# Cosmology with Photometric Redshifts


Zahra Gomes

Somerville College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Michaelmas 2019

# Contents

# Abstract

Upcoming large scale photometric surveys will require accurate photometric redshifts (photo-zs) to optimally extract astrophysical information. This thesis investigates methods of making and improving cosmological measurements using photo-zs.

Gaussian Processes for Photometric Redshift Estimation (GPz) has been proven to provide accurate estimates and reliable uncertainty estimates. In Chapter 2 I evaluate the effects of adding near-IR magnitudes and angular size as features for training and find improvements in accuracy of $\sim 15-20$ per cent. A method of shifting the photo-zs based on Quantile-Quantile plots is also investigated and improves the bias by $\sim 40$ per cent. In Chapter 3 I use GPz based photo-zs to measure the baryon acoustic oscillations with SDSS galaxies. Two galaxy samples are used: one with all galaxy types and another with luminous galaxies. Using multi-wavelength photometry and morphological data, the errors on the photo-z estimates for the luminous galaxies showed an improvement on previous studies. Angular correlation functions are measured and BAO peaks are detected for the luminous galaxies at positions consistent with expected results. The whole galaxy sample, with a magnitude cut of $i < 21$ results in BAO peaks for the bins $0.5 < z < 0.6$ and $0.6 < z < 0.7$ at $4.09 \pm 0.16°$ and $3.44 \pm 0.13°$. This result implies that the photo-zs are sufficiently accurate to make BAO measurements without being restricted to luminous galaxies.

In Chapter 4 I forecast the constraints on the local primordial non-Gaussianity parameter $f_{\rm NL}$ that can be obtained with future radio surveys and cross-matched photo-zs. I utilize the multi-tracer method, and improvements on previous work include the use of observational bias and halo mass estimates, updated simulations and realistic photo-z expectations. In the most realistic case, with photo-zs up to $z = 2$ the 1-$\sigma$ error falls between 4.07 and 6.58, rivalling the tightest constraints currently available. If photo-zs are available to $z = 5$ this improves to between 1.5 and 2.

# Declaration

This thesis is being submitted for the degree of Doctor of Philosophy at the University of Oxford in the month October 2019. The work presented was carried out in the Department of Astrophysics, University of Oxford between October 2016 and October 2019 under the supervision of Professor Matt Jarvis. Apart from where otherwise stated or the work of others is referenced the content of this thesis is my own, original work. No part of this thesis has been submitted in support of another degree, diploma or other qualification at any higher learning institute.

---

Zahra Gomes (October 2019)

# Acknowledgements

My first thanks goes to my supervisor Prof. Matt Jarvis for his regular guidance with and oversight of my thesis. He had the necessary experience and knowledge to steer my thesis work along a sturdy path. In addition, his encouragement and reassurance at each meeting played an important role in maintaining my mental well-being throughout my DPhil experience, particularly in the more stressful times.

I would also like to thank others who I have collaborated with over the years: Prof. Stefano Camera who hosted me at the University of Turin for a few weeks, Catherine Hale, Jośe Fonseca, Ibraham Almosallam, Prof. Steve Roberts and David Alonso. Other members of my group who provided a source of friendship and assistance when necessary were Peter Hatfield, Nathan Adams, Rebecca Bowler, Leah Morabito, Corentin Schreiber and Josie Peters.

My greatest source of emotional support came from my partner Stefan Hosein who stood beside me throughout my time at Oxford providing a stable source of comfort and support and extra encouragement in particularly difficult times. I would also like to thank Stefan for always believing that I can achieve more than I think and for encouraging me to apply for a Rhodes scholarship to attend Oxford in the first place. I am tremendously grateful to my parents, Vanda and Stanley Gomes and sister, Zandra Gomes for supporting me throughout the years, both during and long before this DPhil and I would like to give a special thanks to my mother for settling me in at Oxford when I first arrived.

My time at Oxford would have been significantly more difficult without the support and love I experienced thanks to the friendships I made over my time in Oxford. Some of these friends are Jenna Hebert, Farah Shammout, Darsh Kodwani, Richard Grumitt and my entire community of Caribbean friends who truly made me feel at home.

My DPhil experience was funded and supported by a Rhodes Scholarship granted by the Rhodes Trust. I would like to express my deepest gratitude to the Rhodes Trust, and in particular those who worked directly with the scholar community: Mary Eaton, Nadiya Figueroa and both wardens: Charles Conn and Dr. Elizabeth Kiss. They were able to provide for me a physical space and community where I was able to feel comfortable and supported, make many friends from around the world and engage in discussions, projects and extra-curricular activities covering a wide range of fields, truly enhancing my Oxford experience.

Zahra Gomes (October 2019)

# List of Figures

# Chapter 1

# Introduction

Astronomy has intrigued humanity for centuries. People of all ages have pursued an understanding of the nature and evolution of the space in which we exist. Theories have taken many different forms: from simple theories of gods making and controlling the earth and the sky, to the Greek theory of matter involving the four elements: fire, water, earth and air which separated spatially leading to the structure of the earth and sky, and finally to the modern cosmological model that describes the entire Universe and the nature of space-time. Our scientific understanding of the Universe has come a long way and scientists are always exploring new theories, developing new analysis techniques and building larger, more sophisticated telescopes. Encouraged by significant achievements and advancements in our theoretical understanding and observational abilities, particularly over the last century, we continue to search for answers.

This thesis is my contribution to this quest. This work explores some methods of making cosmological measurements using photometric redshifts. All of the necessary background is provided in this introduction: the first section will introduce the study of cosmology, describe the best current cosmological model and discuss some common cosmological probes, while the second section will focus on photometry and techniques for estimating photometric redshifts.

## 1.1 Cosmology

Cosmology is the study of the entire Universe as a whole, its origin and evolution, as opposed to the study of individual objects. Humanity's view of cosmology has morphed over time: from believing that the Earth was the centre of the Universe, to thinking that the Milky Way contained the entire expanse of the static Universe, and finally to discovering that we live in an expanding Universe, which started with a Big Bang and in which our galaxy is just one of billions. These discoveries, made in the $20^{th}$ century together with

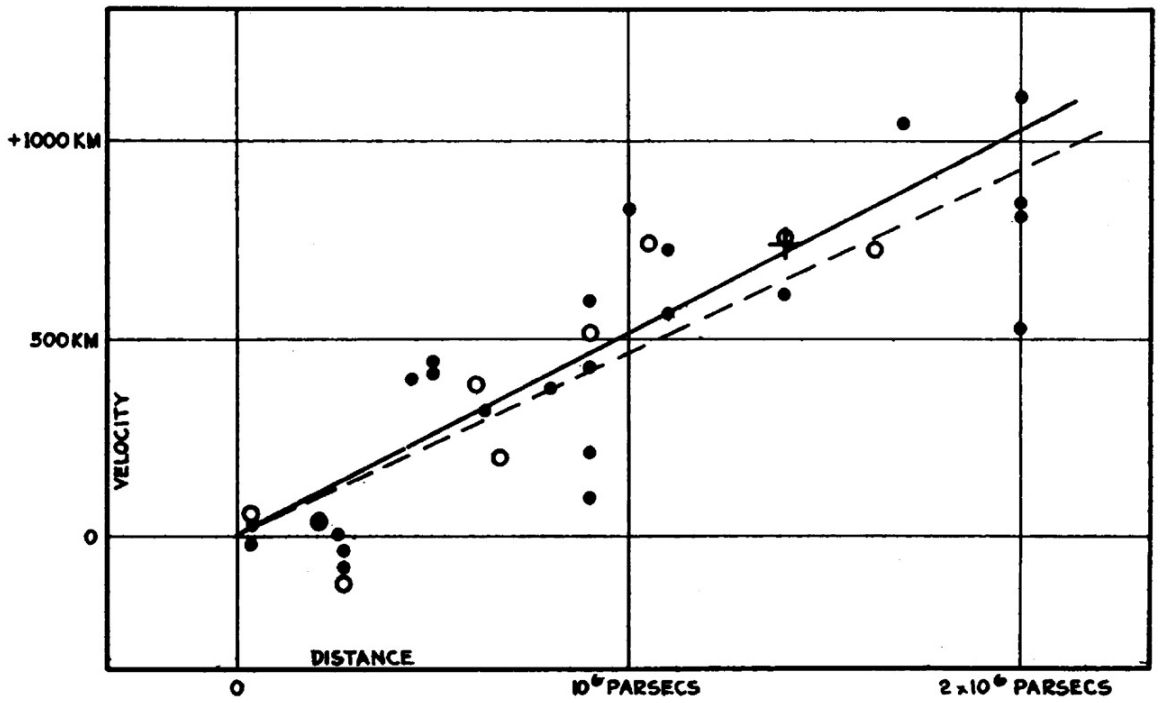



Figure 1.1: Plot of velocity versus distance for 24 extra-galactic nebulae. Taken from Hubble (1929). The value of $H_0$ inferred from this plot is $500$ kms$^{-1}$Mpc$^{-1}$, a significantly higher value than today's measurements of $\sim 70$ kms$^{-1}$Mpc$^{-1}$. This is due to the fact that all the galaxies used were located very nearby and their distances were incorrectly calculated by Hubble.

the development of General Relativity, the framework to understand space-time, led to the birth of modern day cosmology.

