf{k}} \delta^*_{\mathbf{k}} \exp(-i\mathbf{k} \cdot \mathbf{x}) \tag{2.6}$$

where the assumption of periodic boundary conditions forces the wavevector $\mathbf{k}$ to take the values

$$k_x = n\frac{2\pi}{L}, \quad n = 1, 2..., \tag{2.7}$$

with similar expressions for k_y and k_z.

The Fourier coefficients $\delta_{\mathbf{k}}$ are complex quantities given by

$$\delta_{\mathbf{k}} = \frac{1}{V_U} \int_{V_U} \delta(\mathbf{x}) \ \exp(-i\mathbf{k} \cdot \mathbf{x}) \ dx \tag{2.8}$$

The conservation of mass in V_U means that $\delta_{\mathbf{k}=0} = 0$ and the reality of $\delta(\mathbf{x})$ means that $\delta_{\mathbf{k}} = \delta^*_{\mathbf{k}}$. Each new volume that we choose will have different values of $\delta_{\mathbf{k}}$. Therefore, for a large number of volumes realisations $\delta_{\mathbf{k}}$ will vary in both amplitude and phase. Following standard models of inflation, we can assume Gaussian statistics with the phases of $\delta_{\mathbf{k}}$ random both across the ensemble of realisations and from node to node within single realisations. With this being the case, the mean value of the perturbation $\delta(\mathbf{x}) \equiv \delta$ across the statistical ensemble is zero by definition. Its variance σ^2 is not;

$$\sigma^2 \equiv \langle \delta^2 \rangle = \sum_{\mathbf{k}} \langle |\delta_{\mathbf{k}}|^2 \rangle = \frac{1}{V_U} \sum_{\mathbf{k}} \delta_k^2 \tag{2.9}$$

where the average is taken over an ensemble of realisations. In Eq. 2.9, $\langle |\delta_{\mathbf{k}}|^2 \rangle$ represents the contribution made to the variance due to waves of wavenumber $\mathbf{k}$. Taking the limit where $V_U \rightarrow \infty$ under the assumption that the density field is statistically homogeneous and isotropic ie. no spatial dependence, we can show that

$$\sigma^2 = \frac{1}{V_U} \sum_{\mathbf{k}} \delta_k^2 \rightarrow \frac{1}{2\pi} \int_0^\infty P(k) k^2 dk \tag{2.10}$$

where $P(k)$ is the power spectrum. The amplitude of the perturbations evolve with time. Therefore, the variance only provides us with information about the amplitude of perturbations and not about their spatial structure.

The power spectrum can be defined in a similar way to $\xi(r)$ in Eq. 2.2 as

$$P(\mathbf{k}, \mathbf{K}) = \frac{1}{(2\pi)^3} \langle \delta(\mathbf{k}) \delta(\mathbf{K}) \rangle \tag{2.11}$$

with statistical isotropy and homogeneity giving

$$P(\mathbf{k}, \mathbf{K}) = \delta_D(\mathbf{k} - \mathbf{K}) P(k) \tag{2.12}$$

where δ_D is the Dirac delta function.

The correlation function and power spectrum form a Fourier pair

$$P(k) \equiv \frac{1}{V}\int \xi(r)\exp(i\mathbf{k}\cdot\mathbf{r})\ d^3r \tag{2.13}$$

$$\xi(\mathbf{r}) = \frac{1}{(2\pi)^3}\int P(k)\exp(-i\mathbf{k}\cdot\mathbf{r})\ d^3k \tag{2.14}$$

and so provide the same statistical information. Given isotropy of $P(k)$, we can reduce Eqns. 2.14 & 2.13 to 1D integrals of the form

$$\xi(r) = \frac{V}{(2\pi)^3}\int P(k)\frac{\sin kr}{kr}4\pi k^2 dk, \tag{2.15}$$

and

$$\Delta^2(k) \equiv \frac{V}{(2\pi)^3}4\pi^3 P(k) = \frac{2}{\pi}k^3\int_0^\infty \xi(r)\frac{\sin kr}{kr}r^2 dr, \tag{2.16}$$

where Eq. 2.16 is the dimesionless form of the power spectrum. These solutions are complete for a Gaussian random field.

2.1.3 Gaussian Random Field

The assumption that the primordial density field is Gaussian in its linear regime means that its *n*-point joint probability distribution obeys that multi-variate Gaussian

$$P(\delta_1, \delta_2, ..., \delta_n)d\delta_1 d\delta_2 ... d\delta_n = \frac{1}{\sqrt{(2\pi)^n det(M)}}\exp\left[-\sum_{i,j=1}^{n}\frac{1}{2}\delta_i(M^{-1})_{i,j}\delta_j\right]d\delta_1 d\delta_2 ... d\delta_n \tag{2.17}$$

for an arbitrary positive integer *n*, where $M_{ij} \equiv \langle\delta_i\delta_j\rangle$ is the covariance matrix, and M^{-1} is its inverse. We have already shown that $M_{ij} = \xi(x_i, x_j)$, therefore Eq. 2.17 implies that the statistical nature of the Gaussian density field is completely specified by the two-point correlation function ξ and its linear combination (Bardeen et al., 1986).

From the definition of the Gaussian distribution we can see that Eq. 2.17 formally assumes a symmetrically distributed density field in the range $-\infty < \delta_i < \infty$. In reality, the density field cannot be less than -1, which means that it only preserves its Gaussian nature in its linear evolution stage, and not in its nonlinear stage. In the case when fluctuations are infinitesimally small this assumption makes no practical difference. However, in the nonlinear regime, where the typical amplitude of fluctuations exceed unity, it is no longer valid.

We can describe the linear theory of cosmological density fluctuations using Eq. 2.8, where $\delta_{\mathbf{k}}$ is a complex variable that can be decomposed into a set of two real variables: amplitude $D_{\mathbf{k}}$ and phase $\phi_{\mathbf{k}}$, such that

$$\delta_{\mathbf{k}} \equiv D_{\mathbf{k}} \exp(i\phi_{\mathbf{k}}). \tag{2.18}$$

The linear perturbation equation now becomes

$$\ddot{D}_{\mathbf{k}} + 2\frac{\dot{a}}{a}\dot{D}_{\mathbf{k}} - (4\pi G\bar{\rho} + \dot{\phi}^2)D_{\mathbf{k}} = 0 \tag{2.19}$$

$$\ddot{\phi}_{\mathbf{k}} + 2\left(\frac{\dot{a}}{a} + \frac{\dot{D}_{\mathbf{k}}}{D_{\mathbf{k}}}\right)\dot{\phi}_{\mathbf{k}} = 0 \tag{2.20}$$

From Eq. 2.20 we get $\dot{\phi}(t) \propto a^{-2}(t)D_{\mathbf{k}}^{-2}(t)$, and $\phi(t)$ rapidly converges to a constant value. Therefore, $D_{\mathbf{k}}$ evolves following the growing solution in linear theory.

The most popular statistic of choice to measure clustering in the Universe is the power spectrum of density fluctuations,

$$P(t, \mathbf{k}) \equiv \langle D_{\mathbf{k}}(t)^2 \rangle, \tag{2.21}$$

which measures the amplitude of the mode of the wavenumber $\mathbf{k}$. As shown in Section 2.1.2, this is the Fourier transform of the two-point correlation function ξ. Although these clustering statistics are both very powerful, they have one fundamental flaw; they contain no information about the phase $\phi_{\mathbf{k}}$. Thus, in principle two clustering patterns may be completely different even if they have identical two-point correlation functions.

In the Gaussian field however, we can show that Eq. 2.17 reduces to the probability distribution function of $\phi_{\mathbf{k}}$ and $D_{\mathbf{k}}$, which can be explicitly written as

$$P(|\delta_{\mathbf{k}}|, \phi_{\mathbf{k}})d|\delta_{\mathbf{k}}|d\phi_{\mathbf{k}} = \frac{2|\delta_{\mathbf{k}}|}{P(k)} \exp\left(-\frac{|\delta_{\mathbf{k}}|^2}{P(k)}\right) d|\delta_{\mathbf{k}}|\frac{d\phi_{\mathbf{k}}}{2\pi}, \tag{2.22}$$

and is mutually independent of k. The phase distribution is uniform and therefore does not carry information, so Eq. 2.22 is completely fixed if $P(k)$ is specified. Consequently, the Gaussian field is completely specified by the two-point correlation function in real space. The most fundamental statistic for characterizing the large-scale structure of the Universe can be found in the form of a probability distribution function of the cosmological density fluctuations (see Section 2.1.4).

2.1.4 Log-Normal Distribution

The probability density function (PDF) of the cosmological density fluctuations deviate significantly from Gaussianity when they are in the nonlinear regime. This is due to the strong nonlinear mode-coupling and the non-locality of the gravitational dynamics. Kayo et al. (2001) showed that a one-point log-normal PDF of the functional form

$$P_{LN}^{(1)}(\delta) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{[\ln(1+\delta)+\sigma_1^2/2]^2}{2\sigma_1^2}\right) \frac{1}{1+\delta} \tag{2.23}$$

describes the cosmological density distribution very accurately, even in the nonlinear regime.

From an empirical perspective, it was Hubble (1934) who first noted that the galaxy distribution in angular cells on the celestial sphere may be approximated by a log-normal distribution, rather than a Gaussian. In theory, the log-normal function described by Eq. 2.23 can be obtained from the mapping between the linear random-Gaussian and the non-linear density fields. Defining a linear density field g smoothed over R obeying the Gaussian PDF we have,

$$P_G^{(1)}(g) = \frac{1}{\sqrt{2\pi\sigma_{lin}^2}} \exp\left(-\frac{g^2}{2\sigma_{lin}^2}\right), \tag{2.24}$$

where the variance is computed from its linear power spectrum,

$$\sigma_{lin}^2(R) \equiv \frac{1}{2\pi^2}\int_0^\infty P_{lin}(k)\bar{W}^2(kR)k^2dk. \tag{2.25}$$

If we introduce a new field δ from g as

$$1+\delta = \frac{1}{\sqrt{1+\sigma_{nl}^2}} \exp\left(\frac{g}{\sigma_{lin}}\sqrt{\ln(1+\sigma_{nl}^2)}\right), \tag{2.26}$$

the PDF for δ is given by $(dg/d\delta)P_G^{(1)}(g)$, which reduces to Eq. 2.23.

2.1.5 Choosing an Estimator

Our ability to accurately quantify the amount of clustering in the Universe depends crucially on our ability to understand where we should have been able to observe galaxies. An easy way to do this is to create an unclustered random catalogue with which we can contrast a galaxy catalogue. We can achieve this by matching the angular and radial properties of the random catalogue to that of the galaxy catalogue under analysis, in the absence of clustering. Typically, this is done by generating smooth model distributions of the galaxy catalogue, which can then be Poisson sampled to compose an unclustered random catalogue with the same spatial properties of the galaxy catalogue.

Correlation function estimators generally contrast combinations of galaxy-galaxy (DD), galaxy-random (DR), and random-random (RR) pair counts, in and between galaxy and random catalogues. For statistical purposes, random catalogues typically have much larger number densities than galaxy catalogues. Because of this, pair counts are suitably normalised to match the expected pair counts for each catalogue, ie. $DD = 2DD/N_g(N_g - 1)$, where N_g is the number of galaxies in the catalogue. A number of estimators for the 2PCF have been constructed over the years for use in clustering analyses, each with its own advantages and disadvantages. Some of the most popular estimators for $\hat{\xi}$ include:

$$\widehat{\xi}_N = \frac{DD}{RR} - 1, \tag{2.27}$$

$$\widehat{\xi}_{DP} = \frac{DD}{DR} - 1, \tag{2.28}$$

$$\widehat{\xi}_{He} = \frac{DD - DR}{RR}, \tag{2.29}$$

$$\widehat{\xi}_{Ha} = \frac{DD\ RR}{DR^2}, \tag{2.30}$$

$$\widehat{\xi}_{LS} = \frac{DD - 2DR + RR}{RR}, \tag{2.31}$$

where subscripts denote natural (N), Davis & Peebles (1982) (DP), Hewett (1982) (He), Hamilton (1993) (Ha) and Landy & Szalay (1993) (LS).

Naturally, we want to choose an estimator that will give us the most robust results possible. Kerscher et al. (2000) compared the performance of nine of the most important estimators known for the 2-pt correlation function using a predetermined and rigorous criterion and found that the Landy-Szalay estimator out-performed the others on most counts. Let us consider how this estimator is derived.

Combining Eqns. 2.1 and 2.2 we get

$$\xi(r) = \left\langle \left(\frac{\rho(\mathbf{x_1}) - \bar{\rho}}{\bar{\rho}} \right) \cdot \left(\frac{\rho(\mathbf{x_2}) - \bar{\rho}}{\bar{\rho}} \right) \right\rangle, \tag{2.32}$$

where $r = |\mathbf{x_1} - \mathbf{x_2}|$. Expanding this expression gives

$$\xi(r) = \frac{\langle \rho(\mathbf{x_1})\rho(\mathbf{x_2})\rangle - \langle \rho(\mathbf{x_1})\bar{\rho}(\mathbf{x_2})\rangle - \langle \rho(\mathbf{x_2})\bar{\rho}(\mathbf{x_1})\rangle}{\langle \bar{\rho}(\mathbf{x_1})\bar{\rho}(\mathbf{x_2})\rangle} + 1, \tag{2.33}$$

where we now include a scale dependence on the mean space density $\bar{\rho}(x)$ to ensure that the correlation function is not contaminated by smooth large-scale variations in the mean density of the field. Dropping the angular brackets and explicitly averaging each term gives,

$$\xi(r) = \frac{\frac{1}{\alpha_{DD}}\sum_i^{N_D}\sum_j^{N_D}\rho(\mathbf{x_i})\rho(\mathbf{x_j}) - \frac{2}{\alpha_{DR}}\sum_i^{N_R}\sum_j^{N_D}\bar{\rho}(\mathbf{x_i})\rho(\mathbf{x_j})}{\frac{1}{\alpha_{RR}}\sum_i^{N_R}\sum_j^{N_R}\bar{\rho}(\mathbf{x_i})\bar{\rho}(\mathbf{x_j})} + 1, \tag{2.34}$$

where N_D and N_R are the number of random positions x_i sampled in the data and random fields, respectively. α_{DD}, α_{DR}, and α_{RR} are normalisation terms that match the average number densities of each field.

Now consider a discrete denisty field rather than a continuous one and say that the number of objects in two volumes separated by r is given by $n(x) = V\rho(x)$. Allowing $V \to 0$ means that the number of cells occupied will be equivalent to the number of points, that is, N_D and N_R represent the number of data and random points exactly. Hence, the Landy-Szalay estimator can be defined

$$\xi = \frac{DD - 2DR + RR}{RR}. \tag{2.35}$$

In the above expression we have adopted the following definitions:

$$DD = \frac{1}{N_D(N_D-1)}\sum_i^{N_D}\sum_j^{N_D}\sqcap(|\mathbf{x}_i^{data} - \mathbf{x}_j^{data}|), \tag{2.36}$$

$$DR = \frac{1}{N_D N_R}\sum_i^{N_R}\sum_j^{N_D}\sqcap(|\mathbf{x}_i^{rand} - \mathbf{x}_j^{data}|), \tag{2.37}$$

$$RR = \frac{1}{N_r(N_R-1)}\sum_i^{N_R}\sum_j^{N_R}\sqcap(|\mathbf{x}_i^{rand} - \mathbf{x}_j^{rand}|), \tag{2.38}$$

where x^{data} and x^{rand} are the position vectors of the data and random points, respectively, and $\sqcap$ is the rectangular step function defined as,

$$\sqcap(t) = \begin{cases} 0|t-r| > dr/2, \\ 1|t-r| < dr/2. \end{cases} \tag{2.39}$$

A simple test can be conducted to show the merits of using the Landy-Szalay estimator. Consider a simple argument where we have an overdense region A and an underdense region B. In order to quantify the total clustering signal we need to contrast the density field with an underlying Poisson distribution. We do this by counting data and random pairs. For example, suppose there are 12 galaxies in region A ($D_A = 12$) and 8 galaxies in region B ($D_B = 8$) where we expect there to be 10 galaxies per region on average ($R_A = R_B = 10$). The counts of small-scale pairs are:

$$DD = \frac{1}{2}\left(A_D^2 + B_D^2\right) = 104, \tag{2.40}$$

$$DR = \frac{1}{2}\left(A_D B_R + A_R B_D\right) = 100, \tag{2.41}$$

$$RR = \frac{1}{2}\left(A_R^2 + B_R^2\right) = 100. \tag{2.42}$$

Plugging these values into the estimators introduced in Section 2.1 gives: $\widehat{\xi}_N = DD/RR - 1 = 0.04$, $\widehat{\xi}_{DP} = DD/DR - 1 = 0.04$, $\widehat{\xi}_{He} = (DD - DR)/RR = 0.04$, $\widehat{\xi}_{Ha} = (DD)(RR)/DR^2 - 1 = 0.04$ and $\widehat{\xi}_{LS} = DD - 2DR + RR/RR = 0.04$. Each estimator gives the same result, as should be the case.

Now let us consider an example where we only have region B. We have that

$$DD = B_D^2 = 64, \tag{2.43}$$

$$DR = B_D B_R = 80, \tag{2.44}$$

$$RR = B_R^2 = 100, \tag{2.45}$$

so that $\widehat{\xi}_N = -0.36$, $\widehat{\xi}_{DP} = -0.2$, $\widehat{\xi}_{He} = -0.16$, $\widehat{\xi}_{Ha} = 0.0$ and $\widehat{\xi}_{LS} = 0.04$. The Landy-Szalay estimator is the only one that remains accurate. To understand where this discrepancy arises let us consider the two scenarios in more detail.

In the first case, we are considering both the clusters *and* the voids in the sample. Therefore, the average overdensity $\langle\delta\rangle \simeq 0$. In the second case however, even though we are probing the same density field, we are only considering the voids in the sample, leading to an average overdensity $\langle\delta\rangle < 0$. The latter is equivalent to introducing *fake* clustering into the sample, because what we are seeing is a mismatch between the data density field and the expected random density field. All estimators, with the exception of Landy-Szalay, fail to recover the true clustering signal in this case when $\langle\delta\rangle \neq 0$. This implies that the Landy-Szalay estimator is the only one that is robust in all scenarios.

To demonstrate this effect on an actual data-set, we have considered the SDSS DR7 Luminous Red Galaxy sample. The SDSS DR7 sample consists of 80,046 LRGs as de-

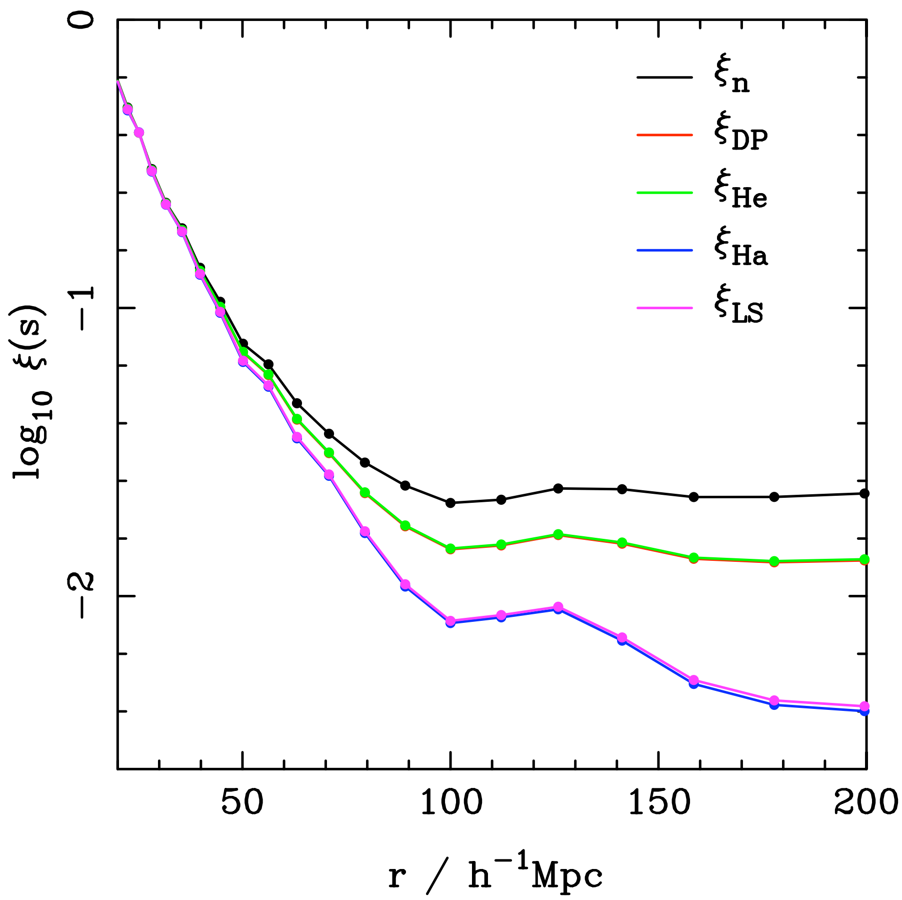

Figure 2.1: The redshift-space correlation function for Luminous Red Galaxies in the SDSS DR7 sample, calculated from the initial draft of the catalogues as defined in Percival et al. (2010). We plot 5 different estimators of the correlation function: $\xi_n = DD/RR - 1$, $\xi_{DP} = DD/DR - 1$ (Davis & Peebles, 1982), $\xi_{He} = DD - DR/RR - 1$ (Hewett, 1982), $\xi_{Ha} = DD\ RR/DR^2 - 1$ (Hamilton, 1993), and $\xi_{LS} = DD - 2DR + RR/RR$ (Landy & Szalay, 1993). There is an obvious discrepancy between estimators on large-scales. Although the variance of each estimator is expected to differ, the estimators should provide us with an unbiased result since they are describing the same clustering signal. This suggests discrepancies between the galaxy and random catalogues.

fined by Eisenstein et al. (2001). An additional 30,530 main galaxies are classified as LRGs, bringing the total to 110,576. An unclustered random catalogue was constructed with the same angular distribution and $10\times$ the number density of the galaxy catalogue. The radial distribution was created such that it only approximately matched the galaxy catalogue, and was created via a smooth spline-fit to the data using 30 spline nodes as described in Eisenstein et al. (2005). Using a simple N^2 analysis, we calculated the number of galaxy-galaxy (DD), galaxy-random (DR), and random-random (RR) pairs in 70 $\log$-normal bins of 3D separation, r. Normalised pair-counts were used to calculate the correlation function using the estimators defined in § 2.1.5. Results are plotted in Fig. 2.1.

There is an obvious discrepancy between estimators on large-scales. Although the variance of each estimator is expected to differ, the estimators should provide us with an unbiased result since they are describing the same clustering signal. What we are seeing is the mismatch between the galaxy and random catalogues introducing a spurious clustering signal that is biasing the average overdensity such that $\langle\delta\rangle \neq 0$, as we saw before in our simple example. Modelling the radial distribution of the random catalogue is a non-trivial task, since the typical radial distribution of an LRG sample has a complicated form. Our results suggest that much more care is required in the modelling of the radial distribution in order to avoid these discrepancies. Without this, use of the Landy-Szalay estimator can alleviate these issues.

Kazin et al. (2010) and Percival et al. (2010) conducted clustering analyses on the final SDSS DR7 LRG sample using more accurate radial distribution modelling techniques, thus avoiding mismatches between galaxy and random catalogues and mitigating this effect.

2.1.6 Calculating Pair-Counts

There are a number of different approaches we can take when computing pair-counts. The simplest is to construct a double integral over the catalogues that we are searching for pairs within. Computationally, this scales as $\sim N_p^2$, where N_p represents the number of data points in a catalogue. Modern sky surveys have the capability of producing large amounts of data. For example, the final version of DR7 for the SDSS contains $\sim 930,000$ galaxies. Given that random catalogues are typically constructed with around 10 to 100 times the number density of galaxy catalogues, the use of this technique becomes computationally unfeasible. Constant effort is being made to alleviate this problem.

One solution is to assign the data field to a density grid (Barriga & Gaztanaga, 2002; Eriksen et al., 2005), which is relatively easy to implement in practice. The nearest grid point (NGP) mass assignment scheme is used to place the data field onto N_{grid} cells. The correlation function can then be estimated using the equation

$$\hat{\xi}(r) = \frac{1}{N_{pairs}} \sum_{ij} \delta_i \delta_j, \tag{2.46}$$

where $\delta_i = (n_i - \langle n \rangle)/\langle n \rangle$ is the density contrast in the i^{th} bin of the grid and the sum extends over the N_{pairs} cells separated by distances between $r - \Delta r/2$ and $r + \Delta r/2$. This process scales as N_{grid}^2, providing a vast reduction in computation time since generally $N_{grid} \ll N_p$. This method provides an accurate estimation of the correlation function on scales larger than a few grid cells. Getting down to smaller scales to observe non-linear effects requires a finer resolution grid, which can often result in $N_{grid} \rightarrow N_p$.

The symmetrical nature of the density grid can be utilised to further speed up this process. The indices of the nearest neighbour N_{neigh} grid-cells for a given bin of separation can be calculated and stored. This list of indices can then be translated to different locations on the density grid so that no CPU time is wasted in re-computing the grid-cells that contribute to the correlation function in a given bin of separation. Processing time is now reduced from N_{grid}^2 to $N_{grid}N_{neigh}$(Sanchez et al., 2008).

In Table 2.1 and Fig. 2.2, we present the average CPU time required for the N_{grid}^2 and the $N_{grid}N_{neigh}$ processes, assuming a density field with volume $V = (3000\,h^{-1}\,\mathrm{Mpc})^3$. Since results can only be trusted on scales larger than a few grid cells (d_c), we started at a resolution of $N_{grid}^{1/3} = 300$, which gives a minimum pair separation of $5d_c = 50\,h^{-1}\,\mathrm{Mpc}$. The results from this code-based test were then used to infer results for lower and higher resolution grids using the formulae:

$$N_{neigh} = 2\left(\frac{r_{max}}{d_c}\right)^3, \tag{2.47}$$

$$t(n) = t(n-1)\left(\frac{N_{grid}(n)}{N_{grid}(n-1)}\right)^2 \qquad \text{for } N_{grid}^2, \tag{2.48}$$

$$t(n) = t(n-1)\left(\frac{N_{grid}N_{neigh}(n)}{N_{grid}N_{neigh}(n-1)}\right) \qquad \text{for } N_{grid}N_{neigh}, \tag{2.49}$$

where $r_{max} = 150\,h^{-1}\,\mathrm{Mpc}$. The reduction in CPU time for the $N_{grid}N_{neigh}$ regime is vast, and is optimum on lower resolution grids since $N_{neigh} \ll N_{grid}$. It is important to reiterate that the number density of the data-set is irrelevant, as it is pairs of grid-cells that are being counted. A typical N_p^2 process for $\sim 1 \times 10^6$ galaxies takes ~ 3 hours for the data-pairs alone. For approximately the same CPU time, we can construct a full correlation function calculation on a grid for *any* number of galaxies. However, getting

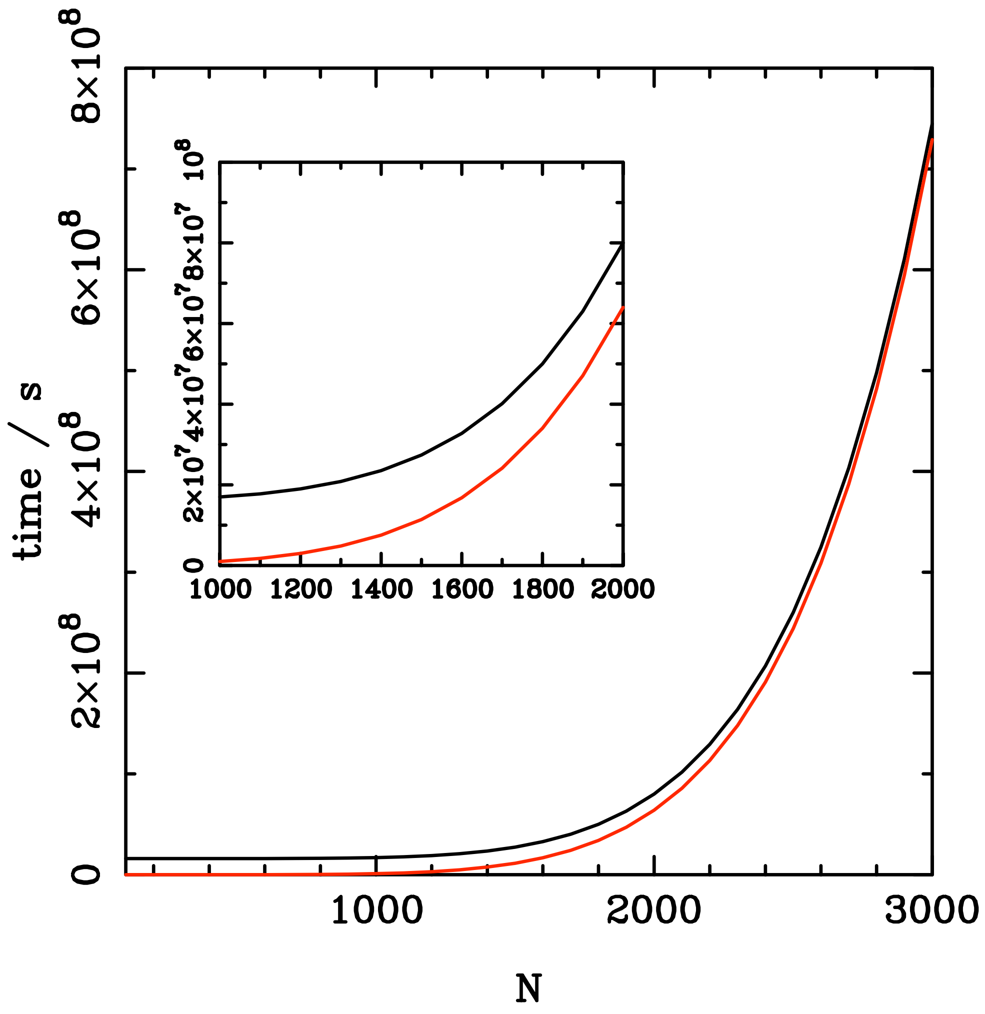

Figure 2.2: Average CPU time required for N_{grid}^2 (black) and $N_{grid}N_{neigh}$ (red) processes, plotted as a function of $N_{grid}^{1/3}$. The nearest neighbours process reduces the average CPU time drastically at low grid resolutions, as shown inset. The N_{grid}^2 process scales exactly as the traditional N_p^2 process if $N_{grid} = N_p$. From this we can deduce that for grid sizes $d_c < 1\,h^{-1}\,\mathrm{Mpc}$ the nearest neighbours process $N_{grid}N_{neigh} \rightarrow N_{grid}^2 = N_p^2$, and so we do not gain in CPU time by moving the data onto a grid. However, if $N_p \gg N_{grid}$, there is a clear optimisation achievable for $N_{grid} < 3000^3$ in a volume $3000\,h^{-1}\,\mathrm{Mpc}$, as explored in this example. See text for more details.

.

down to small scales by employing higher resolution grids results in $N_{neigh} \rightarrow N_{grid}$. For this reason, it is advisable to run a simple N_p^2 process on a sub-sample of the data to obtain robust small-scale clustering signals.

In Fig. 2.3 we present the correlation function $\xi(r = \sqrt{\sigma^2 + \pi^2})$ for the DES Large-Scale Structure Working Group MICE simulations (see § 5.3.1 for details), calculated

$N_{grid}^{1/3}$	N_{grid}	d_c	$\sim N_{neigh}$	$\sim$CPU time/s (N_{grid}^2)	$\sim$CPU time/s ($N_{grid}N_{neigh}$)
300	2.7×10^7	10	6.75×10^3	1.6×10^7	8.3×10^3
400	6.4×10^7	7.5	1.6×10^4	8.9×10^7	4.7×10^4
500	1.25×10^8	6	3.125×10^4	3.4×10^8	1.8×10^5
1000	1.0×10^9	3	2.5×10^5	2.2×10^{10}	1.2×10^7

Table 2.1: Table comparing the relative CPU time required for both N_{grid}^2 and $N_{grid}N_{neigh}$ processes. See text for details.

on a coarse grid with $d_c = 10\,h^{-1}\,\mathrm{Mpc}$. The baryonic ridge can be clearly detected at $\sim 100\ h^{-1}\,\mathrm{Mpc}$ in both planes.

2.1.7 Error and the Covariance Estimation

Error on clustering measurements may be estimated using a variety of techniques. In this section we will discuss three of the most commonly utilised "internal" methods, namely sub-sampling, jackknife resampling, and bootstrap resampling, all of which use the data itself to derive an estimate of the error on a measurement.

All of these techniques work by making copies of the observed data in order to sample the underlying probability density function of the quantity we are trying to measure. There are three main assumptions that are made for "internal" approaches:

1. The data provides an accurate representation of the underlying probability distribution.

2. The number of sub-samples the data is split into is sufficient to allow accurate estimates of the errors.

3. The volume of each sub-sample is sufficiently large to be representative.

The violation of the first assumption would indicate that the data-set is subject to cosmic variance, in which case the clustering signal would be distorted. The second and third

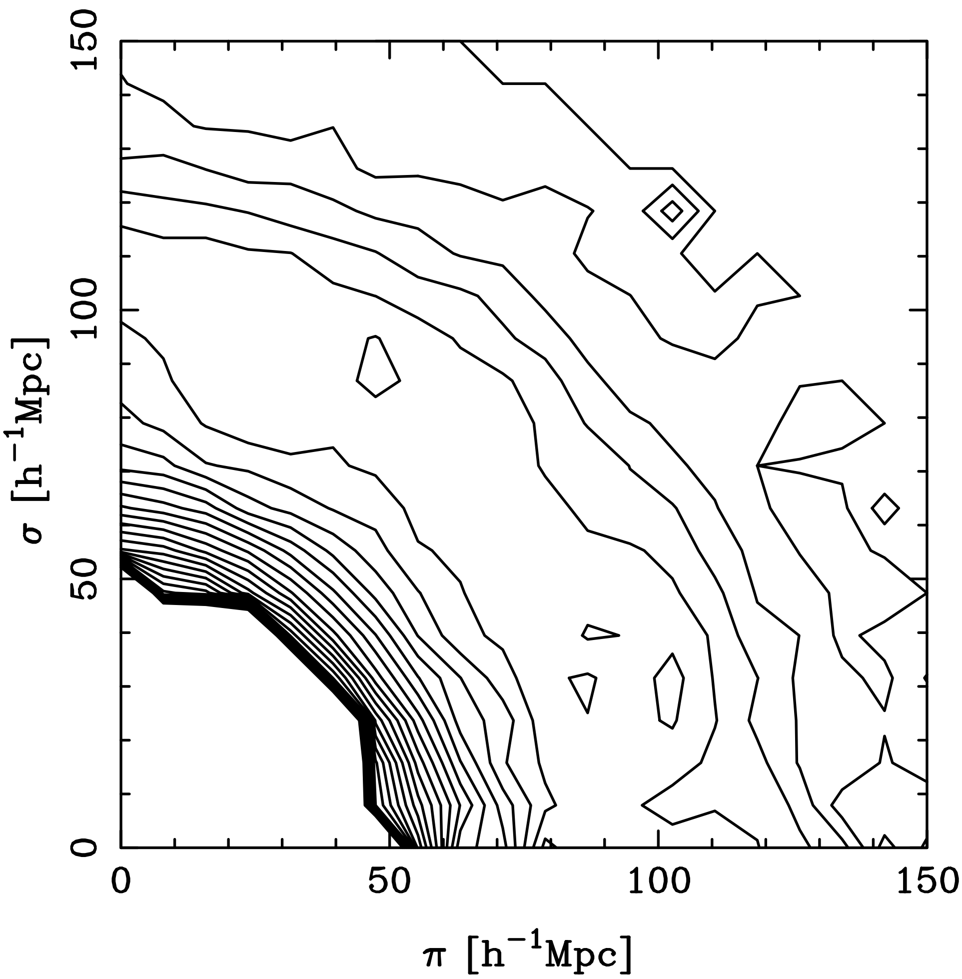

Figure 2.3: Split 2PCF $\xi(\sigma, \pi)$ calculated on a grid using the MICE simulations from the DES LSS WG Simulation Challenge 1. The grid based technique has a lower separation limit of $\sim 5d_c = 50\,h^{-1}\,\mathrm{Mpc}$. The baryonic ridge can be clearly detected both in the radial and tangential direction at $\sim 100\,h\,\mathrm{Mpc}^{-1}$.

.

assumptions are strongly related. The number of sub-samples that we use is dependent on what we want to find out and the form of the underlying probability distribution that we are probing. However, care needs to be taken that the sub-samples are not so small that they become strongly correlated.

Sub-sampling

The simple sub-sampling method consists of splitting the data-set into N independent samples and estimating the covariance matrix using the equation

$$C(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N}(x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j), \tag{2.50}$$

where the clustering statistic is estimated for each sub-sample separately. The assumption here is that the mean expectation value $\bar{x}_i$ is not estimated from the $\{x_i^k\}_{k=1}^{N}$ samples, but from an independent realisation of the data. For N independent sub-samples, this returns the correct covariance for a sample of volume $1/N$ of the original volume.

Jackknife Resampling

The jackknife resampling method consists of splitting the data-set into N sub-volumes and then systematically omitting each one in turn (Shao, 1986). The resampling of the data-set consists of the $N-1$ remaining sub-volumes, with volume $(N-1)/N$ times the volume of the original data-set. The clustering measurement is then repeated on the resampling of the original data-set. The covariance matrix for N jackknife resamplings is estimated using

$$C_{jk}(x_i, x_j) = \frac{(N-1)}{N}\sum_{k=1}^{N}(x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j), \tag{2.51}$$

where x_i is the i^{th} measure of the statistic of interest, and it is assumed that the mean expectation value is given by

$$\bar{x}_i = \sum_{k=1}^{N}\frac{x_i^k}{N}. \tag{2.52}$$

The factor of $N-1$ that appears in Eq. 2.51 (Tukey, 1958; Miller, 1974) accommodates the lack of independence between the N resamplings of the data, since only 2 sub-volumes are different from one resampling to the next.

Bootstrap Resampling

The bootstrap resampling method consists of selecting N random sub-samples, with replacement, from the original data-set (Efron, 1979). Each sub-volume in the original data-set has equal weight. As the data-set is resampled, a new weight is generated for each sub-volume corresponding to the number of times that sub-volume has been selected. The clustering measurement is then repeated for each resampled data-set. For a given N, the mean fractional effective volume of the resampled data-sets tends to a fixed fraction of the original sample volume. Therefore, for $N_{bootstrap} = N_{jackknife}$, the mean effective volume is less than the volume of each of the jackknife resamples.

The covariance matrix for N bootstrap resamplings is estimated by using

$$C_{boot}(x_i, x_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j), \tag{2.53}$$

where it is assumed that the mean expectation value is given by Eq. 2.52. There is no $N-1$ factor here as there is for the jackknife method. This is because the resamplings are considered to be more independent for the bootstrap method.

Advantages and Disadvantages for each Method

Sub-sampling is considered one of the easiest error estimation techniques to implement. However, in the context of galaxy clustering studies the sub-samples are never fully independent of each other, due to the presence of long-range modes in the density fluctuations. This means that sub-samples are always correlated with each other to some extent, which violates one of the fundamental basic assumptions in this approach.

Both the jackknife and bootstrap resampling techniques account for the lack of independence between sub-volumes, with the $N-1$ term in Eq. 2.51 and the randomisation of resampling, respectively. The bootstrap technique, in principle, has no limit on the number of N resamplings that it can use for error estimates. In practice however, the rate of convergence of the variance on a measurement is relatively slow for an increasing number of trials, and the computational cost of analysing resamplings increases dramatically with N.

Each of the error estimation techniques that we have discussed here are calculated directly from the data-set that we are analysing. Consequently, all systematics and biases within the data-set are accounted for, which is particularly important for clustering analyses since the errors on the 2-point correlation function depend on the higher order clustering of the data. However, internal error estimates are generally severely limited by the size of the data-set. To avoid such cosmic-variance limited sampling, we can consider using an "external" error estimator via Monte Carlo realisations. This method consists of creating N statistically equivalent versions of the data-set under analysis, and conducting a full analysis on each. However, this method requires the user to define the exact statistics that need to be included in the Monte Carlo realisation, and so misses all systematics and biases that may be present in the original data-set.

One of the main focusses of this thesis is to understand the systematic effects arising in clustering analysis techniques for future photometric redshift surveys. As a consequence of this, I will be using the internal jackknife resampling method to estimate errors on my clustering analyses.

Chapter 3

Redshift-Space Distortions & Binning Techniques

Redshift-space distortions can alter the angular clustering in a redshift slice because the distortions are correlated across the direction of projection. Although redshift-space distortions are sub-dominant compared with photometric redshift uncertainties, they give rise to a systematic effect, which needs to be included when photometric redshift surveys are analysed (Padmanabhan et al., 2007; Blake et al., 2007). This can complicate the analysis as the size of the redshift-space distortions, and therefore of this effect, is dependent on the cosmological model. Consequently, for every model to be tested against the data, we need to make a revised estimate of the redshift-space effect, thus significantly complicating an analysis.

3.1 Projected 2-pt Statistics of the Overdensity Field

3.1.1 Correlation Function

In order to simplify the problem, we assume that the clustering strength does not change across the samples under consideration and make the plane-parallel (distant observer) approximation, with redshift-space distortions along the z-axis of a Cartesian basis. In the absence of redshift distortions the projected correlation function is given by:

$$\xi_p(d_p) = \langle \delta_p(\mathbf{r_p}) \delta_p(\mathbf{r'_p}) \rangle, \tag{3.1}$$

$$= \int \int dr_z dr'_z \phi(r_z) \phi(r'_z) \xi \left[d(r_z, r'_z, d_p) \right] \tag{3.2}$$

where $d(r_z, r'_z, d_p) = \sqrt{(r_z - r'_z)^2 + d_p^2}$, subscripts x, y and z denote the direction along each Cartesian axis, and p denotes projected quantities $p \equiv xy$. $\phi(r_z)$ is the radial galaxy selection function, normalised such that $\int dr_z\ \phi(r_z) = 1$, and $\xi(d) = \langle\delta(\mathbf{r})\delta(\mathbf{r}+\mathbf{r}')\rangle$, where $\delta(\mathbf{r})$ is the overdensity of galaxies at real-space position $\mathbf{r}$. Throughout our analysis we use $\mathbf{r}$ to describe a galaxy position and $\mathbf{d}$ to describe the distance between two galaxies, so, for example, r_z is the position of a galaxy along the z-axis, while d_p is the amplitude of the separation between two galaxies when projected into the x, y-plane.