### 1.1.1 An Expanding Universe

Until the $20^{th}$ century the standard belief among astronomers was that our Universe was static and its age infinite. This changed when Edwin Hubble discovered a linear relationship between the recessional velocity, $v$ of galaxies and their distance from Earth, $d$ (Hubble 1929). This is shown in Figure 1.1 and the law is given below:

$$v = H_0 d, \tag{1.1.1.1}$$

or

$$z \approx \frac{H_0}{c} d \tag{1.1.1.2}$$

where $z \simeq v/c$ for $v \ll c$. $H_0$ is the Hubble constant which was later generalised to the Hubble parameter with $H_0$ representing the present day value. The Hubble parameter, $H$, can be expressed in terms of the scale factor, $a$: $H = \frac{\dot{a}}{a}$. The scale factor is the ratio of the

distance between two objects at some time in the past, $t$ to the distance now and therefore represents the expansion of space. Redshift, $z$, represents the change in wavelength of photons as they travel through an expanding space. This is a general relativistic effect that can be interpreted as the stretching of the electromagnetic waves (increase in wavelength) as they travel though a stretching space. It is given as:

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{1}{a(t)} \tag{1.1.1.3}$$

where $\lambda_o$ is the observed wavelength of a feature on the spectrum of an object and $\lambda_e$ is the wavelength of the same feature when it was emitted (the rest-frame wavelength). $a(t)$ is the scale factor at the time at which the photons were emitted.

### 1.1.2 FLRW Metric

Einstein's General Relativity (Einstein 1915) is the currently accepted theory of gravity. In this theory, matter and energy cause space-time to curve and space-time determines how matter moves. This relationship is described by the Einstein field equation below:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R - g_{\mu\nu}\Lambda = \frac{8\pi G}{c^4}T_{\mu\nu} \tag{1.1.2.1}$$

where $R_{\mu\nu}$ is the Ricci tensor that describes the curvature of space-time, $R$ is the Ricci scalar, $T_{\mu\nu}$ is the stress-energy tensor that represents the matter content of the universe and its properties, $G$ is the gravitational constant $6.6710^{-11}$ $\mathrm{m^3kg^{-1}s^{-2}}$ and $g_{\mu\nu}$ is the metric (the measurement used for distance in curved space-time). $\Lambda$ is the cosmological constant which was originally introduced to allow a static Universe but was removed when the Universe was found to be expanding. In the late 1990's the discovery that the Universe's expansion is accelerating (Riess et al. 1998; Perlmutter et al. 1999) led to the reintroduction of this constant and an introduction to the concept of dark energy.

The metric solution to this equation for a universe that is isotropic and homogeneous is the Friedmann-Lemaître-Robertson-Walker (FLRW) metric:

$$ds^2 = dt^2 - a^2(t)(\frac{dr^2}{1-kr^2} + r^2d\theta^2 + r^2\mathrm{sin}^2\theta d\phi^2) \tag{1.1.2.2}$$

where $s$ is the proper time/distance, $t$ is the time coordinate, $r$ is the comoving radial coordinate and $\theta$ and $\phi$ are spherical coordinates. $k$ is the constant spatial curvature of the universe where $k = 0$ corresponds to a flat universe, $k = 1$ represents a positively curved (elliptic) universe and $k = -1$ is a negatively curved (hyperbolic) universe. The behaviour

of an isotropic and homogeneous universe within general relativity (using Einstein's field equations and the FLRW metric) is described by the Friedmann equations below:

$$\frac{\dot{a}^2 + kc^2}{a^2} = \frac{8\pi G\rho + \Lambda c^2}{3} \tag{1.1.2.3}$$

and

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\Big(\rho + \frac{3p}{c^2}\Big) + \frac{\Lambda c^2}{3} \tag{1.1.2.4}$$

where $p$ and $\rho$ are the pressure and density of the universe.

A number of measurements can be made using these Friedmann equations. For instance, the age of the Universe can be calculated. It is found to depend on the Hubble constant, $H_0$ along with a correction based on the density of the various physical components of the Universe:

$$t_0 = \frac{1}{H_0}\int_0^\infty \frac{dz}{(1+z)E(z)} \tag{1.1.2.5}$$

where E(z) is a useful function that is defined as:

$$E^2(z) = \Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_k(1+z)^2 + \Omega_\Lambda. \tag{1.1.2.6}$$

$\Omega_x$ is the fraction of the critical density that is formed of component $x$ and $m$ is matter, $r$ is radiation, $k$ is the spatial curvature and $\Lambda$ is the cosmological constant or dark energy. These components will be discussed in more detail in Section 1.1.3. The critical density, $\rho_c = \frac{3H_0^2}{8\pi G}$ is obtained using the first Friedmann equation and assuming that there is no cosmological constant and the universe is flat.

Distance measures can also be defined. The proper distance between two objects is the distance between them at a given time. As the Universe expands, this measure will change. For a flat universe the proper distance is just the coordinate distance between the objects, $a(t)r$. See all cases below:

$$s(t) = a(t) \cdot \begin{cases} \frac{1}{\sqrt{k}}\sin^{-1}(r\sqrt{k}) & \text{for } k > 0 \\ r & \text{for } k = 0 \\ \frac{1}{\sqrt{|k|}}\sinh^{-1}(r\sqrt{|k|}) & \text{for } k < 0. \end{cases} \tag{1.1.2.7}$$

The comoving distance between two objects that are both moving with the Hubble flow removes the effect of the expansion and results in a measurement that is constant over time. The line-of-sight comoving distance is calculated using:

$$d_C(z) = \frac{c}{H_0}\int_0^z \frac{dz'}{E(z')}. \tag{1.1.2.8}$$

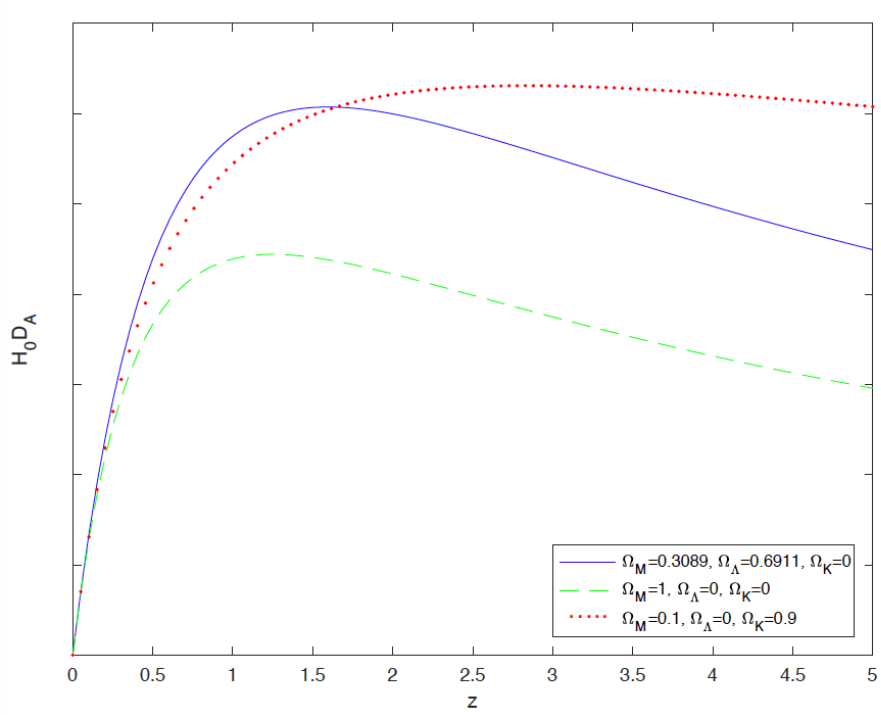



Figure 1.2: Angular diameter distance as a function of redshift for different cosmologies.

While the transverse comoving distance depends on the curvature term and is equivalent to the line-of-sight distance for a flat universe, see below:

$$d_M(z) = \begin{cases} \frac{c}{H_0\sqrt{\Omega_k}}\sinh(\frac{c}{H_0}\sqrt{\Omega_k}d_C(z)) & \text{for } k > 0 \\ d_C(z) & \text{for } k = 0 \\ \frac{c}{H_0\sqrt{|\Omega_k|}}\sin(\frac{c}{H_0}\sqrt{|\Omega_k|}d_C(z)) & \text{for } k < 0. \end{cases} \tag{1.1.2.9}$$

Another distance measure, the angular diameter distance is the ratio of the length of an object (perpendicular to the observer's line of sight), $x$, to the angle subtended by the object, $\theta$:

$$d_A(z) = \frac{x}{\theta} = \frac{d_M(z)}{1+z}. \tag{1.1.2.10}$$

This measure can be used to obtain physical distances from angular distances measured by telescopes and is used when measuring the Baryon Acoustic Oscillation (BAO) standard ruler, which will be discussed in further detail in Section 1.1.5.4 of this chapter. Angular diameter distance is also interesting because there is a turnover at a redshift between 1 and 2 (see Figure 1.2). This means that at redshifts greater than this, objects start to appear bigger (have a larger angular size). This is due to the expansion of the Universe and the fact that distant objects were much closer to us when they first emitted the light that is now reaching us. Luminosity distance is another very useful measure and is defined as the distance to an

object with luminosity $L$ that provides a measured flux $F$ at the observer's location if the Universe were flat and static. That is, it is the distance, $d_L$, that satisfies:

$$F = \frac{L}{4\pi d_L^2}. \tag{1.1.2.11}$$

It can be calculated using the following equation:

$$d_L(z) = (1+z)d_M(z). \tag{1.1.2.12}$$

### 1.1.3 ΛCDM Model

Most of modern day cosmology can be summarised as a search for, fine-tuning and understanding of a time-evolving model that describes our Universe. The model that currently best satisfies our observations is the ΛCDM model of the Universe. Λ, the cosmological constant or dark energy, represented in the Friedmann and Einstein equations as a positive constant, is a very poorly understood form of energy that is thought to have a negative pressure ($P = \omega\rho$) where $\omega$, the dark energy equation of state, is believed to be $-1$ and a constant energy density as the Universe expands. It is also known as vacuum energy. Cold dark matter (CDM) is a massive, but very weakly interacting (with baryonic matter or electromagnetic radiation) and slow moving (compared to light) type of matter with a nature that is yet to be understood. Its presence is detected via its gravitational interaction with baryonic matter and radiation. A number of ideas for the nature of this matter have been proposed; these include axions (Primack et al. 1988; Duffy & van Bibber 2009), Weakly Interacting Massive particles (WIMPs), which are a collection of fundamental particles that have the appropriate properties (see Primack et al. 1988; Jungman et al. 1996) and Massive Compact Halo Objects (MACHOs), which are baryonic objects such as brown dwarfs and black holes that we are simply unable to detect via electromagnetic radiation (Griest 1991; Alcock et al. 2001).

This model is described in $\mathcal{O}(10)$ important numbers called cosmological parameters from which the other parameters can be derived:

- Expansion:
  - The Hubble constant $h$, represents the expansion rate of the Universe and is defined and discussed in section 1.1.1. The dimensionless quantity $h$ is defined as $H_0/100$ kms$^{-1}$Mpc$^{-1}$
- The components:

  - The dark energy density $\Omega_\Lambda$, the fraction of the critical density that is dark energy
  - The matter density $\Omega_m$, the fraction of the critical density that is baryonic matter or cold dark matter
  - The baryon density $\Omega_b$, the fraction of the critical density that is baryonic matter
  - The neutrino density $\Omega_\nu$, the fraction of the critical density formed by neutrinos
  - The radiation density $\Omega_r$, the fraction of the critical density that is radiation

- Initial fluctuations:
  - Amplitude of the primordial density perturbations $A_s$. This can also be represented as $\sigma_8$ which represents the rms fluctuation in the mass distribution in spheres of $8h^{-1}$Mpc.
  - Spectral index of the primordial density fluctuations $n_s$

- Reionization:
  - The reionization optical depth $\tau$ which is linked to the redshift of reionization.

This model describes a flat ($k = 0$) universe with baryonic matter ($\sim 5\%$), cold dark matter ($\sim 25\%$), radiation, and predominantly, dark energy ($\sim 70\%$). The total energy density of the universe is given by the sum of these components: $\Omega_{tot} = 1 = \Omega_m + \Omega_r + \Omega_\Lambda + \Omega_k$ where $\Omega_k = 0$ for $k = 0$ and $\Omega = \rho/\rho_c$. These energy densities have changed over time because as the universe expands, the number of radiation and matter particles remain the same and therefore the density decreases. Matter density decreases in a manner that is proportional to the inverse volume ($\propto a^{-3}$) as the number of particles is fixed and the volume increases while the decrease in radiation energy density ($\propto a^{-4}$) has an additional factor of $a^{-1}$ due to the decrease in energy caused by the redshifting of the photons as the volume increases, thus, radiation density decreases faster than matter density. On the other hand, the dark energy density remains constant with expansion. It follows that the Universe underwent eras of domination by different components. Our expanding Universe is believed to have been radiation dominated in early times, immediately after inflation, and this led to an expansion following $a(t) \propto t^{1/2}$. It then became matter dominated with $a(t) \propto t^{2/3}$ at ~50000 years after the Big Bang and finally it became dark energy dominated and will remain so indefinitely with $a(t) \propto e^{Ht}$ (where $H$, the Hubble constant becomes a constant over time).

This Universe and the space-time we know is believed to have started in the Big Bang, which resulted in a very homogeneous Universe with tiny quantum fluctuations in temperature and density. Soon after the big bang the fundamental forces took the form we know and, in the inflationary scenario, this was followed by a state of exceptionally rapid expansion of space called inflation which led to the growth of the quantum fluctuations and resulted in a Gaussian random field in the density field. These small primordial density fluctuations are believed to seed the formation of all the large-scale structure present in the Universe today as well as the cosmic microwave background (CMB) temperature anisotropies that we observe (Figure 1.3). Sub-atomic particles then began to form as matter and anti-matter. Matter/ anti-matter pairs annihilated leaving a small amount of excess matter. Simple composite particles such as protons and neutrons quickly formed and this was followed by the process of Big Bang nucleosynthesis which led to the formation of nuclei that were heavier than the hydrogen nucleus (a proton) such as helium nuclei and small amounts of lithium nuclei. At this point most of the universe is composed of protons, electrons and photons and the photons and electrons are coupled via Thomson scattering so the photons are unable to escape, that is, until the Universe cooled sufficiently, to $\sim 3000$K, at which point protons and electrons combined and the photons streamed away. This resulted in the CMB which was discovered by Penzias & Wilson (1965). Subsequent structure formation is believed to occur via a bottom-up hierarchical model in which cold dark matter collapses under the effect of gravity into the denser regions and cluster together forming 'halos', larger gravitational potential wells that cooled baryonic matter and led to its collapse followed by star formation and eventually galaxy formation within the halos (White & Rees 1978). These halos then continued to collide and merge over time, forming larger galaxies.

In the time before the first stars had formed, the Universe was in a stage called the 'dark ages' as there was no visible light (the CMB had become infrared). In order for stars to form, the dark matter halos needed to be large enough that their mass was larger than the Jeans mass of the gas, which depends on the temperature and density of the surrounding region and a radiative cooling mechanism to cool the gas and make it more concentrated was also necessary. When the first stars eventually formed, they emitted large quantities of photons with energies greater than 13.6 eV and therefore able to ionize the H atoms, that were the predominant component of the Universe, into ionized HII particles. Stars first ionized their immediate surroundings (gas within their halos) and then the nearby intergalactic region and 'bubbles' of ionized regions started to appear and grow until eventually, all of the intergalactic medium was ionized. This process is called reionization and it was complete by a redshift of about 6.5 (Ota et al. 2017; Mason et al. 2018; Planck Collaboration et al.