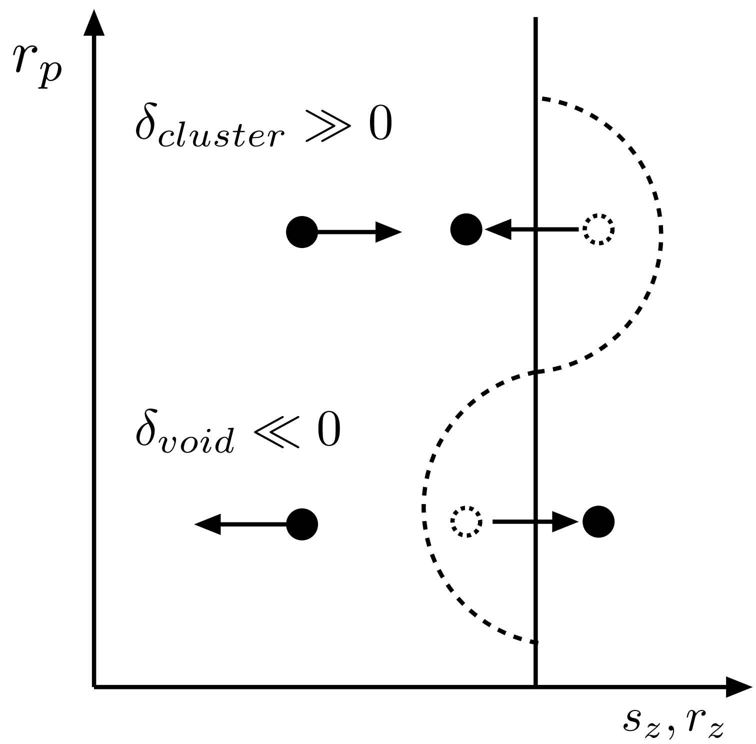

Figure 3.1: Schematic showing the boundary of a region selected in redshift-space (solid line) compared with the boundary of the same region in real-space (dashed line). The boundary is distorted in real-space around an overdensity and an underdensity. The positions of two galaxies whose apparent motion crosses the boundary are shown in redshift-space (solid circles) and in real-space (dashed circles). Note that, in this simplified picture where the under and over-densities have the same amplitude, the galaxy pair lost and the galaxy pair gained would contribute the same amount to the 3D real-space correlation function, following the dashed boundary. However, the projected clustering is different because we do not know the shape of the dashed line, and instead assume that the projection length is the same for all r_p. It is the 2D clustering strength of the boundary that is important, rather than the loss or gain of particular galaxy pairs.

In reality, our radial position is determined via a redshift. In this case, Eq. (3.2) must be altered to

$$\xi_p^s(d_p) = \langle\delta_p(\mathbf{s_p})\delta_p(\mathbf{s'_p})\rangle. \tag{3.3}$$

The weighted, projected overdensity field $\delta_p(r_p)$ can now be written

$$1 + \delta_p(\mathbf{r_p}) = \int ds_z \; \phi(s_z)[1 + \delta(\mathbf{s})], \tag{3.4}$$

where $\mathbf{s} = (\mathbf{r_p}, s_z)$ is the redshift-space position of each galaxy and $\phi(s_z)$ gives the galaxy selection function along the line of sight corresponding to s_z (e.g. Peebles 1980).

The difference between the projection in redshift-space and real-space is shown schematically in Fig. 3.1. An edge to a window function (or a contour of constant galaxy density) that is straight in redshift-space is systematically distorted in real-space. The edge of the bin is itself clustered with a non-negligible projected correlation function, i.e. the real-space boundary has a correlation function that depends on $\mathbf{r_p}$. The inclusion or exclusion of galaxies is balanced in terms of the 3D correlation function within the boundary; while we lose voids, we gain clusters and these give the same clustering signal. However, we assume that the projected field has a constant projection length, and this implies that the underdensity of the void will become larger (since we include less of the galaxies) and the overdensity of the cluster becomes larger (since we will include more of its galaxies). Thus the overall clustering signal becomes stronger.

By construction, $n_s(\mathbf{s})d^3\mathbf{s} = n_r(\mathbf{r})d^3\mathbf{r}$, where $n_s(\mathbf{s})$ is the redshift-space number density and $n_r(\mathbf{r})$ is the real-space number density. Therefore, if the perturbations induced in the density field by peculiar velocities are small compared to the volume of the field, we can treat the apparent shift in galaxy positions caused by moving from real to redshift space $(s_z - r_z)$ via a Taylor expansion of the selection function (Fisher et al., 1994a). To first order this gives

$$\phi(s_z) = \phi(r_z) + \frac{d\phi(r_z)}{dr_z}(s_z - r_z). \tag{3.5}$$

We consider this to be an Eulerian picture as it is based on apparent galaxy motions. We can write

$$\delta_p(\mathbf{r_p}) = \int dr_z \left[\phi(r_z)\delta(\mathbf{r}) + (s_z - r_z)\frac{\partial\phi(r_z)}{\partial r_z}\right] \tag{3.6}$$

to first order in $\delta(\mathbf{r})$. Following linear theory (see Appendix. B), $(s_z - r_z)$ can be written as a function of the overdensity field,

$$(s_z - r_z) = -\beta\frac{\partial}{\partial r_z}\nabla^{-2}\delta(\mathbf{r}), \tag{3.7}$$

where $\beta \equiv f/b$, with f being the logarithmic derivative of the linear growth rate with respect to the logarithm of the scale factor, and b the galaxy bias. We therefore have that

$$\delta_p(\mathbf{r_p}) = \int dr_z \left[\phi(r_z) - \beta \frac{\partial \phi(r_z)}{\partial r_z} \frac{\partial}{\partial r_z} \nabla^{-2}\right] \delta(\mathbf{r}). \tag{3.8}$$

If we think of $\phi(s_z)$ as setting up boundaries in s_z, then substituting Eq. (3.8) into Eq. (3.3) shows that we can expect coherent apparent galaxy motion across these boundaries. Correlations between galaxies moved into the sample by the redshift-space distortions, and those already within the sample, give rise to cross terms from the two terms in Eq. (3.8). The second term in Eq. (3.8) also adds a component to the projected correlation function from the coherence of the velocities at different points on the boundary. We see that, even with constant $\phi(s_z)$ within a fixed interval, redshift-space distortions can still affect the correlation function of the volume within the sample due to the motion of galaxies across the boundary.

In addition to the Eulerian picture given by Eq. (3.8), we can also consider a Lagrangian picture based on the redshift-space overdensity field that we wish to project. Following this equivalent picture, we can work directly with redshift-space overdensities using Eq. (3.3).

$$\xi_p^s(d_p) = \int \int ds_z ds_z' \phi(s_z)\phi(s_z')\xi^s \left[d(s_z, s_z', d_p)\right]. \tag{3.9}$$

In the plane-parallel approximation, we can use the redshift-space correlation function of equation 5 of Hamilton (1992) as input into the projection equation.

$$\xi^s(\mathbf{d}) = \xi_0(d)P_0(\mu) + \xi_2(d)P_2(\mu) + \xi_4(d)P_4(\mu), \tag{3.10}$$

where

$$\xi_0(d) = (b^2 + \frac{2}{3}bf + \frac{1}{5}f^2)\xi(d), \tag{3.11}$$

$$\xi_2(d) = (\frac{4}{3}bf + \frac{4}{7}f^2)[\xi(d) - \xi'(d)], \tag{3.12}$$

$$\xi_4(d) = \frac{8}{35}f^2[\xi(d) + \frac{5}{2}\xi'(d) - \frac{7}{2}\xi''(d)], \tag{3.13}$$

P_i are the standard Legendre polynomials, and

$$\xi'(d) \equiv 3d^{-3}\int_0^d \xi(d')(d')^2 dd', \tag{3.14}$$

$$\xi''(d) \equiv 5d^{-5}\int_0^d \xi(d')(d')^4 dd'. \tag{3.15}$$

b is the large-scale bias of the galaxy population being considered, f is the standard dimensionless linear growth rate, ξ is the 3-dimensional real-space correlation function, and μ is the cosine of the angle between the separation along the line of sight and the transverse separation, $\mu \equiv |s_z - s'_z|/d$.

3.1.2 Modelling ξ_p using Eulerian and Lagrangian Frameworks

In the Eulerian picture, the effect of redshift-space distortions on projected angular clustering is caused by the apparent movement of galaxies into and out of the sample. Eqns. (3.2) & (3.8) describe the redshift-space density field within the window $\phi(s_z)$. The problem with applying this is that we measure galaxy positions, rather than observing the overdensity field directly. Galaxy motions, which cause the difference between real- and redshift-space density fields, will also move galaxies into and out of the window. In general, distortions that decrease pair separation will tend to bring galaxies into a sample. Motion to increase pair-separation will tend to move galaxies out of a given sample. If we use a method such as counting pairs of galaxies, we do not observe the full field within the window. Note that counts-in-cells techniques, which evenly weight by volume element would trace the overdensity field as described in Section 3.1.

We can try to consider the effect of redshift-space distortions based on predicting galaxy motions (e.g. Regos & Szalay 1995), since we know that galaxy motion is correlated with overdensity. For example, if we denote galaxies in a pair with subscripts 1 and 2 then:

- If $\delta_1 >> 0$, the velocity vector v_2 will generally be orientated towards galaxy 1; galaxies tend to fall towards overdensities.
- If $\delta_1 << 0$, v_2 will generally be directed away from galaxy 1; galaxies tend to move away from underdensities.

Fig. 3.2 shows how the overdensity at one of the galaxies δ_1 affects the change in pair separation as we move from real to redshift-space. This was calculated from a Monte-Carlo realisation of 10^8 pairs of points with a randomly selected projected separation of up to $150\,h^{-1}\,\mathrm{Mpc}$ within top-hat redshift-space windows of size $50\,h^{-1}\,\mathrm{Mpc}$ and $100\,h^{-1}\,\mathrm{Mpc}$ along the line-of-sight direction in a ΛCDM (flat, $\Omega_m = 0.3$) Gaussian random density field. The correlated distribution of overdensities and velocities of the two points were calculated using the formulae of Regos & Szalay (1995), who show how a 26×26 covariance matrix can be constructed for a multi-variate gaussian distribution

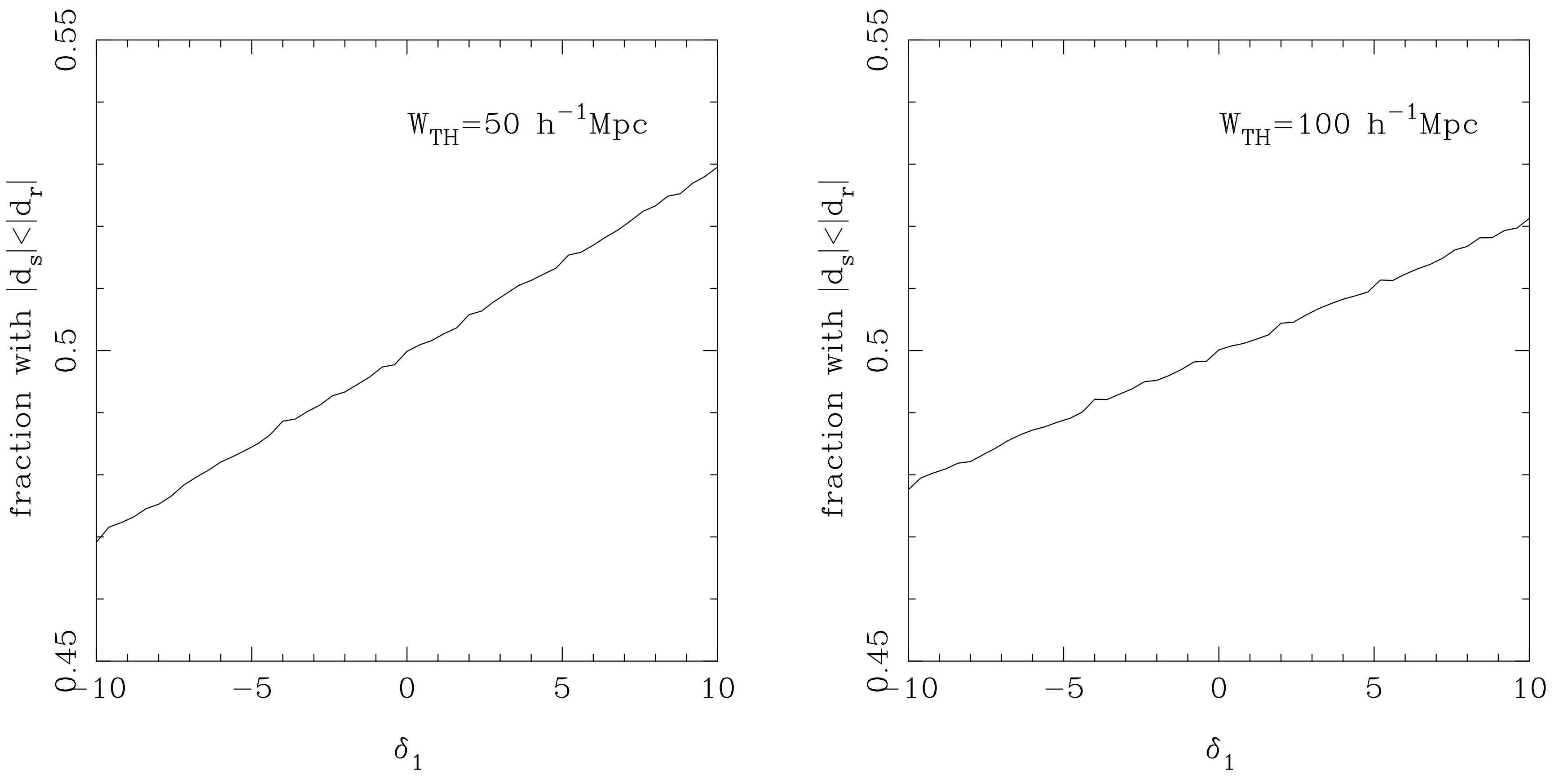

Figure 3.2: The fraction of galaxy pairs with randomly selected projected separations up to $150\,h^{-1}\,\mathrm{Mpc}$ whose redshift-space radial separation $|d_s|$ is less than that in real space $|d_r|$, for top-hat projection windows of width 50 $h^{-1}\,\mathrm{Mpc}$ (left) and 100 $h^{-1}\,\mathrm{Mpc}$ (right). Although, the redshift-space apparent galaxy motion in high and low density regions is opposite, the number of galaxies moving into and out of the sample is the same, ie. galaxies close to higher densities are preferentially moved into the window whereas galaxies in low-density regions move out. This can set up a balance where we lose and gain pairs that have the same correlation function. We see that the apparent redshift-space galaxy motion is reduced for the larger projection window.

for the properties of pairs of points in a smoothed Gaussian random field. The matrix depends upon the power spectrum moments given by

$$\gamma_\nu = \frac{\sigma_0^2}{\sigma_{-1}\sigma_1}, \quad \text{and} \quad \sigma_j^2 = 4\pi \int_0^\infty dk\ k^{2j+2} P(k) \tag{3.16}$$

and the functions of the pair separation r given by

$$K_{lm} = 4\pi \int_0^\infty dk\ k^m j_l(kr) \bar{P}(k). \tag{3.17}$$

These covariance matrices can be used to draw random realisations of pair properties of points in the field, assuming they follow a multi-variate Gaussian distribution, as described by Percival & Schaefer (2008).

The redshift-space apparent motion of galaxies in high and low density regions are opposite, although the real- to redshift-space increase in the correlation function is the same. Galaxies close to higher densities are preferentially moved into the window but galaxies in low-density regions move out. This can set up a balance where we lose and gain pairs that have the same correlation function. However if, for example, galaxies only form at peaks in the density field, and no galaxies move out of a window because they are in voids, then the peculiar velocities reduce the average separation of galaxy pairs, leading to an enhancement of the correlation function. This suggests that the effect of redshift-space distortions on projected clustering measurements depends strongly on the way in which galaxies sample the density field. Fig. 3.3 shows the expected real- and redshift-space projected correlations functions, calculated for the ΛCDM model, with a top-hat redshift-space window of size $50\,h^{-1}\,\mathrm{Mpc}$ (left) and $100\,h^{-1}\,\mathrm{Mpc}$ (right) along the line-of-sight direction. The solid red triangles show the expected redshift-space projected correlation function if galaxies sample the volume uniformly (i.e. in the linear limit $|\delta| << 1$). The solid blue squares show the expected redshift-space projected correlation function if we only select pairs where $\delta_1 > 0$; here we pick up the enhancement due to overdensities, but not the reduction in pair counts caused by regions with $\delta_1 < 0$ and $\delta_2 < 0$. In this case, the redshift-space projected correlation function is significantly enhanced compared with that where the galaxies uniformly sample in real-space.

Fig. 3.3 clearly shows that the way in which the galaxies trace the overdensity field is critical in determining the effect of redshift-space distortions on projected clustering measurements. Simply calculating the correlations between overdensity and velocity fields (e.g. Regos & Szalay 1995) is not sufficient. In order to utilise this approach, we would need to correlate multiple points on the boundary and internal locations within the bin,

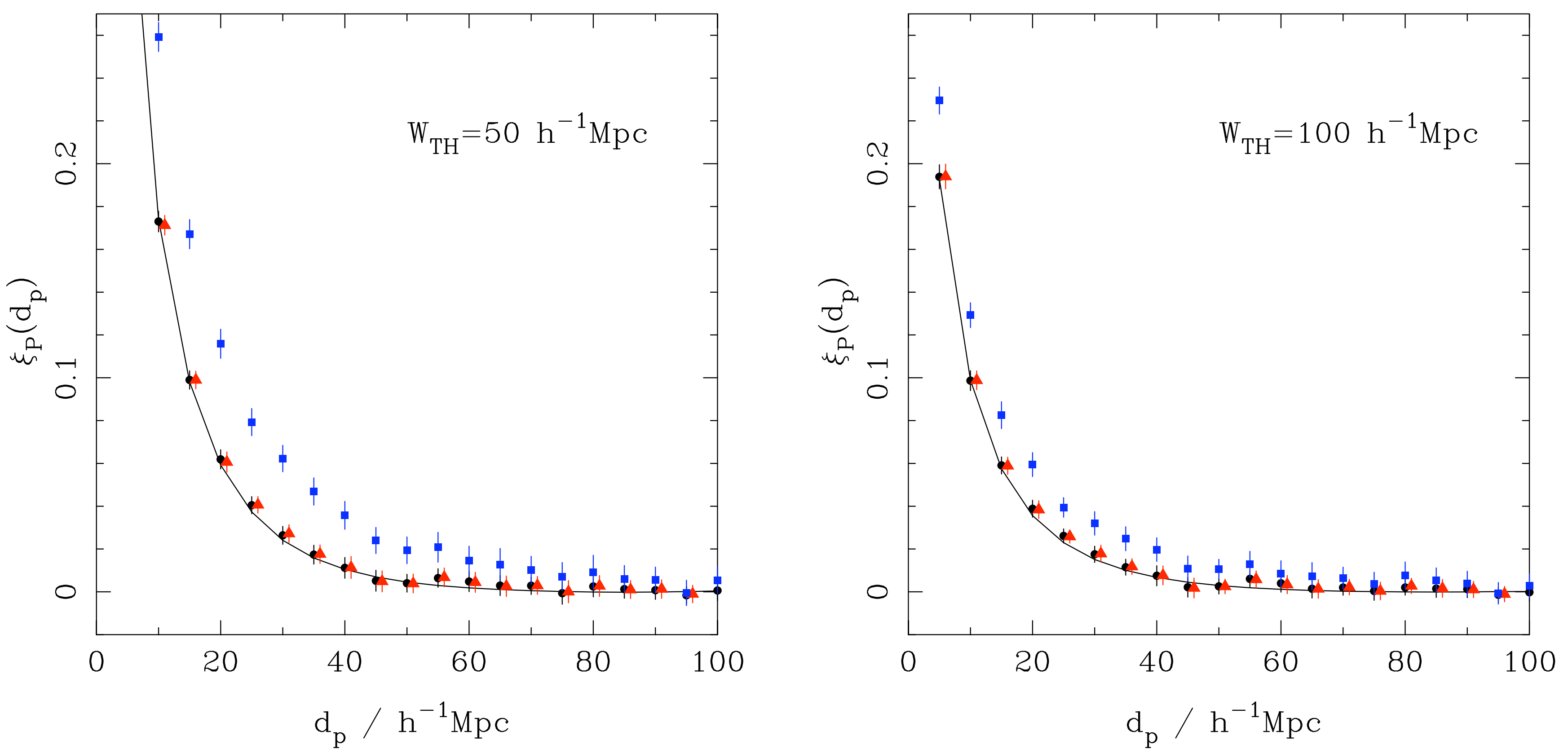

Figure 3.3: Projected correlation functions expected for top-hat windows of width $50\,h^{-1}\,\mathrm{Mpc}$ (left) and $100\,h^{-1}\,\mathrm{Mpc}$ (right) along the line-of-sight. These were calculated from a Monte-Carlo distribution of 5×10^8 pairs of points in a Gaussian random field based on a ΛCDM cosmology (see text for details). The solid line was calculated from the expected 3D correlation function for the pairs, using the pairs only to calculate the projection (effectively working as points in a Monte-Carlo integration of the true 3D correlation function). The projected correlation function calculated using both the overdensities $\delta_1\delta_2$, and positions of the pairs, is also plotted (black solid circles with 3-σ, 99% error bars). The redshift-space correlation function calculated in the same way (red solid triangles), offset in d_p for clarity, and the correlation function if we only accept pairs of points with $\delta_1 > 0$ (blue solid squares) are also shown. The effect of redshift-space distortions on the projected correlation function is much greater for the narrower projection window.

which is very difficult to do in practice. In this analysis we only consider a uniform sampling of the density field, where such effects are negligible.

As mentioned in Section 3.1.1, the Lagrangian picture, based on the redshift-space overdensity field that we wish to project, allows us to work directly with redshift-space overdensities. One strong advantage of the Lagrangian framework is that it is straightforward to determine the projected correlation function, even when the galaxy selection function is discontinuous. This allows simple comparison between the results one expects to obtain with and without redshift-space distortions. Therefore, we will use this approach throughout this analysis.

3.1.3 The Limber Approximation

For pairs of galaxies, we can define the mean $m_z \equiv (r_z + r_z')/2$ and separation $d_z \equiv r_z - r_z'$ along the z-axis. For a survey whose depth is larger than the correlation length, and with a slowly varying selection function, so that $\phi(r_z) \simeq \phi(r_z') \simeq \phi(m_z)$, Eq. 3.2 reduces to the Limber equation in real-space ($s_z - r_z = 0$)

$$\xi_p(d_p) = \int_{-\infty}^{+\infty} dm_z\, \phi^2(m_z) \int_{-\infty}^{+\infty} dd_z\, \xi\left(\sqrt{d_p^2 + d_z^2}\right). \tag{3.18}$$

We see that, for the Limber approximation, ϕ is a function of m_z alone, and the integrals over dm_z and dd_z in Eq. (3.19) are separable. In redshift-space, a similar reduction of Eq. (3.9) gives

$$\xi_p^s(d_p) = \int_{-\infty}^{+\infty} dm_z \int_{-\infty}^{+\infty} dd_z \left[\phi(m_z) - \beta\frac{\partial\phi(m_z)}{\partial m_z}\frac{\partial}{\partial r_z}\nabla^{-2}\right]^2$$
$$\times\, \xi\left(\sqrt{d_p^2 + d_z^2}\right), \tag{3.19}$$

if we expand redshift-space distortions in $(s_z - r_z)$, or

$$\xi_p^s(d_p) = \int_{-\infty}^{+\infty} dm_z\, \phi^2(m_z) \int_{-\infty}^{+\infty} dd_z\, \xi^s\left(\sqrt{d_p^2 + d_z^2}\right), \tag{3.20}$$

in the Lagrangian picture. Because no galaxies are lost or gained moving from real-space to redshift-space, the result of the integral over d_z is the same in real or redshift space, so we see that in this approximation there are no redshift-space effects. But, as we show later, this picture is too simplistic to be applied to the analysis of future data sets.

3.1.4 Power Spectrum

In Padmanabhan et al. (2007), the projection of the 2-pt clustering was analysed through the power spectrum. We now consider such an approach in the plane-parallel approximation and for a Cartesian basis. Taking the Fourier transform of $\delta(\mathbf{s})$ in Eq. (3.4) gives

$$\delta_p(\mathbf{r_p}) = \int ds_z\ \phi(s_z) \int \frac{d^3k}{(2\pi)^3} \delta(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{s}}. \tag{3.21}$$

We now define a window function

$$W(k_z) = \int ds_z\ \phi(s_z) e^{-ik_z s_z}, \tag{3.22}$$

and use statistical isotropy and homogeneity within the definition of the power spectrum $\langle\hat{\delta}(\mathbf{k})\hat{\delta}^*(\mathbf{k}')\rangle = P(k)\delta_D(\mathbf{k}-\mathbf{k}')$, where δ_D is the Dirac delta function. We assume that the power spectrum does not evolve over the volume covered by the window[1]. Taking the 2-point function of the projected overdensity (Eq. 3.21) gives

$$\xi_p(d_p) = \langle\hat{\delta}_p(\mathbf{r_p})\hat{\delta}(\mathbf{r_p'})\rangle \tag{3.23}$$

$$= \int \frac{dk^3}{(2\pi)^3} W^2(k_z) P(\mathbf{k}) e^{-i\mathbf{k_p}\cdot(\mathbf{r_p}-\mathbf{r_p'})} \tag{3.24}$$

The projected overdensity can be written in terms of a 2D power spectrum $P_p(k_p)$,

$$\xi_p(d_p) = \int \frac{dk_x\, dk_y}{(2\pi)^2} P_p(k_p) e^{-i\mathbf{k_p}\cdot(\mathbf{r_p}-\mathbf{r_p'})}. \tag{3.25}$$

If we compare Eqns. (3.24) & (3.25), we see that

$$P_p(k_p) = \int \frac{dk_z}{(2\pi)} W(k_z)^2 P\left(\sqrt{k_p^2+k_z^2}\right). \tag{3.26}$$

Note that the power $P(\mathbf{k})$ depends on the amplitude of the full 3-dimensional wavevector, and so is dependent on k_p.

Using Eq. (3.5) to include redshift-space distortions, the window $W(k_z)$ has an extra term,

$$W(k_z) = \int dr_z \left[\phi(r_z) + (s_z - r_z)\frac{d\phi(r_z)}{dr_z}\right] e^{-ik_z r_z}. \tag{3.27}$$

[1]This is true if analysing a single time slice from a simulation.

In Fourier space, $(s_z - r_z) = -\beta(k_z^2/k^2)\delta(\mathbf{r})$, so we can expand $\delta(\mathbf{s})$ to 1st order in $\delta(\mathbf{r})$, leaving a new window function for Eq. 3.26

$$W(k_z) = \int dr_z \left[\phi(r_z) - \beta\left(\frac{k_z}{k}\right)^2 \frac{d\phi(r_z)}{dr_z}\right] e^{-ik_z r_z}. \tag{3.28}$$

If we drop the plane-parallel approximation and expand in Spherical Harmonics, the standard result (Peebles, 1973) is

$$\langle |a_{lm}|^2 \rangle = \frac{1}{2\pi^2} \int dk\; k^2 P(k) W^2(k), \tag{3.29}$$

where

$$W(k) = \int dr\; \phi(r) j_l(kr) + \frac{\beta}{k}\frac{d\phi(r)}{dr} j_l'(kr), \tag{3.30}$$

(Fisher et al., 1994a). Here the l dependence is contained within $W(k)$, while in Eq. (3.26), it was the power that depended on k_p. Eq. (3.26) could have been rewritten by changing the variable of the convolution integral k to match.

3.1.5 Monte-Carlo Simulations of the Projection Effect

In order to test the projection formulae presented in Sections 3.1.1 & 3.1.4, ie. without redshift-space distortions, we have used Monte-Carlo realisations of δ-function real-space correlation functions in a similar vein to that of Simpson et al. (2009). We work in a plane parallel approximation throughout and construct a real-space 3D δ-function correlation function at an arbitrary location d_0 such that

$$\xi(d) = \delta_D(d - d_0)\, \xi_0, \tag{3.31}$$

where δ_D is the standard Dirac delta function. We do this by introducing a pre-determined excess of data pairs at the location d_0. The number of excess pairs we introduce depends on the value of $\xi(d_0)$ we require and is determined using the natural estimator $\xi = DD/RR - 1$. For example, if we have a uniform distribution of data and random pairs with 100,000 pairs per bin of separation, we would require an excess of 10,000 data pairs at the location d_0 for $\xi(d_0) = 0.1$. In doing this we create an unnormalised 3D δ-function correlation function.

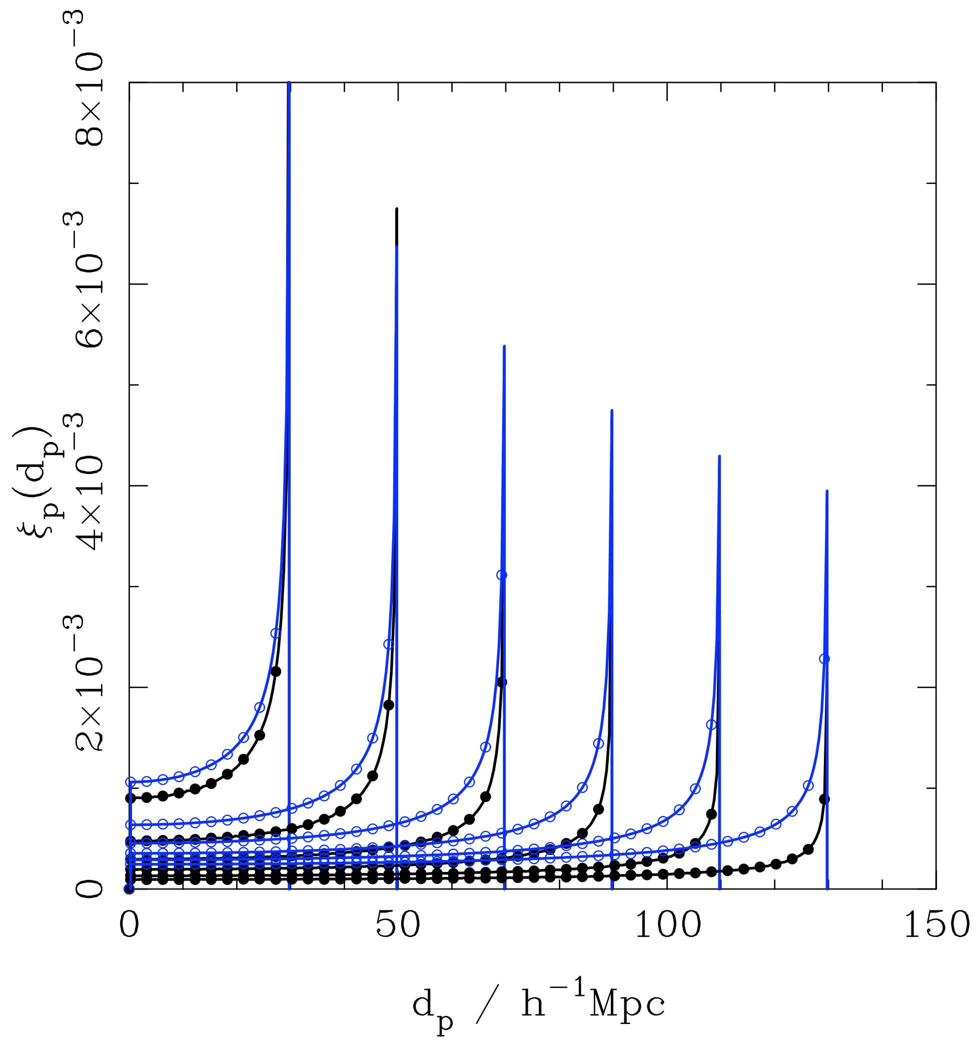

Figure 3.4: Projected correlation functions calculated for 3-dimensional δ-function correlation functions at locations $d_0 = 30, 50, 70, 90, 110$ and $130\,h^{-1}\,\mathrm{Mpc}$, with no radial window (solid symbols) and with a top-hat window in radial distribution of width $100\,h^{-1}\,\mathrm{Mpc}$ (open symbols). Models calculated using Eq. (3.33) are shown by the solid lines.

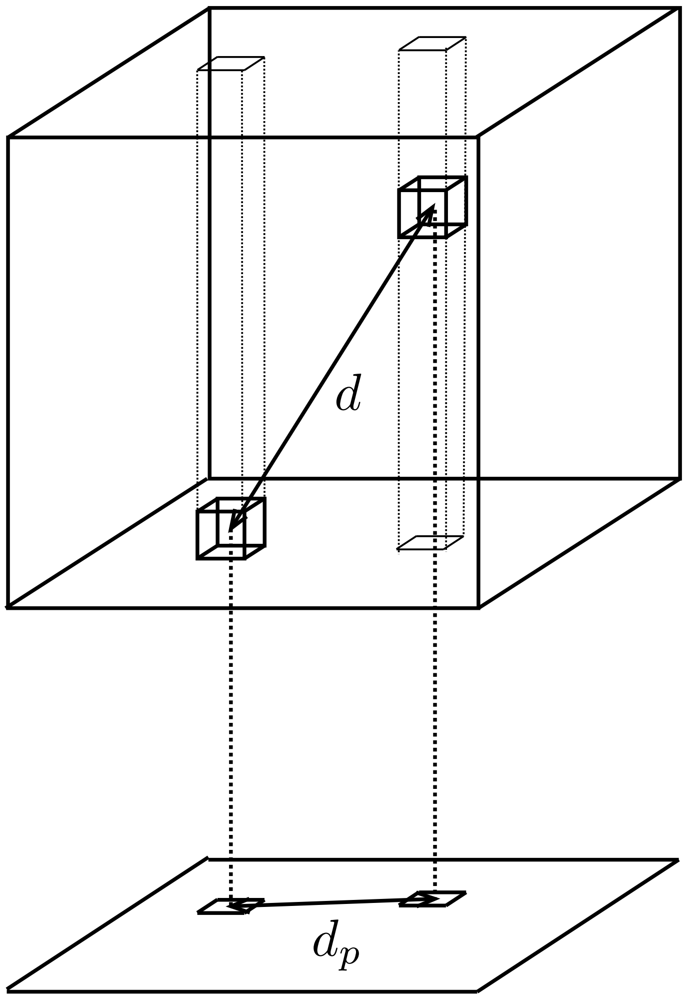

Figure 3.5: Schematic showing how the clustering signal from a range of 3D separations are projected to a single separation in 2D. This means that potentially uncorrelated pairs in 3D will contribute signal to the projected correlation function.

Changing the variables in the inner integral of Eq. (3.18) to be a function of 3D pair separation d gives

$$\xi_P(d_{xy}) = \int \int_V dm_z \, dd \;\; \phi^2(m_z) \, \frac{2\xi(d)\, d}{\sqrt{d^2 - d_{xy}^2}}, \tag{3.32}$$

and is simplified for the δ-function case such that

$$\xi_P(d_{xy}) = \frac{2}{\pi d_0} \int dm_z \; \phi^2(m_z) \, \xi_0 \frac{d_0}{\sqrt{d_0^2 - d_{xy}^2}} \tag{3.33}$$

The factor $1/\pi d_0$ accounts for the fact that the δ-function real-space correlation function was unnormalised.

By introducing a radial window, we are preferentially selecting pairs of galaxies from the sample. A further volume reduction normalisation is required in Eq. (3.32) to account for this. The excess probability of finding two galaxies in areas δA_1 and δA_2 with a 2D projected separation d_{xy} is the sum of all the probabilities of finding two galaxies in volumes δV_i and δV_j along the radial axis at *all* 3D separations d. That is,

$$1 + \xi_P(d_{xy}) = \frac{\bar{n_V}^2}{\bar{n_A}^2} \frac{1}{\delta A_1 \delta A_2} \left(\sum_i \sum_j [1 + \xi(d_{ij})] \delta V_i \delta V_j \right) \tag{3.34}$$

In Fig. 3.4 we show the clustering expected for projections of density fields created from δ_D-function 3D correlation functions in the case where there is no window function (solid symbols) and for a window function of width $100\,h^{-1}\,\mathrm{Mpc}$ (open symbols). The excess of pairs that exists at a single scale in 3D is projected onto a range of scales, up to and including this scale, in 2D. There is a damping of power on all scales in each case, which increases as we project from larger scales. This effect depends upon the window size. When we consider wide projection windows we allow wide 3D separation pairs, that are most likely uncorrelated, to contribute to the projected clustering signal. This means that the projected correlation function will become negative at increasingly smaller scales, and will approach ξ at much larger scales, for wider projection windows (see Fig. 3.5). As we move to smaller projection windows the inclusion of uncorrelated, wide-separation pairs is decreased and $\xi_{2D} \to \xi_{3D}$. The projection of a more general density field, where there is clustering on a range of scales, can be considered as the linear combination of the projections of a series of δ_D-function 3D correlation functions. The trends observed in this analysis will help us to interpret the behaviour of the projected correlation function in the more general situation analysed in later sections.

3.1.6 Binning Galaxy Samples

Future surveys will automatically have a standard selection function caused by the changing cosmological volume, the number density of galaxies as a function of redshift, and selection effects such as a magnitude limit below which we cannot observe galaxies or obtain accurate photometric redshifts. In addition to this distribution we will wish to bin galaxies based on their photometric redshifts in order to analyse the evolution of galaxy properties and/or cosmology across the sample. We now consider how the way in which this sub-division is applied affects the importance of redshift-space distortions.

One simple approach would be to bin galaxy positions in redshift, equivalent to a **top-hat binning**. Such galaxy selection means that galaxy pairs, where galaxies lie in different bins, are not included in the estimate of the correlation function. This exclusion of pairs leads to the observed difference between the projected real-space and redshift-space correlation function, as described in Section 3.1.1. An alternative to this approach, considered here, would be to bin galaxy pairs rather than individual galaxies.

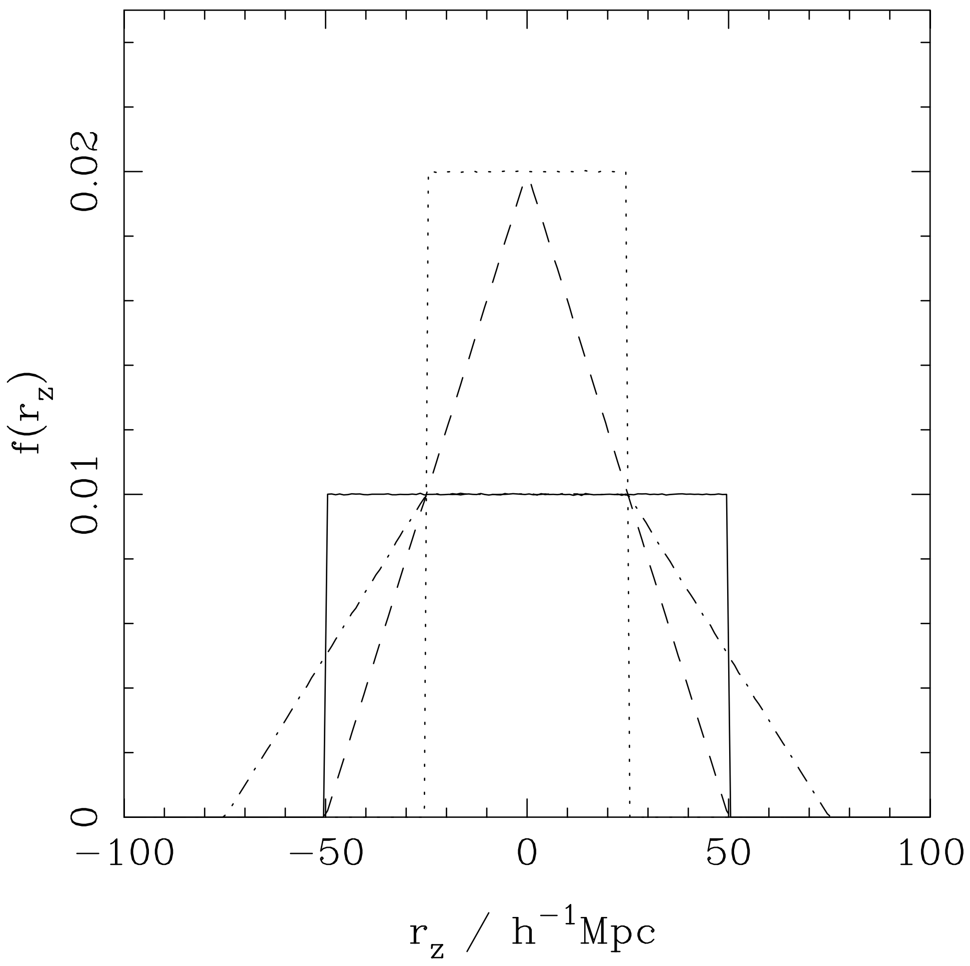

Figure 3.6: The normalised radial distribution of galaxies (solid line) and pair-centres (dashed line) for the distribution of galaxies in a top-hat bin of width $100\,h^{-1}\,\mathrm{Mpc}$. These are compared with the distributions of galaxies (dot-dash line) and pair centres (dotted line) for galaxies whose pair-centre is within a $50\,h^{-1}\,\mathrm{Mpc}$ bin, and with $d_z < 100\,h^{-1}\,\mathrm{Mpc}$.