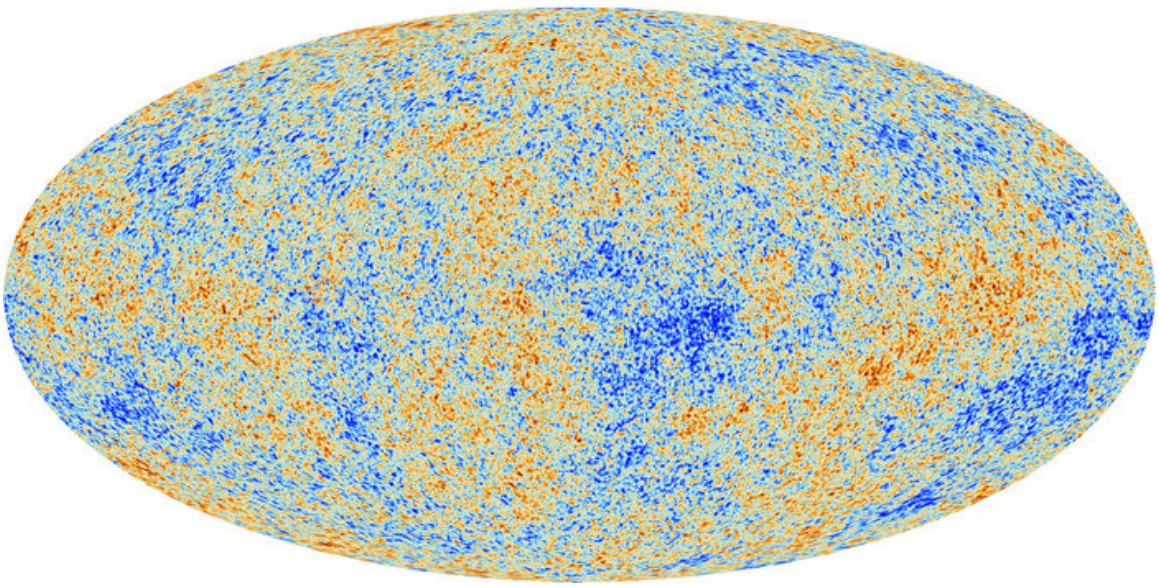


Figure 1.3: Temperature map of the CMB. The galactic plane has been filled in with a realization of a Gaussian random field. Taken from Planck Collaboration et al. (2014).

2018). The large number of free electrons present due to reionization led to further scattering of the CMB photons. The reionization optical depth, $\tau$, one of the cosmological parameters listed above, is a measure of the mean free path of a CMB photon before encountering a free electron and Thomson scattering off it, immediately after reionization (assuming it occurs instantaneously). It is linked to the redshift of reionization because the expansion of the Universe and correspondingly the density of electrons depends on the redshift at which the reionization occurred, therefore, a larger $\tau$ corresponds to a higher redshift of reionization. See Miralda-Escudé (2003) for more details on reionization.

In the last few decades this model has been tested with a range of cosmological probes and it has succeeded in explaining a large number of observations (these will be discussed further in section 1.1.5). This field is therefore now focused on the fine-tuning and sub-percent precision measurements of cosmological parameters. This level of precision has led to some inconsistencies or tensions between different probes and these have caused significant debate about whether these are due to systematics in the measurement techniques or whether they imply deviations from the $\Lambda$CDM Model.

### 1.1.4 Large Scale Structure

On its largest scales, the Universe is believed to be homogeneous. This seems to be the case for the CMB temperature maps, which show that in the early stages of the Universe there were no large features and only small temperature fluctuations. Large scale structure can be

measured at more recent times as well, by studying the distribution of galaxies at various redshifts. Such studies have found structure throughout the Universe on large scales. These structures take the form of galaxy clusters, voids (regions of space without any galaxies) and filaments (lines of galaxies). Together, these structures form what is known as the 'cosmic web' with clusters at the intersection of filaments and voids forming in the spaces in between. Galaxy clustering analyses can also elucidate other features of the galaxy distribution, such as the shape of the power spectrum or the presence of large-scale peaks in the correlation function (due to Baryon Acoustic Oscillations which will be discussed in Section 1.1.5.4).

This observed large scale structure depends both on cosmological parameters and on the formation and evolution of galaxies. Using the two-point correlation function, one can trace the dependence of large scale structure on galaxy properties such as luminosity, colour, stellar mass, and track its evolution with redshift. Comparison of the observed galaxy clustering signatures with dark matter simulations allows one to model and understand the clustering of galaxies and their formation and evolution within their parent dark matter halos. Clustering measurements can be used to determine the parent dark matter halo mass of a given galaxy population, connect observed galaxy populations at different epochs, and constrain cosmological parameters and galaxy evolution models.

#### 1.1.4.1 Galaxy Clustering

The primordial fluctuations present after inflation can be described in the form of a power spectrum, a function that describes the amplitude of the fluctuations as a function of scale (length or mass scale), see the definition below:

$$(2\pi)^3 P(\vec{k})\delta_D(\vec{k}+\vec{k'}) = \langle \delta(\vec{k})\delta(\vec{k'})\rangle \tag{1.1.4.1}$$

where $k$ is the wave number and $\delta_D$ is the Dirac delta function. The primordial power spectrum is assumed to follow a power law with a power of 1 i.e. $P(k) \propto k$. This is a scale invariant spectrum called the Harrison-Zeldovich spectrum and it has no preferred length, meaning that fluctuations of all scales had the same amplitude when they first entered the horizon. The matter power spectrum that we measure differs from this primordial spectrum by a function called the transfer function, $T(k)$, which is due to various physical effects, therefore the power spectrum takes the form $P(k) = Ak^n T(k)$. One of the main deviations from the Harrison-Zeldovich power law shape is the turnover from $n = 1$ to $n = -3$ at the scale of the horizon at the time of matter-radiation equality (the crossover from a radiation dominated to a matter dominated Universe which occurred at $z \approx 2700$) and the accompanying suppression on small scales (large $k$), see Figure 1.4. This is due to the

fact that fluctuations that are in the horizon in the time of radiation domination are frozen and unable to grow due to the pressure created by the dense radiation. On the other hand, fluctuations outside of the horizon continue to grow until they are encompassed by the growing horizon. After matter-radiation equality the dark matter is able to collapse into the over-densities, but baryonic matter is coupled to the radiation by Thomson scattering and therefore is still prevented from collapsing under gravity by the pressure of the radiation. This is the case until $z \approx 1100$ when recombination occurs. The decoupling of the baryons and photons and free-streaming of the photons allows the neutral baryonic matter to collapse.

There are other effects that alter the shape of the power spectrum that can be measured (via the CMB or the large scale structure). A major effect is that of the BAO. These are the oscillations that were present in potential wells before the photons and electrons decoupled due to the competing forces of gravity and radiation pressure. At recombination the baryon waves freeze in space leaving an over-density at a scale of the maximum distance the sound waves could have travelled before recombination. The dark matter was not influenced by these acoustic waves and therefore was still concentrated in the initial over-density. Over time it attracts most of the baryons while some move towards the BAO peak. This creates a series of small peaks in the matter power spectrum (Figure 1.4) and corresponding peaks are also formed in the CMB angular power spectrum (Figure 1.5). This will be discussed in further detail in section 1.1.5.4. Another effect is Silk damping (Silk 1968), which is the damping of small scale fluctuations due to the free streaming photons dragging baryons along and out of the over-densities.

The growth of fluctuations after this point follows linear perturbation theory until the over-densities become large $\delta \approx 1$. Linear perturbation theory models small inhomogeneities on large scales by introducing small deviations to the FLRW metric:

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}, \tag{1.1.4.2}$$

where $\bar{g_{\mu\nu}}$ is the FLRW metric and $\delta g_{\mu\nu}$ is the small perturbation. In this theory fluctuations on different scales grow independently of each other and their evolution can be calculated independently. When the over-densities become too large, the fluctuations on different scales become coupled and non-linear perturbation theory is needed (see the modifications produced in Figure 1.4). Within the non-linear regime, dark matter collapses into halos within overdensities and stars and galaxies begin to form. The processes involved in galaxy formation: hydrodynamical effects, heating and cooling of gas etc. lead to complex relationships between the distribution of galaxies and that of the dark matter. Despite this, on large scales, fluctuations are small and the relationship between the over-densities of

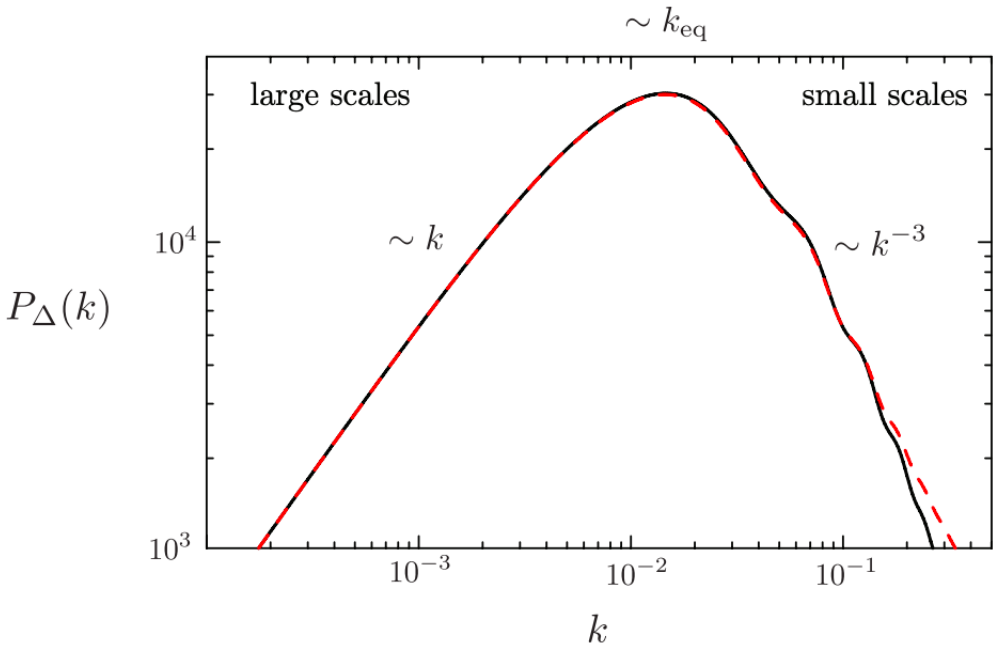



Figure 1.4: The present day ($z = 0$) matter power spectrum from linear theory (black solid line) and with non-linear corrections (red dashed line). The $P(k) \propto k$ relationship is seen on large scales along with the $P(k) \propto k^{-3}$ relationship on small scales after the turn-over at matter-radiation equality. The small fluctuations are the baryon acoustic oscillations. Figure taken from Daniel Baumann Part III Cambridge Cosmology notes.

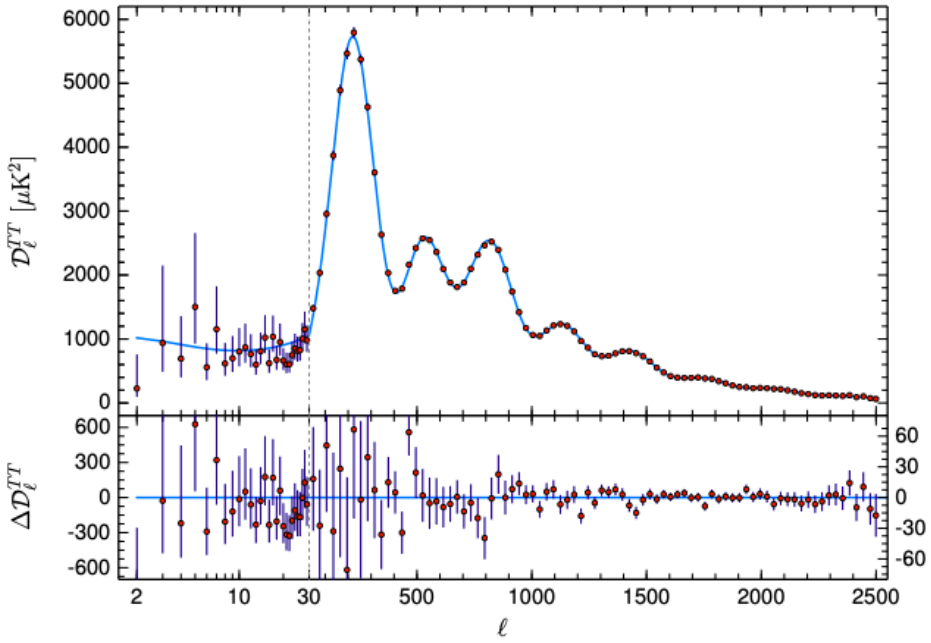



Figure 1.5: Temperature angular power spectrum measurement made using the Planck CMB telescope. The solid line is the best fit model assuming a ΛCDM cosmology. The residuals of this fit are given in the lower panel. Figure taken from Planck Collaboration et al. (2018).

galaxies and those of dark matter is often estimated as a linear one:

$$b = \frac{\delta_{galaxy}}{\delta} \tag{1.1.4.3}$$

where $\delta$ is the density contrast field or overdensity field. This overdensity field is given by the ratio of the mass density to the mean mass density on a given scale $(\frac{\rho}{\bar{\rho}} - 1)$ where $\rho$ is density. Correspondingly, the power spectra are related by:

$$P_{\text{galaxy}} = b^2 P_{\text{matter}}. \tag{1.1.4.4}$$

The parameter $b$ is the bias and describes the relationship between the clustering of the galaxy population and that of the underlying dark matter distribution. In a similar manner, dark matter halos which form due to spherical collapse of dark matter in dense regions do not exactly mimic the distribution of the underlying dark matter and there is also a halo bias that connect these density distributions.

The galaxy bias can also be expressed as the square root of the ratio of the correlation functions of galaxies to dark matter:

$$b = \left(\frac{\xi_{\text{galaxy}}}{\xi_{\text{dark matter}}}\right)^{\frac{1}{2}}. \tag{1.1.4.5}$$

The correlation function, $\xi(s)$ is the Fourier transform of the power spectrum:

$$P(\vec{k}) = \int d^3 s e^{-i\vec{k}\cdot\vec{s}} \xi(\vec{s}). \tag{1.1.4.6}$$

$\xi(\vec{s})$ measures the excess probability of finding a galaxy inside a volume element $dV$ separated from another galaxy by a distance $\vec{s}$ compared to a random Poisson distribution (Peebles 1980):

$$dP = n[1 + \xi(\vec{s})]dV, \tag{1.1.4.7}$$

where n is the mean number density of the galaxy sample. $\xi(s)$ is given by the following:

$$\xi(\vec{s}) = \langle \delta(\vec{x})\delta(\vec{x}+\vec{s})\rangle = \left\langle \left(\frac{n(\vec{x})-\bar{n})}{\bar{n}}\right\rangle \left\langle \frac{n(\vec{x}+\vec{s})-\bar{n})}{\bar{n}}\right)\right\rangle \tag{1.1.4.8}$$

where $\delta$ represents the over-density of galaxies in a particular volume, $n$ is the number of galaxies in a volume $x + \delta x$ while $\bar{n}$ is the average number of galaxies in every such volume and $\vec{s}$ is a vector to another position distance $|\vec{s}|$ away and the angle brackets represent the expectation value or ensemble average. If we assume the volume of interest is isotropic then we only care about distance $s = |\vec{s}|$. Note here that if we are considering the matter power spectrum instead of the galaxy power spectrum the number functions $n$, are replaced by the density functions $\rho$.

There are a number of variations on the measurement of the two-point correlation function and they are all useful in different cases depending on the data that is available and the information that is required. When the three-dimensional positions of galaxies are well known two types of measurements can be made: the monopole correlation function and the three-dimensional correlation function. The monopole correlation function involves simply binning galaxies by the three-dimensional distance between them, the three-dimensional correlation function on the other hand involves binning galaxies by both their radial and transverse separations. This full three-dimensional correlation function requires a complex covariance matrix that necessitates a large number of mock catalogs for calculation. Therefore, some methods of compressing this information have been suggested. A multipole expansion can be applied to the correlation function and the first 2 or 3 even multipoles can be used (Padmanabhan & White 2008). Another method is called 'clustering wedges', this involves integrals over the correlation function over a range of angles (Kazin et al. 2013). These methods also involve the use of a fiducial cosmology to allow for the conversion from distances in redshift space to co-moving real space. Angular correlation functions, $\omega(\theta)$ are used when there is limited redshift information and therefore no information on radial separation is available. In this case, galaxies are binned solely based on their angular

separations. $\omega(\theta)$ is the excess probability of finding a galaxy in an angle element $d\Omega$ at an angle $\theta$ away from another galaxy:

$$dP = n[1 + \omega(\theta)]d\Omega. \tag{1.1.4.9}$$

It is the 2PCF of galaxies projected onto the sphere, and is thus calculated within redshift shells. Finally, the radial correlation function is calculated by binning galaxies that are approximately collinear (within a small angular separation from each other) by their redshift separation. Both of these measurements are model independent.