A simple argument shows that in an ideal situation, applying a binning based on the centre of galaxy pairs in the radial direction, which hereafter we refer to as **pair-centre binning**, can completely remove the effect of redshift-space distortions while retaining information about the evolution of the correlation function. A schematic representation

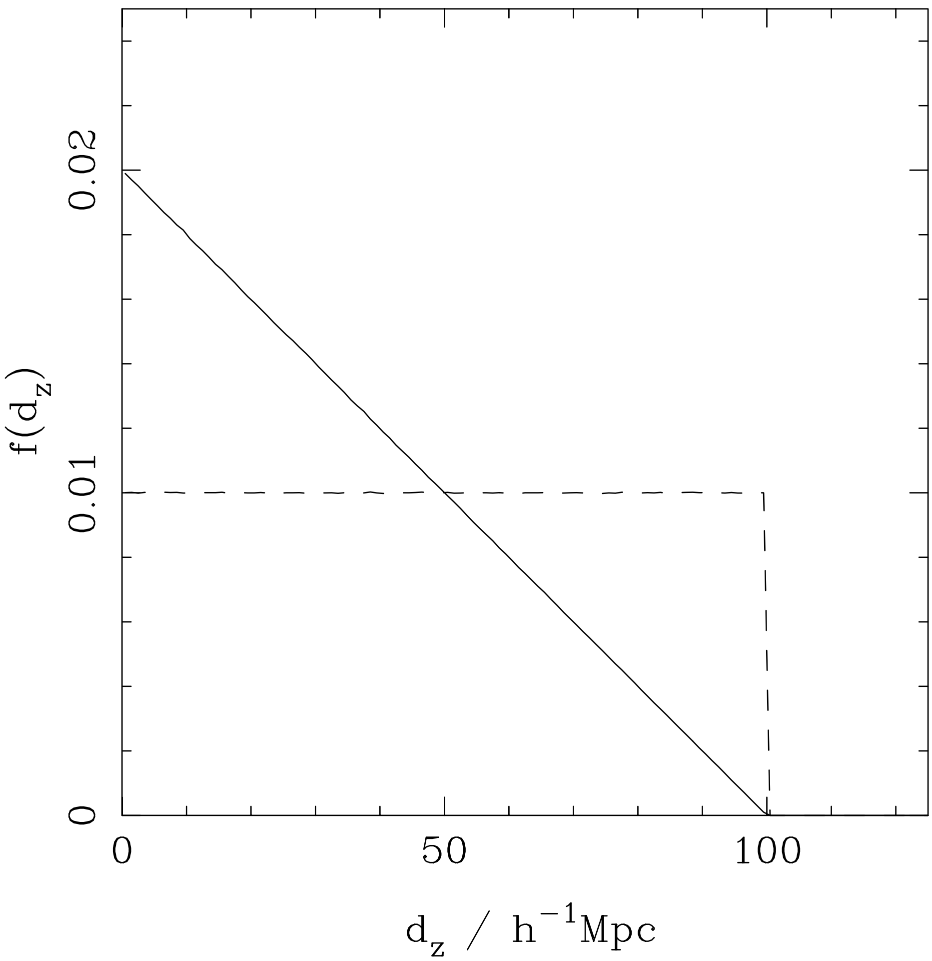

Figure 3.7: Comparison of the radial pair separations (d_z), between top-hat (solid line) and pair-centre (dashed line) binning.

of this binning scheme is shown in Fig. 3.8. Consider the galaxy pair defined by galaxies A and B: the positions of both galaxies and their pair-centre are within redshift slice 2. This pair would therefore be included in analyses conducted on this slice in both top-hat and pair-centre binning schemes. The positions of galaxies C and D span two separate redshift slices and therefore the pair they define would not be included in an analysis of either slice 2 or 3 when using a top-hat binning scheme. However, this pair would be included in an analysis of slice 2 when using the pair-centre binning scheme. This schematic demonstrates both the pair-centre binning scheme and the fact that it prevents the loss of pairs from an analysis.

Suppose that we have a clustered distribution of D galaxy pairs of separation r with

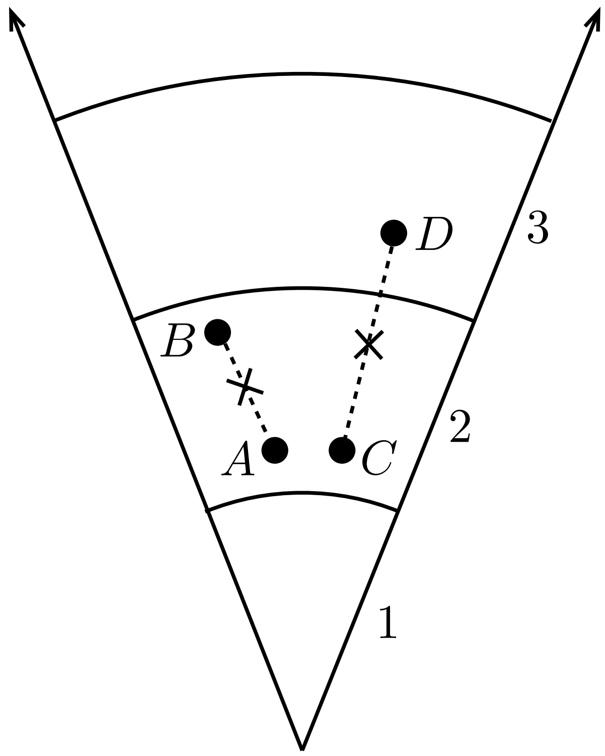

Figure 3.8: A schematic representation showing how galaxy pairs are selected using top-hat and pair-centre binning schemes. Using a top-hat binning scheme, where galaxy pairs are selected according to the position of each individual galaxy, pair AB would be placed in redshift bin 2, whereas pair CD would not be placed in any bin, and would simply not be counted in an analysis. In contrast, the pair-centre binning scheme would place both pairs in bin 2.

a uniform sampling function along the z-axis in a large volume that would contain R pairs if galaxies were randomly distributed. Because of the large volume assumption, we can assume that boundary effects for this sample are negligible. Therefore, redshift-space distortions have no effect for the full catalogue for which our estimate of $\xi_p(d_p)$ is $\hat{\xi}_p(d_p) = D/R - 1$. Now suppose the sample is split into n sub-samples, based on the redshift-space positions of the centres of the pairs within equal volumes, chosen independently of the observed galaxy distribution. Then all pairs are still counted in some bin; none are lost or gained as opposed to galaxy based selection functions. For the sub-samples, $\langle D' \rangle = D/n$, $R' = R/n$, and $\langle \xi_p(d_p) \rangle$ is unchanged from the value for the full sample. This is true regardless of bin size. The key difference here, compared with considering a set of bins based on galaxy selection, is that no pairs are left out, so the expected correlation function has to be the same for all bins.

For a sample where we do not know the true distance to each galaxy, but instead rely on photometric redshifts, binning based on apparent pair centre will also remove redshift-space distortions. The above argument based on pair conservation will also hold in this situation.

We therefore see that we can add boundaries based on pair-centres and analyse projected clustering in bins without being affected by redshift-space distortions. However, there are two problems with applying this approach in practise:

1. Galaxy pairs of wide separation now have to be included.

2. Galaxy surveys typically have flux limited boundaries, which will cause redshift dependent effects that cannot be removed by any binning. However, this effect can be removed by k-correcting the observed luminosities and cutting the sample at a more stringent k-corrected luminosity limit. We now investigate this further.

3.1.7 Flux-Limited Selection Functions

Peculiar velocities can directly influence galaxy brightness through relativistic beaming, but such effects are small for typical galaxy peculiar velocities. Redshift distortions would additionally change the apparent magnitudes through the k-correction, potentially causing galaxies to either enter or exit flux-limited samples. The change in apparent magnitude will correlate with bulk-flow motions and thus the boundary of the survey in real-space will fluctuate in a manner analogous to that described in Fig. 3.1. In this situation, the amplitude of the effect and whether it enhances or reduces the real-space clustering signal will depend on galaxy type and the band used for detection, but for a homogeneous sample of galaxies (e.g. Luminous Red Galaxies) one would expect that this effect will be significant.

This redshift-space effect is simple to remove; k-corrections derived by fitting to galaxy spectra will correct for spectral shifts caused by both the Hubble flow and any peculiar velocities. It therefore makes sense to select galaxy samples after applying the k-correction, and cutting back from survey boundaries based on apparent magnitude, until no galaxies outside the original sample would be expected to pass the revised boundary. This is not as onerous as it sounds as one has to do this to create true volume-limited catalogues.

The wavelength of light emitted by an object at redshift z will have increased by a factor $1 + z$ by the time the light reaches the observer. k-corrections transform the observed wavelength of light emitted by an object at redshift z into a standard measurement at redshift zero. The exact nature of the calculation that needs to be applied in order to perform a k-correction depends upon the type of filter used to make the observation and the shape of the galaxy's spectrum. If multi-color photometric measurements are available for a given object, thus defining its spectral energy distribution (SED), k-corrections can be computed by fitting it against a theoretical or empirical SED template. Uncertainties on photometric redshift measurements however, mean that k-corrections are unreliable for individual galaxies. For this reason, and the fact that cutting back from the survey boundary removes a large amount of data, k-corrections are not always applied to apparent magnitudes (e.g. Ross & Brunner 2009 select galaxies with de-reddened $r < 21$ for their parent sample). We therefore consider the amplitude of the effect.

One can express the fluctuation in magnitude, δm, as

$$\delta m = \mathrm{d}k_{corr}/\mathrm{d}z\delta z \tag{3.35}$$

where δz is the magnitude of the redshift distortion. This will cause fluctuations in the effective depth of the survey such that

$$DM(z_{eff}) - DM(z) = \delta_m \tag{3.36}$$

where $DM(z)$ is the distance modulus, z_{eff} is the effective depth and z would be the predicted depth. The SDSS DR7 photometric redshift table includes r-band k-corrections for every galaxy. Studying galaxies with type-value equal to 0 (the most early-type), one can determine that $\mathrm{d}k_{corr}/\mathrm{d}z \sim 3.3$ at $z = 0.4$. For an arbitrary δz, this $\mathrm{d}k_{corr}/\mathrm{d}z$ yields $z_{eff} - z = 0.5\delta z$. For example, assuming bulk flows have a velocity $\sim 10^3$km/s — thereby imparting redshift distortions at the $\sim 1\%$ level ($\delta z = 0.004$) — they impart coherent fluctuations in apparent magnitude equivalent to 0.013 magnitudes (in the r-band). At $z = 0.4$, these fluctuations in magnitude imply a change in the survey depth of $z_{eff} - z = 0.002$ (0.5%). Thus, the redshift distortions caused by selecting a flux-limited sample of galaxies can be as large as 50% of those caused by selecting a sample in redshift. Therefore, even for a flux limited selection function, redshift distortions may be important. The size of the effect depends on the slope of $k_{corr}(z)$, and one can minimise the effect by carefully choosing the band used for selection and the type(s) of galaxies included in the sample. (One can envision cases where slope of the average k-correction is zero, thus removing any effect.)

3.2 Analysis of the Hubble Volume Simulations

In order to test the effect of redshift-space distortions on the projected correlation function for a realistic non-linear distribution of galaxies, we have analysed results from the ΛCDM Hubble Volume (HV) simulations (Evrard et al., 2002). The ΛCDM HV simulation, covering a $(3000\,h^{-1}\,\mathrm{Mpc})^3$ box, assumes a cosmological model with $\Omega_m = 0.3$, $\Omega_{CDM} = 0.25$, $\Omega_b = 0.05$, $\Omega_\Lambda = 0.7$, $h = 70$, $\sigma_8 = 0.9$, & $n_s = 1$. Fig. 3.9 shows a density slice from the HV simulation with volume $V_s = 1000^3$, where every 100^{th} galaxy is sampled.

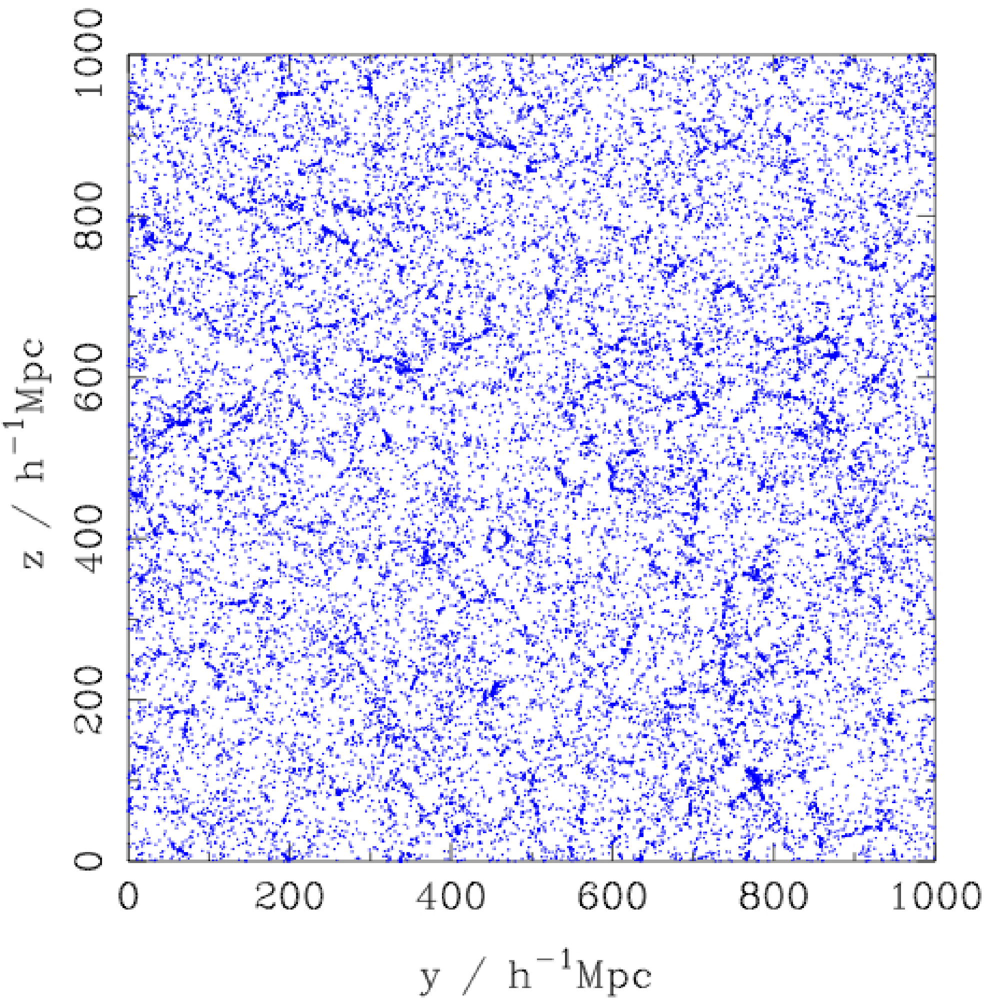

Figure 3.9: Density slice taken from the ΛCDM Hubble Volume Simulation.

We make a number of simplifications in order to help with the calculation of projected real-space and redshift-space correlation functions. For each sample to be analysed, along the two non-projection axes, we use the periodic nature of the numerical simulation to eliminate boundaries. This means that we can confidently use the natural estimator

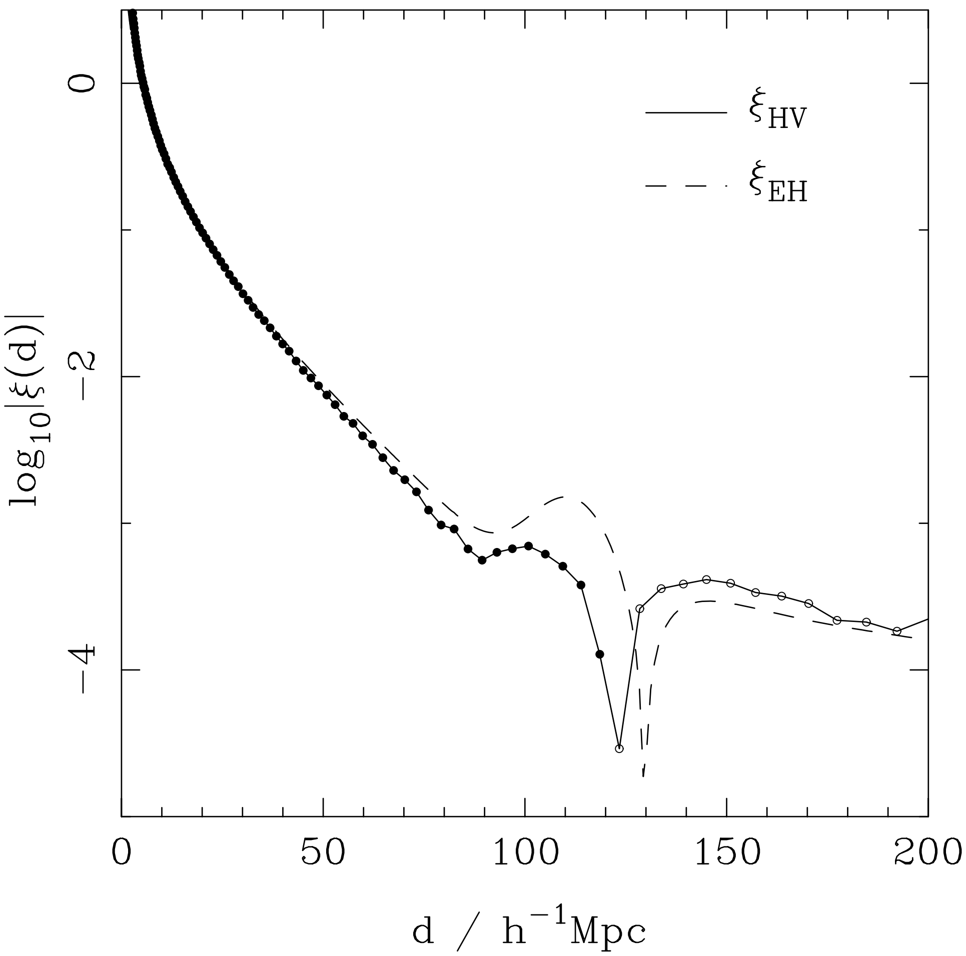

Figure 3.10: The 3D correlation function calculated from HV data in boxes of volume $V = 3000(\,h^{-1}\,\mathrm{Mpc})^3$, averaged over 75 random realisations, each containing $\sim 10^6$ galaxies. The dashed line shows a model correlation function calculated via the Eisenstein & Hu (1998) transfer function, with input cosmology as defined above.

$\xi + 1 = D/R$, where the expected number of galaxy pairs in the absence of clustering R can be calculated analytically. We also do not introduce a galaxy-bias model, and assume that galaxies Poisson sample the matter particles. The inclusion of such a model would not alter the conclusions of this work.

We start by measuring the 3D correlation function of the HV data. Fig. 3.10 shows the correlation function averaged over 75 random realisations of the HV, each containing $\sim 10^6$ galaxies. The dashed line represents a model correlation function calculated via the standard Fourier transform of the power spectrum obtained using the Eisenstein & Hu (1998) transfer function, with input cosmology as defined above. The model fails to

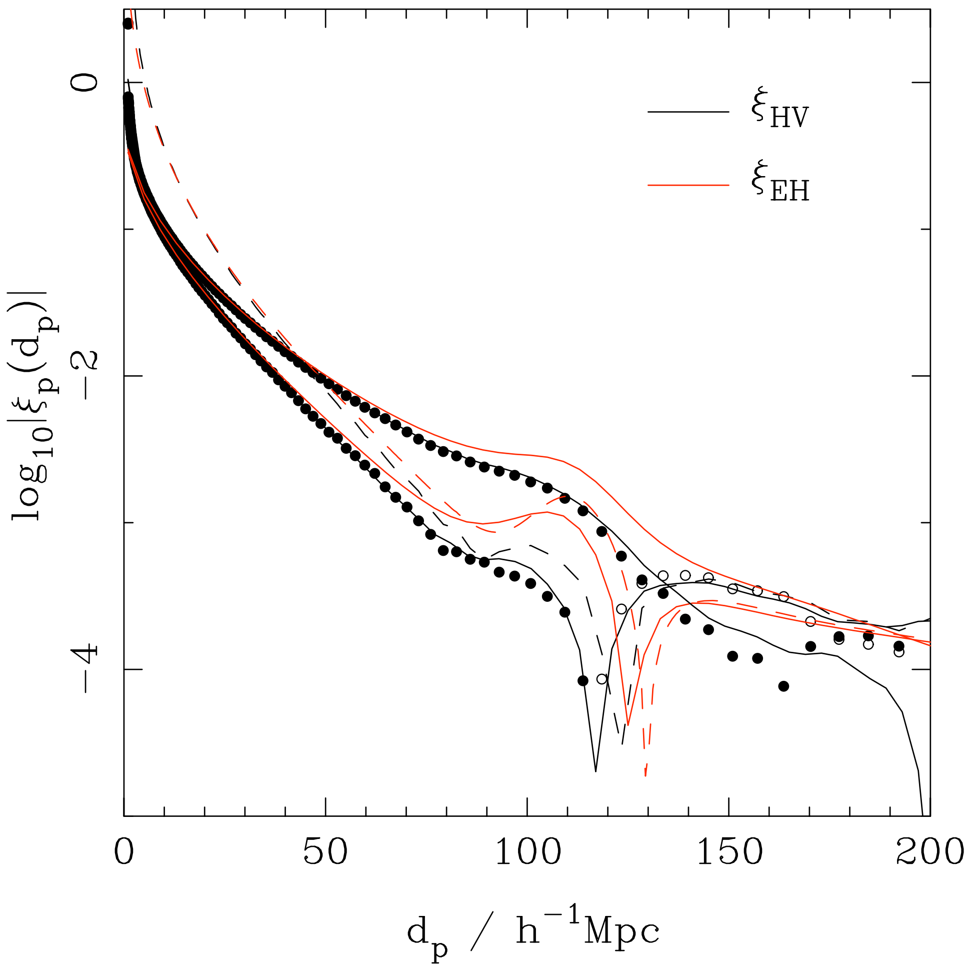

Figure 3.11: The projected correlation function calculated from HV data in galaxy density slices of width $100\,h^{-1}\,\mathrm{Mpc}$. Solid symbols are plotted where the correlation function is positive, while open symbols show where the correlation function is negative. Solid lines represent projected correlation function models calculated via Eq. (3.10) (with or without the expected redshift-space distortion anisotropic shift) with input 3D correlation functions $\xi_{EH}(d)$ (red) and $\xi_{HV}(d)$ (black), respectively. $\xi_{EH}(d)$ was calculated using the Eisenstein & Hu (1998) transfer function, whilst $\xi_{HV}(d)$ is the measured 3D correlation function averaged over 75 random realisations of the HV, each with $\sim 10^6$ galaxies. $\xi_{EH}(d)$ and $\xi_{HV}(d)$ are plotted as dashed lines (assuming $d = d_p$).

fit the data on all scales and deviates significantly around the BAO scale. The accurate modelling of the projected correlation function is strongly dependent on the $\xi(d)$ input into Eq. (3.10). At this point, we can appeal to the shear size and periodic nature of the HV. The high number density of the simulation means that we can draw many random realisations of the volume without hitting a shot-noise limit. The periodic nature of the simulation also means that we are not cosmic variance limited. Therefore, it is safe to assume that the 3D correlation function that we have measured is a true representation of the overall clustering signal present. Consequently, we can use this measured 3D correlation function $\xi_{HV}(d)$ in Eq. (3.10) to calculate the projected correlation function. To ensure that this was the right decision, we investigated upon the discrepancy between the models by checking the positioning of the BAO peak in the model correlation function calculated from the HV CAMB input power spectrum. We found that the position of the BAO peak in our averaged 3D correlation function matched that of the input correlation function very well, and so feel confident with our choice of model.

As a test, we apply a top-hat selection function to the galaxy positions, calculating the projected correlation function for a window of width $100\,h^{-1}\,\mathrm{Mpc}$. Fig. 3.11 shows the correlation function after reducing noise by averaging over samples. Projected models calculated using Eq. (3.10) with input 3D correlation functions $\xi_{EH}(d)$ (red) and $\xi_{HV}(d)$ (black) are plotted as solid lines. The corresponding 3D correlation functions are shown as dashed lines (plotted assuming $d = d_p$). As predicted, the use of $\xi_{HV}(d)$ in Eq. (3.10) to calculate the expected projected correlation function provides the best fit to the data. Therefore, we will use $\xi_{HV}(d)$ as the model 3D correlation function throughout this analysis, unless otherwise stated.

We continue to apply top-hat selection functions to the galaxy positions, calculating projected correlation functions for window widths 50 $h^{-1}\,\mathrm{Mpc}$, 100 $h^{-1}\,\mathrm{Mpc}$, 250 $h^{-1}\,\mathrm{Mpc}$ and 500 $h^{-1}\,\mathrm{Mpc}$ in real and redshift space. Figs. 3.12 & 3.13 show the averaged correlation function. In real-space the projected correlation function tends towards the 3D correlation function at large scales, as expected. In line with the analysis presented in §3.1.5, the scale at which ξ_p becomes $\sim \xi_{3D}$ is larger for the 500 $h^{-1}\,\mathrm{Mpc}$ bin. For each bin size, the inclusion of redshift-space distortions clearly has a strong effect and this effect grows dramatically as the scale gets larger. Notably, it is larger even than the effect of redshift-space distortions on the 3D spherically averaged correlation function (or power spectrum). The effect is enhanced in the narrower projection window. As well as increasing the amplitude of the projected correlation function, we see that redshift-space distortions also act to wash out the baryon acoustic oscillation signal.

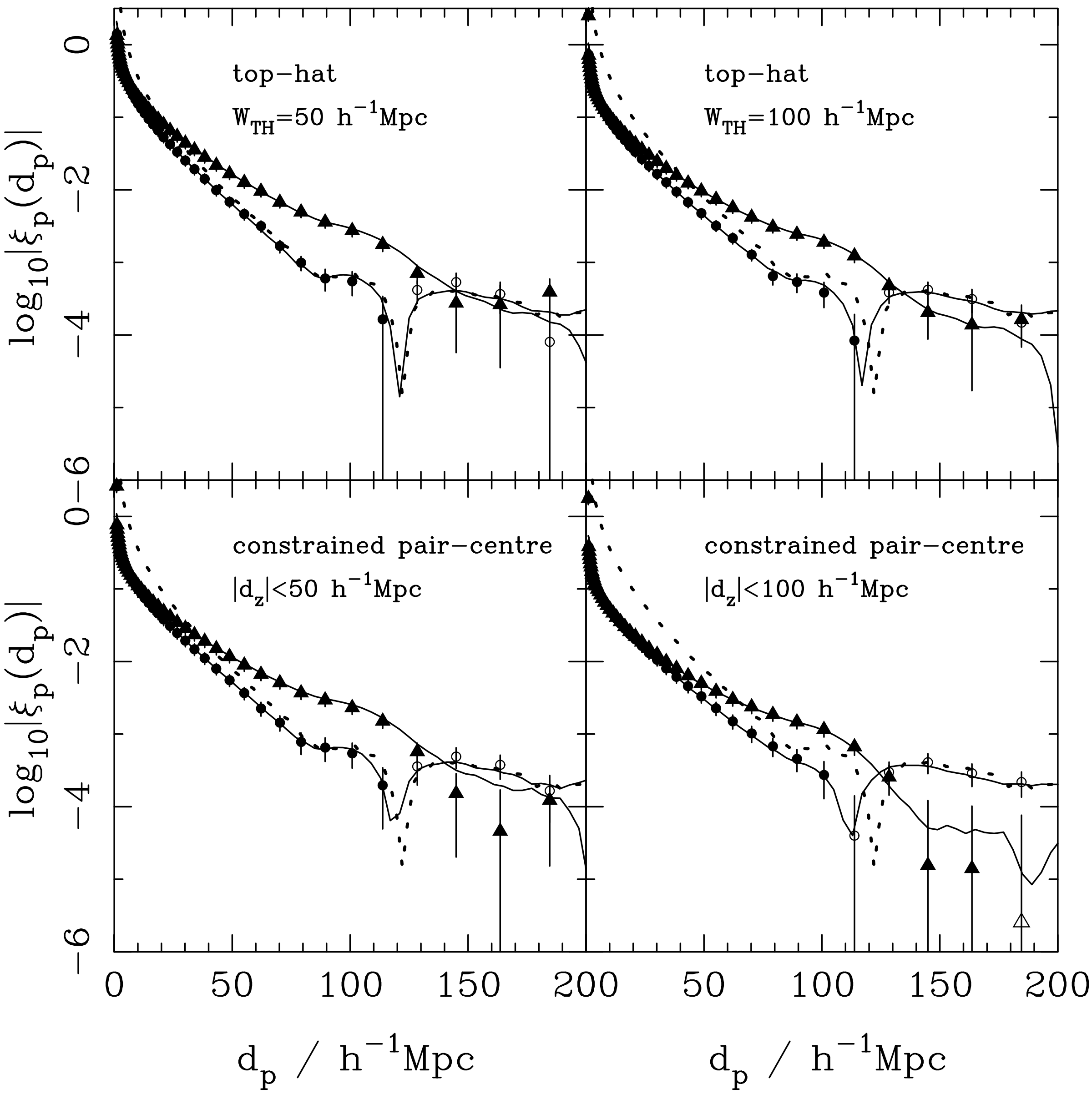

Figure 3.12: Top row: Correlation functions calculated from HV data in galaxy density slices of width $50\,h^{-1}\,\mathrm{Mpc}$ and $100\,h^{-1}\,\mathrm{Mpc}$. Solid symbols are plotted where the correlation function is positive, while open symbols show where the correlation function is negative. Bottom row: Correlation functions calculated from HV data for galaxy pairs selected based on the constrained pair centre binning scheme, and with radial separation less than $50\,h^{-1}\,\mathrm{Mpc}$ or $100\,h^{-1}\,\mathrm{Mpc}$. 1σ error bars are plotted in both cases, assuming that the slices analysed draw correlation functions from a Gaussian distribution. The dotted line gives the 3D HV correlation function (plotted assuming $d = d_p$) as measured from the simulation. Models calculated using Eq. (3.10) with this 3D correlation function as the input (with or without the expected redshift-space distortion anisotropic shift) are shown by the solid lines.

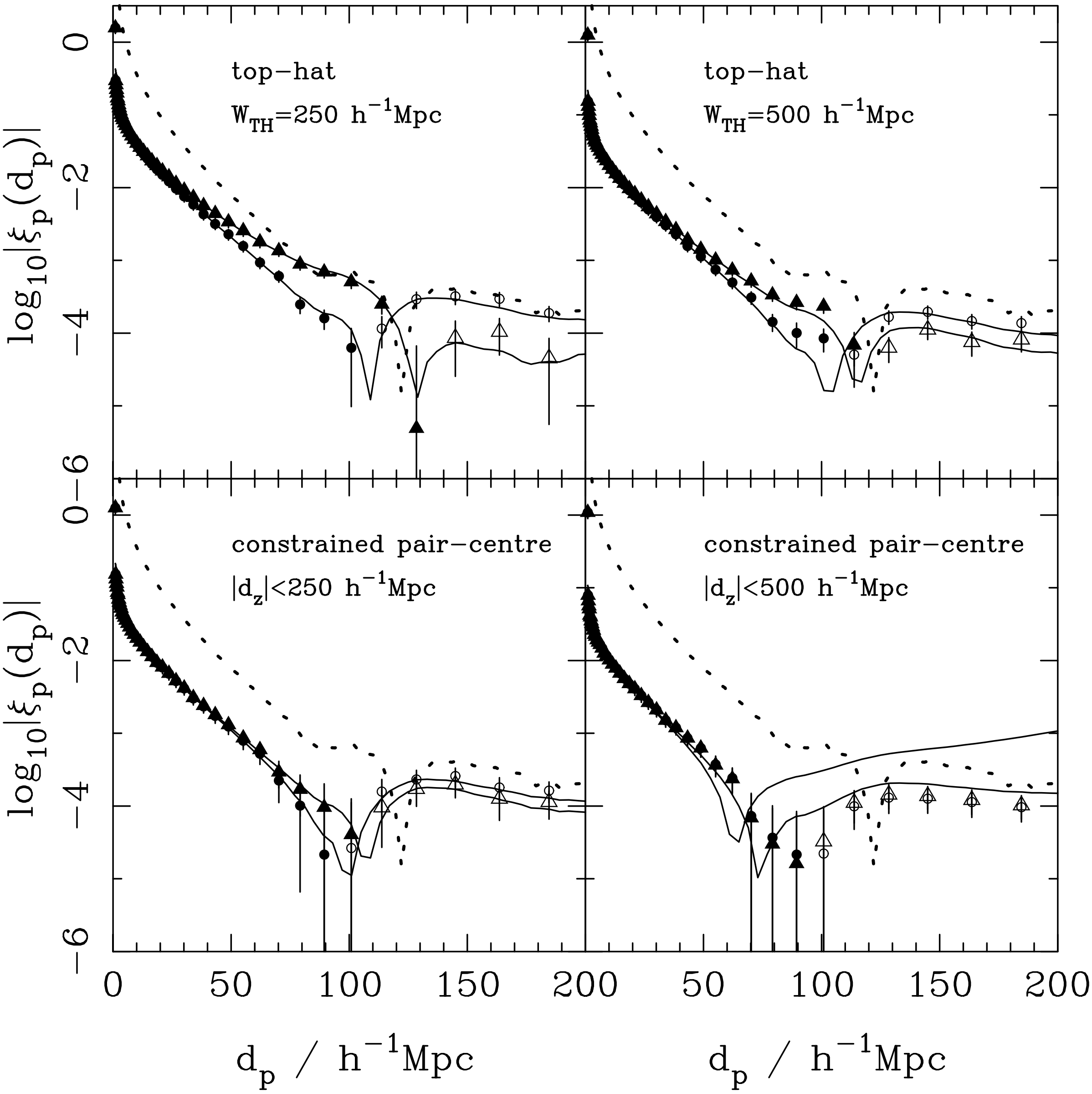

Figure 3.13: As fig. 3.12. Top row: Galaxy density slices of width $250\,h^{-1}\,\mathrm{Mpc}$ and $500\,h^{-1}\,\mathrm{Mpc}$. Bottom row: Pair-centre selection with radial separation less than $250\,h^{-1}\,\mathrm{Mpc}$ and $500\,h^{-1}\,\mathrm{Mpc}$.

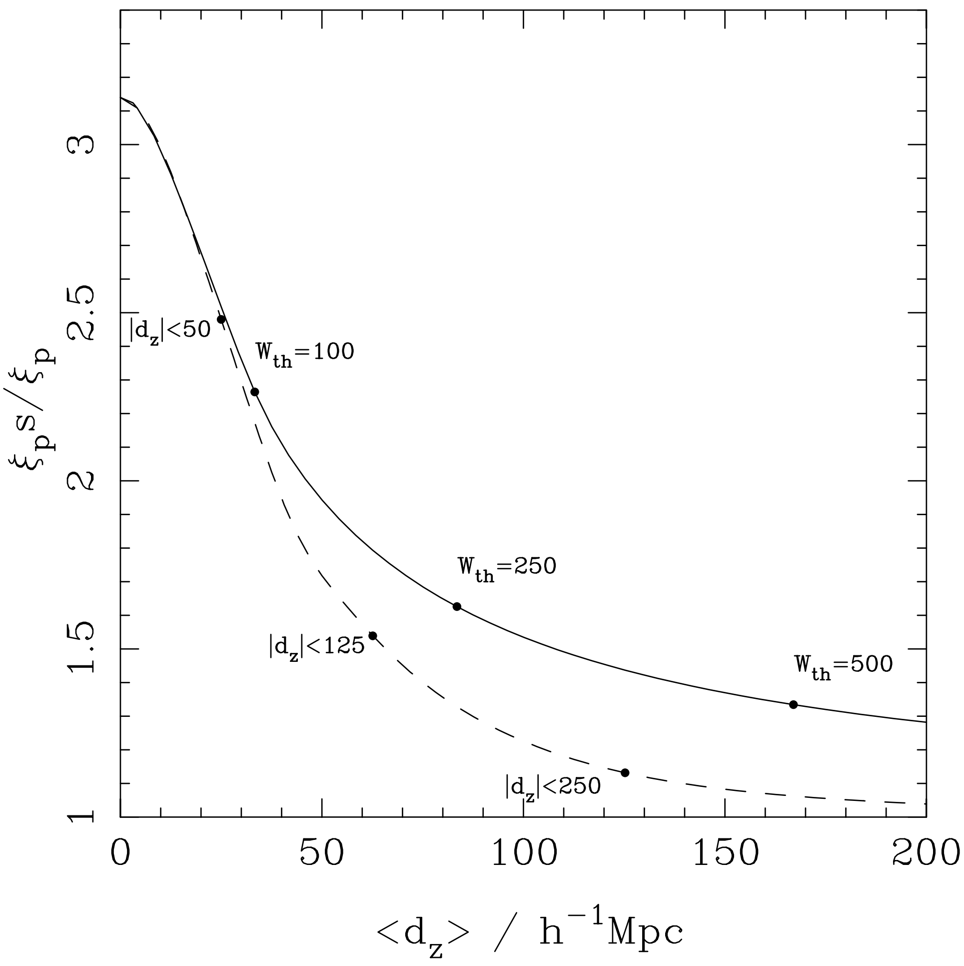

Figure 3.14: The expected ratio of the projected correlation functions in redshift-space and in real-space, averaged for "angular" separations between $40\,h^{-1}\,\mathrm{Mpc}$ and $80\,h^{-1}\,\mathrm{Mpc}$, as a function of bin width. The solid line show the difference as a function of the width of the top-hat window. The dashed line show the result for constrained pair-centre binning as a function of an additional constraint placed on the radial galaxy separation. We have plotted results (and therefore matched filters) as a function of the mean radial galaxy separation.

Selecting galaxy pairs solely based on the position of their pair-centre removes the effect of redshift-space distortions. To see this, suppose we split along the projection axis into N slices, and average the DD counts over all slices. Then the average is independent of N as all pairs are counted however many bins are selected. In addition, the periodic nature of the simulation means that no pairs are gained or lost between real-space and redshift-space: we always count all pairs of galaxies, so there will be no change in the measured correlation function. As explained in §3.1.6, we cannot apply such a binning in practice as there will always be additional observational constraints such as a magni-

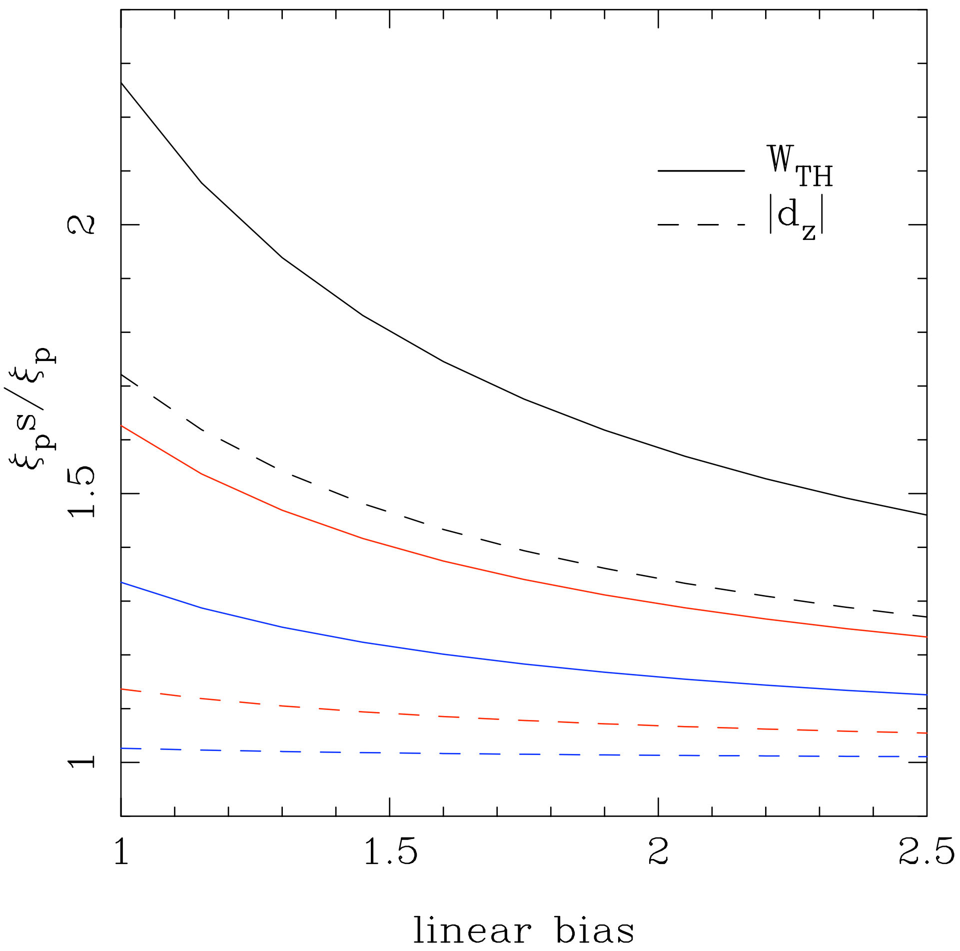

Figure 3.15: The expected ratio of the projected correlation functions in redshift-space and in real-space, averaged for "angular" separations between $40\,h^{-1}\,\mathrm{Mpc}$ and $80\,h^{-1}\,\mathrm{Mpc}$, as a function of galaxy bias, assuming a ΛCDM cosmology with $\Omega_m = 0.25$. The solid lines show the difference as a function of top-hat window widths $100\,h^{-1}\,\mathrm{Mpc}$ (black), $250\,h^{-1}\,\mathrm{Mpc}$ (red) and $500\,h^{-1}\,\mathrm{Mpc}$ (blue). The dashed line show the result for constrained pair-centre binning as a function of an additional constraint placed on the radial galaxy separation. As in Fig. 3.14, we have plotted results (and therefore matched filters) as a function of the mean radial galaxy separation.

tude limit on the galaxy distribution. As explained in Section 3.1.7, if we select based on an apparent magnitude limit, we can remove redshift distortions by applying a more stringent magnitude limit based on k-corrected luminosities. Here we have to cut the luminosity limit back to make sure that the new sample is complete, and contains all of the possible galaxies. However, there is a further practical problem in that including galaxy pairs with wide radial separation might complicate the modelling of cosmological evolution required to fit the correlation function. Consequently, it might be difficult to analyse the measured correlations function for a pair-centre binned sample in practice.

We therefore introduce a **constrained pair-centre** binning scheme that includes an upper limit on the pair separation along the projection axis, in addition to pair-centre binning. This is equivalent to locating *each* galaxy included in the analysis in the centre of a top-hat bin. We should expect that the effect of redshift-space distortions will be reduced compared with binning galaxy distributions in a top-hat with the same width, as boundaries will only affect galaxy pairs with the maximum radial separation, whereas for top-hat bins they affect galaxy pairs with a range of radial separations (see Figs. 3.6 & 3.7). Results calculated using this binning scheme are shown in Figs. 3.12 & 3.13. Here we see that the effect of redshift-space distortions is reduced, especially for the larger $|d_z|$ limit.

In order to investigate the effect of different binning schemes further, Fig. 3.14 shows a comparison on the large-scale redshift-space and real-space correlation function amplitude. These are averaged for galaxy separations between $40\,h^{-1}\,\mathrm{Mpc}$ and $80\,h^{-1}\,\mathrm{Mpc}$. We have plotted these as a function of average radial galaxy separation, in order to compare filters in an unbiased way. We clearly see that, when binning radially using the constrained pair-centre binning scheme, the effect of redshift-space distortions is significantly reduced.

The relative importance of redshift-space distortions depends on the average galaxy bias of the populations being considered; there is a balance between the impacts of b and f in Eq. (3.10). In order to demonstrate this, Fig. 3.15 shows that the relative effect of redshift-space distortions decreases as the bias of the galaxy sample analysed increases. This explains why the effect of redshift-space distortions was reduced in the work of Baldauf et al. (2010).

In this section, we have considered the cases of a top-hat or pair-centre galaxy selection. We have argued that while, in principle, pair-centre binning removes the effects of redshift-space distortions provided k-corrections are included when magnitude limits are applied, this is affected by physical factors such as magnitude limits, and also that there are good reasons to remove galaxies of wide separation if we are to measure the evolution in the correlation function.

3.3 Dealing with Hybrid Selection Functions

In practice, radial selection functions are dependent on both observational constraints, such as the limiting apparent magnitude of the survey, and additional binning. One expects that the boundary based on observational constraints can be treated as a real-space boundary, though, as shown in §3.1.7, this is not always so straightforward. Consequently, when one applies a top-hat selection in redshift to an observed sample of galaxies, the resulting boundaries of the selection function will include both real-space and redshift-space components.

Fig. 3.16 shows a schematic representation of a top-hat selection in redshift made at positions s_{z1} and s_{z2} along a non-uniform real-space radial selection function. It shows that we can split galaxies within this bin into three sub-samples, with different boundaries:

- A_s (redshift-redshift): Selected with both boundaries in redshift-space.
- B_h (redshift-real): Selected with one boundary in real-space and one boundary in redshift-space (*hybrid*-space).
- C_r: Selected with both boundaries in real-space.

The real-space and redshift-space boundaries of Fig. 3.16 are represented by solid and dotted lines respectively. Any auto-correlation of galaxies with this selection function will essentially be a weighted sum (based on the amplitude of the selection function) of the auto-correlations of galaxies within the individual subsamples and the cross-correlations of galaxies in different subsamples.

In order to investigate the projected clustering of these different subsamples, we have drawn samples of particles from the HV simulation (see Section 3.2), created in top-hat bins of width $100\,h\,\mathrm{Mpc}^{-1}$. Sample A_s has top-hat selection boundaries in redshift-space, sample B_h has one real-space and one redshift-space boundary, while sample C_r has both boundaries in real-space. These samples cover the same region of the simulation.

Fig. 3.17 shows the projected auto-correlation functions for these subsamples. The measured ξ_p for the C_r and A_s samples are essentially the same as those shown in the top-right panel of Fig. 3.12, and just as before they return the expected real and redshift-space correlation functions calculated via Eq. 3.10. However, the hybrid-space correlation function, ξ_p^h of sub-sample B_h has an amplitude that lies in-between those of the pure real

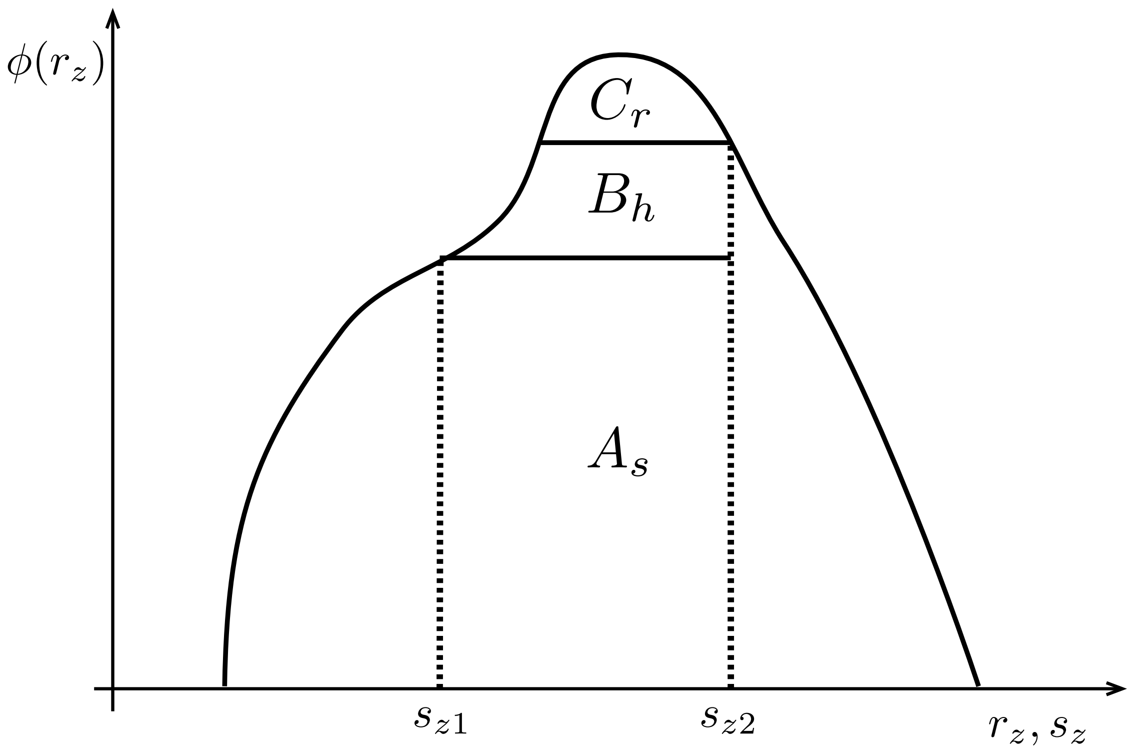

Figure 3.16: Schematic representation of an evolving real-space radial selection function with populations A_s, B_h and C_r defined according to where a top-hat bin with redshift-space boundaries at s_{z1} and s_{z2} intersect the radial selection. Populations have boundaries in: A_s redshift-space, B_h hybrid-space and C_r real-space.

and redshift-space correlation functions. We find that we can effectively model ξ_p^h by assuming the underlying 3D overdensity field has a correlation function ξ^hgiven by

$$\xi^h + 1 = \sqrt{(1+\xi^r)(1+\xi^s)}, \tag{3.37}$$

for ξ^s in Eq. (3.10). Note that we are using ξ^r to represent the real-space 3-dimensional correlation function. As can be seen in Fig. 3.17, this model is well-matched to the measured ξ_p. The justification for this model is that the multiplicative boost to the projected density fluctuations (R if we consider that $\xi = D/R - 1$) can be decomposed into multiplicative contributions from each boundary.

For example, population B_h has only one redshift-space boundary, therefore we only

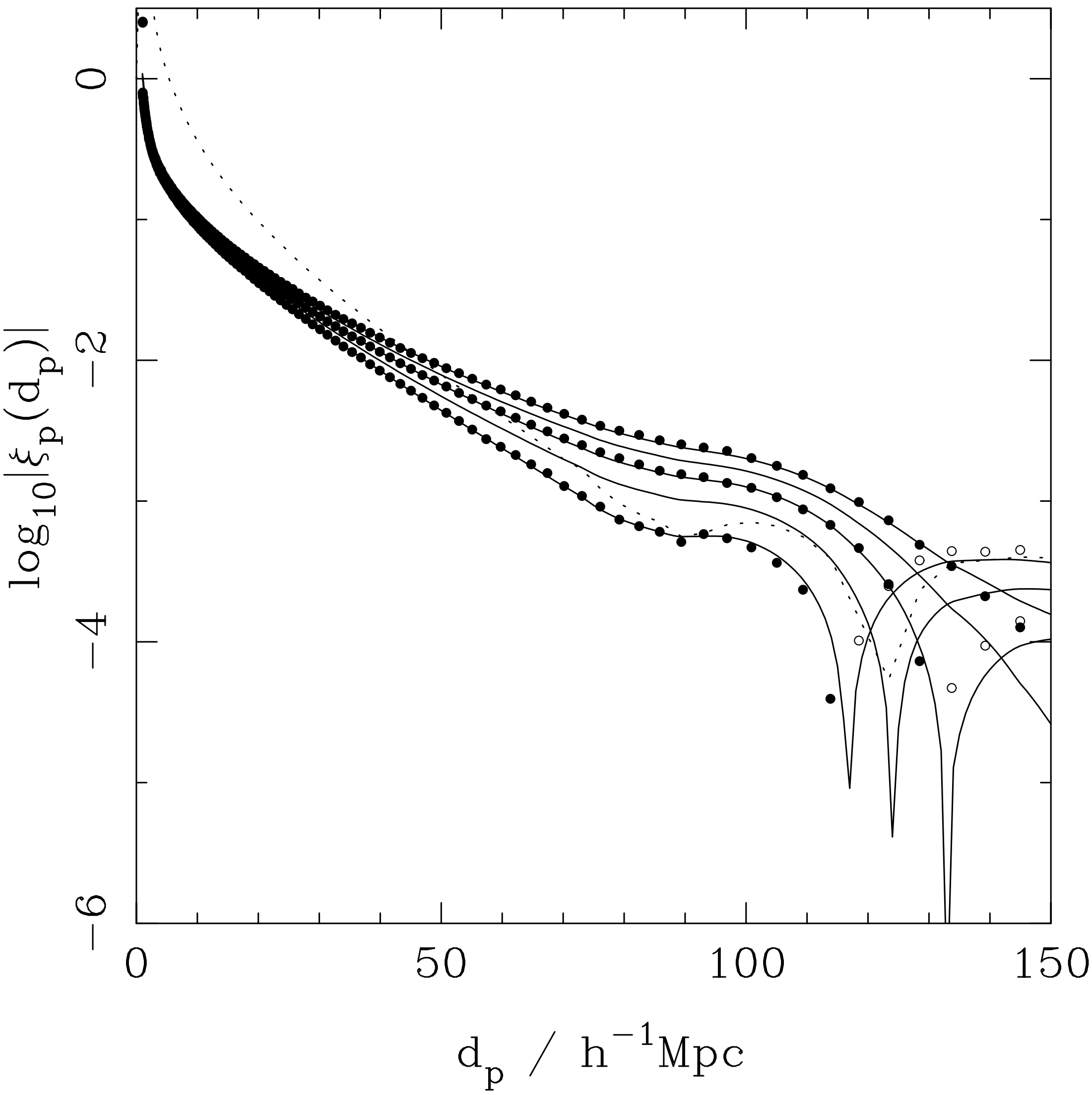

Figure 3.17: The average recovered auto-correlation function (solid circles) for galaxies from 90 samples drawn from the Hubble Volume simulation using three different radial selections, each with top-hat width 100 h^{-1}Mpc. These are compared against model correlation functions calculated for different galaxy samples Eq. 3.39. The three radial selections are: 1) two real-space boundaries (lowest points), which best matches the model calculated using the real-space correlation function, 2) two redshift space boundaries (highest points), which best matches the model calculated using the redshift-space correlation-function and, 3) a real-space boundary on one side and a redshift space boundary on the other side (points in the middle), which best matches the model calculated using the geometric mean of the real- and redshift-space correlation functions.

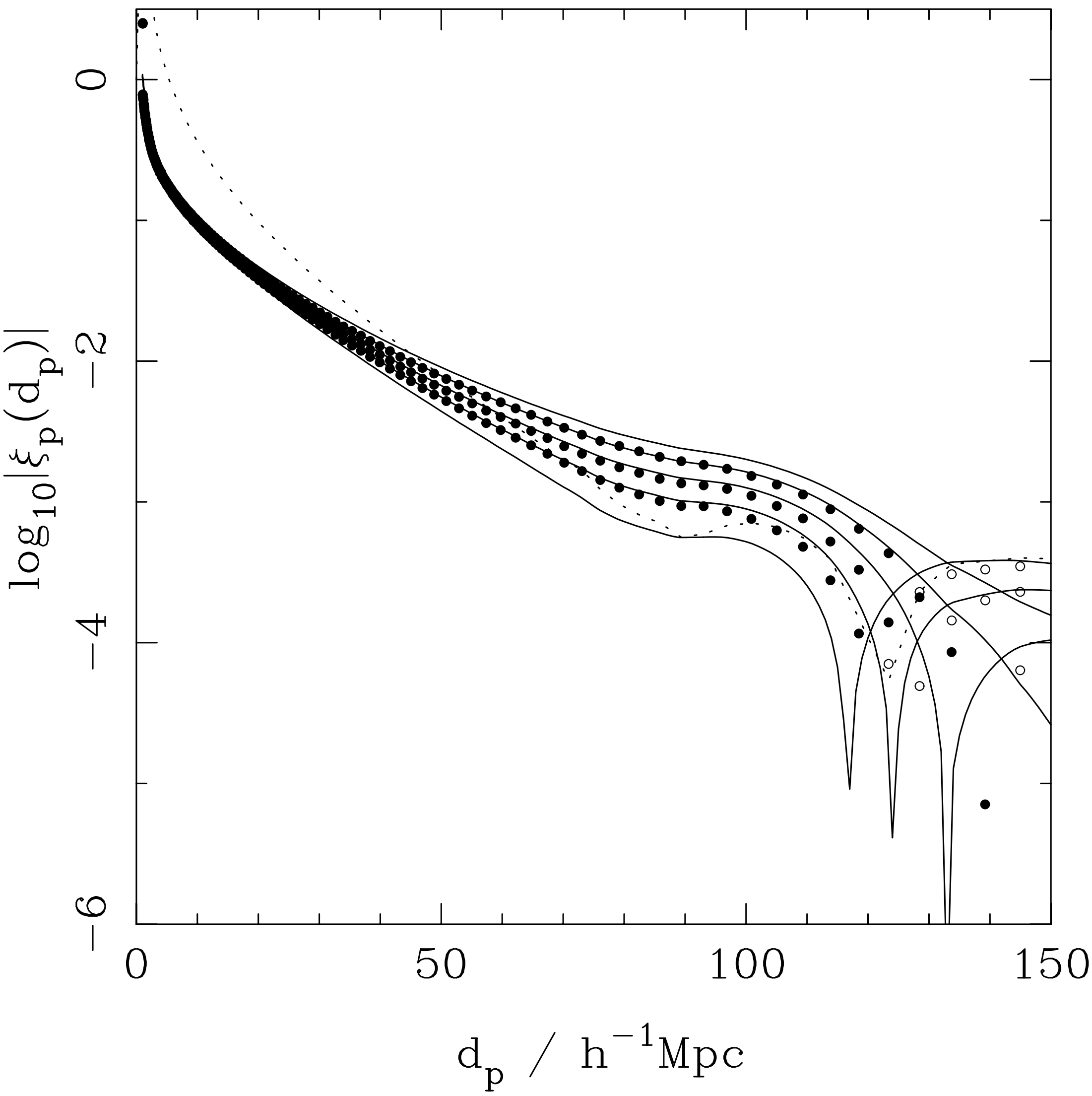

Figure 3.18: The average measured cross-correlation functions from 90 radial slices of width 100 h^{-1}Mpc in real-space, redshift-space or a hybrid with one real-space and one redshift-space boundary, each containing 10^6 galaxies (solid circles). These are compared against the model ξ_p^h of Eq. (3.39) (solid lines), for different total numbers of redshift boundaries. The amplitude of both model and data correlation functions increase with increasing dependence on the redshift-space correlation function.

pop'n	A_s	B_h	C_r
A_s	$\xi^h + 1 = \sqrt[4]{(\xi^s+1)^2(\xi^s+1)^2}$	$\xi^h + 1 = \sqrt[4]{(\xi^s+1)^3(\xi^r+1)}$	$\xi^h + 1 = \sqrt[4]{(\xi^s+1)^2(\xi^r+1)^2}$
B_h	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)(\xi^s+1)^3}$	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)^2(\xi^s+1)^2}$	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)^3(\xi^s+1)}$
C_r	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)^2(\xi^s+1)^2}$	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)^3(\xi^s+1)}$	$\xi^h + 1 = \sqrt[4]{(\xi^r+1)^2(\xi^r+1)^2}$

Table 3.1: Table showing expected correlation functions for different combinations of selection boundaries. Populations A_s, B_h and C_r in fig. 3.16 represent 3 distinct regimes where galaxy pairs are selected in redshift-space, hybrid-space and real-space, respectively. Here, hybrid-space refers to galaxy pairs with one galaxy selected in real-space and the other in redshift-space.

have half the number of galaxies moving into or out of the sample compared with population A_s. Following this theory, and considering Eq. 3.37, we should find that the relative effect of redshift-space distortions on each population, and their cross-correlations, are simply proportional to the number of redshift-space boundaries present. In this particular example, one would expect the cross-correlation between samples A_s and B_h to include the effect of three redshift-space boundaries. Using Eq. 3.37 we can express this as

$$\xi^{sssr} + 1 = \sqrt[4]{(1+\xi^s)^3(1+\xi^r)}. \tag{3.38}$$

Table 3.1 summarises the ξ we expect to use in Eq. 3.37 for all of the auto- and cross-correlation functions for populations A_s, B_h and C_r.

We can condense these relationships into one simple equation as follows: if we choose galaxies from a sample with $m \in \{0, 1, 2\}$ redshift-space boundaries, and another from a sample (possibly the same one) with $n \in \{0, 1, 2\}$ redshift-space boundaries, then the expected correlation function is given by

$$\xi^h + 1 = (1+\xi^r)^{1-l/4}(1+\xi^s)^{l/4}, \tag{3.39}$$

where $l = m + n$.

Fig. 3.18 displays the cross-correlations between our three HV subsamples. As expected, the model calculated using the appropriate ξ^h from Eq. (3.39) is the closest match to the measured cross-correlation in every case. All of the models do over-predict all three measurements at large scales, but we believe this is reflective of the error associated with our measurements (one would expect it to be covariant between each sample as they all sample the same density field). It is possible that we are seeing effects caused by the coherence of the boundaries with each other that would be removed for wider bins, such as those we consider in the next chapter (DES).

Given a hybrid selection function such as that shown in Fig. 3.16 we must split the sample into populations where we can assume simple boundary conditions for each. In fact, we can consider solving the projection equation (e.g. Eqns. 3.2 & 3.9 in real-space and redshift-space) by Monte-Carlo integration over pairs of radial galaxy locations. For each pair of locations we can determine the relative contributions from galaxies in each of the subsamples, and therefore construct a full model for the correlation function.

3.4 Constraining w

The ultimate aim of any clustering analysis is to place constraints upon cosmological parameters, in particular the dark energy equation of state w. Although a full clustering analysis is beyond the scope of this thesis, we consider here how accurately w may be determined using the acoustic signal in the correlation function as a standard ruler, following the methodology of Blake & Glazebrook (2003). We highlight limits imposed on our projected correlation function as well as various alternative clustering analysis techniques.

The apparent scale of features in the galaxy correlation function may prove to be a powerful method for constraining certain cosmological parameters including the matter density parameter Ω_m, the dark energy density parameter Ω_{DE}, the dark energy equation of state parameter w, and the Hubble constant H_0. In order to make such an approach work we need to either know the true physical scale of particular features in the correlation function beforehand, or compare the relative features when measured parallel and perpendicular to the line-of-sight (Alcock & Paczynski, 1979). Assuming all cosmological parameters are well constrained, excluding the dark energy equation of state, we can use the scale of the BAO as a standard ruler.

The conversion of the redshift data into real space requires us to assume values for cosmological parameters; an incorrect choice can lead to a distortion of the correlation function and the appearance of the acoustic signal in the wrong place. If dark energy is neglected at high redshift, the comoving sound horizon size at last scattering s is given by

$$s = \frac{1}{H_0 \Omega_m^{1/2}} \int_0^{a_r} \frac{c_s}{(a + a_{eq})^{1/2}} da, \tag{3.40}$$

redshift	tangential x'/x	radial dx'/dx	ruler % re-scaling
0.3	0.988373	0.979556	1.403322
0.5	0.984084	0.974568	1.860739
1.0	0.979681	0.974882	2.180711
2.0	0.979625	0.984826	1.846872
3.0	0.981184	0.991051	1.468723

Table 3.2: Table detailing the length distortion of a series of rulers located at redshifts $z = 0.3, 0.5, 1.0, 2.0, 3.0$ in the cases where it is oriented radially, dx'/dx, and tangentially, x'/x. The percentage re-scaling, relating to the level of precision required in the determination of the length of the ruler to obtain a measurement of $\Delta w = 0.1$, is calculated via the relation $100 \times (x'/x)^{2/3}(dx'/dx)^{1/3}$.

where a_r and a_{eq} are the values of the scale factor $a = 1/(1+z)$ at recombination and matter radiation equality, respectively. The sound speed $c_s \sim c/\sqrt{3}$ over the interval of integration. The theoretical value of $s \sim 100\,h^{-1}\,\mathrm{Mpc}$ and is set by fundamental CMB physics which depend strongly on Ω_m, weakly on Ω_b, and negligibly on dark energy. Thus, s is our *standard ruler*. In a redshift survey at intermediate redshift z, the apparent size of s will depend on the cosmological geometry, now including the effects of dark energy. The effects of assuming an incorrect world model would include a distortion of the measured value of s. To zeroth order, the precision with which we can empirically measure s tells us how accurately we can measure the geometrical distance to the redshift z and hence how accurately we can measure the equation of state w.

Let us now consider how accurately a detailed measurement of the BAO ruler can constrain the dark energy parameter w, assuming it does not vary with redshift. In order to measure the correlation function from a galaxy redshift survey we must convert redshifts to comoving coordinates, assuming values for Ω_m and w. As we have shown, the expected BAO signal scale is determined by the sound horizon before recombination; it is a function of Ω_m, Ω_b and h. The result of assuming an incorrect set of cosmological parameters would be a distortion of the measured correlation function so that the derived value of the length of the ruler is inconsistent with that expected from theory. For a flat geometry, the fundamental relation between comoving distance x and redshift z can be written in the form

$$\frac{dx}{dz} = \frac{c}{H_0}\frac{1}{\sqrt{\Omega_m(1+z)^3+\Omega_{DE}(1+z)^{3(1+w)}}} \tag{3.41}$$

$$= \frac{c}{H_0\Omega_m^{1/2}}\frac{1}{\sqrt{(1+z)^3+(\Omega_m^{-1}-1)(1+z)^{3(1+w)}}}. \tag{3.42}$$

For values of $w \approx -1$, the second term inside the square-root in Eq. 3.42 is small for $z \gtrsim 1$ and the zeroth-order dependence is $x \propto H_0^{-1}\Omega_m^{-1/2}$. This cancels the zeroth-order dependence of the sound horizon scale on Ω_m and H_0 in Eq. 3.40. Thus this cosmological test has reduced sensitivity to uncertainties in Ω_m and H_0.

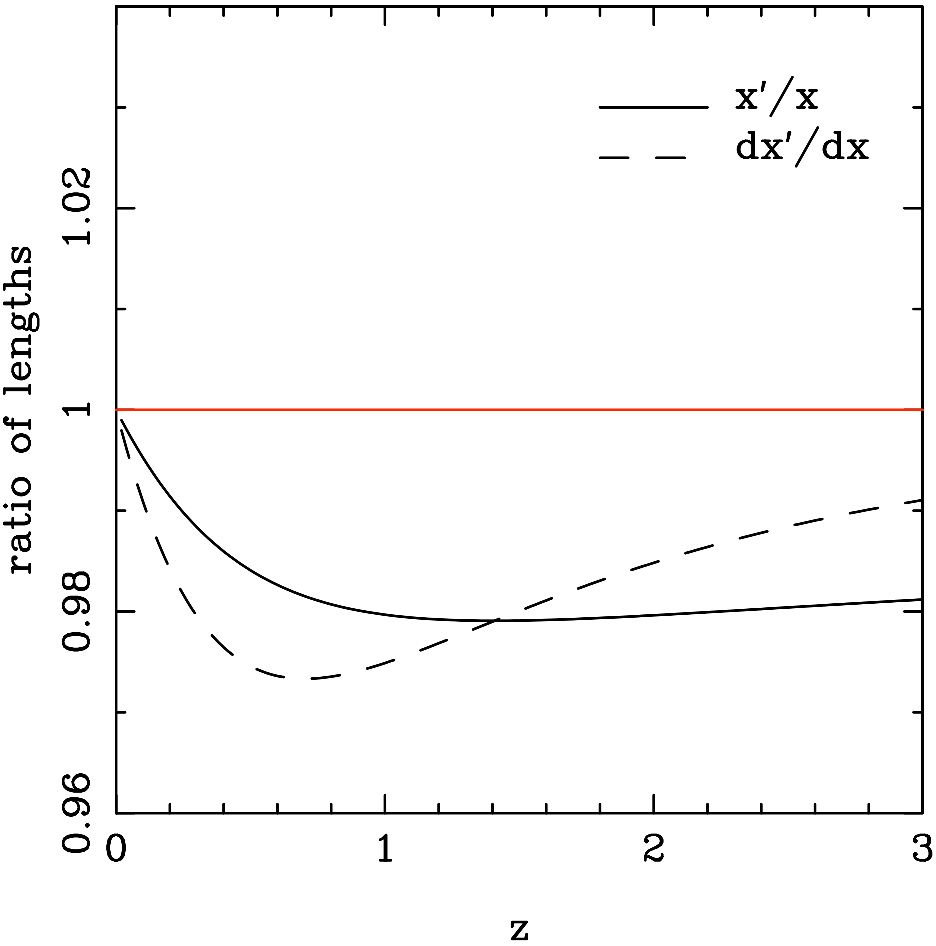

Figure 3.19: Figure showing the length distortion of a ruler in the cases where it is oriented radially, dx'/dx, and tangentially, x'/x, for a cosmology where w is perturbed away from it's true value such that $\Delta w = 0.1$. We show the relationship over the full redshift range $0 < z < 3$. The percentage re-scaling of the ruler is calculated assuming Cartesian coordinates. See main text for details.

Suppose $\Omega_m = 0.3$ and the true value of the dark energy parameter is $w_{true} = -1$. With an assumed cosmology of $w_a = 0.9$, we can use Eq. 3.42 to calculate the length distortion of the ruler in the cases where it is oriented radially, dx'/dx, and tangentially, x'/x. The zeroth order effect is a re-scaling in the length of the ruler and the first order effect is a radial/transverse shear. Table 3.2 summarises these effects for a series of rulers located at redshifts $z = 0.3, 0.5, 1.0, 2.0, 3.0$, with Fig. 3.19 showing the relationship over

method	advantages	disadvantages
$\xi(r)$, $P(k)$	• Retains *all* clustering signal. • Well understood theory. • Well understood and tested methodologies with proven results.	• Cosmological model dependence in translation from redshift- to real-space in both tangential and radial components of distance. • Sensitive to inaccuracies in redshift measurements. • Does not retain phase information.
$\xi_p^{TH}(d_p)$, $P_p^{TH}(k_p)$	• Reduced effects of redshift-space distortions over infinite radial windows. • Reduced level of precision required in real-space distance determinations via removal of radial component.	• Affected by redshift-space distortions over narrow projection windows. • Cosmological model dependence in translation from redshift- to real-space in the tangential component of distance. • Loss of signal in the radial direction. • Does not retain phase information.
$\xi_p^{PC}(d_p)$, $P_p^{PC}(k_p)$	• No redshift-space distortion effects across all projection window sizes for ideal surveys. • Drastically reduced redshift-space distortion effects across all projection window sizes for realistic surveys. • Reduced level of precision required in real-space distance determinations via removal of radial component.	• Cosmological model dependence in translation from redshift- to real-space in the tangential component of distance. • Loss of signal in the radial direction. • Diluted BAO signal due to inclusion of pairs with wide-separation in the radial component. • Increased level of cross-correlation between radial bins. • Does not retain phase information.
$w(\theta)$, C_l	• No translation from redshift- to real-space required, therefore no cosmological model dependencies. • Abundance of data due to large sizes of 2D surveys. • Can be used to reconstruct the 3D power spectrum.	• Reconstruction of 3D power spectrum introduces cosmological model dependencies in the translation from redshift- to real-space in the radial component of distance. • Does not retain phase information.

Table 3.3: Table summarising some of the advantages and disadvantages of various clustering analysis techniques.

the full redshift range $0 < z < 3$. The percentage re-scaling of the ruler is calculated assuming that the tangential and radial components are Cartesian, that is, they comprise $2/3$ and $1/3$ of the length respectively. Therefore, this distortion of spatial scale suggests that a full 3D measurement of the ruler with $\sim 2.2\%$ precision at $z \sim 1$ translates into a measurement of w with an accuracy $\Delta w \approx 0.1$. If we consider the projected correlation function, we only need to calculate the re-scaling of the ruler in the tangential direction, since we are removing the radial component. Following the assumption of Cartesian coordinates, we see that this results in a measurement of w with an accuracy $\Delta w \approx 0.1$ for a measurement of the tangential component of the ruler with only $\sim 2\%$ precision at redshift $z \sim 1$. Alternatively, we can choose to measure the galaxy angular two-point correlation function, which is defined by the joint probability δP of finding two galaxies occupying each of the elements of solid angle $\delta\Omega_1$ and $\delta\Omega_2$ separated by angle θ. This method eradicates the need to translate between redshift and real-space entirely, and is therefore model independent. There are numerous clustering techniques that can be employed and the choice of which to use is essentially down to the requirements of the user and the data that is available. We summarise the advantages and disadvantages of various clustering analysis techniques in Table 3.3.

As an aside, it is interesting to contrast the BAO method used here with the Alcock-Paczynski test. The latter does not utilise a standard length-scale that is pre-known, but rather compares quantities parallel and perpendicular to the line-of-sight. Physically, it is sensitive to distortions in dx/x. This is essentially the difference between the solid and the dashed lines in Fig 3.19, where $xdx'/x'dx - 1 \approx dx'/dx - x'/x$ for small distortions.

3.5 Discussions and Conclusions

Redshift distortions produce a strong effect on projected clustering measurements — one that is far stronger than the redshift-space distortion effect on the 3D clustering signal for galaxy samples with low bias and a narrow radial window. It is clear that redshift distortion effects must be included when modelling the projected galaxy clustering in redshift slices.

If we consider the apparent motion of galaxies as we move from real- to redshift-space, then redshift-space distortions cause an apparent coherent motion of galaxies into and out of samples. This is true whether samples have sharp boundaries, or if the selection function changes more gradually with distance. In fact, we have argued that such motion does not in itself alter the projected correlation function — we would recover the real-space projected correlation function if we could correct for the movement of the boundary (i.e. allow for the depth of the survey to change with the distortions). However, this is not easy to do, although it is theoretically possible and is an interesting alternative approach. The effect of redshift-space distortions is due to the redshift-space boundaries themselves having an angular clustering signal, and their correlation with the overdensity field. We can alternatively view the effect from a Lagrangian standpoint, where we have to consider that the projection does not remove redshift-space effects from the anisotropic correlation function.

We have used Hubble Volume simulations to show that the projected correlation function can be modelled most easily by integrating the redshift-space correlation function over the radial selection function. Galaxy selection will often be a mix of real and redshift-space constraints, and we have shown that this can be modelled by splitting the population into samples that can be considered to have top-hat windows in either real-space, redshift-space or a hybrid of the two. In the hybrid situation, the projected correlation function can be modelled using both the real-space and redshift-space correlation function over the radial selection function, and that more complicated selection functions can be effectively modelled in a similar manner. Prior to the publication of this work (Nock et al., 2010), no-one has considered how these hybrid selection functions affect the recovered projected clustering signal.

We have presented a new measurement technique, *pair centre* binning, and shown that it minimises the effects of redshift space distortions. In this new scheme, we only include galaxies where their apparent *pair centres* lie within a given radial bin, whereas traditional methods select pairs where both galaxies lie within the bin. The new scheme includes individual galaxies that lie *outside* the traditionally applied top-hat boundaries. This simple modification acts to reduce the effect of the coherent movement of galaxies between slice boundaries on projected correlation function clustering analyses. It is important to note that this new technique does not *prevent* the movement of galaxies between slices; redshift-space distortions due to peculiar velocities will always exist in the radial direction. It simply makes sure that they do not produce a coherent effect on the measurements.

There are two potential disadvantages of the pair-centre binning scheme. One is the fact that the same galaxy may be included in multiple radial bins — thus introducing a correlation between radial bins. Another is the fact that such a scheme results in necessarily wider radial bins, which causes the clustering signal to be diluted. We do not feel that either is a large problem. Applying the more traditional top-hat binning scheme to photometric surveys necessarily results in overlapping radial bins (due to photometric redshift errors) and there will always be considerable covariance between radial bins selected with photometric redshifts — we do not think that pair-centre binning will make this problem considerably worse. The dilution effect can be mitigated by imposing a maximum separation between the pairs included in a pair-centre bin: we call this constrained pair-centre binning. Imposing such a constraint increases the expected signal while not causing a significant change in the effects of redshift-space distortions. More detailed studies of these effects are warranted, but we are confident that the reduction in the redshift distortion effect we observe when utilising pair-centre binning will make this scheme considerably preferable to a top-hat binning scheme.

Pair-centre binning completely removes the effect of redshift distortions when given a uniform galaxy distribution. Such perfect distributions do not exist — most galaxy samples selections are based on an apparent magnitude limit — and thus realistic radial distributions of galaxies are more complicated. However, we have argued that if galaxy samples selected based on an apparent magnitude limit are cut back so that no k-corrected galaxies are missing from the sample, then this does not matter: the boundaries of the bins are either in real-space, or based on pair-centres, neither of which introduces redshift distortion effects.

We have argued, and it is clear from previous work, that any interpretation of projected clustering measurements must account for redshift space distortions. In fact, comparing correlation functions calculated using different binning schemes might actually prove to provide a mechanism for measuring the amplitude of the redshift-space distortions. This is beyond the scope of this thesis, and we leave this for subsequent work.

A brief consideration of how accurately the projected correlation function can provide constraints on the dark energy equation of state has been shown, following the methodology of Blake & Glazebrook (2003). We found that the lack of a radial component in the projected correlation function acts to reduce the level of precision required in the measurement of our standard ruler by $\sim 0.2\%$ for $\Delta w = 0.1$ at a redshift $z \sim 1$, compared to a full 3D analysis. This is not a significant gain and suggests that the use of the projected correlation function is only preferable to the 3D correlation function where the radial clustering signal is sufficiently inadequate; for example, if the level of uncertainty on redshift measurements is high. The advantages and disadvantages of using various alternative clustering analysis techniques have been presented.

Chapter 4

Impact of Photometric Redshift Systematics

The negative effects of photometric redshift uncertainties on clustering analyses are well known and discussed widely throughout the literature. We have already provided a solution to overcome redshift-space distortion effects for future photometric redshift surveys by means of a new pair-centre binning scheme for projected 2D clustering analyses in redshift slices (see Chapter 3). In this chapter, we investigate additional systematic uncertainties arising in the 3D clustering signal due to different photometric redshift estimation techniques. We do this empirically by comparing the recovered clustering signal for a single data-set with redshifts obtained from various estimation techniques via spectroscopy and photometry.

4.1 Photometric Redshifts

There are two main techniques used to calculate photometric redshifts:

- Template Fitting from Spectral Energy Distributions (SED): A model fitting technique that fits empirical or theoretical template spectra to observed photometric SEDs, with an efficiency based upon overall shape and strong spectral properties.

- Spectroscopic Training Sets: An empirical training method that derives a relationship between magnitudes and redshifts using a subsample of objects with measured spectroscopic redshifts.

The template fitting technique works by matching target galaxies to a set of reference SEDs that cover a range of galaxy types, luminosities and redshifts, specific to the sample for which photometric redshifts are required. The photometric redshift of any given target corresponds to the template that minimises the χ^2 between the template and the actual magnitudes. Typically, small sets of SEDs that represent a variety of galaxy types at redshift $z = 0$ form the basis for a set of template spectra. These SEDs are then usually redshifted by hand to create a discrete sampling along the redshift axis.

The ability of template fitting methods to produce accurate photometric redshift estimations depends strongly on how well the template spectra represent the target population. For example, high redshift populations will be poorly matched to low redshift templates, and vice-versa. An improvement in accuracy may be gained by an increase in number of templates used, or by matching templates to target populations more carefully, ie. so that they are more fully representative. An advantage of using empirical template spectra is that they give by definition a physically consistent picture of real galaxies. However, at earlier cosmological epochs evolution could play a substantial role in changing both morphological and spectrophotometric properties of distant galaxies. Since theoretical models try to take galaxy evolution into account, the use of them may be preferable. Whichever the user decides to utilise, the success of the template fitting method can be tested with a comparison of photometric and spectroscopic redshifts obtained on a restricted sub-sample of bright objects (Bolzonella et al., 2000). This method combined with a Bayesian marginalisation introduces a prior probability on template selection (Benitez 2000, Edmondson et al. 2006). Overall, with careful selection of templates, this method can be very successful at providing accurate photometric redshift estimates.

A more empirical approach is to derive a parameterisation for the redshift as a function of the photometry. The most accurate parameterisations are derived from large and representative training samples that have both photometry and measured spectra. Typically, the redshift polynomial is expressed as a function of galaxy colours and the coefficients are adjusted to optimise the fit between the predicted and measured redshifts. By applying the optimised function to the colours of target galaxies with no spectra, an estimate for the photometric redshift can be acquired (Connolly et al., 1995). It has been shown by Collister & Lahav (2004) that a successful way to parameterise the redshift-photometry relation is via the use of Artificial Neural Networks (ANNs). Alternatively, parameterisations may be derived via template spectra or simulated catalogues (Vanzella et al., 2004).

A stringent requirement for the training sample is that its photometry must have the same filter set and noise characteristics of the target sample. When applied to targets with redshift ranges and spectral type that have been adequately sampled by the training population, this method will generally provide more accurate photometric redshift estimates than the template fitting method. However, the current unavailability of high redshift spectra means that extrapolation beyond the limits of the training set is unreliable.

4.2 Photometric Redshifts in Clustering Analyses

Photometric redshift uncertainty only affects inferred distances in the radial direction. The effect of this on the recovered correlation function is a damping of power on small scales and a smearing of signal such that the BAO peak cannot be detected. Predictions of the level of damping expected for future surveys can be made by considering an anisotropic model of the correlation function. We can start by expressing the direction dependent 3D 2-point correlation function as

$$\xi(\mathbf{r}) = \frac{V}{(2\pi)^3} \int d^3k\ P(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{r}}. \tag{4.1}$$

Using the identity

$$e^{-i\mathbf{k}\cdot\mathbf{r}} = \cos(\mathbf{k}\cdot\mathbf{r}) - i\sin(\mathbf{k}\cdot\mathbf{r}), \tag{4.2}$$

and the fact that the correlation function is real, we can re-write Eq. 4.1 as

$$\xi(\mathbf{r}) = \frac{V}{(2\pi)^3} \int d^3k\ P(\mathbf{k}) \cos(\mathbf{k}\cdot\mathbf{r}). \tag{4.3}$$

Changing to spherical harmonic coordinates (r, θ, ϕ) with $\theta = 0$ along the direction of $\mathbf{r}$ gives

$$d^3k = k^2 \sin(\theta) dk\ d\theta\ d\phi, \tag{4.4}$$

where $0 < r < \infty$, $0 < \theta < \pi$ and $0 < \phi < 2\pi$. Substituting this into Eq. 4.3 and solving for ϕ leaves us with a 2D anisotropic integral such that

$$\xi(r,\theta) = \frac{V}{(2\pi)^2} \int d\theta \int dk\ P(k) \cos(rk_z) k^2 \sin(\theta) \tag{4.5}$$

where $k_z = k\cos(\theta)$ is the radial component of the wavevector k.

Assuming that the photometric redshift uncertainty has a Gaussian distribution, we can say that the radial comoving coordinate x of each galaxy is smeared by an amount δx sampled from a probability distribution of the form

$$f(\delta x) \propto exp\left[-\frac{1}{2}\left(\frac{\delta x}{\sigma_x}\right)^2\right]. \tag{4.6}$$

In practice, we can choose a photometric redshift uncertainty σ_z and obtain σ_x via the coordinate transform equation (Blake & Bridle, 2005)

$$\sigma_x = \delta x \frac{dx}{dz}(z_{eff}) \tag{4.7}$$

$$= \sigma_z(1+z_{eff})\frac{c}{H(z_{eff})}, \tag{4.8}$$

where z_{eff} is the effective redshift of the survey and $H(z) = H_0\sqrt{E(z)}$ is the Hubble constant measured by the observer at redshift z.

The galaxy number distribution is smeared along the radial direction as predicted by the photometric redshift error function. According to the convolution theorem, the resulting power spectrum signal is damped along the radial direction such that

$$P(k_x, k_y, k_z) \rightarrow P(k_x, k_y, k_z) \times \exp[-(k_z\sigma_x)^2]. \tag{4.9}$$

Note that the Fourier transform of a Gaussian is another Gaussian. There is a phase term, corresponding to the position of the centre of the Gaussian, and then the negative squared term in an exponential. The standard deviation has also moved from the denominator to the numerator. This means that as a Gaussian in real space gets broader, the corresponding Gaussian in k-space gets narrower, and vice versa. That is, as the Gaussian in real space gets broader, contributions from points within that Gaussian start to interfere with each other at lower and lower resolutions. Convolution with a Gaussian will shift the origin of the function to the position of the peak of the Gaussian, and the function will be smeared out.