Analogously to the angular correlation function, the angular power spectrum provides a measure of the power as a function of angle on the sky; it is the projection of the power spectrum on a sphere. Angular power spectra, $C_l$ are not expressed as a function of angle but of spherical harmonic multipole, $l$ where $\theta \approx 180°/l$. These angular power spectra, $C_l$ are related to the angular correlation function by the following expression:

$$\omega(\theta) = \frac{1}{4\pi\sigma_0^2}\sum_{l=1}^{\infty}(2l+1)C_l P_l(\cos\theta), \tag{1.1.4.10}$$

where $P_l$ is the $l^{th}$ Legendre polynomial and $\sigma_0$ is the source surface density.

Angular power spectra are commonly used for CMB analyses and are also used for clustering analyses when radial information is limited. When comparing this statistic to the angular correlation function, $C_l$s have much larger signal to noise at large scales while $\omega(\theta)$ have large signal to noise on much smaller scales (smaller angles). One complication involved in the calculation of the $C_l$s is that each multipole $l$ depends on the entire sky and surveys tend to have incomplete sky coverage.

The functions described above are all 2-point statistics, the corresponding three-point statistics are the three-point correlation function and the bispectrum. The 3-point correlation function, $\zeta$ is defined as the function that satisfies the following expression for the probability of finding three galaxies in the three volume elements $dV_1$, $dV_2$ and $dV_3$:

$$dP = [1 + \xi(s_1) + \xi(s_2) + \xi(s_3) + \zeta(s_1, s_2, s_3)]\bar{n}^3 dV_1 dV_2 dV_3. \tag{1.1.4.11}$$

Peebles & Groth (1975) found that the galaxy three-point correlation function takes the simple form:

$$\zeta(s_1, s_2, s_3) = Q[\xi(s_{12})\xi(s_{23}) + \xi(s_{23})\xi(s_{31}) + \xi(s_{31})\xi(s_{12})], \tag{1.1.4.12}$$

where $Q$ is a constant that was found to be $1.3 \pm 0.2$. The bispectrum, $B(k)$ is the Fourier transform of the 3-point correlation function and is defined as below:

$$\langle\delta(\vec{k}_1)\delta(\vec{k}_2)\delta(\vec{k}_3)\rangle = B(\vec{k}_1, \vec{k}_2, \vec{k}_3)(2\pi)^3\delta_D(\vec{k}_1 + \vec{k}_2 + \vec{k}_3). \tag{1.1.4.13}$$

There are a number of different types of galaxies and some are stronger or more biased tracers of the underlying dark matter, i.e. their bias values are larger. For instance, 'normal' star forming galaxies tend to have significantly smaller biases than active galactic nuclei (AGN) which have very luminous central regions dominated by light from the accretion of matter onto the central black hole Fabian 1999; Shields 1999. AGN galaxies tend to be much more massive than star forming galaxies and correspondingly are hosted by more massive dark matter halos and therefore are more biased tracers (McLure et al. 1999; Seymour et al. 2007; Hale et al. 2018). In terms of optically-selected galaxies, a population with large bias that is very useful for clustering analyses is the luminous red galaxy (LRG) population. These are massive ($10^{11-12}M_{\odot}$), very luminous ($\geq 3L*$) elliptical galaxies with an old (red) stellar population and little or no star formation. These properties are in contrast to star-forming galaxies which have younger stellar populations and therefore tend to be blue (star formation emits blue light), they also tend to be spiral galaxies. LRGs reside in larger dark matter halos than their blue, star-forming counterparts and therefore are better tracers of dark matter, this means they are more likely to cluster in the same patterns as the dark matter making them better for studying large scale structure, and they can enhance the measurement of the BAO. Their high luminosity is also an advantage as they can be seen to high redshifts. Another reason why LRGs are good for such analyses is that it is very straightforward to measure their photometric redshifts (photo-zs) as the Balmer break (or 4000Å break) (Eisenstein et al. 2005b) is present in their spectra and there are few emission lines present, making the photo-z estimation task simply a location of the break. This break switches from the g band to the r band (considering SDSS filters) at $z \approx 0.4$. At significantly higher redshifts, near-IR filters are more useful for determining the photo-zs of these LRGs.

### 1.1.5 Cosmological Probes of Large Scale Structure

In order to determine the true cosmological model, experiments that constrain the cosmological parameters must be conducted. The large scale structure contains rich information about these parameters and there are a number of different probes that are able to extract this information. Some of these will be discussed below. Often, the results of some of these probes are combined to provide tighter constraints. In this time of precision cosmology, such combined measurements have provided us with the tightest constraints available (see Hildebrandt et al. 2017b; Abbott et al. 2018a,b; van Uitert et al. 2018; Planck Collaboration et al. 2018 for some recent examples of combined constraints). An important cosmological probe that does not directly utilise the large scale structure but deserves to be mentioned is

type Ia supernovae standard candle measurements. Supernovae are very energetic explosions of stars which appear as very luminous objects that can outshine their host galaxies and that reach a peak brightness then fade slowly. Type Ia supernovae (SNIa) form in binary systems of a star and a white dwarf. If the white dwarf accretes so much gas from the star that its mass increases to the Chandrashekar limit, $1.4M_{\odot}$ then it explodes in a supernova. Since all of these SNIa form under similar conditions, they are expected to have the same absolute luminosity at the peaks of their light curves. Therefore, measuring the apparent luminosity from Earth allows one to determine the distance between the supernova and the Earth.

#### 1.1.5.1 The Cosmic Microwave Background

The CMB radiation that was discussed in Sections 1.1.3 and 1.1.4 is another very valuable source of cosmological information. Maps of both temperature anisotropies (e.g. Figure 1.3) and polarisation are measured and power spectra for both of these can be determined (see Figure 1.5 for an example of a temperature angular power spectrum). In addition, the gravitational lensing (see Section 1.1.5.5) of the CMB can be measured and another power spectrum created. These measurements allow the determination of the Hubble constant $H_0$, dark matter density, $\Omega_{CDM}$, the baryon density $\Omega_b$, the scalar spectral index $n_s$ and the optical depth $\tau$. These CMB measurements also provide the size of the acoustic scale which is important for BAO measurements. The most recent and accurate measurements of the CMB are provided by the Wilkinson Microwave Anisotropy Probe (WMAP; Bennett et al., 2003) and the Planck mission (Tauber et al. 2010). WMAP focused on measuring the temperature anisotropies and operated from 2001 to 2010 and produced improved results over the 9 years of observations (Bennett et al. 2013; Komatsu et al. 2014). Planck made measurements of both temperature and polarisation anisotropies and made observations between 2009 and 2013 providing significant improvements over WMAP with improved sensitivity and a broader range of frequencies allowing the separation of components and extraction of foregrounds (Planck Collaboration et al. 2014, 2018).

#### 1.1.5.2 The Alcock-Paczynski Cosmological Test

The Alcock-Paczynski (AP) test (Alcock & Paczynski 1979) is a geometrical test of the ratio of angular size to redshift size of a spherical region. The comoving distance between two galaxies is calculated using the galaxy redshifts, the angle between them and an assumed cosmology. Therefore, if the assumed cosmology is incorrect, these measurements will also be incorrect. In particular, two galaxies separated by comoving distance $d$ should be measured to be separated by distance $d$ independent of their separation angle compared

to the line of sight. This also means that the clustering of a group of galaxies should be spherically symmetric. This implies that for an isotropic distribution of galaxies, the change in the correlation function with distance should be the same in both along the line of sight and perpendicular to this. This test then involves trying different cosmologies to see which ones lead to this equivalence. Alcock & Paczynski (1979) first used this test to determine whether or not the cosmological constant is equal to 0. Other cosmological methods can also be tested for (e.g. López-Corredoira 2014), or a model can be assumed (such as ΛCDM) and its parameters constrained (e.g. Li et al. 2014; Nusser 2005).

An advantage of this method, compared to many other cosmological measurements is that it is not influenced by galaxy evolution. On the other hand, it is affected by redshift space distortions (Kaiser 1987; Hamilton 1998) which are caused by the peculiar velocities of galaxies which also produce apparent anisotropies in galaxy distributions. These will be further discussed in Section 1.1.5.3. Understanding of these distortions has allowed researchers to account for these effects in their analyses.