In Fig. 4.1 we show the effects of uncertainty on radial distance measurements in the 3D correlation function. We use a grid based method to calculate the correlation function, $\xi(\sigma, \pi)$, for 5×10^9 particles in the MICE simulations (ΛCDM cosmology with $\Omega_m = 0.25$), where π and σ are components of pair separation measured parallel and perpendicular to the line-of-sight, respectively. The left panel shows the true correlation function recovered when there are no uncertainties on the measured radial distances. The baryonic ridge can be clearly detected at $\sim 100\ h^{-1}\,\mathrm{Mpc}$ in both planes. The introduction of uncertainties on radial distance measurements act to *wash-out* the BAO signal in the radial direction as shown in the middle and right panels, where typical photometric redshift uncertainties of $\sigma_z = 0.01(1+z)$ and $\sigma_z = 0.03(1+z)$ have been considered.

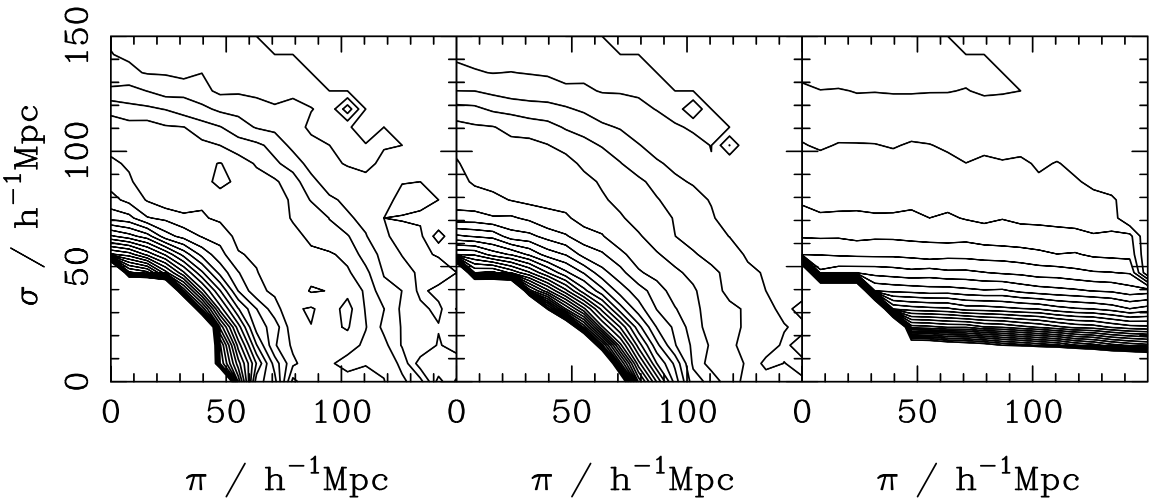

Figure 4.1: Contours show the amplitude of the correlation function, $\xi(\sigma, \pi)$, calculated on a coarse grid for 5×10^9 particles in the MICE simulations, where π and σ are components of pair separation measured parallel and perpendicular to the line-of-sight, respectively. The left panel shows the true correlation function recovered when there are no uncertainties on the measured radial distances. The baryonic ridge can be detected at $\sim 100\ \ h^{-1}\,\mathrm{Mpc}$ in both planes. The middle and right panels show the effect of typical photometric redshift uncertainties on radial distance measurements for $\sigma_z = 0.01(1+z)$, and $\sigma_z = 0.03(1+z)$, respectively. The BAO signal is washed out in the radial direction in both cases.

4.3 Spectroscopic Redshift vs. Photometric Redshift: Comparison of $\xi(r)$ for SDSS S82

Existing models describe *how* photometric redshift uncertainties affect 3D clustering analyses, but they do not consider additional systematic uncertainties arising from differing photometric redshift estimates. In order to better understand these systematic uncertainties arising in the correlation function we can perform an empirical test that contrasts recovered clustering signals coming from a variety of different photometric redshift estimators. This simple test provides a general approach to better understanding these errors using real data-sets, allowing us to separate systematic uncertainty from statistical uncertainty. Hence, we start by constructing a data catalogue from the SDSS Stripe 82 sample that contains 3 photometric redshift measurements with a complimentary set of spectroscopic redshift measurements. As well as this, we generate a further 2 sets of photometric redshift estimations via template fitting and neural network training methods. This gives us a total of 5 different photometric redshift estimators with which we may work. Details of how these 5 estimators were calculated are given below.

Throughout this analysis we assume that the spectroscopic redshift measurements within the sample have negligible uncertainty. Although this is not entirely true it is a fair assumption, since uncertainties on spectroscopic redshift measurements are typically $\sim 100\times$ smaller than those for photometric redshift measurements, and we only wish to contrast photometric redshift measurements. Therefore, we refer to any results obtained using spectroscopic redshifts as *true*.

4.3.1 The Sloan Digital Sky Survey and Stripe 82

The Sloan Digital Sky Survey ran from 2000-2008. During this time, $\sim 10,000$ deg^2 of sky was observed to obtain deep, multi-colour images in 5 passbands (*ugriz*) for $\sim 930,000$ galaxies and $\sim 120,000$ quasars in total. Located at Apache Point Observatory, New Mexico, a dedicated 2.5-meter telescope, equipped with a 120-megapixel camera and a pair of spectrographs fed by optical fibers, measured spectra of more than 600 galaxies and quasars in a single observation.

Photometrically observed galaxies in the SDSS were targeted for spectroscopic follow-up in 2 ways:

- Main Galaxy Sample: The main galaxy sample target selection algorithm (Strauss et al., 2002) targeted galaxies brighter than the r-band Petrosian limit $r = 17.77$, resulting in a sample surface density of ~ 90 per deg^2.

- LRG Sample: LRGs were selected according to colour and magnitude to yield an intrinsically luminous and red sample that extends beyond the main galaxy sample (Eisenstein et al., 2001). A cut at $z \sim 0.4$ was introduced due to the movement of the $4000\AA$ break from the g to r band at this redshift.

For the purpose of this analysis we only consider galaxies from the main sample, since it spans a wide colour and luminosity space similar to that proposed for the DES. In particular we construct a data-set from the Stripe 82 (S82) database. The S82 database contains all imaging from SDSS S82 along the Celestial Equator in the Southern Galactic Cap. The addition of Data Release 7 (DR7) data means that the S82 database now includes a total of 303 runs (plus 2 coadds), covering any given piece of the ~ 270 deg^2 area (where $-50 <$RA< 59 and $-1.25 <$DEC< 1.25 deg) approximately 80 times.

4.3.2 Data Catalogues

The raw galaxy catalogue is taken from the S82 M-table generated from all spectroscopic galaxies in standard DR7 using the CasJobs neighbours feature, which contains all neighbouring objects in S82 within some short distance. We pick out all objects in the S82 PhotoObjAll table P, restricting selections to coadd runs 106 & 206. Note that coadding is the stacking of 50 scans into one image. The photometry code (Edmondson et al., 2006) is run on the stacked image. It gives a catalogue that runs about 2 magnitudes deeper than the non-coadds with smaller photometric error - the latter being the relevant part here. We are left with huge catalogue with multiple matches around each galaxy. Firstly, the entire table is sorted by single scan object ID, then by the separation of 2 matches. The latter selects the best match for each galaxy leaving a nearest neighbour table.

We want to compare the clustering signal we obtain for a single data-set using various different redshift estimators. The SDSS CasJobs database provides us with a spectroscopic and three photometric redshift estimates. In addition to this we obtain two further photometric redshift estimates via template fitting and neural network training techniques. In total, we have six redshift estimates for the single data-set. Each redshift estimation technique is described in more detail below.

4.3.3 Estimating Photometric Redshift Uncertainties for S82

Estimates of the uncertainty on our photometric redshifts can be obtained by considering their scatter around the spectroscopic sample (which we now label A). In this section we define our photometric samples.

SDSS Photometric Redshifts

There are two main sources of photometric redshift estimations in the SDSS CasJobs database:

- Photoz table - Photometric redshifts are calculated via a hybrid method, which uses a combination of neural network training and template fitting. The DR7 spectroscopic redshift set (including Main, LRGs, special photometric redshift plate survey of high redshift non-LRGs, low redshift plates) is used as a reference set for redshift estimation. This sample contains $\sim 700,000$ galaxies spanning the entire colour region, which means that no synthetic spectroscopic redshifts are required, or used. The estmation method searches ubercal *ug, gr, ri, iz* colour space for k nearest neighbours of every object in the estimation set. Each redshift is estimated by fitting a local low order polynomial on these points. Neighbours with redshift values too far from the fitted hyperplane are excluded. Roughly 5% of outliers are a result of extrapolation and should not be trusted. K-corrections, distance modulus, absolute magnitudes and rest-frame colours are calculated by combining the neural nets method with template fitting. A search for the best match of measured and synthetic colours from empirical template spectra at a given redshift is conducted from a local nearest neighbour fit. This also gives an estimate of the spectral type of the object. Error propagation from magnitude errors does not give a reliable estimate of redshift errors. Instead, when fitting the linear polynomial, the mean deviation of redshifts for the reference objects are calculated (see Budavari et al. 2001 and Csabai et al. 2003 for details).

- Photoz2 table - Photometric redshifts are calculated via a pure neural network training method. The photometric redshift training and validation sets consist of $\sim 551,000$ unique spectroscopic redshifts mathched to $\sim 640,000$ photometric measurements. The spectroscopic training set comes from a plethora of sources including: SDSS, 2dF-SDSS LRG and QSO (2SLAQ), Canada-France Redshift Survey (CFRS), Canadian Network for Observational Cosmology 2 (CNOC2), Team Keck Treasury Redshift Survey (TKRS), Deep Extragalactic Evolutionary Probe (DEEP1 and DEEP2). Photometric redshift estimates are obtained for ~ 77.4 million DR6 primary objects, classified as galaxies by the SDSS PHOTO pipeline (TYPE=3), with dereddened model magnitudes $r < 22$, and with none of the flags BRIGHT, SATURATED, SATUR_CENTER set (Oyaizu et al., 2008). There are two regimes used:

  - D1 - This technique uses *galaxy magnitudes* in the photometric redshift fit, as well as a concentration index (ratio of PetroR50 and PetroR90). A smaller photometric redshift error is obtained than the CC2 method (see next point), and is recommended for bright galaxies $r < 20$ to minimize overall photometric redshift scatter and bias.
  - CC2 - This simple technique uses *galaxy colours* (magnitude differences), as well as a concentration index. It provides larger photometric redshift errors than D1, but is recommended for faint galaxies $r > 20$ as it gives more accurate photometric redshift distributions.

Fig. 4.2 shows the scatter between spectroscopic and photometric redshifts for all SDSS galaxies, where panels from left to right represent hybrid, CC2, and D1 photometric redshift estimation techniques, respectively. All three of these photometric redshift estimates coming from the SDSS CasJobs database are used in our analysis. We label each sample B, C and D, for the hybrid, CC2 and D1 estimates, respectively.

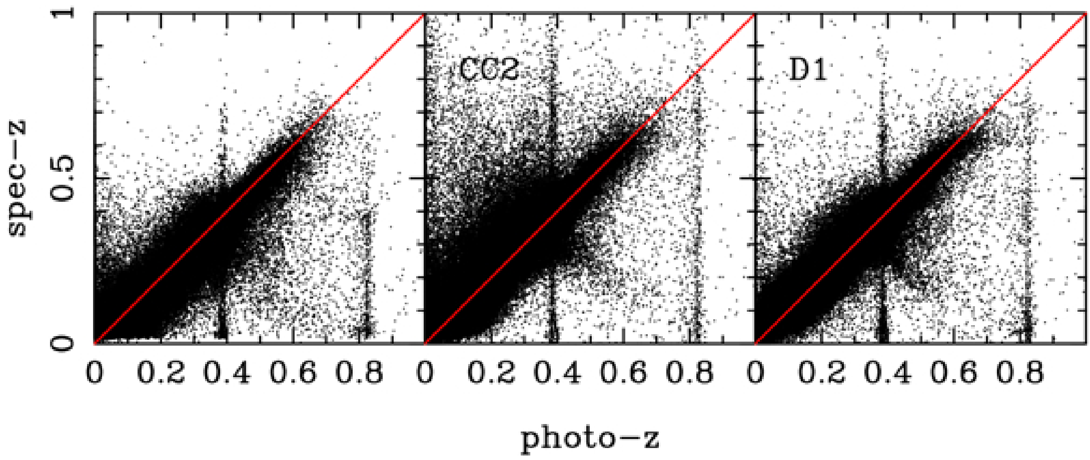

Figure 4.2: Panels show the scatter between distances derived from photometric redshifts and their corresponding true distances for all SDSS photometric redshifts.

Neural Networks

We have used the public software package *ANNz* (Firth et al. 2003, Collister & Lahav 2004) to calculate a set of photometric redshift estimates for our galaxy catalogue (labelled E), using *ugriz* filters and the spectroscopic redshifts from sample A as a training set. Although the training set is not particularly large, it is entirely representative since it contains only the galaxies that we wish to test. A quarter of the training set was selected randomly to become a validation set, which is required for the optimisation of the neural network generalisation performance. To maximise the overall accuracy of our photometric redshift estimates we created a committee of three networks for use in the testing phase. The program reported a slightly negative mean deviation between the estimated photometric redshifts and the input spectroscopic redshifts for our committee of three networks. This suggests that on average, ANNz underestimated the photometric redshift. The returned rms deviation from the ANNz program was ~ 0.018, which is much smaller than our calculated value from the final sample ~ 0.054. This discrepancy is most likely caused by the fact that our training set is small, which is a problem highlighted by the authors of the algorithm in Collister & Lahav (2004), and is to be expected.

Template Fitting with Bayesian Priors

Following the methodology of Edmondson et al. (2006) we used Bayesian template fitting techniques to create a final set of photometric redshift estimates. Typically, the Bayesian approach considers a posterior prior of the form $p(z|C,\mathcal{P}) \propto \mathcal{L}(C|z)p(z|\mathcal{P})$ for some set of colours, C, where $\mathcal{P}$ denotes prior knowledge, and $p(z|\mathcal{P})$ is the corresponding prior probability distribution for redshift, z. In this case, the prior information is in the form of a galaxy luminosity function, ϕ, determined from spectroscopic redshift surveys, giving

$$p(z|\mathcal{P}) = p(z|m, S, \phi(m, S, z)), \tag{4.10}$$

where m corresponds to galaxy magnitudes and S is the spectral type.

We used COMBO 17 luminosity functions (Wolf et al., 2003) calculated for 3 broad galaxy types, and galaxy spectral models from PEGASE stellar population synthesis models. Template wavelengths were shifted and passed through SDSS filter models to obtain colour templates. In total, 360 SEDs were shifted over 177 intervals equidistant in $\log(1+z)$ for the redshift range $0 < z < 1.4$. Templates were fit according to template flux ratios by evaluating the equation

$$p(z, S|f, \mathcal{P}) \propto \int_0^\infty p(z, S|\mathcal{P}) \exp\left[-\frac{1}{2}\left(\frac{\mathbf{f} - \hat{\mathbf{f}}}{\boldsymbol{\sigma_f}}\right)^2\right] d\hat{f}_R, \tag{4.11}$$

where $\mathbf{f}$ are observed fluxes and $\hat{\mathbf{f}} = \hat{f}_R \hat{\boldsymbol{\lambda}}$ are template fluxes (equivalent to flux ratios). The results were then marginalised over the SED range, S, to obtain estimates for the photometric redshift.

Scatter

As a simple initial test of the robustness of the photometric redshift estimates we plot their derived distances as a function of their true distances in Fig. 4.3. We assume each sample has Gaussian statistical properties and calculate their mean and standard deviations accordingly. A random selection of error bars are plotted. Inset we plot the 1D distribution of distance errors and highlight the position of the mean. Table 4.1 shows the mean, μ_d, standard deviation, σ_d, sample minimum and maximum galaxy positions, and percentage of galaxies lying within 1σ of the mean for each sample. The latter measurement provides us with a simple test of non-Gaussianity; if the percentage of galaxies within 1σ of the mean is not equal to 68% there must be outliers in the tails of the distribution that are becoming statistically important. This makes sense when we consider the range of distances each individual sample covers, that is, the larger the deviation from the true sample, the larger the deviation from Gaussianity. Panels B, C, and D represent samples with photometic redshifts coming from the SDSS CasJobs database. They all have very similar scatter and errors, which is to be expected as they have all been created from the same reference set of galaxies from DR7. Sample E, which has been created via a neural network fit using the true sample as a training set, has the least scatter and

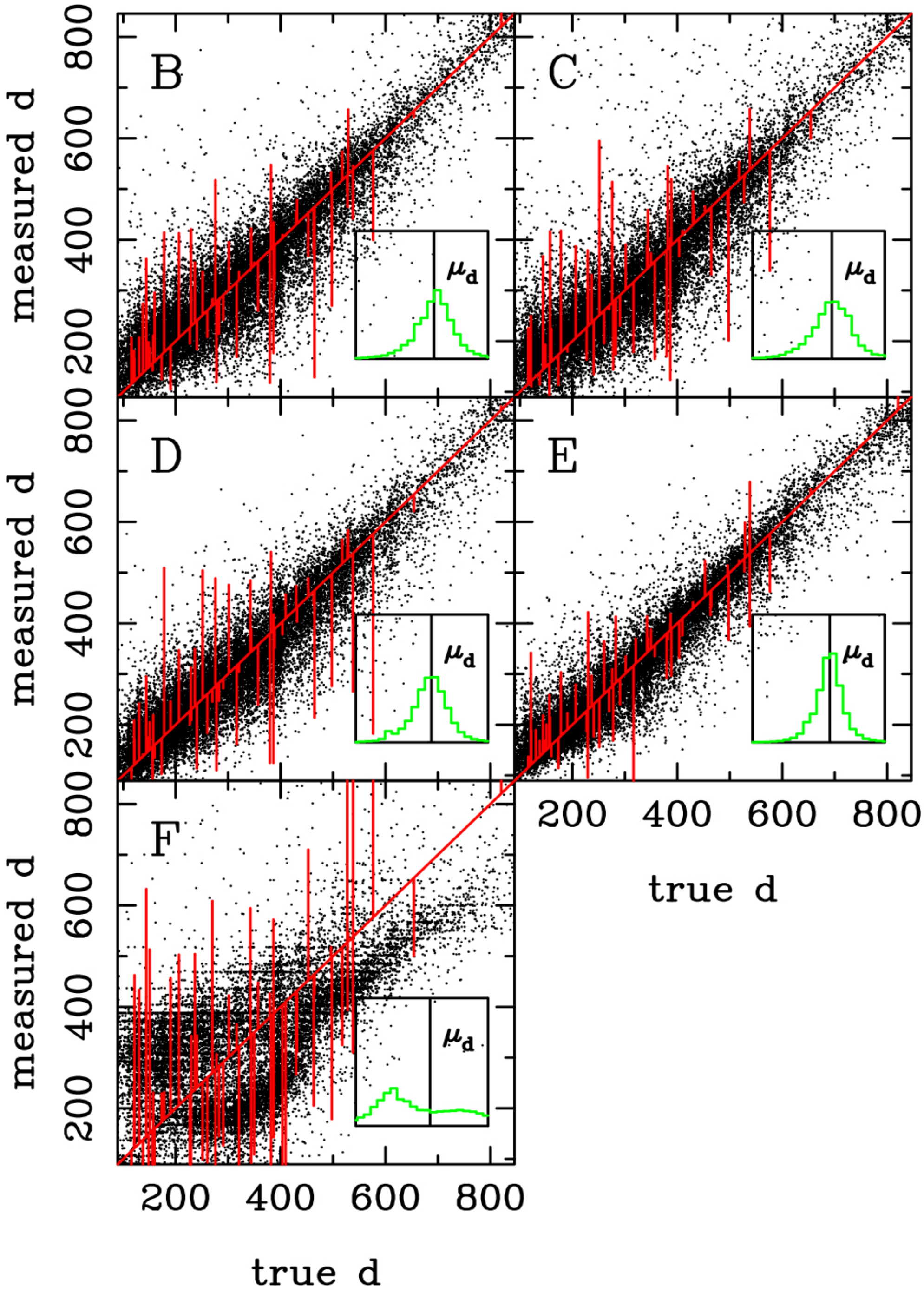

Figure 4.3: Panels show the scatter between distances derived from photometric redshifts and their corresponding true distances. A random selection of error bars are plotted, calculated assuming Gaussianity. The 1D error distribution is plotted inset, where μ_d is the mean galaxy position per sample. Table 4.1 details the mean, standard deviation, minimum and maximum galaxy positions and percentage of galaxies lying within 1σ of the mean of each sample that correspond to each panel. Details of each sample can be found in the text.

	A	B	C	D	E	F
μ_d	334.559	336.957	339.911	328.809	333.474	324.845
σ_d	0.0	156.143	169.617	154.609	151.818	238.623
min d	89.4894	51.695	34.8376	32.7495	63.5391	0.0
max d	845.976	1291.71	2373.51	1172.79	872.967	3046.39
$-\sigma_d < d < \sigma_d$ (%)	n/a	72.1273	74.8845	70.2477	69.1936	82.2005

Table 4.1: Table of statistical properties for distances derived from each redshift sample, calculated assuming Gaussian statistics. Sample A represents the true distance distribution and the properties of the other samples are calculated accordingly.

smallest errors. Again, this is expected since we have trained on the exact set with which we are comparing. Note that although this error is larger than that calculated by the ANNz program with which the catalogue was created, we will use this estimate of the error throughout the analysis. Sample F, that was created via a template fitting technique, has a terrible fit to the true data. There is a clear bi-modality in the scatter. This may be as a result of using the wrong luminosity prior in the template fit. That is, galaxies are assumed to be brighter than they actually are and so are shifted to higher volumes. In this particular case a COMBO 17 (Classifying Objects by Medium Band Observations – a spectrophotometric 17-filter survey) luminosity prior was used, which is set up for a small area, deep survey, and as such struggles to fit low redshift galaxies. This overestimation of photometric redshift at small z forces the fit to correct around a mean value for the rest of the sample, thus underestimating the photometric redshift at larger z. The use of an optimised SDSS template would no doubt correct for these effects, but we did not undertake this task in this analysis. Because of these problems we will not use sample F in any further analyses.

4.3.4 Random Catalogue

In order to estimate the correlation function accurately we need to compare the clustering of galaxies to that of a random field. It is crucial to correct for any artificial clustering that may be introduced into a random catalogue via galaxy survey selection effects. The main issues to consider include: angular completeness, evolving radial number density and galaxy bias. If we consider all of these effects in the construction of our random catalogue, we should recover an accurate clustering signal. Details of how we achieve this are given in the following section.

Angular Selection

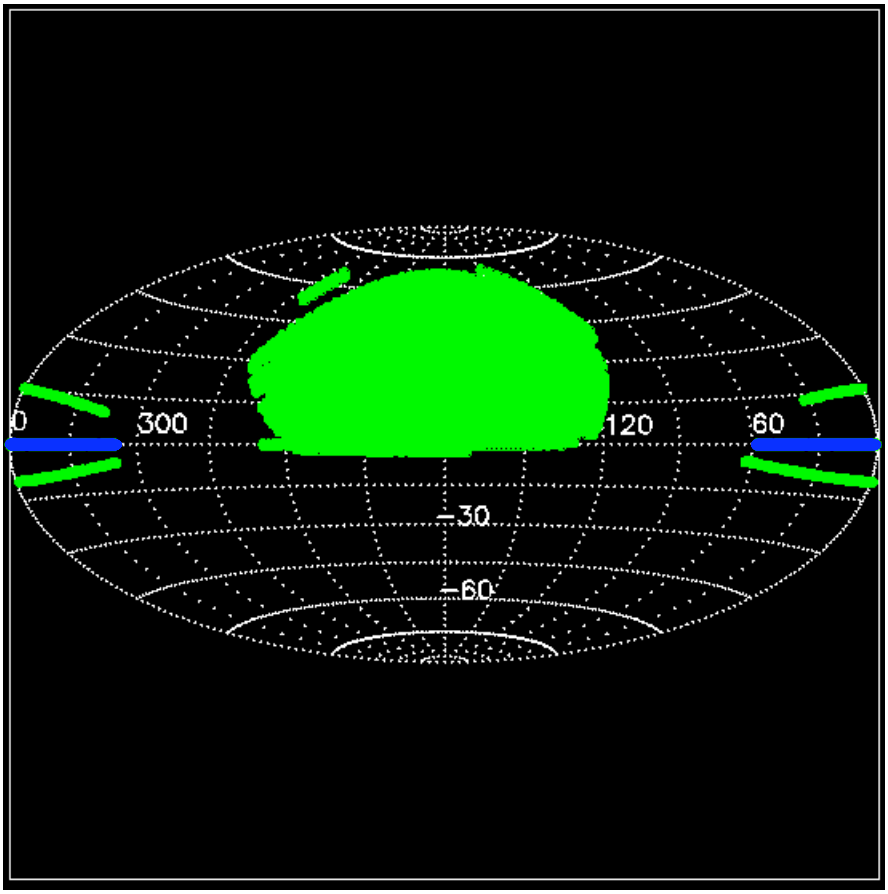

Figure 4.4: Aitoff projection of the spectral sky coverage for DR7 of the SDSS. Stripe 82 is highlighted in blue. This is where the sample we are using is drawn from. Credit: `http://www.sdss.org/DR7/`.

Although the six different redshift estimators we have selected will alter the radial distribution of each sample, the angular selection will be the same across all samples. This angular selection was created via a HEALPix decomposition of the sphere into equal-area pixels. A total of 3145728 pixels were created across the sphere, each of size 0.013 deg^2. This means that each SDSS plate is covered by 532 pixels, and that the S82 sample covers 23910 pixels in total corresponding to an area ~ 313 deg^2. Since the sample covers quite a small angular scale it is possible that the effect of the pixelisation scheme may show up on the resulting correlation function.

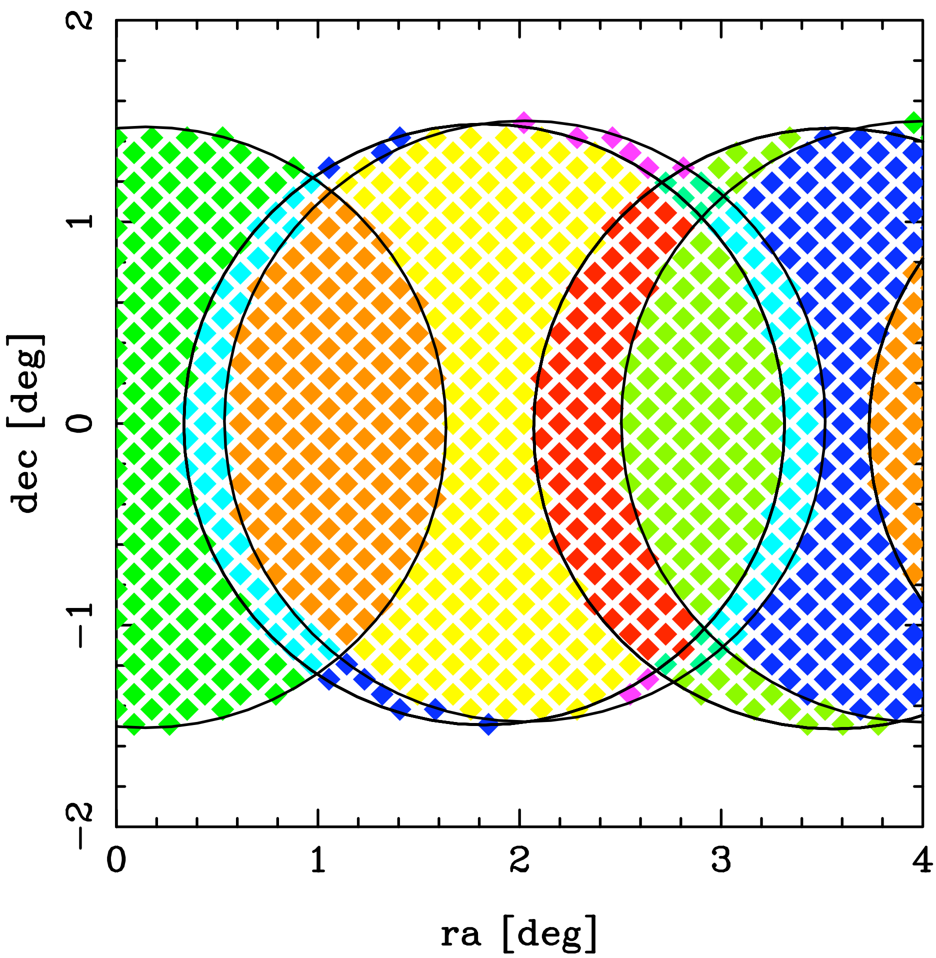

Figure 4.5: A section of the angular mask for the SDSS Stripe 82 sample. Large black circles represent spectroscopic plates. Coloured regions denote groups of galaxies with the same targeting information, ie. plate number, tile number. The angular completeness of each group is calculated by considering the ratio of observed galaxies to target galaxies. Each galaxy is assigned the completeness of the region it resides within.

Firstly, we group pixels according to their spectroscopic targeting information, eg. pixels with the same plate number, tile number etc. Then, the angular completeness of each group is calculated by considering the ratio of spectroscopically observed galaxies to spectroscopically targeted galaxies. Fig 4.6 shows an example region from the S82 sample. Large open circles denote spectroscopic plates. Small circles denote photometrically observed galaxies that have been targeted for spectroscopic follow-up. Circles are filled where the target has been successfully observed. Groups where all targeted galaxies are observed will be assigned a completeness of 1, whereas groups where no targeted galaxies have been observed will be assigned a completeness of 0. Group completeness is then assigned to individual pixels within the group. Individual pixels are grouped together to prevent large statistical fluctuations across the mask.

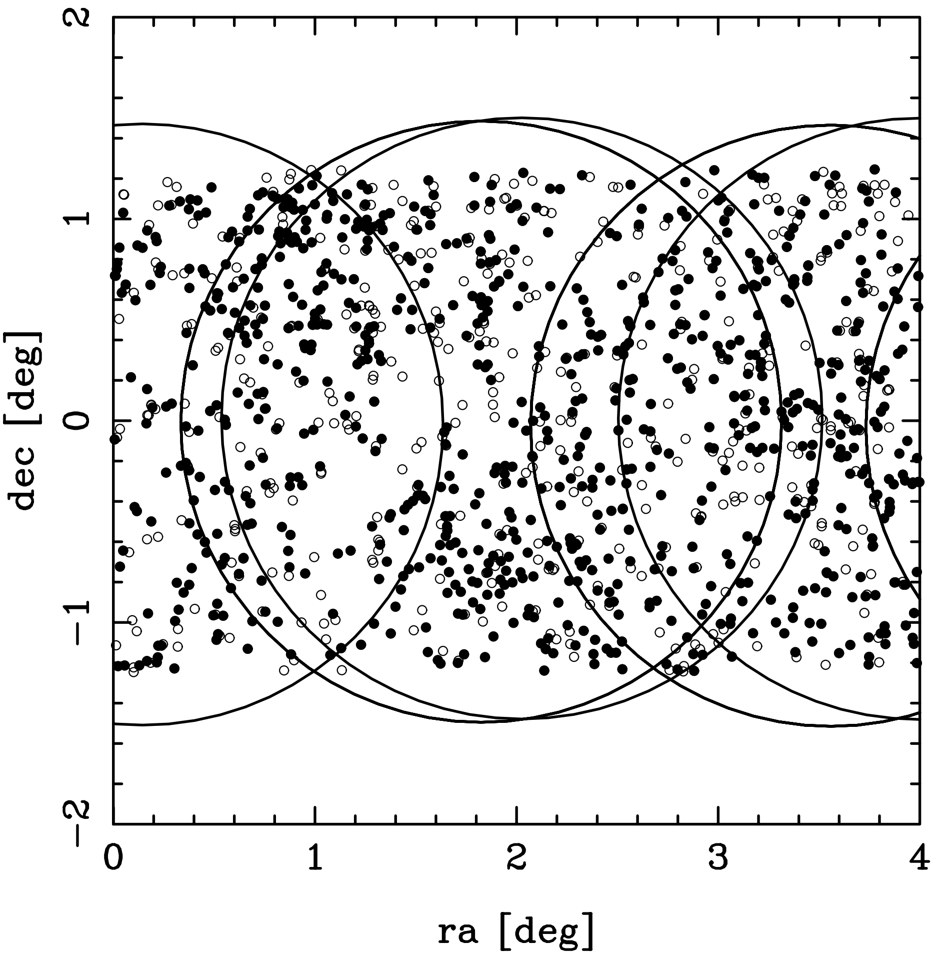

Figure 4.6: A section of the SDSS Stripe 82 sample. Large black circles represent spectroscopic plates that are capable of obtaining spectra for ~ 600 galaxies in a single observation. Galaxies that were observed photometrically and then targeted for spectroscopic follow-up are plotted as open circles. Target galaxies that a good spectra was obtained for are plotted as filled circles. The ratio of observed target galaxies to target galaxies is used to determine the angular mask for the sample.

An example region from the S82 angular mask is shown in Fig. 4.5. Large black circles denote spectroscopic plates. Coloured regions show groups of pixels with varying completeness across the sample. The unclustered random sample, with which we contrast the galaxy sample, is created in accordance with this angular selection function, so that we are not introducing artificial clustering signal due to selection effects.

Radial Selection

The radial selection function of each sample will vary due to the redshift measurement uncertainty. In order to fully quantify the expected clustering signal we need to model the radial selection function of each sample. Since we are dealing with a main galaxy sample, we do this by fitting a model distribution of the functional form

$$f(z) = z^g \exp\left[-\left(\frac{z}{c}\right)^b\right] \tag{4.12}$$

where z is the redshift, and g, b, and c are parameters to be fit. We minimise Eq. 4.12 via Powell's method as described in Press et al. (1992). Best-fit values for each sample are

	A	B	C	D	E
g	13.6097	19.1089	15.8645	11.5627	33.2291
c	7.86194e-07	1.0981e-10	2.91666e-09	1.61991e-06	8.44279e-19
b	0.322868	0.218811	0.243757	0.329565	0.13983

Table 4.2: Table of best fit parameters for the modelling of the radial distributions for each redshift sample, calculated via a Powell minimisation of Eq. 4.12. The smooth model radial distribution is used in the creation of an unclustered random catalogue.

listed in Table. 4.2. Fig. 4.7 shows the normalised radial distributions for each sample. Galaxy, model and random distributions are plotted in black, blue and red, respectively. The effect of uncertainty on photometric redshifts causes the radial distribution to smear out, stretching out over a larger redshift range than that of the true distribution.

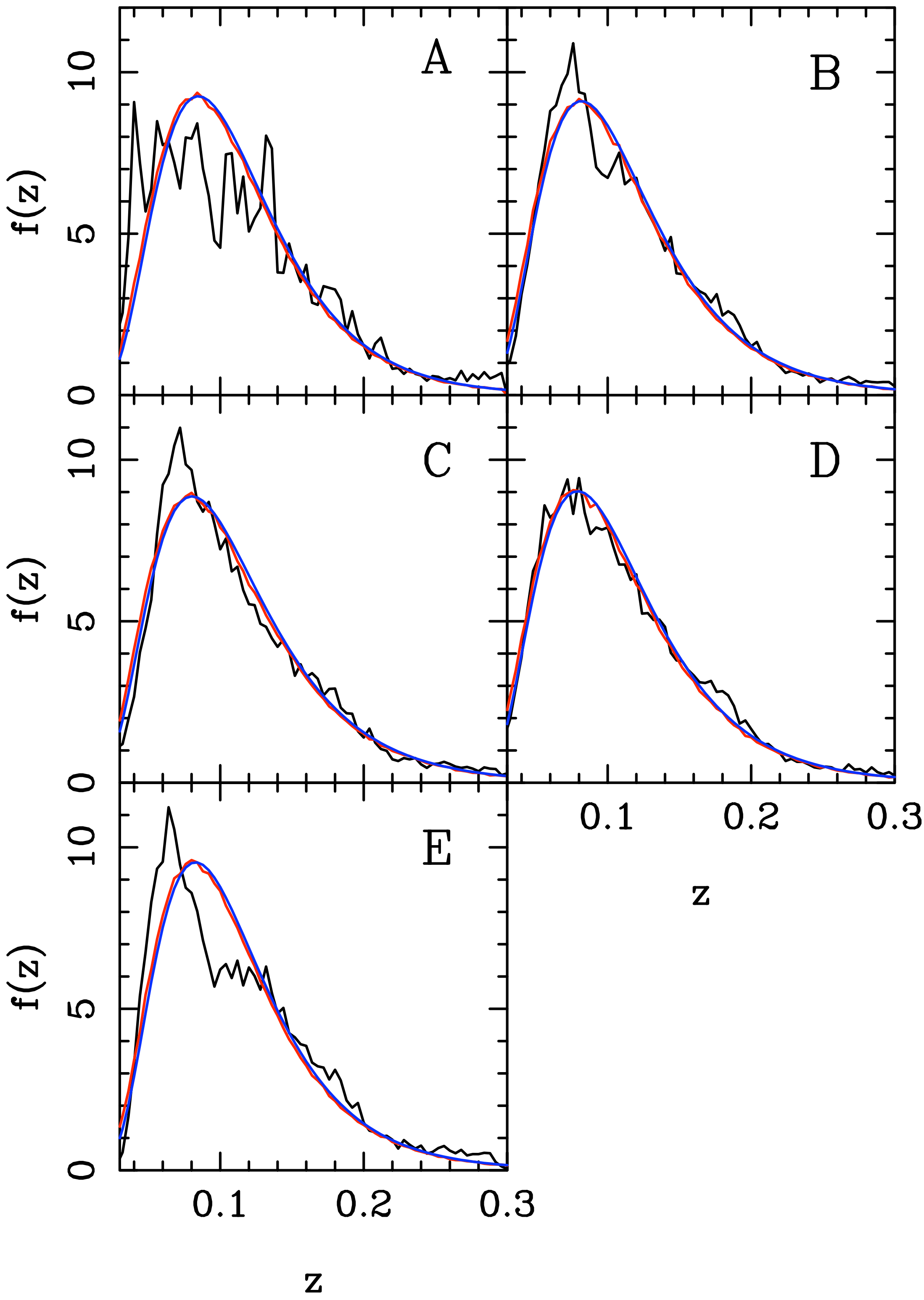

Figure 4.7: Normalised radial distributions for the SDSS Stripe 82 samples as a function of redshift. Three lines per panel represent galaxy (black), model (blue) and random (red) radial distributions, respectively. Model distributions were calculated by minimising the function $f(z) = z^g \exp -(z/c)^b$, where g, c and b are free parameters to be fit. Random distributions were created via a Monte-Carlo sampling of the model distributions with $\sim 30\times$ the number of galaxies. The effect of photometric redshift uncertainty acts to smear the radial distribution.

Galaxy Bias

The astrophysical process of galaxy formation means that galaxies do not trace the underlying dark matter distribution directly. Instead we need to relate the two quantities via a galaxy bias correction term. It is common to assume a local linear bias where $\delta_{gal} = b\delta_{mass}$ leading to a simple power spectrum relationship $P_{gal} = b^2 P_{mass}$.

Galaxy bias is not independent of scale, since galaxies of different type have different clustering strengths. It is however, possible to correct for the effects of galaxy bias in clustering analyses. Percival et al. (2004) show that, given an accurate linear bias model for each type of galaxy in the sample to be analysed, it is possible to multiply the contribution of each galaxy to the estimate of the overdensity field by the inverse of an expected bias, thus removing any systematic offset in the recovered correlation function caused by galaxy bias. The galaxy bias of the SDSS galaxies are best fit with a model proposed by Tegmark et al. (2004) of the form

$$\frac{b}{b_*} = 0.85 + 0.15\frac{L}{L_*} - 0.04^{M_* - M_{0.1r}} \tag{4.13}$$

where b is the galaxy bias, L is the luminosity, $M_* = -20.44$ as defined by Blanton et al. (2003), and $M_{0.1r}$ is the $z = 0.1$ shifted r-band k-corrected absolute magnitude.

The galaxy bias correction term is usually calculated as a function of galaxy redshift and attributed to the random catalogue. In this analysis we have varying redshift ranges and smeared distributions due to the photometric redshift uncertainties, which makes the accurate modelling of galaxy bias difficult. Because of this we do not try to correct for galaxy bias effects here, which is fine since we are conducting a comparative study of the clustering in our samples so any effects caused by galaxy bias will be present in all samples and will not affect our overall result.

Weights

We apply the standard Feldman-Kaiser-Peacock (FKP) (Feldman et al., 1994) weights $w(\mathbf{r})$ to each sample, where

$$w(\mathbf{r}) = \frac{1}{1 + \bar{n}(\mathbf{r})P(k)}, \tag{4.14}$$

$P(k)$ is some input power usually set to $P(k) = 5000\,h^{-1}\,\mathrm{Mpc}$ and $\bar{n}(\mathbf{r})$ is the mean comoving number density. Since $\bar{n}(r)$ is a rapidly decreasing function we have 2 distinct regimes:

- Small r: We have many galaxies per unit volume V, which means that the error on the measured power will be determined by a finite number of realisation volumes. In this case we weight each galaxy equally by volume (ie. $\propto 1/\bar{n}$).
- Large r: We have few galaxies per unit volume and the measured power is dominated by shot-noise. In this case we weight each galaxy equally.

These optimization weights depend on spatial frequency. Power decreases with frequency, and so if we want to measure longer wavelengths we need to give greater weight to the more distant galaxies.

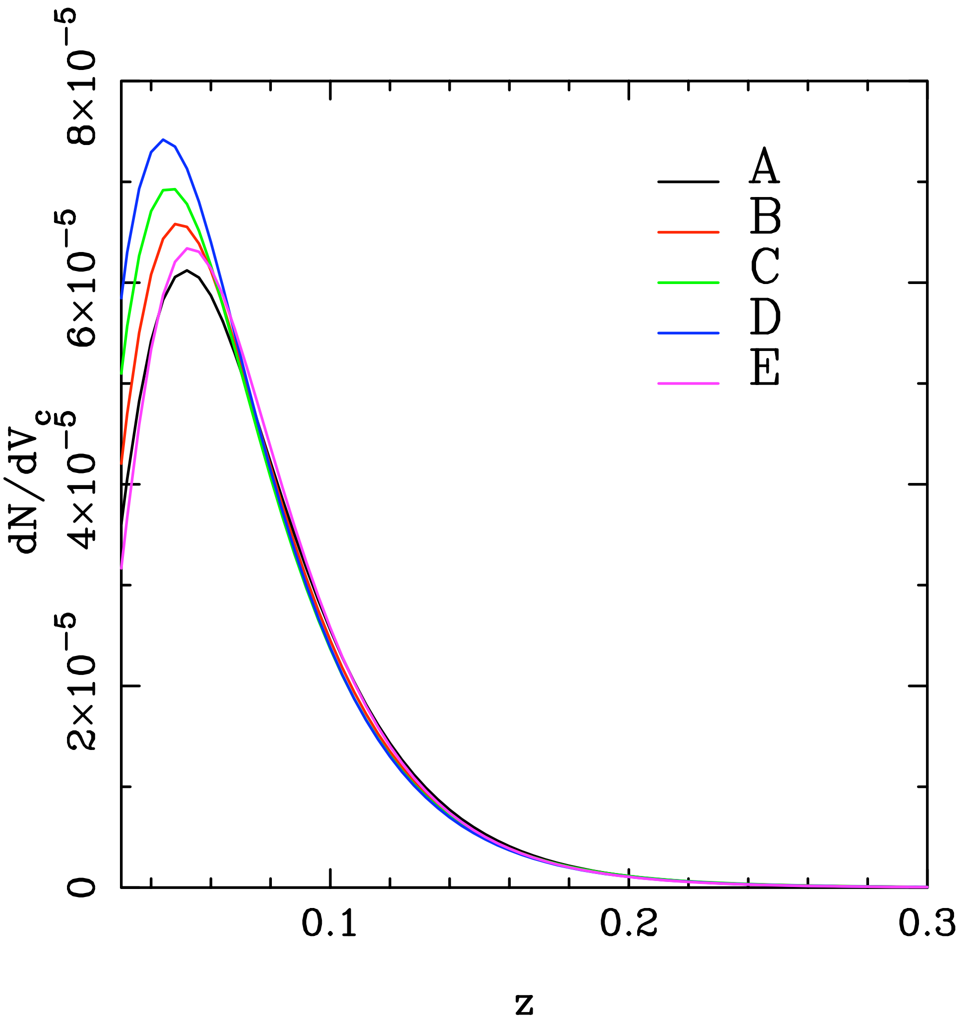

Figure 4.8: The mean comoving number density of galaxies as a function of redshift for the SDSS Stripe 82 samples. Each sample is selected according to a hard limit in spectroscopic redshift, effectively creating a top-hat selection in the radial direction. Uncertainty in the photometric redshift measurements smooth this bin in the radial direction causing the peak comoving number density to shift away from the true value (spectroscopic).

The comoving number density for each sample is plotted in Fig. 4.8. The result of imposing a hard limit on the spectroscopic redshift for each sample means that we are selecting a top-hat bin in the radial direction. The uncertainty on photometric redshift measurements effectively smooth this top-hat. This effect is evident here as we see a shift in the peak comoving number density for the various photometric redshift estimates. This will change the weights attributed to each galaxy sample.

4.3.5 Modelling

In section 4.2 we introduced a model to describe the effect of photometric redshift uncertainties on the correlation function. In practice, we calculate the expected damping of the clustering signal via a Monte Carlo integration over pairs of points.

We start by randomly sampling pairs according to their 3D separation r from a distribution $r^3 \in U[d^3_{min} - d^3_{max}]$. The true correlation function, ξ_{true}, is calculated per pair of points at r via a spline fit to an input 3D spherically averaged correlation function model. In this case this input model is calculated using the transfer function of Eisenstein & Hu (1998). The positions of each point in the pair are determined for random realisations of the azimuthal, θ, and zenith, ϕ, angles in a spherical coordinate system, where $\theta \in U[0 - 2\pi]$ and $cos(\phi) \in U[0 - 1]$. We then convert to Cartesian coordinates via the standard transformation where

$$x = r\ cos(\theta)sin(\phi) \tag{4.15}$$

$$y = r\ sin(\theta)sin(\phi) \tag{4.16}$$

$$z = r\ cos(\phi). \tag{4.17}$$

Radial positions are taken to be along the z-axis in the Cartesian framework. Galaxy positions are smeared in the radial direction via a random Gaussian deviation σ_0, where $\sigma_0 \in N[0 - \sigma_d]$. The true value of the correlation function is binned at the new pair separation r_{photo}.

Hubble Volume Tests

To test the validity of our method, we create a toy photometric data-set using the Hubble Volume simulations (Evrard et al., 2002). We incorporate uncertainties into the galaxy positions along one of the box axes using Eq. 4.7, where $\sigma_0 = 0.0$, 0.01, 0.02, 0.03, 0.04 and 0.05, to obtain six different estimates of the correlation function. Using a similar method as in the previous chapter, we measure the average photometric redshift-space[1] correlation function over 25 boxes of volume $V = (3000\,h^{-1}\,\mathrm{Mpc})^3$, each containing $\sim 1M$ galaxies. We use the periodic nature of the simulation to wrap around all box axes and use the natural estimator $\xi(r_{photo}) = DD/RR - 1$ throughout. The input correlation function in the model is the measured 3D true correlation function averaged over 225 random boxes. Results are plotted in Fig. 4.9.

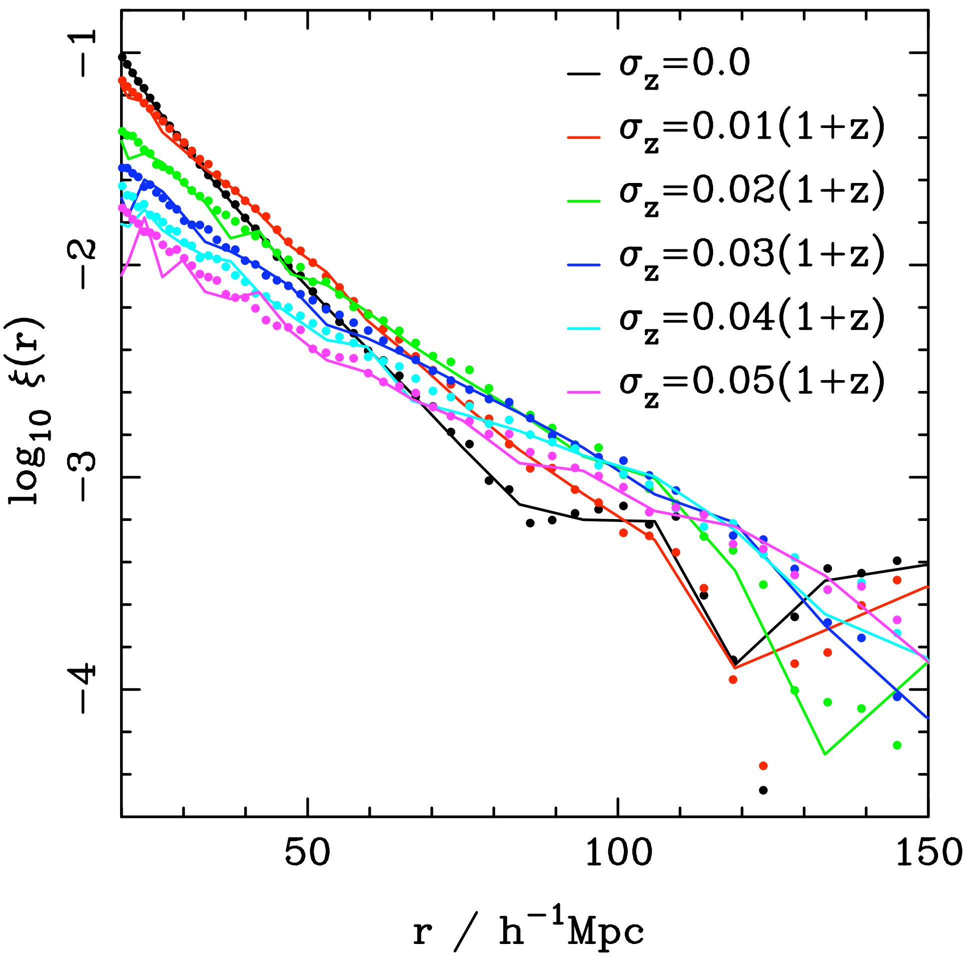

Figure 4.9: Filled circles represent the average 3D correlation function for the Hubble Volume simulation with varying uncertainty on the radial distance measures. Lines show the predicted correlation function for each sample, calculated via a Monte Carlo integration over 10^9 pairs of points in photometric redshift space. Models all agree with the data very well and confirm that our MC modelling technique can successfully predict the effect of uncertainties on radial distances on the recovered clustering sample.

[1]We do not include the effect of redshift-space distortions in this analysis.

The effect of increasing uncertainty in radial distances is an increasingly damped power on small scales, and a loss of the BAO signal. Models were created via a MC integration over $\sim 10^9$ pairs of points using the method described above and agree very well with the data. It is worth noting here that future photometric sky surveys predict redshift uncertainties of the order $\sigma_0 = 0.03$. We can already see from this toy model that it will be difficult to conduct accurate clustering analyses with this level of uncertainty in our distance measurements.

When using this technique to predict the clustering signal of our S82 data-sets, we will need to consider the radial distributions of each sample within the analysis. This is easy to do and should not provide any large difficulties in the method.

4.3.6 S82 Correlation Function: Results

After selecting galaxies in the spectroscopic redshift range $0.03 < z < 0.3$, our final data catalogue contained 18949 galaxies, each with a spectroscopic redshift and 4 different photometric redshift estimates. Random catalogues were created with $30\times$ the number of galaxies according to the angular and radial masks we calculated per redshift estimate. We conducted an N^2 Monte-Carlo integration over data-data (DD), data-random (DR) and random-random (RR) pairs, binning in pair separation r. An estimate of the correlation function for each sample was obtained via the Landy & Szalay estimator

$$\xi = \frac{nDD - 2nDR + nRR}{nRR}. \tag{4.18}$$

Normalised pair counts $nDD = 2DD/n_g(n_g - 1)$, $nDR = DR/n_g n_r$, and $nRR = 2RR/n_r(n_r - 1)$ were calulated to match the expected number of pairs for each catalogue, where n_g is the number of galaxies and n_r is the number of randoms. Models were created via a MC integration over 10^9 pairs of points as described in Section. 4.3.5. Pairs were sampled according to the model radial distribution for each sample. Points in each pair were moved to a photometric redshift space position via a random realisation of a Gaussian distribution with a standard deviation, σ_d, given in Table. 4.1.

Results are plotted in Fig. 4.10. Filled circles show where the correlation function is positive, whilst open circles show where it goes negative. The mean correlation function $\langle\xi_{pz}\rangle$ of the four photometric samples is plotted in black. Errors are calculated via the standard Jackknife resampling technique. We split each catalogue into $N = 15$ equal

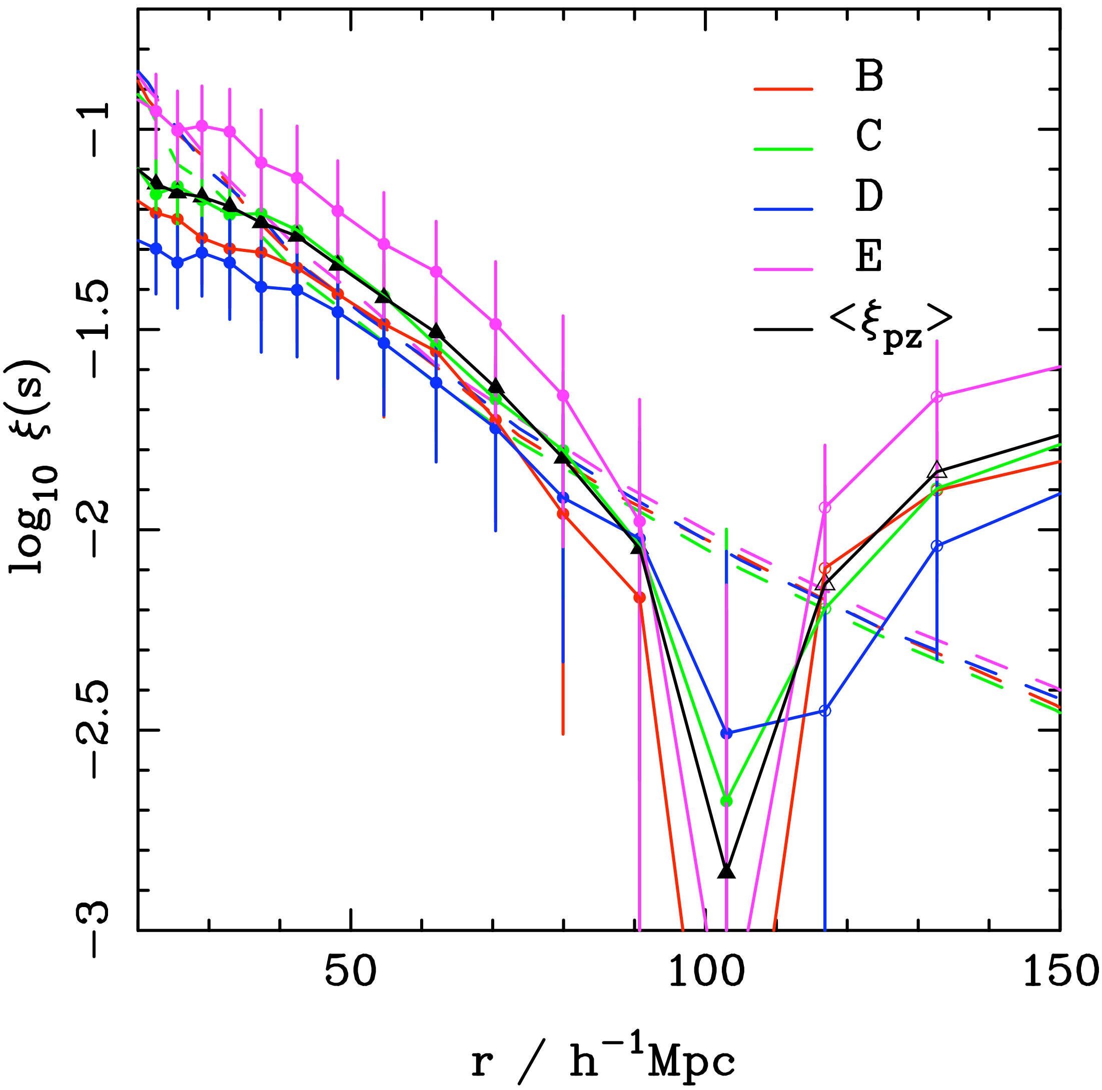

Figure 4.10: 3D correlation functions calculated for 4 different photometric redshift estimates of a singular data-set taken from the SDSS S82. Filled points show where the correlation function is positive, whilst open points show where it goes negative. Triangles denote the mean correlation function. Errors are calculated via a jackknife resampling method over 15 equal area regions per sample. Models are plotted as dashed lines, calculated using a Monte-Carlo integration over 10^9 pairs of points in photometric redshift space (see text for details). Each measurement is normalised to model B so that the dispersion around the mean is purely systematic.

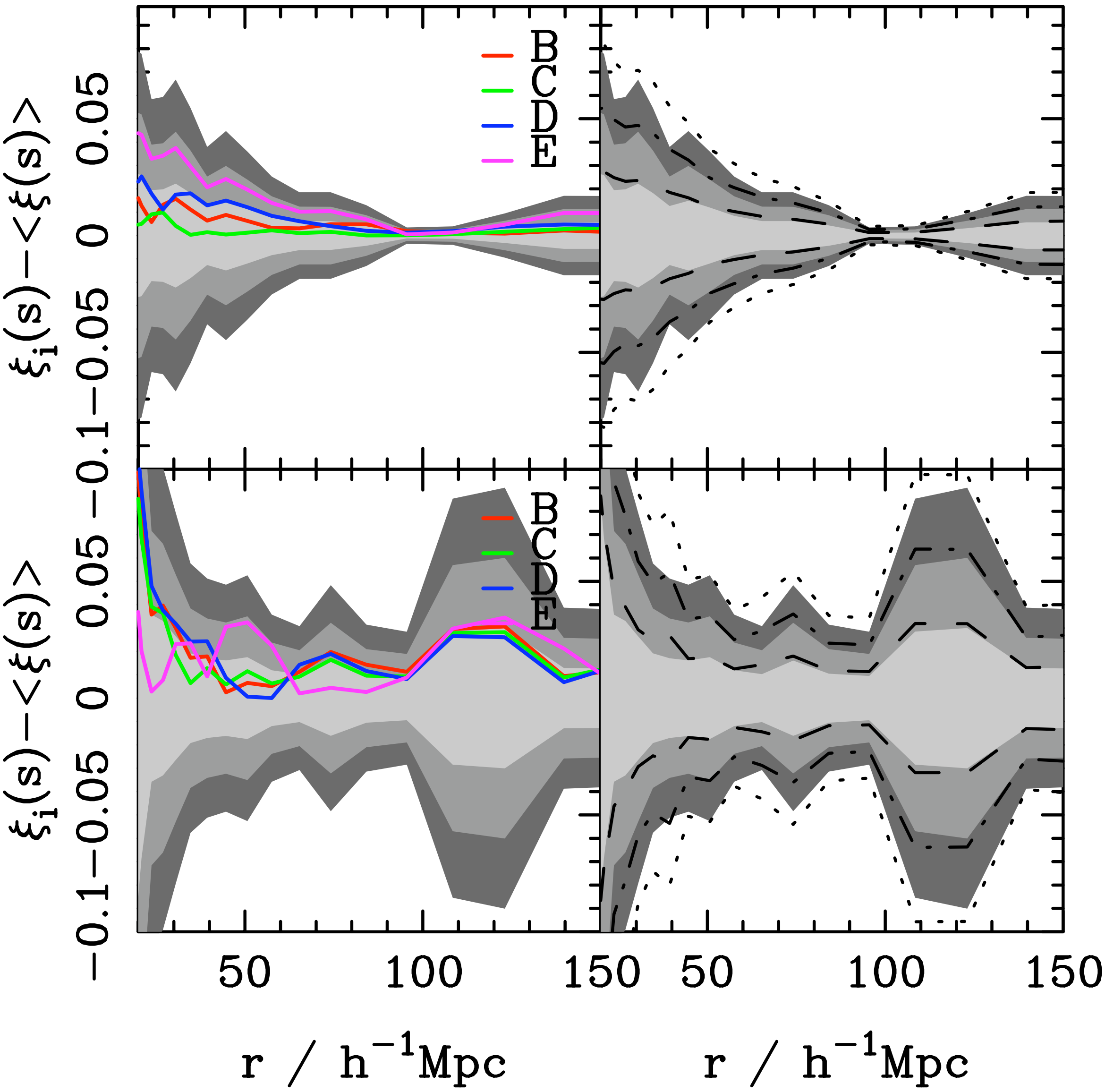

Figure 4.11: Left panels: Scatter around the mean (top) and true (bottom) correlation function is plotted per photometric redshift sample, with shaded grey regions representing 1σ (light grey), 2σ (mid-grey) and 3σ (dark-grey) systematic confidence limits. 3 out of the 4 samples lie within the 1σ confidence limit in both cases. Samples B, C, and D follow same trends, which is not surprising since they all come from SDSS and were created using large and representative training sets. Sample E, calculated using ANNz with the true sample as a training set, lies within the 2σ confidence limit. This is expected since the mean correlation function is dominated by the SDSS samples. It is more surprising that sample E does not sit prominently within the 1σ bounds around the true sample, since it is trained on this specific sample and should be entirely representative. The fact that it does not may be due to discrepancies in the data and random radial distributions (see text for further discussion). Right panels: Shaded grey regions as left-hand panels. Black lines show 1σ (dashed), 2σ (dot-dashed) and 3σ (dotted) statistical confidence limits. Statistical errors on the correlation function for our SDSS S82 photometric redshift samples are dominated by systematic errors coming from uncertainties on photometric redshift estimates. This suggests that more effort needs to be made to ensure that we understand where these systematic uncertainties are arising for future surveys.

area regions and then systematically removed 1 region from each, calculating the correlation function in the remaining sample. Jackknife errors are calculated via the equation

$$\sigma^2_{\xi(r)} = \frac{N-1}{N} \sum_{i=1}^{N} [\bar{\xi}(r) - \xi_i(r)]^2. \tag{4.19}$$

We correct for statistical errors by normalising each recovered correlation function to that expected from model B. The scatter around the mean in the plotted correlation functions is therefore purely systematic.

We show the extent of this scatter in Fig. 4.11. Top and bottom panels on the left show the scatter around the mean and true correlation functions for each photometric sample, respectively. Shaded grey regions denote 1σ (light-grey), 2σ (mid-grey) and 3σ (dark-grey) systematic error confidence limits, calculated over the four samples assuming Gaussian statistics. Statistical error confidence limits are plotted in the right-hand panels as dashed (1σ), dot-dashed (2σ), and dotted (3σ) lines. In the top panel we see that all 3 of the SDSS samples lie within 1σ of the mean, whereas the ANNz sample lies within 2σ of the mean. This is expected since the SDSS samples are created with similar and highly representative training sets, thus following the same trends and dominating the mean. In the bottom panel we see a similar scenario, but with the ANNz sample fluctuating between the 1σ and 2σ limits. This makes sense because the scatter around the true sample will not be biased by averaging, in contrast to the scatter around the mean.

One would expect in this case that the ANNz sample, which was trained on the true sample, would sit prominently within the 68% confidence interval. The fact that it does not may be attributed to differences between the data and random radial distributions. Panel E (relating to the ANNz sample) of Fig. 4.7 shows a prominent deficit of galaxies around a redshift of $z = 0.1$. The corresponding model distribution however, does not fit to this feature. This would have created a mismatch between data and random catalogues. Panels A, B, and D show a similar, albeit more subtle, data-random mismatch. Returning to the top panel of Fig. 4.11, we can see that samples B, D and E exhibit similar clustering signals, which further supports this theory.

The right-hand panels compare the expected confidence limits on the correlation function coming from systematic (shaded grey regions) and statistical (black lines) uncertainties. Statistical errors on the correlation function for our SDSS S82 photometric redshift samples are dominated by systematic errors coming from uncertainties on photometric redshift estimates. This suggests that more effort needs to be spent to ensure that we understand where these systematic uncertainties are arising.

4.4 Discussions and Conclusions

Uncertainties on photometric redshift estimates induce errors in radial distance measurements. These errors cause a damping effect on the correlation function and dilute the BAO signal. Existing models predict the effect of these uncertainties on the 3D correlation function, but do not account for systematic uncertainties arising from differing photometric redshift estimates. Many future surveys, including the DES, PanStarrs and LSST, will fully utilise photometric redshifts estimation techniques. Therefore in this chapter, we have investigated how different photometric redshift estimation techniques affect 3D clustering analyses. We have achieved this with a simple empirical test to compare the recovered 3D correlation function for a single sample of galaxies from the Sloan Digital Sky Survey Stripe 82 data-set, where each galaxy had a true spectroscopic redshift and a number of different photometric redshift estimates.

In all cases where radial distances were derived from photometric redshift estimates we saw a damping of the correlation function on all scales, and an absence of the BAO signature expected at $100\,h^{-1}\,\mathrm{Mpc}$. To quantify the level of statistical uncertainty on our measurements we conducted Jackknife resampling over 15 equal area regions for each of our data-sets. In the case where we calculated the scatter around the mean value of the correlation function for the photometric redshift samples we found that 3 out of 4 lay within the 1σ confidence interval. These 3 samples all came from the SDSS CasJobs database and were trained on extensive and representative training sets. Subsequently, it is not surprising that they display similar results and weight the mean such that the sample trained with the spectroscopic data via artificial neural network techniques (E) lies outside the 1σ confidence limit.

In calculating the scatter around the true spectroscopic sample for the correlation function we find that the errors on the measurements increase, thus allowing sample E to spend more time within the 1σ limit. We can attribute this increase in statistical uncertainty to the poor quality of the data-set, since we are both shot-noise and cosmic variance limited.

Our next step was to understand how systematic uncertainties, coming from varying photometric redshift estimations, contribute to the overall statistical uncertainty on our recovered correlation functions. To achieve this, we started by calculating expected models for our anisotropic correlation functions using Monte-Carlo integration techniques with the required level of damping set by the variance in galaxy distances across our samples. After normalising each measured correlation function to one of the expected models, to ensure that we were dealing with pure systematic uncertainties, we repeated our calculation of the scatter around the mean and true values. From this exercise we found that systematic uncertainties, arising from differences between photometric redshift estimation techniques, dominates the overall statistical uncertainty on the measurement. This result implies that future measurements of the 3D clustering signal will be impaired by our ability to estimate photometric redshifts and not by our ability to accurately measure the covariance matrix. As such, it would be wise for us to concentrate more effort into understanding the origins of these systematic uncertainties if we want to obtain accurate measurements of the 3D correlation function using data from future photometric sky surveys.

Chapter 5

Future Surveys: The Dark Energy Survey

A number of extremely wide angle imaging surveys are planned over the next few years including: the Dark Energy Survey (DES), the Panoramic Survey Telescope & Rapid Response System (Pan-Starrs) and the Large Synoptic Survey Telescope (LSST). The primary goal of these surveys is to constrain the current acceleration of the Universe. Inparticular, the DES has been designed to extract cosmological information on dark energy via 4 complimentary methods:

- Weak lensing tomography - Weak lensing measurements will be made in several redshift shells out to $z \sim 1$.
- Galaxy clusters - Measurements of the spatial distribution of clusters and cluster counting will be conducted at $0.1 < z < 1.4$.
- Galaxy clustering - The shifting of the galaxy spatial angular power spectrum will be measured in redshift shells (considered here).
- Supernova - The luminosity distances for 2000 supernovae will be collected at $0.3 < z < 0.8$.

For these experiments, radial distances to galaxies will be estimated from photometric redshifts. In the context of BAO measurements we have already shown that:

- Uncertainties on photometric redshift estimates can wash out information in the radial direction, leaving little information on the scale of BAO.
- Differences in photometric redshift estimators introduce a systematic uncertainty on the recovered 3D clustering signal.

As a consequence of the first point, analyses will tend to rely on making projected galaxy clustering measurements in redshift slices that are sufficiently narrow to be able to reveal cosmological acceleration. In this chapter we assess the effect of redshift-space distortions on such measurements by considering one of these surveys, the DES, in more detail. We also consider how 3D clustering measurements in the DES will be affected by systematic uncertainties coming from photometric redshift estimation techniques.

5.1 Introduction to the Dark Energy Survey

Background

The Dark Energy Survey (DES) collaboration formed in response to an Announcement of Opportunity (AO) made by the National Optical Astronomy Observatory (NOAO) in December 2003. NOAO were launching a competition to find a partner with which they could build an advanced instrument to be employed on the Blanco-4 meter telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile. In reward, the partner would receive 30% of observing time over a 5-year period for use in a compelling science project. The DES collaboration proceeded to make DECam, an instrument made to assist in addressing the nature of dark energy, and secured the partnership with NOAO.

Instrument and Survey

DECam is a 519 Megapixel optical CCD camera with a 1 meter diameter and a wide 2.2 degree field of view optical corrector. It is extremely red sensitive using a 4-band filter system with SDSS g, r, i, and z filters, and utilises a very fast data acquisition system enabling images to be taken every 17 seconds. The camera will be mounted at the prime-focus of the Blanco-4 meter telescope at CTIO in Chile.

Over a 5 year period the DES project will be allocated 30% of telescope dark time, which amounts to 525 nights. This time will be dedicated to obtaining high precision multi-bandpass photometric redshifts in the range $0.2 < z < 1.4$. Angular coverage amounts to 5000 deg^2 with 4000 deg^2 overlapping with the Sunyeav-Zeldovich CMB survey. The expected redshift distribution of the galaxies will be approximately [1]

$$\phi(z) \propto \left(\frac{z}{0.5}\right)^2 \exp\left(-\frac{z}{0.5}\right)^{1.5}, \tag{5.1}$$

after applying approximate survey depths to basic luminosity functions. This function is plotted in Fig. 5.1.

[1] We thank the DES LSS working group for providing this approximation

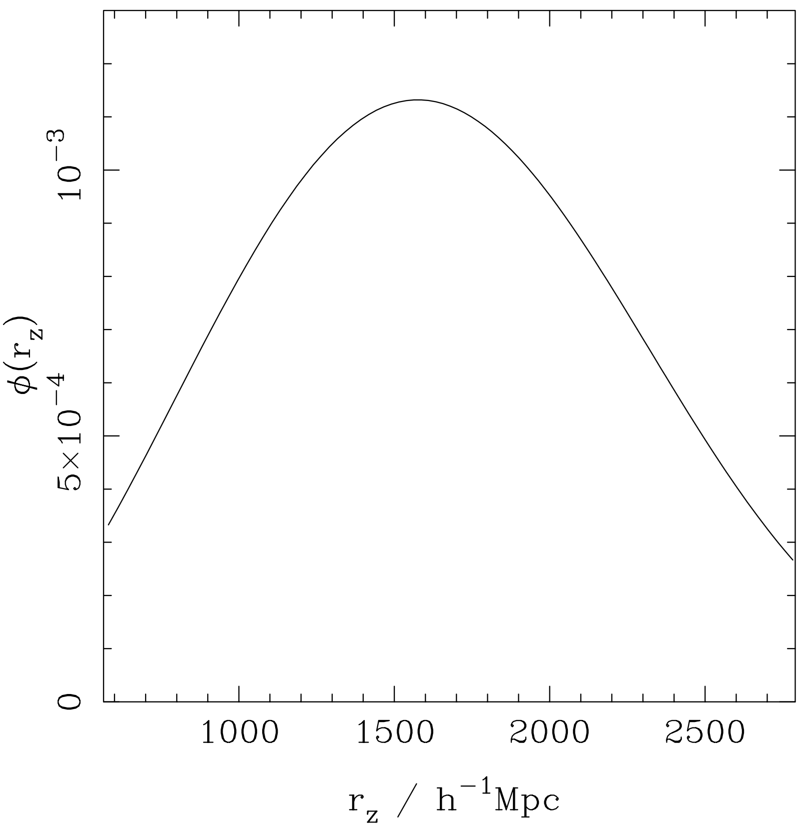

Figure 5.1: Approximate redshift distribution similar to that expected from the Dark Energy Survey. In order to use this distribution of galaxies to easily measure cosmological acceleration using projected clustering measurements, this population will have to be subdivided or binned in redshift.

Science

The DES will aim to improve our understanding of dark energy by focussing on obtaining constraints on the dark energy equation of state parameter w. If dark energy is truly a cosmological constant, the dark energy density should remain constant in an expanding universe. This means that $w = -1$ and $dw/dt = 0$. Each experiment is predicted to provide a 5-15% precision measurement in w and a 30% measurement in w'. Combined constraints will be stronger and will provide checks on systematic errors.

5.2 Projected Clustering for Future Surveys

Photometric redshifts induce uncertainties in derived radial distances to galaxies, leaving little information in the radial direction on the scale of BAO. Consequently, analyses for future photometric surveys such as the DES will tend to rely on making projected galaxy clustering measurements in redshift slices that are sufficiently narrow to be able to reveal cosmological acceleration. In this section we assess the effect of redshift-space distortions on such measurements for a survey like the DES.

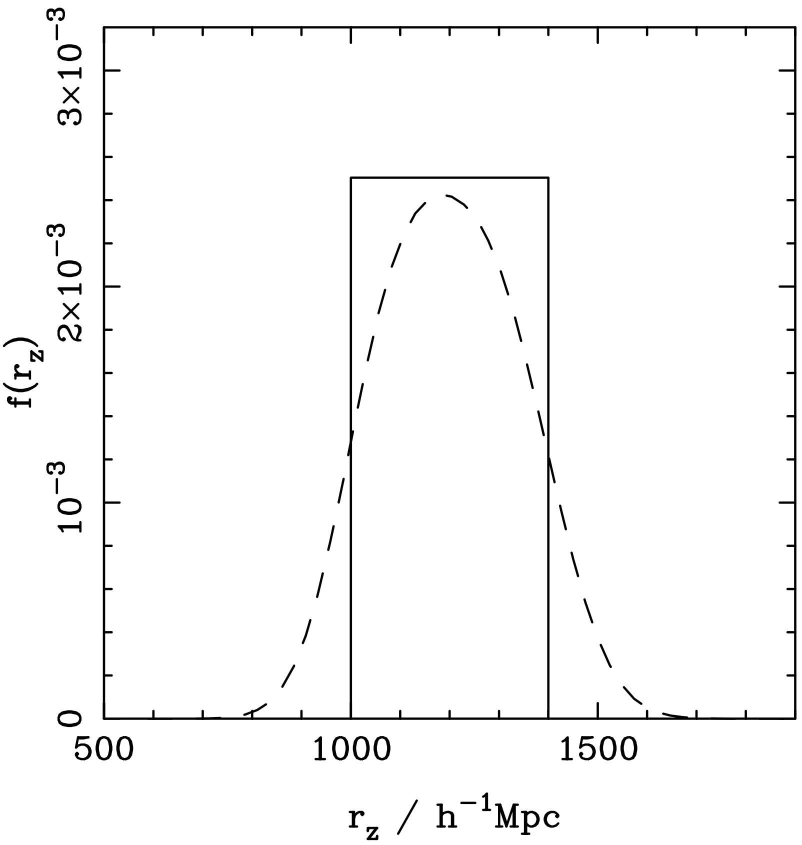

Figure 5.2: The radial distribution of galaxies selected in a bin of width $400\,h^{-1}\,\mathrm{Mpc}$, calculated using photometric redshifts to estimate distances (solid line). This is compared against the distribution of true distances to these galaxies (dashed line) assuming a photometric redshift error of $\sigma_z = 0.03(1+z)$. If the photometric redshifts of different galaxies are independent, then the expected projected correlation function of the photo-z selected sample, and a sample selected applying the dashed line as a selection function based on the true distances, are the same.

We consider the simplified problem in the plane-parallel approximation, and only consider linear redshift-space distortions. Both photometric redshift errors and the random motion of galaxies in clusters provide an additional convolution of the overdensity field along the radial direction. While these effects need to be corrected in any analysis, the required correction is easily modelled and can be separated from the linear redshift-space distortion effects. For a measurement of the projected clustering, including such effects is equivalent to simply broadening the radial window function with which the galaxies were selected. This is demonstrated in Fig. 5.2.

A top-hat bin in photometric redshift gives the same expected projected correlation function as simply applying the convolved version of the bin as a selection function for the true distances. As we have to include a window function anyway, we simply assume in this paper that this window already includes the effects of both photometric redshift errors and the random motion of galaxies in clusters. In the following analysis, we therefore assume that there are no redshift errors without loss of generality.

5.2.1 Selecting DES Samples

As discussed in Chapter 3, measurements of the projected correlation function will be affected by redshift-space distortions, which will increase the signal strength and decrease the importance of BAO features. The distribution of galaxies in the DES, as plotted in Fig. 5.1, will be sub-divided into bins in order to assess the evolution of the BAO scale across the survey. This radial selection function is defined in real-space, whereas observations, and therefore radial binning, will be conducted in redshift-space. By slicing the radial distribution into bins we automatically generate 3 populations A, B, and C, which have two, one, and zero redshift-space boundaries. We label these 3 scenarios as redshift-, hybrid-, and real-space selections, respectively. We now consider how the choice of binning methodology affects the impact of redshift-space distortions.

The hybrid-space correlation function can simply be thought of as a real-space correlation function weighted by the redshift-space correlation function, where the magnitude of the weight is dependent on the number of redshift-space boundaries present in the sample under analysis. Using Eq. 3.39, we can see that there are 9 possible combinations of auto- and cross-correlation functions between these 3 populations: AA ($l = 4$), AB ($l = 3$), AC ($l = 2$), BA ($l = 3$), BB ($l = 2$), BC ($l = 1$), CA, ($l = 2$) CB ($l = 1$), and CC ($l = 0$). Out of these 9 combinations, only 5 are distinct, ie. $l = 0, 1, 2, 3, 4$. This is summarised in Table. 5.1. We can now construct the full equation for the model ξ_p:

$$\xi_p = \sum_{l=0}^{4} \xi_{p,l}, \tag{5.2}$$

where

$$\xi_{p,l=4} = \xi_{p,AA} = \int_{z_1}^{z_2} \int_{z_1}^{z_2} dz\, dz'\ \phi_A(z)\phi_A(z')\xi^s(z, z') \tag{5.3}$$

$$\xi_{p,l=3} = \xi_{p,BB} = \int_{z_1}^{z_2} \int_{z_1}^{z_2} dz\, dz'\ \phi_B(z)\phi_B(z')\xi^h(z, z') \tag{5.4}$$

$$\begin{aligned}\xi_{p,l=2} &= \xi_{p,AC} + \xi_{p,CA} \\ &= 2\int_{z_1}^{z_2}\int_{z_1}^{z_2} dz\, dz'\ \phi_A(z)\phi_C(z')\xi^h(z,z')\end{aligned} \quad (5.5)$$

$$\begin{aligned}\xi_{p,l=1} &= \xi_{p,BC} + \xi_{p,CB} \\ &= 2\int_{z_1}^{z_2}\int_{z_1}^{z_2} dz\, dz'\ \phi_B(z)\phi_C(z')\xi^h(z,z')\end{aligned} \quad (5.6)$$

$$\xi_{p,l=0} = \xi_{p,CC} = \int_{z_1}^{z_2}\int_{z_1}^{z_2} dz\, dz'\ \phi_C(z)\phi_C(z')\xi^r(z,z') \quad (5.7)$$

where $\phi_B(z) = \phi(z)$ if $\phi(z) < \phi(z_2)$ and $\phi(z_2)$ otherwise, and $\phi_C(z) = \phi(z)$ if $\phi(z) > \phi(z_2)$ and zero otherwise.

5.2.2 Binning ϕ_{DES}

The traditional method for dealing with redshift-space distortions is to use the projected correlation function in wide redshift bins $> 800\,h^{-1}\,\mathrm{Mpc}$ (Saunders et al., 1992). We have already shown that the effect of peculiar velocities on projected clustering analyses is not negligible, and that the coherent movement of galaxies into and out of the sample at the edges of radial bins induces a clustering signal in the projected correlation function. This effect diminishes when we project over large radial bins. As a consequence however, we lose information about the radial evolution of BAO. Our new constrained pair-centre binning scheme not only significantly reduces the effects of redshift-space distortions on clustering analyses, but also allows for finer binning in the radial direction thus providing us with a better measurement of cosmological evolution.

We consider splitting the galaxy distribution (as shown in Fig. 5.1) into five redshift slices each of width $400\,h^{-1}\,\mathrm{Mpc}$ for distances estimated from photometric redshifts, assumed to be Gaussian with $\sigma_z = 0.03(1+z)$. These bins cover radial distances of $500 \rightarrow 2500\,h^{-1}\,\mathrm{Mpc}$, related to redshifts $z = 0.15$ to $z = 1.06$ (assuming a flat ΛCDM cosmology with $\Omega_m = 0.25$). The upper panel of Fig. 5.3 shows the distributions of galaxies in these slices. The lower panel of Fig. 5.3 shows the redshift distributions when we bin the galaxies based on the centre of the radial separation, calculated from the photometric redshifts. Because we are using photometric redshifts, there is no way to bin without leaving overlap in the true radial distributions. Consequently, the top-hat binning scheme does not provide an obvious advantage over other schemes in terms of analysing disjoint regions.

pop'n	A	B	C
A	$l=4$ $\xi^h=\xi^s$	$l=3$ $\xi^h+1=(\xi^r+1)^{1/4}(\xi^s+1)^{3/4}$	$l=2$ $\xi^h+1=(\xi^r+1)^{1/2}(\xi^s+1)^{1/2}$
B	$l=3$ $\xi^h+1=(\xi^r+1)^{1/4}(\xi^s+1)^{3/4}$	$l=2$ $\xi^h+1=(\xi^r+1)^{1/2}(\xi^s+1)^{1/2}$	$l=1$ $\xi^h+1=(\xi^r+1)^{3/4}(\xi^s+1)^{1/4}$
C	$l=2$ $\xi^h+1=(\xi^r+1)^{1/2}(\xi^s+1)^{1/2}$	$l=1$ $\xi^h+1=(\xi^r+1)^{3/4}(\xi^s+1)^{1/4}$	$l=0$ $\xi^h=\xi^r$

Table 5.1: Table summarising the hybrid correlation function regime that each auto- and cross-correlation of populations A, B and C corresponds to for a DES-like survey. The hybrid correlation function is denoted here as ξ^h where $l = m + n = 0, 1, 2, 3, 4$. See text for details.