#### 1.1.5.3 Redshift Space Distortions

Redshift space distortions, the apparent squashing and stretching of space when galaxies/clusters are observed in redshift space (see Hamilton 1998 for a review), also require two-point correlation functions (2PCFs) for their measurement. On large scales, the Kaiser effect (Kaiser 1987), which is due to the coherent infall of galaxies toward clusters and the large scale structure, results in an enhanced clustering signal along the line of sight. On the other hand, on relatively small scales ($\lesssim 1\mathrm{h}^{-1}$Mpc), the peculiar velocities of galaxies in bound structures leads to the 'finger of god' effect that makes the galaxies in a cluster appear to be spread out along the line-of-sight forming long, thin structures. This is because as galaxies move around in the group/cluster their motion along the line of sight leads to Doppler shifts, resulting in some galaxies appearing to be further away from us and others appearing to be closer to us than the cluster centre. The effects of the Kaiser effect and the 'finger of god' effect on the correlation function are shown in Figure 1.6. Hamilton (1998) showed that the finger-of-god effect can also occur due to the coherent infall towards group/cluster centres. This happens because as galaxies fall in towards the centres of small clusters, they tend to have large velocities and movement along the observer's line of sight leads to large Doppler shifts that cause the elongation. The coherent infall mechanisms are portrayed in Figure 1.7.

Clearly, RSDs can be a nuisance for measuring galaxy clustering (and therefore affect measurements such as the AP test and BAO measurements) as they introduce falsely enhanced/reduced clustering signals, but they are also a valuable cosmological probe. RSDs

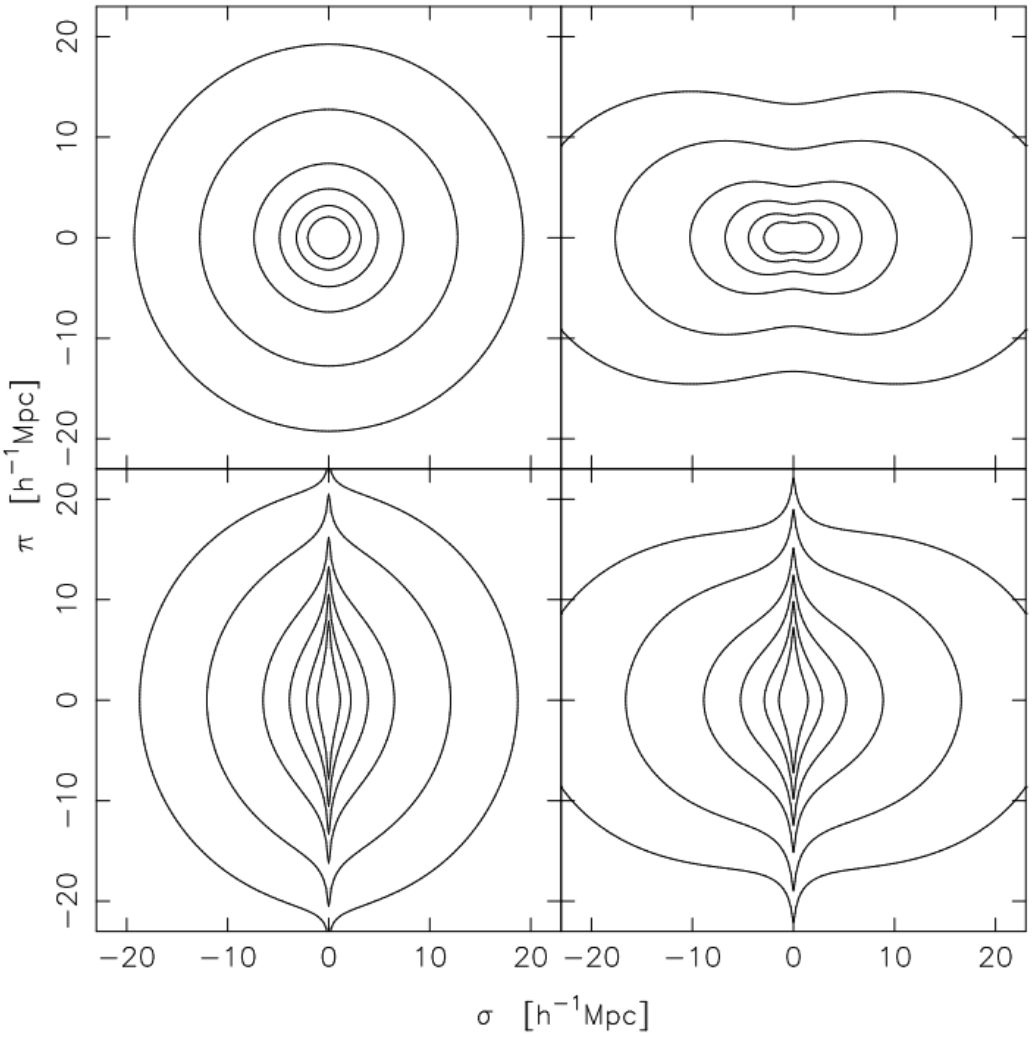



Figure 1.6: Model spatial correlation functions, $\xi(\sigma,\pi)$ where $\sigma$ and $\pi$ are the transverse and radial separations respectively. The lines are contours of constant $\xi(\sigma,\pi)$. The top left panel displays the correlation function with no distortions, the top right panel shows the squashing due to the Kaiser effect and the bottom left panel shows the elongation caused by the (random velocities) finger-of-god effect. The final panel (bottom right) shows the correlation function with both effects. Figure taken from Hawkins et al. (2003)

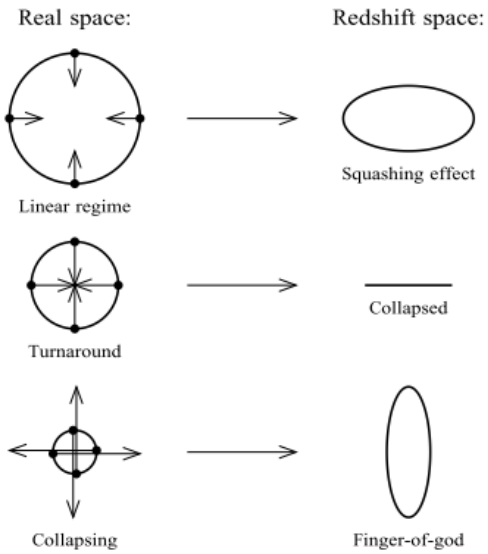



Figure 1.7: Illustration of the mechanisms that lead to the observed redshift space distortions. The first figure represents a large galaxy cluster. The galaxies have peculiar velocities that are relatively small, so along the line of sight the cluster appears squashed. On the other hand, the last case represents a small galaxy group or cluster that has galaxies with large peculiar velocities. The Doppler shifts in this case cause an elongation along the line of sight. In both cases only the line-of-sight positions appear altered as the effects are due to Doppler shifts. Figure taken from Hamilton (1998)

have been used to study the linear growth of structure, allowing measurements of $f(z)\sigma_8(z)$ where $f$ is the rate of change of the linear growth rate and $\sigma_8$ is the root mean square amplitude of fluctuations in the matter over-density field in a sphere of radius $8\mathrm{h}^{-1}$Mpc. RSD measurements normally require spectroscopic clustering measurements that provide measurements in both the transverse direction, in which the effect is not present, and the radial direction, in which the effect is maximised (see Percival et al. 2004; Okumura et al. 2008; Blake et al. 2011a; Reid et al. 2012). Recently, studies have shown that it is possible to measure RSDs from the angular correlation function obtained using photometric surveys (Nock et al. 2010; Crocce et al. 2011a) and this has been demonstrated using LRG samples (Blake et al. 2007; Padmanabhan et al. 2007; Crocce et al. 2011b; Thomas et al. 2011).

#### 1.1.5.4 Baryon Acoustic Oscillations

In the pre-recombination era (z>1100), baryons and photons were coupled by Thomson scattering, and the competing forces of gravity and radiation pressure resulted in acoustic waves propagating out of overdensities. Soon after recombination (when the protons and electrons combined and became neutral), photons decoupled from baryons and streamed away forming the cosmic microwave background (CMB). This left the baryons frozen in space at the characteristic distance called the sound horizon (the largest co-moving distance a sound wave could have travelled by the time of recombination). This overdensity of baryons led to the seeding of galaxy formation at the scale of the sound horizon, resulting in an overdensity of galaxies, which expanded with the Universe over time. This is demonstrated in Figure 1.8 which shows the evolution of the radial mass profiles of dark matter, gas, photons and neutrinos. We see the propagation of the coupled baryons and photons followed by decoupling and the free streaming of the photons while the baryon peak stays in place at a comoving distance of $\sim 150$Mpc ($105\mathrm{h}^{-1}$Mpc). The dark matter and baryons interact gravitationally and thus the baryon peak shifts the dark matter peak towards it. At later times, the dark matter pulls the baryons back towards the dark matter peak at the centre, shrinking the baryon peak, until the dark matter and baryonic matter settle in the same mass profiles, with a large peak at the centre and a small one at the sound horizon (also called the baryon acoustic oscillation (BAO) peak).