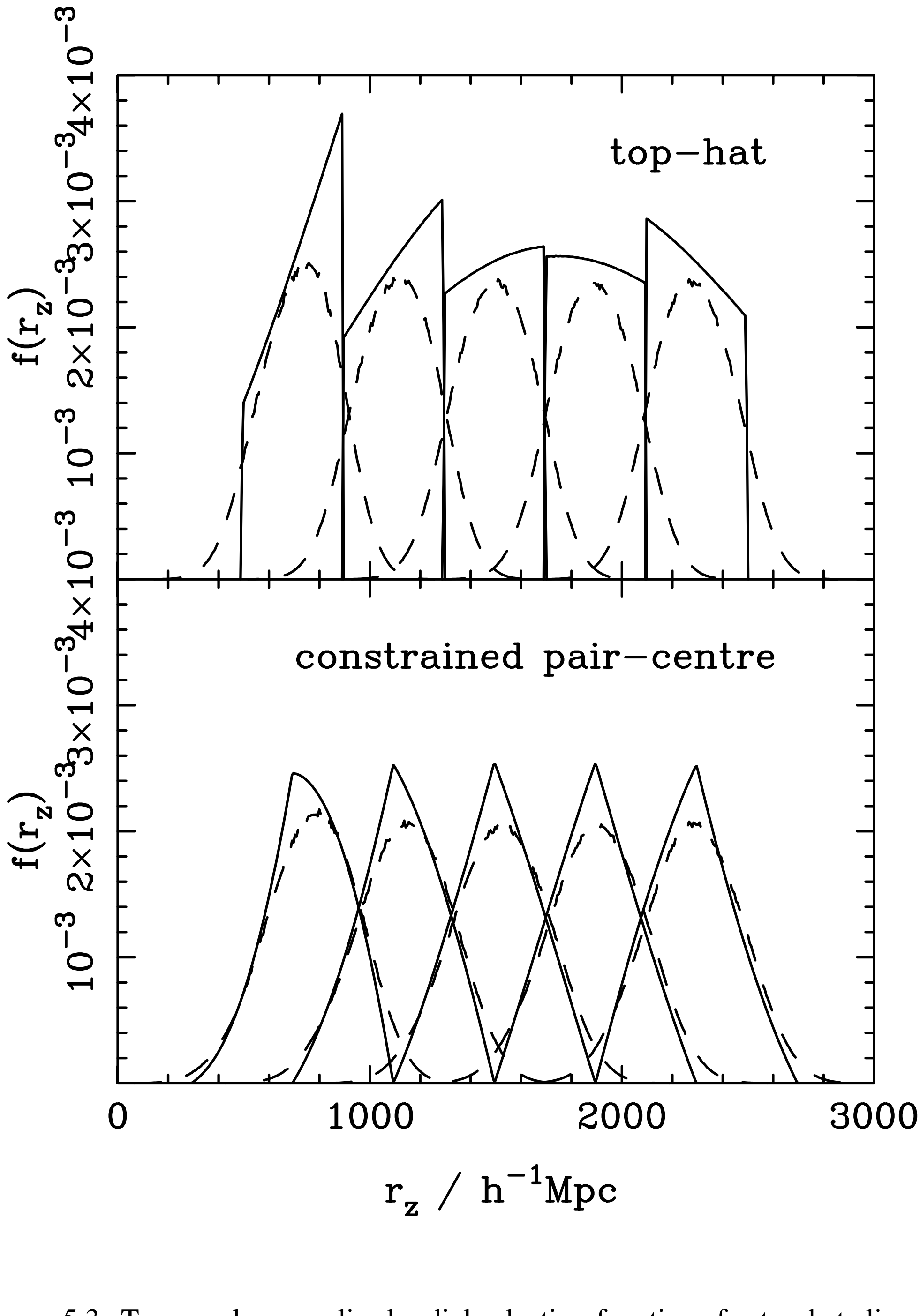

Figure 5.3: Top panel: normalised radial selection functions for top-hat slices of width $400\,h^{-1}\,\mathrm{Mpc}$ created from a DES-like distribution. Bottom panel: We also consider bins in radial galaxy pair centre of the same width $400\,h^{-1}\,\mathrm{Mpc}$. While we bin in distances derived from photometric redshifts (solid lines), the true distribution of radial galaxy distances is shown by the dashed lines.

It is clear from Fig. 5.3 that, even without the broadening of the radial bins due to photometric redshift uncertainties, the pair-centre bins extend beyond the boundaries of the top-hat bins. Practically, this means an increase in the number of galaxies that will be analysed in each bin, and therefore an increase in CPU time required to analyse each bin. In Table. 5.2 we summarise these expected increases, in comparison to a traditional top-hat binning scheme.

Using a Monte-Carlo realisation of 10^{12} pairs of points in our five top-hat bins as defined above, we calculated the corresponding expected number of pairs for the pair-centre and constrained pair-centre binning schemes. We placed an upper limit on radial pair separation for the constrained pair-centre binning scheme of $r_z < 400\,h^{-1}\,\mathrm{Mpc}$. The expected number of pairs, N_p, are presented in Table. 5.2. Comparing these results to that expected for the top-hat binning scheme we see that there are $\sim 14\times$ more pairs in the pair-centre binning scheme, and $\sim 2\times$ in the constrained pair-centre scheme. This corresponds directly to the increase in CPU time. We calculate the expected number of galaxies, N_g, per bin assuming that $N_g \sim \sqrt{2N_p}$. From this we see that there is an expected increase in N_g of $\sim 373\%$ for the pair-centre scheme, and $\sim 141\%$ for the constrained pair-centre scheme. This makes the pair-centre binning scheme unfeasible in practice, since the CPU time required to conduct a full analysis on a data-set so large would be too long. Placing an upper constraint on the radial pair separation however, means that this number is reduced dramatically, further supporting the validity of this scheme.

5.2.3 Predictions for DES

In light of the discussion in § 3.1.7, we consider both the case in which ϕ_{DES} is treated as a real-space boundary, and the case in which it is treated as a redshift-space boundary (as may be the case when the slope of $k_{corr}(z)$ is especially large). The former scenario means that we are dealing with hybrid boundary selections, so we can employ the techniques described in § 3.3 to determine the full form of the projection. When we treat ϕ_{DES} as a redshift-space boundary, we can simply use Eqns. 3.2 & 3.9 to determine ξ_p in real and redshift-space.

We work in photometric redshift-space throughout, and therefore need to consider how the resulting correlation function is convolved due to photometric redshift uncertainties on each radial distance measurement. Implementation of this convolution in practice is relatively straightforward. In the simple case where we have a redshift-space radial selection function, all that is required is a single additional Monte-Carlo integration over

SCHEME	BIN	$N_p(\times 10^{12})$	$N_g(\times 10^5) \sim \sqrt{2N_p}$	% of gals	CPU
top-hat	1	1	1.414214	100	$1\times$
	2	1	1.414214	100	$1\times$
	3	1	1.414214	100	$1\times$
	4	1	1.414214	100	$1\times$
	5	1	1.414214	100	$1\times$
pair-centre	1	13.8784	52.68472	372.54	$14\times$
	2	13.9142	52.75263	373.02	$14\times$
	3	13.8851	52.69743	372.63	$14\times$
	4	13.9007	52.72703	372.84	$14\times$
	5	13.9039	52.73310	372.88	$14\times$
constrained pair-centre	1	2.00504	2.002518	141.6	$2\times$
	2	2.0075	2.003747	141.67	$2\times$
	3	2.0003	2.00015	141.43	$2\times$
	4	1.99888	1.99944	141.38	$2\times$
	5	1.99861	1.999305	141.37	$2\times$

Table 5.2: Table showing the expected increase in galaxy number and CPU time for pair-centre and constrained pair-centre binning schemes, in comparison to the traditional top-hat binning scheme. We calculate the expected number of pairs in five redshift bins each of width $400\,h^{-1}\,\mathrm{Mpc}$ for distances estimated from photometric redshifts, assumed to be Gaussian with $\sigma_z = 0.03(1+z)$. These bins cover radial distances of $500 \rightarrow 2500\,h^{-1}\,\mathrm{Mpc}$, related to redshifts $z = 0.15$ to $z = 1.06$ (assuming a flat ΛCDM cosmology with $\Omega_m = 0.25$) and are shown in Fig. 5.3. We do this for the pair-centre and constrained pair-centre ($< 400\,h^{-1}\,\mathrm{Mpc}$) binning schemes using Monte-Carlo realisations of a fixed number of pairs of points in the top-hat binning scheme. We assume that the number of galaxies, N_g is approximately $\sqrt{2N_p}$, where N_p is the number of pairs. From this, we have calculated the percentage increase in the number of galaxies in the pair-centre and constrained pair-centre regimes. We see a $\sim 370\%$ increase in N_g when we use the pair-centre binning scheme, and a $\sim 140\%$ increase for the constrained pair-centre binning scheme. This corresponds to a CPU time increase of $14\times$ and $2\times$ that of the standard top-hat procedure for pair-centre and constrained pair-centre schemes, respectively.

the total photometric redshift uncertainty for the pair, convolved with a Gaussian. Eq. 3.9 now becomes

$$\xi_p^s(d_p) = \int_{s_{z1}}^{s_{z2}} \int_{s_{z1}}^{s_{z2}} ds_z ds'_z \phi(s_z)\phi(s'_z)\xi^s[d(s_z, s'_z, d_p] \int_{-3\sigma_{d_p}}^{+3\sigma_{d_p}} d\sigma_{d_p} \exp\left[-\frac{\sigma_{d_p}^2}{2}\right], \tag{5.8}$$

where $\sigma_{d_p}^2 = \sigma_{r_{z1}}^2 + \sigma_{r_{z2}}^2$ and $\sigma_{r_z} = 0.03(1+z)dr_z/dz$ is the photometric redshift uncertainty. The same theory applies when we have a real-space selection function. In this case however, the selection function has a more complex form and is determined by the number of redshift-space boundaries present (as described in § 5.2.1 - see Eq. 5.2).

Fig. 5.4 shows the expected projected correlation functions when a top-hat binning scheme is applied with width 400 h^{-1} Mpc, for real-space (left) and redshift-space (right) selection functions. Even for this large bin width, in every radial bin there is a significant difference between the result obtained using the redshift-space correlation function and the real-space correlation function. The difference is made clear by observing the ratios between the two, displayed in the bottom panels. The ratios are slightly higher in the case where we treat ϕ_{DES} as a redshift-space boundary. The difference between the two treatments is largest for the lowest redshift bin. If we refer back to Fig. 5.3, we can see that this particular bin is most affected by the overall DES selection, thus acting as a reminder that the evolution of the radial selection function also influences the recovered correlation function. Care must be taken to accurately model evolving radial selections even when predicting projected correlation functions.

The effects of redshift distortions are completely removed when a pair-centre binning scheme is employed and the ϕ_{DES} boundary is assumed to be real-space, as made clear in the left hand panel of Fig. 5.5. Based on the discussion in § 3.1.6, we can simply use Eq. 3.2 for both and thus their ratio is identically 1. Even when the ϕ_{DES} boundary is assumed to be in redshift-space, as displayed in the Fig. 5.5, the difference between the redshift-space and real-space model is considerably smaller than for the top-hat binning. Fig. 5.6 shows that even if one applies the constraint that the separation between pairs be less than $400\,h^{-1}\,\mathrm{Mpc}$ to be included in a pair-centre bin, redshift space distortions introduce a much smaller effect than for a top-hat binning scheme.

In all cases, with the exception of real-space pair-centre binning (left hand panel of Fig. 5.5), there is a large spike in ratio around $120\,h^{-1}\,\mathrm{Mpc}$. This is the scale at which the projected correlation function goes negative. Consequently, measurements of ξ_p at this scale will be tiny and differences between the two treatments will be boosted. If we cut out this scale in all regimes we can obtain a better idea of the average ratio. In the case where the traditional top-hat binning method is used, the ratio is substantial ($\sim$ 1.5) around the BAO scale ($\sim 100\,h^{-1}\,\mathrm{Mpc}$), and the shape of the predicted ξ_p and ξ_p^s measurements differ substantially. Our new constrained pair-centre binning scheme significantly reduces the effect of redshift-space distortions, bringing the ratio down to $\sim$1.1 at the BAO scale.

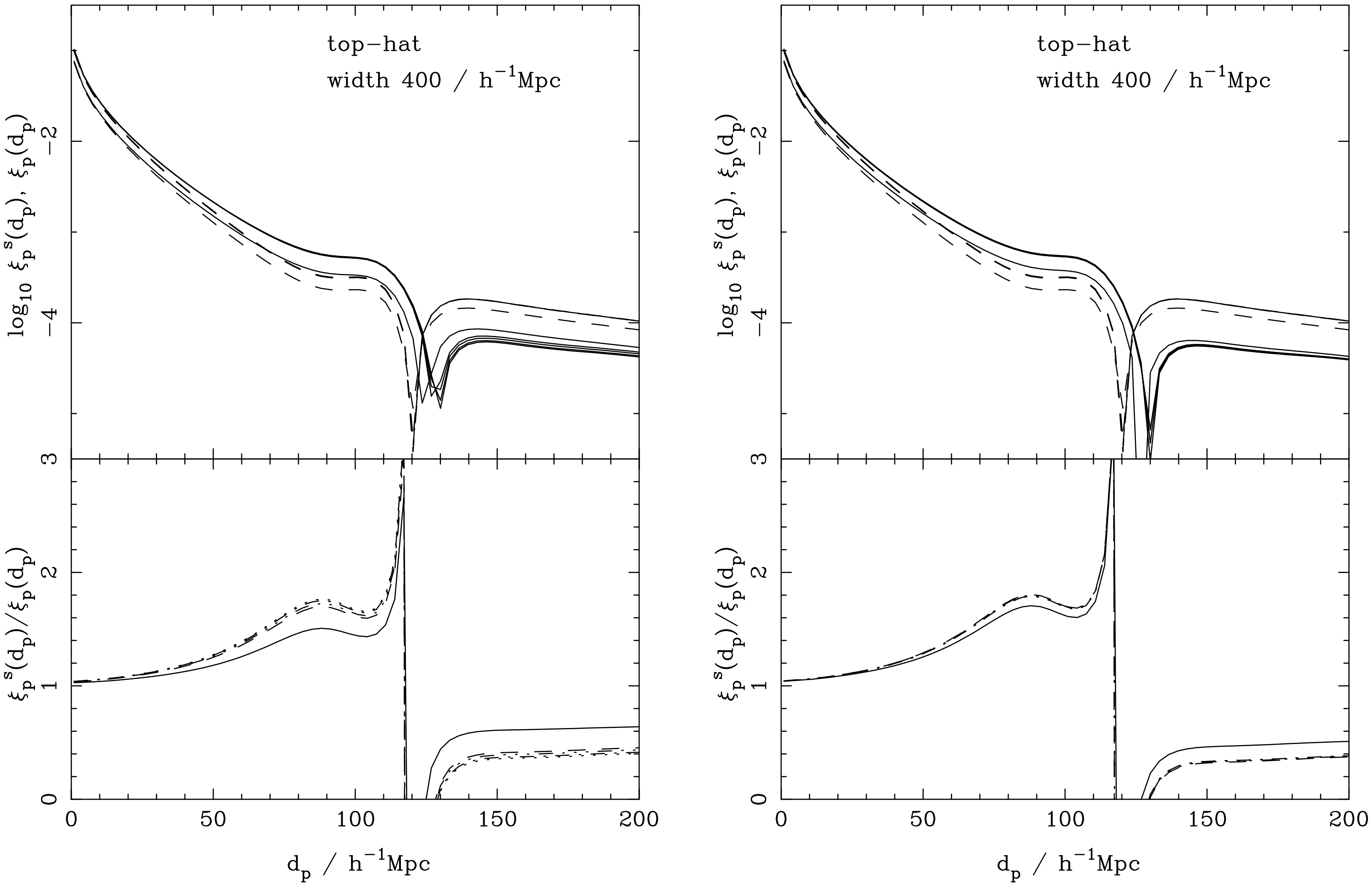

Figure 5.4: Top panel: Real-space (dashed lines) and redshift-space (solid lines) correlation functions predicted for the 5 radial bins drawn from the DES-like selection function, assuming it can be treated as a real-space boundary (left) and a redshift-space boundary (right). Bottom panel: The ratio between the redshift-space and real-space projected correlation function. Here different line styles correspond to different bins: in the order of increasing redshift, they are solid, dashed, dot-dash, dotted, dot-dot-dash. Top-hat bins of width $400\,h^{-1}\,\mathrm{Mpc}$ in the radial direction.

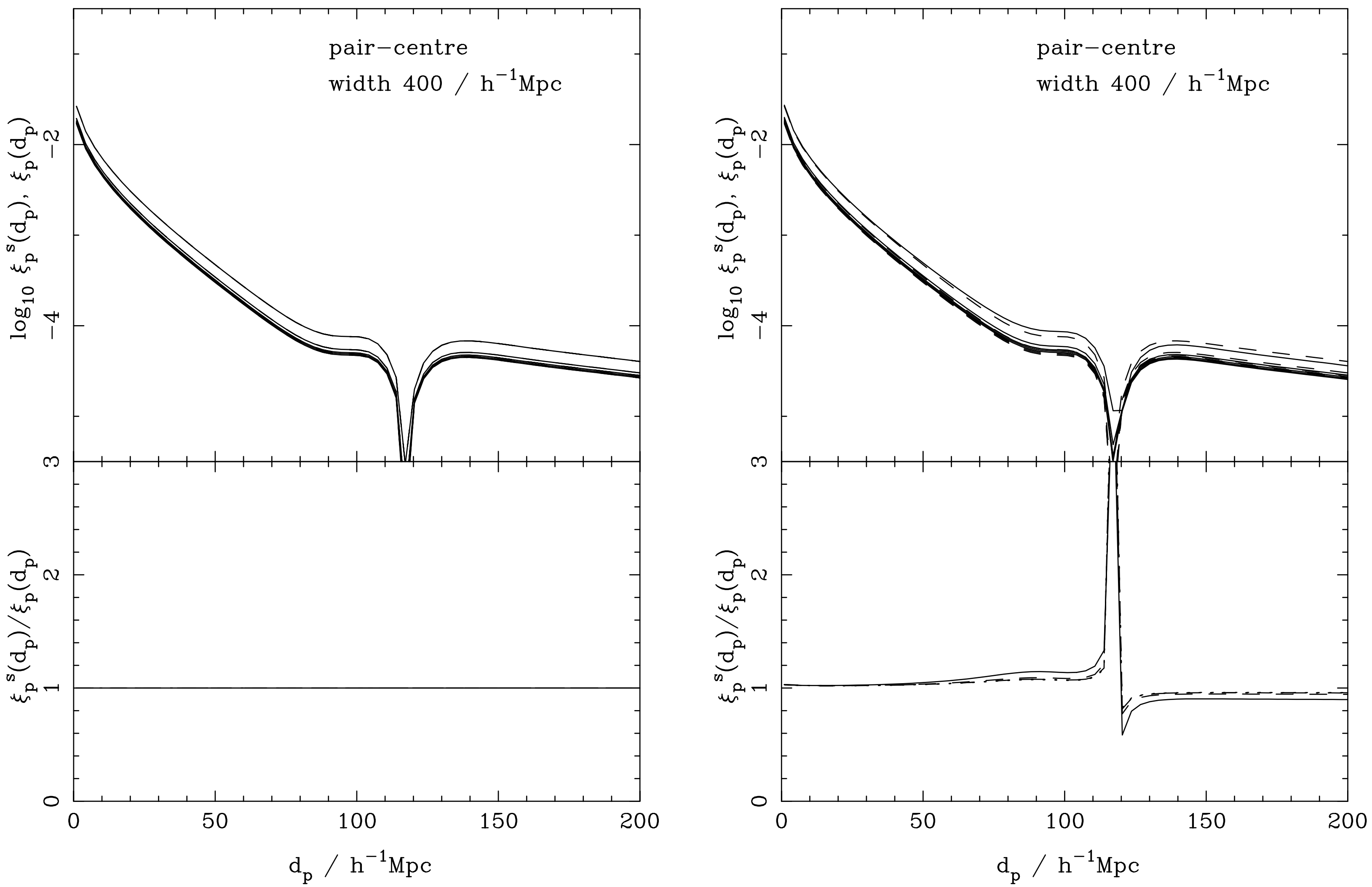

Figure 5.5: As Fig. 5.4 except with pair-centre bins of width $400\,h^{-1}\,\mathrm{Mpc}$. In the case where the radial selection function is in real-space the ratio between the redshift-space and real-space projected correlation functions is exactly 1. The recovered projected correlation function is unaffected by redshift-spae distortions in this regime. Even when the radial selection function is in redshift-space, the ratio is reduced significantly around the BAO scale.

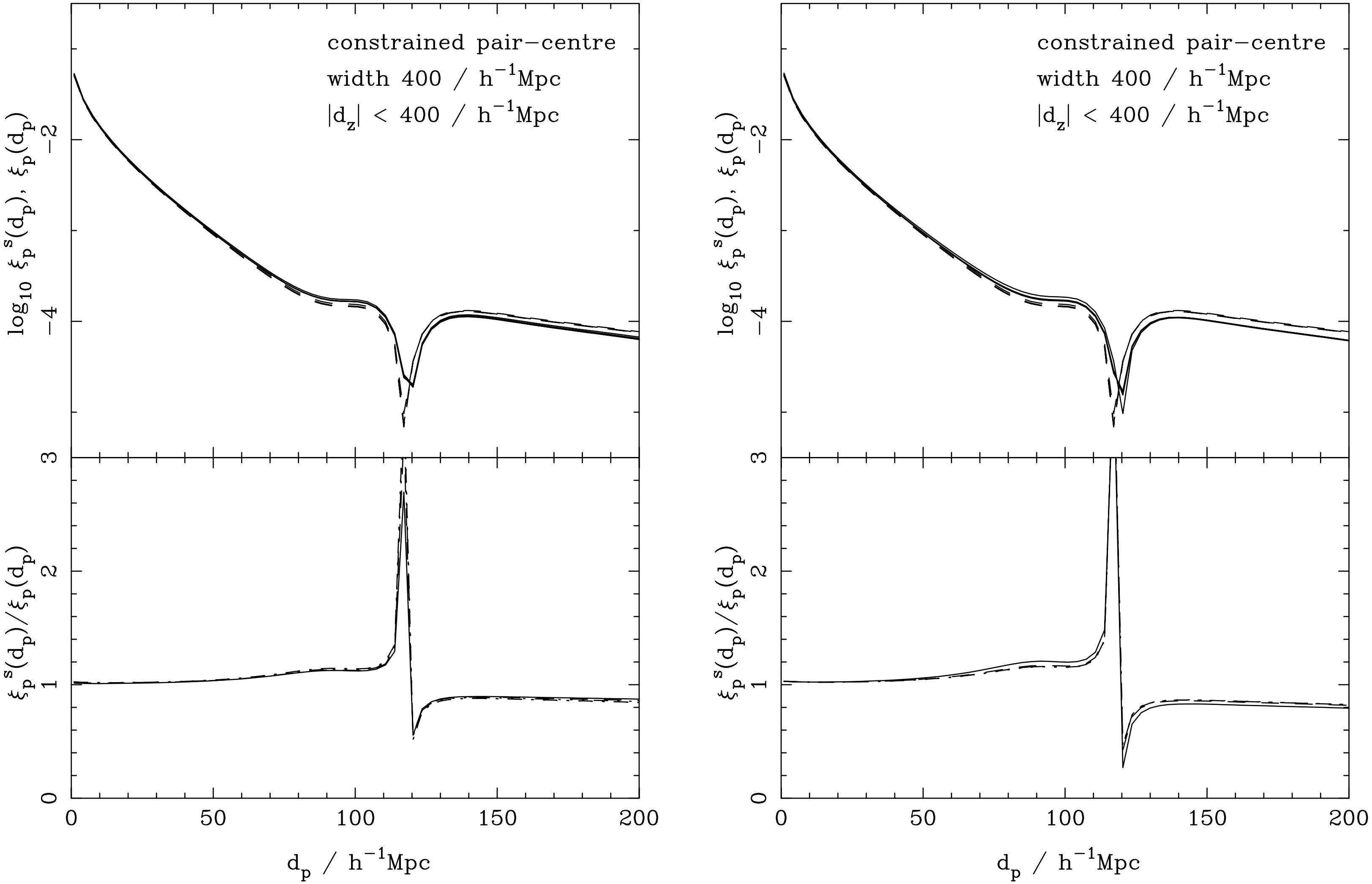

Figure 5.6: As Figs. 5.4 & 5.5 except with pair-centre bins of width $400\,h^{-1}\,\mathrm{Mpc}$, including an additional constraint on the radial separation of $|d_z| < 400\,h^{-1}\,\mathrm{Mpc}$. The ratio between redshift-space and real-space projected correlation functions is reduced significantly ($\sim 80\%$) at the BAO scale, compared to the top-hat binning scheme, for both cases where we have a real-space (left) and redshift-space (right) radial selection function.

5.3 3D Clustering for Future Surveys

Our ability to conduct precision cosmology with 3D clustering analyses depends crucially on our ability to derive accurate distance estimates to galaxies. We have shown in the previous chapters that this is hindered in the radial direction by photometric redshift uncertainties. In Chapter. 4 we showed that systematic uncertainties, arising due to discrepancies between different photometric redshift estimators, were the dominant source of error in a typical clustering analysis. In this section, we calculate the expected statistical errors on the correlation function for a survey like the DES, using simulation data created by the DES Large-Scale Structure (LSS) Working Group (WG). We compare these results with the systematic errors that we calculated in Chapter. 4 to further understand how 3D clustering analyses will be affected by photometric redshift uncertainties in a DES-like survey.

5.3.1 MICE Simulations

Throughout this analysis we use the Marenostrum Institut de Ciéncies de l'Espai (MICE) Simulations that were created for Simulation Challenge 1 as part of the DES LSS WG. The rules for Simulation Challenge 1 were laid out as follows:

- Parent simulation:
  1. Use particles from a Dark Matter (DM) simulation with concordance model.
  2. Use a comoving output at $z = 0$ (no redshift-space distortions yet).
  3. Box size must be $L \geq 1000\,h^{-1}\,\mathrm{Mpc}$ (use periodicity to get to $z = 1.4$).
- Mock catalogue:
  1. Mask should be an octant of the sky: $0 < RA < 90$ deg and $0 < DEC < 90$ deg.
  2. Convert comoving radius r to redshift z using $r(z) = \int_0^z c\,dz/H(z)$.
  3. Add gaussian photometric redshift error to z with $\sigma_z = \sigma_0(1 + z)$ where $\sigma_0 = 0.01, 0.03$.
  4. Use z range $0.2 < z < 1.4$.
  5. Dilute particle density so that $dN/dz = 1.5N_{total}(2z)^2 \exp\left[-(2z)^{1.5}\right]$.
  6. Normalise above so that total number of particles is $10^7 < N_{total} < 10^8$.

Following these guidelines, the Barcelona group at ICE created the MICE simulation. The MICE simulation contains $\sim 5 \times 10^7$ galaxies in real-space (ie. no redshift-space distortions) and covers an octant of the sky over a redshift range $0.2 < z < 1.4$. Cosmological parameters $\Omega_m = 0.25$ and $\Omega_\Lambda = 0.75$. Photometric redshifts are provided where Gaussian uncertainties of order $\sigma_z = 0.01(1+z)$ or $\sigma_z = 0.03(1+z)$ are added to galaxy positions, with the latter representing the expected level of uncertainty for the DES.

Fig. 5.7 shows the angular positions of galaxies in the MICE simulation, plotted in comoving coordinates away from the observer, with no uncertainty on radial distance measurements (left), with photometric redshift uncertainties $\sigma_z = 0.01(1+z)$ (middle) and $\sigma_z = 0.03(1+z)$ (right). In the case where there is no uncertainty on radial measurements due to photometric redshift errors, the clustering of galaxies is very clear. As we move to greater levels of uncertainty on radial distance measurements we can see the formation of filament-like structures pointing towards the observer as a result of the clustering signal being smeared out. Fig. 5.8 shows the comoving radial distribution of galaxies in the MICE simulation for varying levels of uncertainty on photometric redshift estimates, created via the sampling of a Gaussian distribution (as outlined above). We use these photometric redshift estimates throughout the analysis.

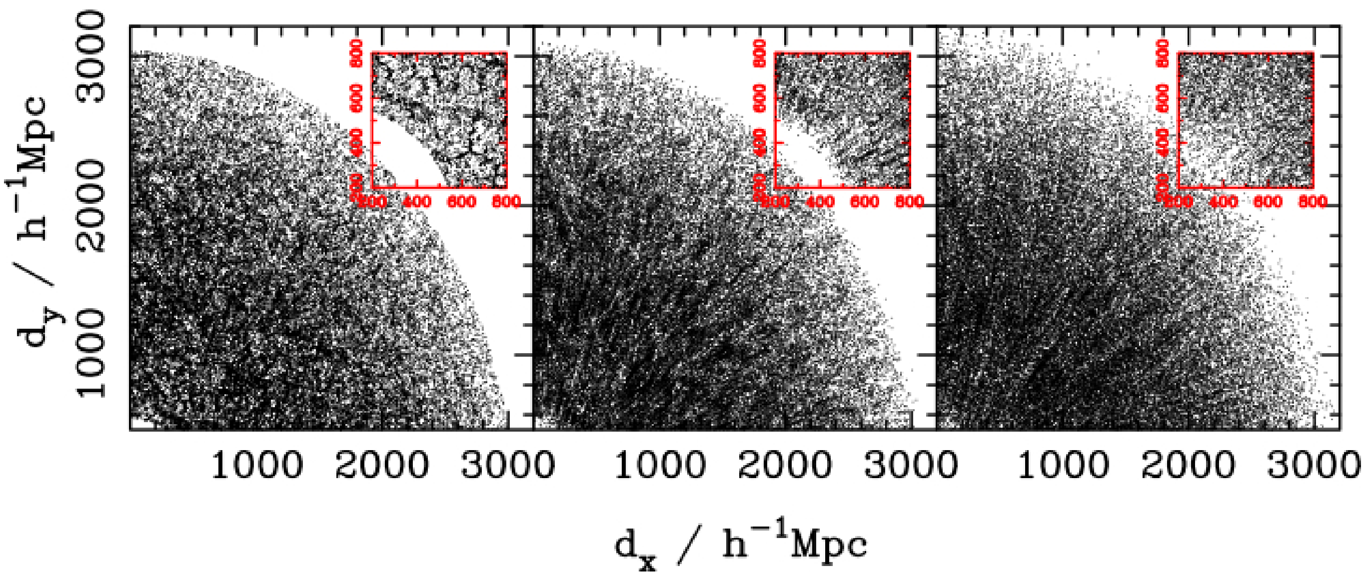

Figure 5.7: Angular distribution of DES LSS-WG MICE simulation data with varying photometric redshift uncertainties

5.3.2 Grid-Based Method

The high resolution of the MICE simulation prevented us from using traditional Monte-Carlo N^2 integration techniques, as it would have be computationally very expensive. Instead, we calculated the correlation function by assigning the data field to a density grid, as described in § 2.1.6. Using the nearest grid point (NGP) mass assignment scheme, we placed the data field onto $N_{grid} = 300^3$ cells, and estimate the correlation function using Eq. 2.46. One drawback to this method is that it only provides an accurate estimation of the correlation function on scales larger than a few grid cells. Getting down to smaller scales to observe non-linear effects requires a finer resolution grid, which can often result in $N_{grid} \rightarrow N_p$. In this analysis, we use a cell size $d_c \sim 10\,h^{-1}\,\mathrm{Mpc}$, which means we can rely on the accuracy of all measurements above $\sim 5 \times d_c = 50\,h^{-1}\,\mathrm{Mpc}$. This is fine, since we are interested in observing the effects of photometric redshift uncertainties on the correlation function around the scale of BAO, which is $> 5 \times d_c$.

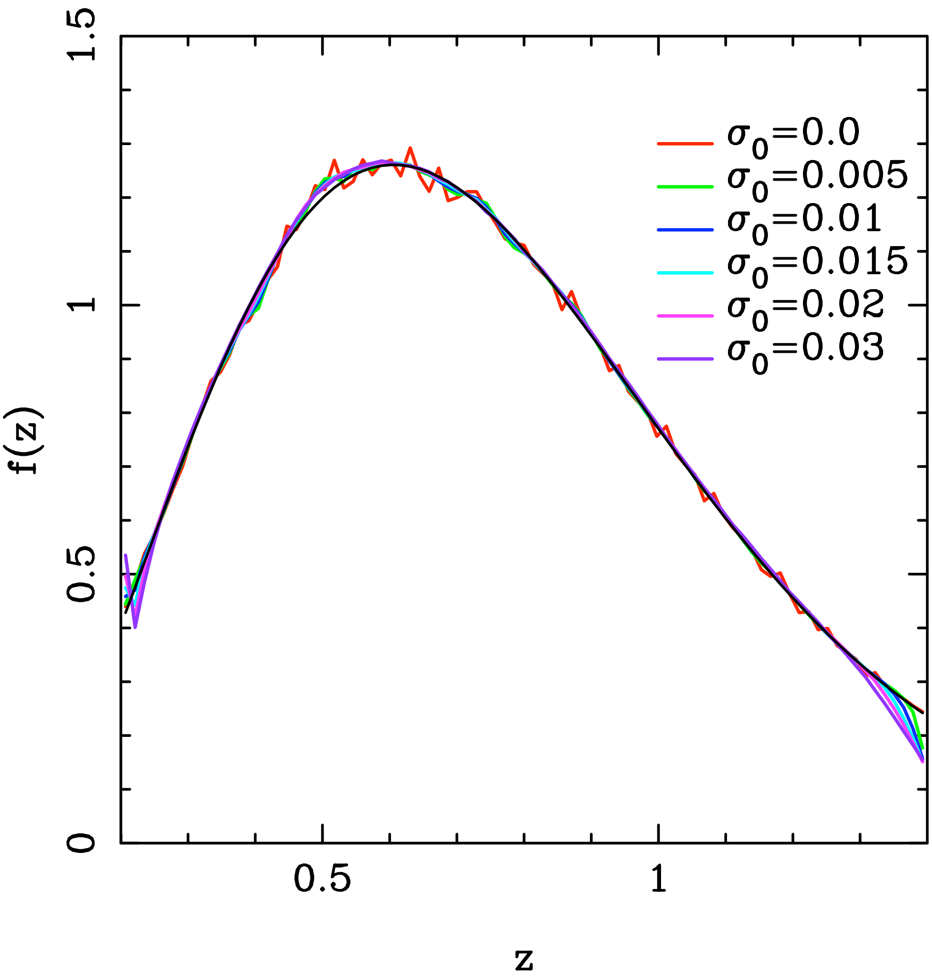

Figure 5.8: Radial distributions of DES LSS-WG MICE simulation data with varying photometric redshift uncertainty.

The symmetrical nature of the density grid was utilised to further speed up these calculations. The indices of the nearest neighbour N_{neigh} grid-cells for a given bin of separation were calculated and stored. This list of indices was then translated to different locations on the density grid. Therefore, no CPU time was wasted in re-computing the grid-cells that contribute to the correlation function in a given bin of separation, and the processing time was reduced from N_{grid}^2 to $N_{grid}N_{neigh}$ (Sanchez et al., 2008).

In total, we calculated the correlation function for 6 different levels of photometric redshift uncertainty: $\sigma_z = \sigma_0(1+z)$ where $\sigma_0 = 0.0,\ 0.005,\ 0.01,\ 0.015,\ 0.02,\ 0.03$, with the latter representing that expected for the DES. Results are plotted in Fig. 5.9. 3D correlation functions, $\xi(r)$, for each regime are plotted outset, with 3σ errors calculated via a jackknife resampling of 10 equal volume regions using the same coarse grid as described above. Models are shown as black lines and were calculated using a Monte Carlo integration over 10^{10} pairs of points, as outlined in § 4.3.5. The input 3D correlation function model was created using the transfer function of Eisenstein & Hu (1998) with a cosmology matching that of the MICE simulation. All models provide a very good fit to the data. Inset, we show the split correlation function $\xi(\sigma, \pi)$, where contours represent the amplitude of the correlation function at transverse, σ, and radial, π, separations. The baryon ridge is highlighted in red. The first panel represents the true correlation function where there is no uncertainty on galaxy positions. The BAO ridge can be clearly detected in both representations of the correlation function at $\xi(r = \sqrt{\sigma^2 + \pi^2}) \sim 100\,h^{-1}\,\mathrm{Mpc}$ in this regime. The BAO signal becomes less pronounced as a result of increasing photometric redshift uncertainty in the radial direction. This effect is particularly prominent in the π-plane of the split correlation function. At the level of photometric redshift uncertainty predicted for the DES we see that the BAO signal is completely degraded, further supporting the requirement for accurate projected clustering analyses much like the ones we have presented previously.

In Fig. 5.10 we compare the relative magnitude of systematic and statistical errors on our recovered 3D correlation function, calculated with photometric redshift uncertainties of $\sigma_z = 0.03(1+z)$, as predicted for the DES. Systematic errors are taken from the analysis on SDSS S82 data, as presented in Chapter. 4, and represent the uncertainty induced on clustering measurements due to discrepancies in photometric redshift values taken from different estimation techniques. In this case we take the result calculated around the mean of the 4 samples that were used. Confidence limits are plotted at the 1σ (light-grey), 2σ (mid-grey), and 3σ (dark-grey) levels. Statistical errors are represented by the results of our jackknife resampling of the MICE simulation. Confidence limits for these measurements are plotted at the 1σ (dark-blue), 2σ (light-blue), and 3σ (green) levels.

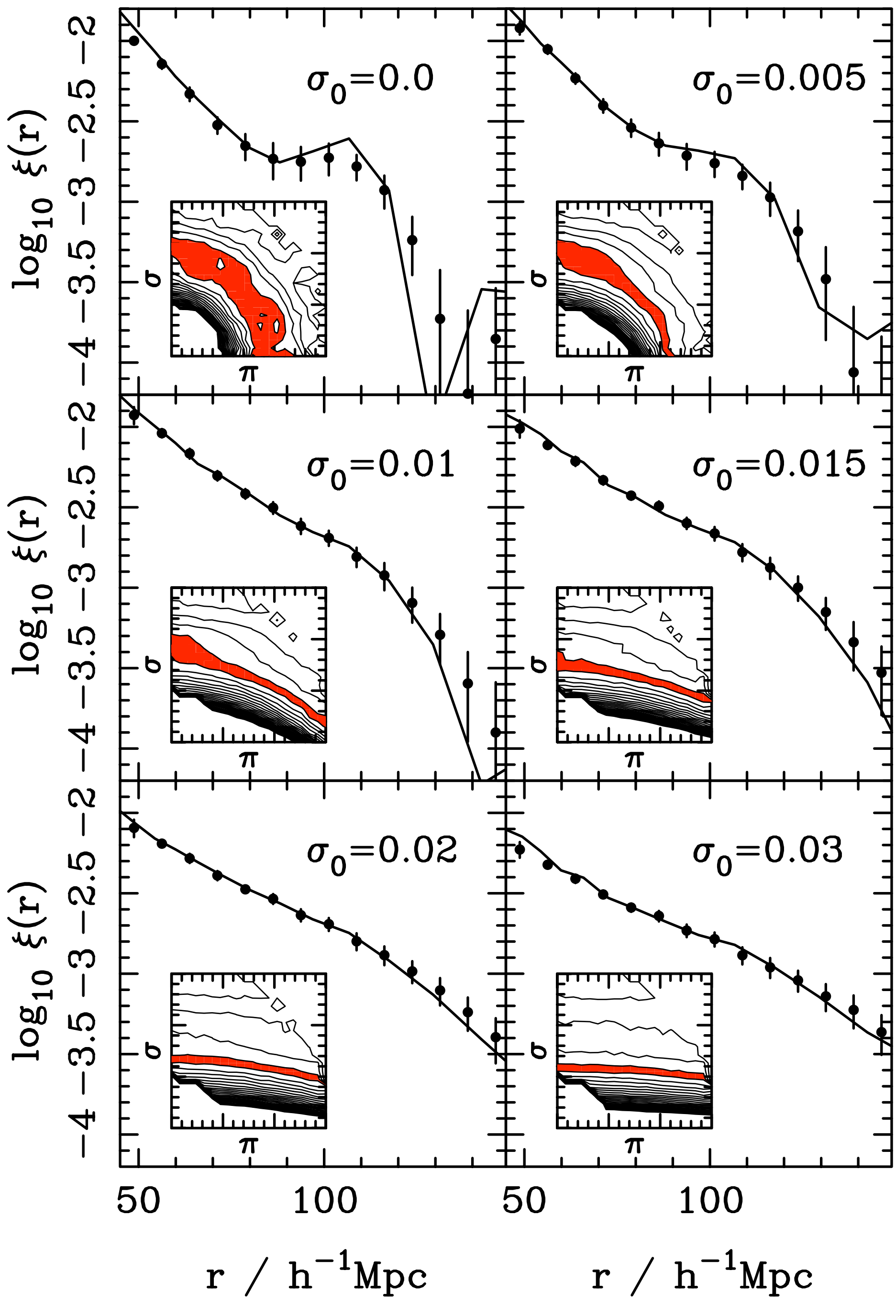

Figure 5.9: Filled circles represent 3D correlation functions calculated using MICE simulation data, where Gaussian photometric redshift uncertainties have been added to galaxy positions along the radial direction according to the relation $\sigma_z = \sigma_0(1+z)$. 3σ jackknife errors are plotted in each case, where jackknife resampling is conducted over 10 equal volume regions. Models are calculated via a Monte Carlo integration over 10^{10} pairs of points, as outlined in § 4.3.5, and show a very good fit to the data. Inset, we show the split correlation function $\xi(\sigma, \pi)$, where σ is the transverse galaxy separation and π is the radial. The baryon ridge is highlighted in red. The first panel represents the true correlation function where there is no uncertainty on radial galaxy positions. The BAO ridge is clearly visible at $\sim 100\,h^{-1}\,\mathrm{Mpc}$ in both the main and inset figures. As we move to larger photometric redshift uncertainties we can see that the BAO signal starts to diminish. This smearing of the BAO signal, due to the photometric redshift uncertainty in the radial direction, can be seen prominently in the $\xi(\sigma, \pi)$ plots.

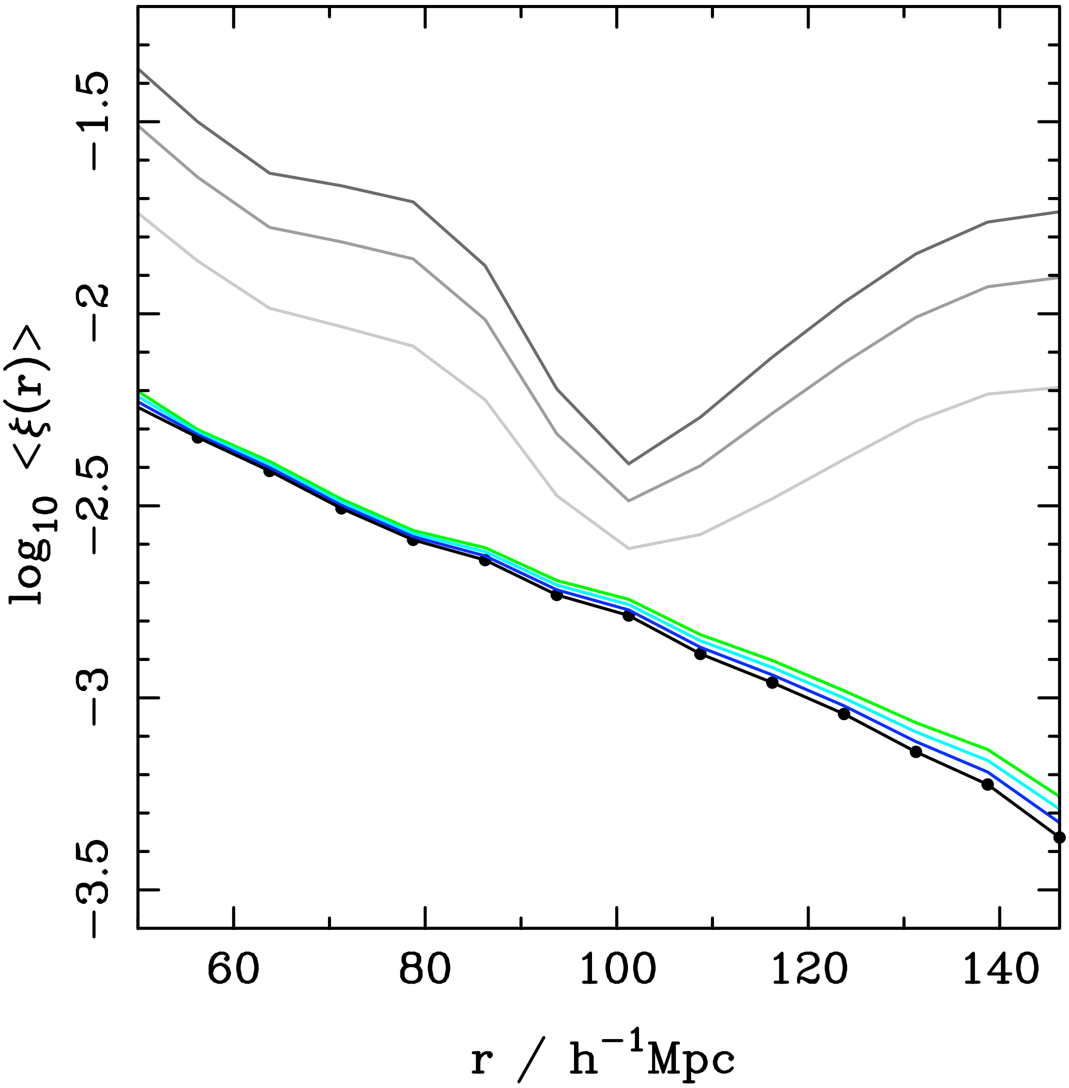

Figure 5.10: A comparison of systematic and statistical errors on the recovered 3D correlation function calculated using the MICE simulation data, where Gaussian photometric redshift uncertainties have been added to galaxy positions along the radial direction according to the relation $\sigma_z = 0.03(1 + z)$, as predicted for the DES. Systematic errors are taken from the analysis conducted in Chapter. 4 on SDSS S82 data and are plotted at the $+1\sigma$ (light-grey), $+2\sigma$ (mid-grey), and $+3\sigma$ (dark-grey) confidence levels. Statistical errors were calculated via a jackknife resampling method, conducted over 10 equal volume regions of the MICE simulation. Confidence limits are plotted at the $+1\sigma$ (dark-blue), $+2\sigma$ (light-blue), and $+3\sigma$ (green) levels. It is clear that systematic uncertainties are the dominant source of error on 3D clustering analyses for a survey like the DES. This implies that we cannot rely on photometric redshifts coming from the DES for 3D clustering analyses. It is likely however, that we have overestimated the systematic errors and underestimated the statistical errors in this analysis, in comparison to what can be expected for future analyses of real DES data. We discuss this in the main text.

The systematic errors clearly dominate in an analysis of this kind, thus implying that we cannot rely on photometric redshifts coming from the DES for 3D clustering analyses. However, it is important to note the following:

- The systematic errors that we have used here come from an analysis of real data, thus imposing a variety of nuisance effects on the result. For example, in this particular case we were hindered by both shot-noise and cosmic-variance limits, which would have boosted the level of uncertainty on the measurement. Another factor to consider is that the mean standard deviation of the photometric redshift samples we analysed was $\bar{\sigma}_z \sim 0.056094$, which is almost double that predicted for the DES. Taking these issues into consideration implies that we have overestimated the effect of systematic uncertainties for this analysis.
- The statistical errors we have calculated here come from an analysis of idealised simulation data, which means that it is likely that they are an underestimate of what we will actually get from DES data.

With these points in mind, it is still very likely that future 3D clustering analyses conducted with DES data will have errors dominated by systematic uncertainties, arising from conflicting estimates of photometric redshifts calculated via different estimation techniques, rather than our ability to construct accurate covariance matrices.

5.4 Discussions and Conclusions

Redshift-Space Distortions To quantify the effect of redshift-space distortions for future surveys, we have used the expected radial selection function and photometric redshift distribution for the Dark Energy Survey to predict the effect of redshift-space distortions on projected clustering measurements. This analysis is also relevant to other planned surveys such as PanStarrs and the LSST, which will have similar radial selection functions. We have contrasted two different types of binning: top-hat - in which we only allow galaxies between a given radial bound to enter our sample - and pair-centre - in which we only count galaxy pairs with an average radial position that lies within our bounds. For typical bin widths that will be applied to these surveys, we find that top-hat binning in the radial direction leaves a strong signal from redshift-space distortions. Using a pair-centre binning scheme can completely remove the effects of redshift-space distortions (see Fig. 5.5). In practice however, this regime presents a computational hurdle due to the increase in the number of galaxy pairs per redshift bin. Also, in using this regime we would be risking the excessive inclusion of galaxy pairs with large 3D separations on small-scales in the projected correlation function. As a consequence of this, we would see the position at which the projected correlation function goes negative shifted to smaller scales.

Introducing an upper limit on radial galaxy separation in the pair-centre binning scheme provides us with:

- A reduction in the redshift-space distortion signal by as much as 80% compared to traditional top-hat binning schemes (see right hand panel of Fig. 5.6), therefore allowing the measurements to be more sensitive to the cosmological parameters one wishes to constrain.

- A reduction in the number of expected galaxy pairs compared to the pair-centre binning scheme, thereby reducing the computational burden.

- A means of preserving the true position of the zero-point crossing of the projected correlation function by preventing the inclusion of uncorrelated galaxy pairs from large 3D scales.

In this analysis, we have only considered the simplified situation where the redshift-space distortions act along one axis of a Cartesian basis. However, the arguments we have put forward in favour of pair-centre binning do not rely on this assumption, and will remain valid even when wide-angle effects are included in any analysis.

Photometric Redshifts To quantify the effect of photometric redshifts for future surveys, we have used the Dark Energy Survey LSS WG MICE simulations, combined with results from the SDSS S82 clustering analysis in the previous chapter, to predict the effect of photometric redshift uncertainties on 3D clustering measurements.

Using the nearest grid-point mass assignment scheme we have calculated the 3D correlation function with increasing levels of photometric redshift uncertainty for the MICE simulation. Positions of galaxies were convolved in the radial direction with Gaussian uncertainties of the form $\sigma_z = \sigma_0(1+z)$, where σ_0 was set to 0, 0.005, 0.01, 0.015, 0.02, and 0.03. The latter represents the level of uncertainty expected for the DES [2]. Jackknife resampling was used to calculate the level of uncertainty on each clustering

[2]It should be noted here that an analysis of the photometric redshift requirements for the DES, conducted by Banerji et al. (2008), found that the optical photometry in the DES grizY bands may be complemented with near infra-red photometry from the planned VISTA Hemisphere Survey (VHS) in the JHK$_s$ bands to give $\sim 30\%$ improvement in the rms scatter on photometric redshift estimates over the redshift range $1 < z < 2$. If such a scheme were to be implemented in the DES, the predictions we have made here would become more optimistic.

measurement. We found the effect of photometric redshift uncertainties on the 3D correlation function acts to reduce the BAO signal in the radial direction. The BAO signal is completely washed out at the level of photometric redshift uncertainty predicted for the DES, indicating a strong requirement for alternative clustering measurements such as the projected correlation function.

Systematic uncertainties on the 3D correlation function, as calculated in the previous chapter using SDSS S82 data, have been combined with the Jackknife uncertainties calculated from the MICE simulations to predict the dominant form of uncertainty expected for 3D clustering analyses in the DES. From this analysis we found that systematic uncertainties due to discrepancies between photometric redshift estimation techniques is the major contributor to the overall level of uncertainty on the correlation function, rather than our ability to accurately construct a covariance matrix. This suggests that we need to concentrate our efforts in the area of photometric redshift estimation, rather than clustering analysis methodology, if we want to reduce the level of uncertainty on 3D clustering measurements for the DES.

It is important to note that the statistical and systematic uncertainties in this analysis were calculated from different data-sets. The systematic uncertainties were calculated for a *real* data-set from the SDSS, whilst the statistical uncertainties were calculated from an idealised *simulation* data-set. Because of this, it is entirely possible that the level of systematic uncertainty on the measurement has been overestimated and the statistical uncertainty has been underestimated. However, combined with the results of the previous section we can confidently predict that 3D clustering analyses conducted with DES data will mainly be limited by systematic uncertainties arising in differing photometric redshift estimation techniques.

Chapter 6

Conclusions

In this thesis, we have considered how the use of realistic photometric redshift estimates to derive radial distances to galaxies will affect the recovered clustering signal. We have shown that the predicted level of uncertainty on photometric redshifts for future experiments such as the Dark Energy Survey will effectively sabotage the acoustic feature in 3D clustering analyses. Consequently, we have argued that careful consideration of alternative clustering measurements - such as the projected correlation function - are required if we want to constrain the evolution of the dark energy equation of state via galaxy clustering. We highlight our main results and conclusions in the following sections.

6.0.1 Clustering Measurement Techniques

In this chapter we introduced the theory and methodology required to conduct accurate clustering analyses. We also carried out simple investigations to ascertain which correlation function estimator and measurement technique is the best to use.

To address the issue of which correlation function estimator to use, we considered a simple toy model that simulated a mismatch between galaxy and random catalogues. From this we concluded that the popular Landy-Szalay estimator provides the most robust measurement of the correlation function, due to its ability to recover the true clustering signal in cases when $\langle\delta\rangle \neq 0$. To test this theory, we compared the recovered correlation function for our suite of estimators using an SDSS LRG galaxy sample, where the radial selection had been deliberately mismatched for galaxies and randoms. The results of this analysis further highlighted where the Landy-Szalay estimator is superior to other estimators, as it was the only measurement that remained robust to the galaxy-random mismatch. In general, these results indicated a need for more accurate radial selection function modelling, which is in agreement with results from Kazin et al. (2010).

In consideration of the fact that the abundance of galaxy data coming from sky surveys is at a peak, we wanted to test by how much we can optimise our analysis by calculating a correlation function on a coarse grid. From a simple speed test we found that the use of a grid is optimal if:

- $N_p \gg N_{grid}$
- The structure of the grid is exploited so that the nearest neighbours may be stored and translated around the grid, thus reducing the algorithm from N_{grid}^2 to $N_{grid}N_{neigh}$, where $N_{neigh} \ll N_{grid}$.
- The size of the grid-cell $d_c > 2\,h^{-1}\,\mathrm{Mpc}$.

6.0.2 Redshift-Space Distortions and Binning Techniques

In this chapter, we used the Hubble Volume simulations to calculate the average projected correlation function in narrow redshift slices, and showed that the effect of redshift-space distortions on projected clustering measurements is far stronger than the redshift-space distortion effect on the 3D clustering signal. For this reason we have concluded that it is vital to account for redshift-space distortions in any interpretation of projected clustering analyses.

From an Eulerian perspective, we can think of redshift-space distortions causing an apparent coherent motion of galaxies into and out of samples at slice boundaries. The motions of the galaxies themselves however, are not responsible for the alteration of the clustering signal, and if we had knowledge of how the boundary changes we could alter the projection length to accommodate the motion. The alteration to the clustering signal is instead caused by the redshift-space boundary, which has an angular clustering signal that is correlated with the overdensity field. From a Lagrangian perspective, we have to consider that the projection simply does not remove redshift-space effects from the anisotropic correlation function. In this regime, we have demonstrated that it is relatively straightforward to accurately model the expected projected clustering signal by integrating the redshift-space correlation function over the radial selection function.

In reality, we often have an evolving real-space radial selection function, which represents the average comoving density of galaxies as a function of comoving radial distance. Applying an additional top-hat selection in redshift-space will result in a galaxy sample with a combination of real- and redshift-space boundaries. We have shown that this can be modelled by splitting the population into samples that can be considered to have top-hat windows in either real-space, redshift-space or a hybrid of the two. In the case of a hybrid-space selection, the projected correlation function can be modelled by using a weighted sum of the real- and redshift-space correlation function over the radial selection function. Prior to the publication of this work (Nock et al., 2010), no-one has considered how these hybrid selection functions affect the recovered projected clustering signal.

In an attempt to reduce the effect of redshift-space distortions on projected clustering analyses we have introduced a new measurement technique called *pair-centre* binning. In this scheme, galaxies are selected according to the apparent position of their pair-centre within a given radial bin, whereas the traditional top-hat binning scheme only selects pairs where both galaxies lie within the bin. This new scheme allows individual galaxies that traditionally lie *outside* the top-hat boundaries into the analysis. This means that we are no longer throwing away information, and that the effect of the coherent movement of galaxies between slice boundaries on the projected correlation function is reduced.

We have highlighted two potential disadvantages arising from the use of the pair-centre binning scheme, including:

1. Radial bin correlation: The same galaxy may be included in multiple redshift slices, which can introduce a correlation between radial bins.

2. Wider bins: This scheme results in necessarily wider radial bins, which causes the clustering signal to be diluted by the inclusion of uncorrelated pairs in the projected clustering signal on small-scales.

However, we do not feel that either effect is a large problem. Photometric redshift errors mean that covariance between overlapping radial bins is unavoidable in photometric surveys, even for the traditional top-hat binning scheme. The pair-centre binning scheme will not make this problem considerably worse. The dilution effect can be alleviated by imposing a maximum separation between the pairs included in a pair-centre bin: we call this constrained pair-centre binning. Applying the constrained pair-centre binning scheme helps to preserve the expected signal whilst simultaneously reducing the effect of redshift-space distortions. Thus, the use of such a scheme is considerably preferable to the traditional top-hat binning method.

6.0.3 Impact of Photometric Redshift Systematics

In this chapter we quantified the level of systematic uncertainty induced in clustering analyses by different redshift estimation techniques. We achieved this by comparing the recovered 3D correlation function for a single sample of galaxies from the Sloan Digital

Sky Survey Stripe 82 data-set, where we had five estimates of the galaxy redshift calculated via different techniques. From this simple empirical test we showed that the overall error on the correlation function measurement increased as a result of conflicting photometric redshift estimates. This extra component of *systematic uncertainty* has previously been ignored in anisotropic correlation function models.

In order to utilise BAO as a precise standard ruler in clustering analyses, we need to obtain accurate distance measurements to galaxies. It has been shown extensively throughout the literature that photometric redshift uncertainties induce errors on inferred radial distances, which act to *wash out* the clustering signal along the radial direction. Existing models incorporate photometric redshift uncertainties via a Gaussian convolution of the power spectrum in the radial direction. They do not incorporate the systematic uncertainty arising from different photometric redshift estimates that we have shown to dominate the total error on typical clustering measurements. Therefore, the use of these models to predict how accurately we can conduct clustering analyses in future photometric redshift surveys is fallible.

The results of this analysis suggest that we should be incorporating empirical test results into our models, when predicting the capabilities of BAO as a standard ruler in future photometric redshift surveys. We have a wealth of *real* data available to us from existing experiments like the SDSS, that we can use in such an analysis. This would not only provide us with more accurate predictions for the future, but would help us to better understand where these systematic uncertainties arise and how we can beat them down.

6.0.4 Future Experiments: The Dark Energy Survey

In this chapter we have predicted the levels of systematic uncertainties induced in projected clustering analyses through redshift-space distortions, and in 3D clustering analyses through different photometric redshift estimation techniques, for a future experiment like the Dark Energy Survey.

In the first case, we contrasted top-hat and pair-centre binning schemes for five redshift slices along an evolving radial selection function, similar to that expected for the DES. We found that the effects of redshift-space distortions can be completely removed when utilising our new pair-centre binning scheme that was introduced in Chapter 3. In this scheme no information is lost in the pair selection, unlike traditional top-hat methods. However, the CPU time required to employ such a scheme is $\sim 14\times$ higher than that required for top-hat binning. Coupled with the dilution effect caused by the necessarily wider bins, this scheme becomes unfeasible practically. Employing constrained pair-centre binning, where an upper limit is placed on radial separation, provides $\sim 80\%$ reduction in redshift-space distortions on the BAO scale and only increases CPU time by factor of 2, compared to the top-hat binning scheme.

This simple modification to the way we select our galaxy sample means that we can obtain accurate measurements of the acoustic signal in *narrow* redshift bins without being hindered by redshift-space distortions, thus enhancing our ability to make robust measurements of the evolution of the dark energy equation of state.

In the second case, we used the DES LSS WG MICE simulations to predict the effect of photometric redshift uncertainties on 3D clustering measurements. To simulate the effect of photometric redshift uncertainty in the analysis we convolved the true positions of the galaxies with Gaussian functions. We found that the acoustic signal was diminished as the level of uncertainty on the radial distance measure was increased, and that it was completely lost at the level of uncertainty predicted by the DES. This strongly indicates that alternative clustering analyses are required for future photometric surveys like the DES.

Another interesting outcome arose from contrasting the level of systematic and statistical error on the correlation function. We did this by combining the results of Chapter 4, where we measured the level of systematic error induced in the correlation function for SDSS Stripe 82 galaxies via discrepancies between photometric redshift estimates, with a simple Jackknife resampling of the MICE simulation data. We found that systematic uncertainties dominate the overall level of error on the correlation function rather than statistical errors, which determine our ability to construct an accurate covariance matrix. This further supports the need to combine empirical test results with existing anisotropic correlation function models, in order to understand where we can reduce these systematic uncertainties for future surveys.

Appendix A

General Relativity

A.1 The Friedmann Equations

In this appendix, we provide a full derivation of the the Friedmann (Eq. 1.4), acceleration (Eq. 1.5) and fluid (Eq. 1.6) equations, as defined in Chapter 1. The Einstein equation describing the relationship between geometry G and matter/energy T is given by

$$G^{\mu}_{\nu} = R^{\mu}_{\nu} - \frac{1}{2}Rg^{\mu}_{\nu} = 8\pi G T^{\mu}_{\nu}, \tag{A.1}$$

where the metric g^{μ}_{ν} in a FLRW universe is defined as

$$g^{\mu}_{\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{a^2}{1-Kr^2} & 0 & 0 \\ 0 & 0 & a^2r^2 & 0 \\ 0 & 0 & 0 & a^2r^2\sin^2\theta \end{pmatrix} \tag{A.2}$$

The Christoffel symbols, Riemann tensor, Ricci tensor and Ricci scalar are calculated via the equations:

$$\begin{aligned}
\Gamma_{\mu\nu\rho} &= \frac{1}{2}(g_{\mu\nu,\rho} + g_{\mu\rho,\nu} - g_{\nu\rho,\mu}) && \text{(A.3)} \\
R_{\mu\nu\rho\sigma} &= \Gamma_{\mu\nu\sigma,\rho} - \Gamma_{\mu\nu\rho,\sigma} + \Gamma_{\mu\tau\rho}\Gamma^{\tau}_{\nu\sigma} - \Gamma_{\mu\tau\sigma}\Gamma^{\tau}_{\nu\rho} && \text{(A.4)} \\
R_{\mu\nu} &= R^{\rho}_{\mu\rho\nu} && \text{(A.5)} \\
R &= g^{\mu\nu}R_{\mu\nu} = g^{ij}R_{ij} + g^{00}R_{00}, && \text{(A.6)}
\end{aligned}$$

with the space-space (i,j) and time-time (0,0) components given by

$$\begin{aligned} R_{ij} &= g_{ij}\left(\frac{\ddot{a}}{a}+\frac{2\dot{a}^2}{a^2}+\frac{2K}{a^2}\right), &\text{(A.7)}\\ R_{00} &= -3\frac{\ddot{a}}{a}, &\text{(A.8)}\\ R=g^{\mu\nu}R_{\mu\nu} &= g^{ij}R_{ij}+g^{00}R_{00} \\ &= 6\left(\frac{\ddot{a}}{a}+\frac{\dot{a}^2}{a^2}+\frac{K}{a^2}\right). &\text{(A.9)}\end{aligned}$$

The Friedmann and acceleration equations come from the G^0_0 and G^i_i components of Eq. A.1:

$$\begin{aligned} G^0_0 = R^0_0 - \frac{1}{2}g^0_0 R &= 8\pi G T^0_0 + \Lambda g^0_0 \\ 3\frac{\ddot{a}}{a} - \frac{6}{2}\left(\frac{\ddot{a}}{a}+\frac{\dot{a}^2}{a^2}+\frac{K}{a^2}\right) &= 8\pi G\rho + \Lambda \\ \frac{\dot{a}^2}{a^2} &= \frac{8\pi G\rho}{3} - \frac{K}{a^2} + \frac{\Lambda}{3} &\text{(A.10)}\end{aligned}$$

$$\begin{aligned} G^1_1 = R^1_1 - \frac{1}{2}g^1_1 R &= 8\pi G T^1_1 + \Lambda g^1_1 \\ \left(\frac{\ddot{a}}{a}+\frac{2\dot{a}^2}{a^2}+\frac{2K}{a^2}\right) - \frac{6}{2}\left(\frac{\ddot{a}}{a}+\frac{\dot{a}^2}{a^2}+\frac{K}{a^2}\right) &= 8\pi G P + \Lambda \\ \frac{-2\ddot{a}}{a} - \left[\frac{8\pi G\rho}{3}+\frac{\Lambda}{3}-\frac{K}{a^2}\right] - \frac{K}{a^2} &= 8\pi G P + \Lambda \\ \frac{\ddot{a}}{a} &= \frac{-4\pi G}{3}(3P+\rho) + \frac{\Lambda}{3}. &\text{(A.11)}\\ & &\text{(A.12)}\end{aligned}$$

The fluid equation can now be found by substituting Eq. A.12 into the time derivative of Eq. A.10, such that

$$\begin{aligned} 2\left(\frac{\ddot{a}}{a}\right)\left(\frac{\dot{a}}{a}\right) - 2\left(\frac{\dot{a}}{a}\right)^2\left(\frac{\dot{a}}{a}\right) - \frac{8\pi G\dot{\rho}}{3} + 2K\frac{\dot{a}}{a^3} &= 0 \\ \dot{\rho} + 3\frac{\dot{a}}{a}(\rho+P) &= 0. &\text{(A.13)}\end{aligned}$$

Appendix B

Dynamics of Structure and the Peculiar Velocity Field

Inflation theory can be used to show that the present structure in the Universe evolved from small fluctuations in the matter density, velocity and temperature fields (eg. BAO, CMB). On scales much less than the curvature of the Universe, we can use a combination of Eularian hydrodynamics and Newtonian mechanics to describe the physics of these cosmological perturbations. Firstly, let us define the equations that describe fluid motion:

$$\frac{D\rho}{Dt} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{B.1}$$

$$\frac{D\mathbf{v}}{Dt} = -\frac{\nabla p}{\rho} - \nabla\Phi, \tag{B.2}$$

$$\Delta\Phi = 4\pi G\rho, \tag{B.3}$$

where $D/Dt = \partial/\partial t + \mathbf{v} \cdot \nabla$ is the convective derivative, ρ is the matter density, $\mathbf{v}$ is the velocity field, Φ is the gravitational potential, Δ is the Laplacian, and G is the gravitational constant. Eq. B.1 is the continuity equation, Eq. B.2 is the equation of state, and Eq. B.3 is the Poisson equation.

In cosmology, it is common to work in comoving coordinates $\mathbf{r}$ that are related to the Eulerian coordinates $\mathbf{x}$ by

$$\mathbf{x}(t) = a(t)\mathbf{r}(t). \tag{B.4}$$

Differentiating Eq. B.4 we get

$$\begin{aligned}\dot{\mathbf{x}} = \mathbf{v} &= a\dot{\mathbf{r}} + \mathbf{r}\dot{a} &\text{(B.5)}\\ &= a\mathbf{u} + \mathbf{x}\frac{\dot{a}}{a} &\text{(B.6)}\\ &= \delta\mathbf{v} + H\mathbf{x}, &\text{(B.7)}\end{aligned}$$

where overdots denote time derivatives. We have introduced the notation $\mathbf{u} = \dot{r}$ here, so that the peculiar velocity $\delta\mathbf{v} = a\mathbf{u}$.

Since the continuity equation has the same form in all units, so we can automatically say that

$$\dot{\rho} + \Delta \cdot (\rho\mathbf{u}) = 0. \tag{B.8}$$

We can transform the equation of motion into comoving coordinates by differentiating $\mathbf{x}$ twice such that:

$$\begin{aligned}\ddot{\mathbf{x}} &= a\dot{\mathbf{u}} + \mathbf{u}\dot{a} + \frac{\dot{a}}{a}\dot{\mathbf{x}} + \mathbf{x}\dot{\mathbf{y}} &\text{(B.9)}\\ &= a\dot{\mathbf{u}} + \mathbf{u}\dot{a} + \frac{\dot{a}}{a}\left(a\mathbf{u} + \frac{\dot{a}}{a}\mathbf{x}\right) + \frac{\ddot{a}}{a}\mathbf{x} - \frac{\dot{a}^2}{a^2}\mathbf{x} &\text{(B.10)}\\ &= a\mathbf{u} + 2\dot{a}\mathbf{u} + \frac{\ddot{a}}{a}\mathbf{x} &\text{(B.11)}\\ &= -\nabla_x\Phi, &\text{(B.12)}\end{aligned}$$

where we have substituted $\mathbf{y} = \dot{a}/a$ at step B.9, and used the Eulerian equation of motion to introduce the last equality. The subscript $_x$ denotes the fact that we are calculating this gradient with respect to the Eulerian coordinates. Eq. B.4 says that

$$\nabla_x = \frac{1}{a}\nabla_r. \tag{B.13}$$

Note: From here onwards we will omit the subscripts to keep the notation simple. If we now use the Eulerian equation of motion together with the Poisson equation for the case of smooth motion ie. $\mathbf{u} = 0$ so that $\mathbf{v} = (\dot{a}/a)\mathbf{x}$, we get

$$\begin{aligned}\dot{\mathbf{v}} + \mathbf{v}\cdot\nabla\mathbf{v} &= \mathbf{y}\dot{\mathbf{x}} + \mathbf{x}\dot{\mathbf{y}} &\text{(B.14)}\\ &= \frac{\dot{a}}{a}\dot{\mathbf{x}} + \mathbf{x}\left(\frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2}\right) &\text{(B.15)}\\ &= \frac{\ddot{a}}{a}\mathbf{x}, &\text{(B.16)}\\ &= -\frac{\nabla p}{\rho} - \boldsymbol{\Delta}\bar{\Phi} &\text{(B.17)}\end{aligned}$$

where $\nabla\bar{\Phi}$ is the solution of the equation

$$\nabla\bar{\Phi} = 4\pi G\bar{\rho}, \tag{B.18}$$

which diverges at infinity. We use this solution to eliminate the $(\ddot{a}/a)\mathbf{x}$ term from Eq. B.12 so that

$$\mathbf{u} + 2\frac{\dot{a}}{a}\mathbf{u} = \mathbf{g}, \tag{B.19}$$

where $\mathbf{g} = -\nabla\phi/a^2$ is the peculiar acceleration, and Φ is now the perturbation of the gravitational potential $\Phi = \Phi - \bar{\Phi}$. Because the Laplace operator is linear, we can write

$$\nabla\Phi = 4\pi G(\rho - \bar{\rho}). \tag{B.20}$$

We will also use the overdensity δ instead of density ρ where

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}}, \tag{B.21}$$

so that the equations of dynamics in comoving coordinates now become:

$$\dot{\delta} + \nabla \cdot [(1+\delta)\mathbf{u}] = 0, \tag{B.22}$$

$$\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\mathbf{u} = \mathbf{g}, \tag{B.23}$$

$$\nabla\Phi = \frac{4\pi G\bar{\rho}\delta}{a}. \tag{B.24}$$

In the linear approximation we can neglect the term $\delta\nabla \cdot \mathbf{u}$ in the continuity equation so that

$$\nabla \cdot \mathbf{u} = -\dot{\delta}, \tag{B.25}$$

when both δ and $\mathbf{u}$ are small. This tells us that it is possible to have vorticity modes with $\nabla \cdot \mathbf{u} = 0$, for which δ vanishes. δ grows or decays as a power of time. These modes require zero density perturbation, which means that the associated peculiar gravity also vanishes. Therefore, these modes are the homogeneous solutions and decay as $\mathbf{v} = a\mathbf{u} \propto a^{-1}$. Taking the appropriate derivatives, Eqns. B.22 and B.23 now become:

$$\frac{d^2\delta}{dt^2} + \nabla \cdot \frac{\partial \mathbf{u}}{\partial t} = 0, \tag{B.26}$$

$$\nabla \cdot \frac{\partial \mathbf{u}}{\partial t} + 2\frac{\dot{a}}{a}\nabla \cdot \mathbf{u} = \mathbf{g}. \tag{B.27}$$

Combining Eqns. B.18, B.25, B.26 and B.27 we get a second-order equation for the evolution of the density contrast:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} - 4\pi G\bar{\rho}\delta = 0. \tag{B.28}$$

Eq. B.28 is a linear second-order ordinary differential equation, and its solution can be written as

$$\delta(\mathbf{r}, t) = A(\mathbf{r})D_1(t) + B(\mathbf{r})D_2(t), \tag{B.29}$$

where the partial solutions $D_1(t)$ and $D_2(t)$ are the *growing* and *decaying* modes, respectively. A solution for the growing modes can be obtained from the linear velocity equation (Eq. B.23), where the gravitational acceleration g can be written

$$\mathbf{g}(\mathbf{x}) = -G\bar{\rho}\nabla \int \frac{\delta(\mathbf{x}')}{|\mathbf{x}' - \mathbf{x}|}, \tag{B.30}$$

which is found by solving the Poisson equation (Eq. B.24). Comparing the velocity equation (Eq. B.23) and the density equation (Eq. B.28), we can see that the velocity component resulting from the first density mode $D_1(t)$ can be written

$$\mathbf{u} = \frac{\mathbf{g}}{4\pi G\bar{\rho}D_1}\frac{dD_1}{dt}. \tag{B.31}$$

Now, since the linear approximation for the density contrast is a function of time only, we can write

$$\frac{d\delta}{dt} = \frac{da}{dt}\frac{d\delta}{da} = \dot{a}\frac{\delta}{a}\frac{d\log\delta}{d\log a} = H\delta\frac{d\log D_1}{d\log a}. \tag{B.32}$$

The continuity equation (Eq. B.25) now gives us

$$\delta = -\frac{1}{Hf(a)}\nabla \cdot \mathbf{u}. \tag{B.33}$$

The formula for the velocity (Eq. B.31) can be rewritten:

$$\mathbf{v} = \frac{2f(\Omega)}{3H\Omega}\mathbf{g}, \tag{B.34}$$

where some good approximations for the function $f(\Omega) = \frac{d\log D_1}{d\log a}$ include:

$$f(\Omega) \approx \Omega_m^{0.6} + \frac{1}{70}\Omega_\Lambda(1 + \Omega_m/2), \tag{B.35}$$

by Lahav et al. (1991),

$$f(\Omega_m) \approx \Omega_m(a)^\gamma, \tag{B.36}$$

by Linder (2005), where γ is the gravitational growth index, and

$$f(\Omega_m) \approx \Omega_m^{\alpha(w)}, \tag{B.37}$$

where

$$\alpha(w) \approx \frac{3}{5 - w/(1-w)} \tag{B.38}$$

by Wang & Steinhardt (1998) for an epoch-independent w.

Working in Fourier space, we can use the fact that $\mathbf{g}$ and $\mathbf{k}$ are parallel to show that

$$\nabla \cdot \mathbf{u} = -i\mathbf{k} \cdot \mathbf{u} = -iku. \tag{B.39}$$

Now we can obtain the velocity equation directly from the continuity equation:

$$\delta \mathbf{v_k} = -\frac{iHf(\Omega)a}{k}\delta_k \hat{\mathbf{k}}. \tag{B.40}$$

References

Alcock, C., & Paczynski, B. 1979, Nature, 281, 358

Amendola, L., Polarski, D., & Tsujikawa, S. 2007, Phys. Rev. Lett., 98, 131302

Armendariz-Picon, C., Mukhanov, V., & Steinhardt, P. J. 2001, Phys. Rev. D., 63, 103510

Baldauf, T., Smith, R. E., Seljak, U., & Mandelbaum, R. 2010, Phys. Rev., D81, 063531

Banerji, M., Abdalla, F. B., Lahav, O., & Lin, H. 2008, Mon. Not. Roy. Astron. Soc., 386, 1219

Bardeen, J. M., et al. 1986, Ap. J., 304, 15

Barriga, J., & Gaztanaga, E. 2002, Mon. Not. Roy. Astron. Soc., 333, 443

Baugh, C. M. 2006, Reports on Progress in Physics, 69, 3101

Benitez, N. 2000, Astrophys. J., 536, 571

Bertotti, B., Iess, L., & Tortora, P. 2003, Nature, 425, 374

Blake, C., & Bridle, S. 2005, Mon. Not. Roy. Astron. Soc., 363, 1329

Blake, C., Collister, A., Bridle, S., & Lahav, O. 2007, Mon. Not. Roy. Astron. Soc., 374, 1527

Blake, C., & Glazebrook, K. 2003, Ap. J., 594, 665

Blanchard, A. 2010, ArXiv e-prints

Blanton, M. R., et al. 2003, Astrophys. J., 592, 819

Bolzonella, M., Miralles, J.-M., & Pello', R. 2000, Astron. Astrophys., 363, 476

Bond, J. R., & Efstathiou, G. 1984, Astrophys. J., 285, L45

Budavari, T., Szalay, A. S. ., Csabai, I., Connolly, A. J., & Tsvetanov, Z. 2001, Astron. J., 121, 3266

Caldwell, R. R., & Linder, E. V. 2005, Phys. Rev. Lett., 95, 141301

Chiba, T. 2003, Phys. Lett. B., 575, 1

Cole, S., et al. 2005, Mon. Not. Roy. Astron. Soc., 362, 505

Collister, A. A., & Lahav, O. 2004, Publ. Astron. Soc. Pac., 116, 345

Connolly, A. J., et al. 1995, Astron. J., 110, 2655

Cooray, A., & Sheth, R. 2002, Phys. Rept., 372, 1

Copeland, E. J., Sami, M., & Tsujikawa, S. 2006, IJMPD, 15, 1753

Csabai, I., et al. 2003, Astron. J., 125, 580

Davis, M., & Peebles, P. J. E. 1982, Astrophys. J., 267, 465

Deffayet, C. 2001, Phys. Lett. B., 502, 199

Dressler, A. 1980, Ap. J., 236, 351

Dvali, G., Gabadadze, G., & Porrati, M. 2000, Phys. Lett. B., 485, 208

Edmondson, E. M., Miller, L., & Wolf, C. 2006, Mon. Not. Roy. Astron. Soc., 371, 1693

Efron, B. 1979, Annals of Statistics, 7, 1

Eisenstein, D. J., & Hu, W. 1998, Astrophys. J., 496, 605

Eisenstein, D. J., Seo, H.-J., & White, M. 2007, Ap. J., 664, 660

Eisenstein, D. J., et al. 2001, Astron. J., 122, 2267

—. 2005, Astrophys. J., 633, 560

Eriksen, H. K., Banday, A. J., Gorski, K. M., & Lilje, P. B. 2005, Astrophys. J., 622, 58

Evrard, A. E., et al. 2002, Astrophys. J., 573, 7

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, Astrophys. J., 426, 23

Firth, A. E., Lahav, O., & Somerville, R. S. 2003, Mon. Not. Roy. Astron. Soc., 339, 1195

Fisher, K. B., Scharf, C. A., & Lahav, O. 1994a, Mon. Not. Roy. Astron. Soc., 266, 219

Fisher, K. B., et al. 1994b, Mon. Not. Roy. Astron. Soc., 267, 927

Hamilton, A. J. S. 1992, ApJL, 385, L5

—. 1993, Astrophys. J., 417, 19

Hamilton, A. J. S. 1997, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185–+

Hewett, P. C. 1982, Mon. Not. Roy. Astron. Soc., 201, 867

Holtzman, J. A. 1989, Astrophys. J. Suppl., 71, 1

Hu, W., & Sugiyama, N. 1995, Astrophys. J., 444, 489

Hubble, E. 1929, Proc. Nat. Acad. Sci, 15, 168

—. 1934, Ap. J., 79, 8

Huetsi, G. 2006, Astron. Astrophys., 449, 891

Kaiser, N. 1984, ApJL, 284, L9

—. 1987, Mon. Not. Roy. Astron. Soc., 227, 1

Kayo, I., Taruya, A., & Suto, Y. 2001, Ap. J., 561, 22

Kazin, E. A., et al. 2010, Astrophys. J., 710, 1444

Kerscher, M., Szapudi, I., & Szalay, A. 2000, Ap. J., 535, L13

Kessler, R., et al. 2009, Astrophys. J. Suppl., 185, 32

Komatsu, E., et al. 2010, arXiv

Kowalski, M., et al. 2008, Astrophys. J., 686, 749

Koyama, K. 2008, Gen. Rel. Grav., 40, 421

Lahav, O., Lilje, P. B., Primack, J. R., & Rees, M. J. 1991, Mon. Not. Roy. Astron. Soc., 251, 128

Landy, S. D., & Szalay, A. S. 1993, Astrophys. J., 412, 64

Larson, D., et al. 2010, ArXiv e-prints

Linder, E. V. 2005, Phys. Rev. D., 72, 043529

Lobo, F. S. N. 2008, ArXiv e-prints, jul

Maartens, R. 2004, Living. Rev. Rel., 7, 7

Miller, R. G. 1974, Biometrika, 61, 1

Mo, H. J., Jing, Y. P., & Borner, G. 1993, Mon. Not. Roy. Astron. Soc., 264, 825

Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J., & Tozzi, P. 1999a, ApJL, 524, L19

Moore, B., Quinn, T., Governato, F., Stadel, J., & Lake, G. 1999b, Mon. Not. Roy. Astron. Soc., 310, 1147

Munshi, D., Valageas, P., Van Waerbeke, L., & Heavens, A. 2008, Phys. Rept., 462, 67

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, Ap. J., 490, 493

Neyman, J., & Scott, E. L. 1952, Ap. J., 116, 144

Neyman, J., Scott, E. L., & Shane, C. D. 1953, Ap. J., 117, 92

Nock, K., Percival, W. J., & Ross, A. J. 2010, Mon. Not. Roy. Astron. Soc., 407, 520

Oyaizu, H., et al. 2008, Astrophys. J., 674, 768

Padmanabhan, N., et al. 2007, Mon. Not. Roy. Astron. Soc., 378, 852

Peacock, J. A. 1999, Cosmological physics (Cambridge, UK: Univ. Pr.)

Peacock, J. A., et al. 2001, Nature, 410, 169

Peebles, P. 1973, Astrophys. J., 185, 413

Peebles, P. J. E. 1980, The Large Scale Structure of the Universe (Princeton University Press)

Peebles, P. J. E., & Yu, J. T. 1970, Astrophys. J., 162, 815

Percival, W. J. 2007, Lect. Notes Phys., 720, 157

Percival, W. J., & Schaefer, B. M. 2008, Mon. Not. Roy. Astron. Soc., 385, L78

Percival, W. J., Verde, L., & Peacock, J. A. 2004, Mon. Not. Roy. Astron. Soc., 347, 645

Percival, W. J., et al. 2007a, Astrophys. J., 657, 51

—. 2007b, Astrophys. J., 657, 645

—. 2010, Mon. Not. Roy. Astron. Soc., 401, 2148

Perivolaropoulos, L. 2008, ArXiv e-prints

Perlmutter, S., et al. 1999, Astrophys. J., 517, 565

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. 1992, Numerical Recipes in C, 2nd edn. (Cambridge, UK: Cambridge University Press)

Randall, L., & Sundrum, R. 1999a, Phys. Rev. Lett., 83, 4690

—. 1999b, Phys. Rev. Lett., 83, 3370

Ratra, B., & Peebles, P. J. E. 1988, Phys. Rev. D., 37, 3406

Regos, E., & Szalay, A. S. 1995, Mon. Not. Roy. Astron. Soc., 272, 447

Riess, A. G., et al. 1998, Astron. J., 116, 1009

—. 2009, Ap. J., 699, 539

Ross, A. J., & Brunner, R. J. 2009, Mon. Not. Roy. Astron. Soc., 399, 878

Sanchez, A. G., Baugh, C. M., & Angulo, R. 2008, Mon. Not. Roy. Astron. Soc., 390, 1470

Saunders, W., Rowan-Robinson, M., & Lawrence, A. 1992, Mon. Not. Roy. Astron. Soc., 258, 134

Shanks, T. 2005, in IAU Symposium, Vol. 216, Maps of the Cosmos, ed. . R. A. S. M. Colless, L. Staveley-Smith, 398–+

Shao, J. 1986, Annals of Statistics, 14, 1322

Silk, J. 1968, Astrophys. J., 151, 459

Simpson, F., Peacock, J. A., & Simon, P. 2009, Phys. Rev., D79, 063508

Smoot, G. F., et al. 1991, ApJL, 371, L1

Spergel, D. N., et al. 2007, Astrophys. J. Suppl., 170, 377

Strauss, M. A., et al. 2002, Astron. J., 124, 1810

Sunyaev, R. A., & Zeldovich, Y. B. 1970, Astrophys. Space Sci., 7, 3

Tegmark, M., et al. 2004, Astrophys. J., 606, 702

Thomas, D., Maraston, C., Schawinski, K., Sarzi, M., & Silk, J. 2010, Mon. Not. Roy. Astron. Soc., 404, 1775

Tukey, J. W. 1958, Annals of Mathematical Statistics, 29, 614

Vanzella, E., et al. 2004, Astron. Astrophys., 423, 761

Wang, L., & Steinhardt, P. J. 1998, Ap. J., 508, 483

Wolf, C., et al. 2003, Astron. Astrophys., 401, 73

Zehavi, I., et al. 2004, Astrophys. J., 608, 16

Zwicky, F. 1937, Ap. J., 86, 217