This signal is now imprinted in the large scale galaxy distribution (as the slight over-density of baryons led to the seeding of galaxy formation) and its scale can be compared to the signal in the CMB, thus allowing the BAO peak to be used as a standard ruler. This allows the independent measurement of the Hubble parameter $H(z)$ and the angular diameter distance $D_A(z)$ as functions of redshift. These can lead to measurements of the

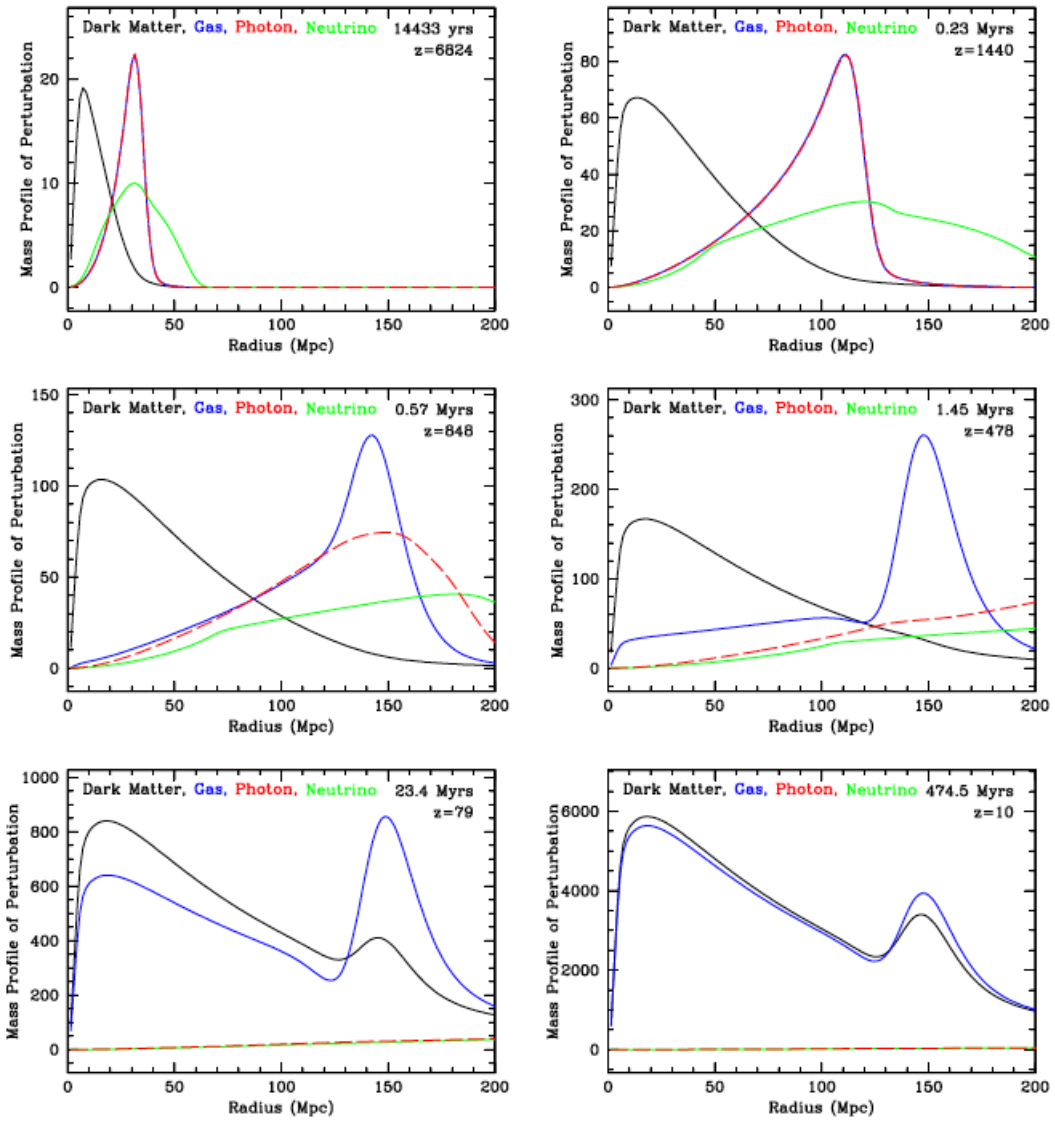



Figure 1.8: Snapshots of the evolution of a spherical density perturbation. The radial mass profiles of dark matter, gas, photons and neutrinos are given as functions of comoving radius from the centre of the overdensity. Figure taken from Eisenstein et al. (2007)

dark energy equation of state parameters $\omega_0$ and $\omega_a$. These are parameters that allow the equation of state to vary over time such that $\omega(a) = \omega_0 + \omega_a(1-a)$ (Linder 2003).

This process implies that this sound horizon can be measured both at the time of recombination via the CMB and at present (and other low-redshift times) via the observed clustering of galaxies, and this allows it to be used as a statistical standard ruler. The sound horizon at the time of recombination is calculated from the CMB power spectrum using the equation:

$$s = \int_0^{t_{rec}} c_s(1+z)dt = \int_{z_{rec}}^{\inf} \frac{c_s dz}{H(z)} \tag{1.1.5.1}$$

where $c_s = [3(1+3\rho_b/4\rho_\gamma)]^{-1/2}$ is the speed of sound. This measurement is assumed to be robust as at such early times, perturbation theory is valid, and the data from Planck is very precise. The BAO peak in the baryon correlation function at the drag epoch—the time at which the baryons stop moving with the photons, which occurs soon after recombination—can be measured from Earth. This measurement can be made in both the radial direction ($\delta z_{BAO}$) and the transverse direction ($\delta\theta_{BAO}$), allowing the independent measurement of the Hubble constant in the radial direction and the angular diameter distance in the transverse direction:

$$H(z) = \frac{c\delta z_{BAO}}{s_\parallel}, D_A(z) = \frac{s_\perp}{(1+z)\delta\theta_{BAO}}. \tag{1.1.5.2}$$

Where angular diameter distance is defined as:

$$D_A = \frac{c}{H_0(1+z)\sqrt{-\Omega_k}} \sin(\sqrt{-\Omega_k}\xi(z)), \tag{1.1.5.3}$$

and if we assume a flat universe: $\Omega_k = 0$, then $D_A$ simplifies to:

$$D_A = \frac{c}{H_0(1+z)} \sin(\xi(z)). \tag{1.1.5.4}$$

The clustering of galaxies on the scale of the BAO can be quantified via the power spectrum or its Fourier transform: the two-point correlation function of galaxies (refer to Section 1.1.4.1). The BAO feature appears as a peak in the two-point correlation function which represents a slight excess of galaxy clustering, and in the power spectrum of galaxies, it manifests as a wiggle.

The most commonly used estimator of the galaxy two-point correlation function is the Landy-Szalay estimator (Landy & Szalay 1993):

$$\xi(s) = \frac{DD(s) - 2DR(s) + RR(s)}{RR(s)}, \tag{1.1.5.5}$$

where DD(s), RR(s) and DR(s) are the normalized numbers of galaxy-galaxy, random-random and galaxy-random pairs with a comoving separation $s$. The random set must have the same sky coverage and redshift distribution as the dataset and larger random samples reduce the errors on the correlation function as they decrease the effect of shot noise from the galaxy sample. The use of photometric redshifts and angular correlation functions will allow the sample of galaxies used to be significantly larger than with spectroscopic redshifts, making the shot noise negligible, and thus also decreasing the necessity of a random sample that is multiple times larger than the galaxy sample.

A number of other estimators have been suggested, but older, simpler estimators such as Peebles & Hauser (1974) and Hewett (1982) are no longer used as Landy-Szalay has been proven to be superior (Kerscher et al. 2000). Estimators such as those suggested by Vargas-Magaña et al. (2013) and Baxter & Rozo (2013) appear to provide more accurate estimations, but are not often used due to the computational difficulties incurred in their implementation.

The three-dimensional 2PCF is normally expressed based on the separation of points in two directions (either the line-of-sight and perpendicular to the line-of-sight ($\xi(\sigma, \pi)$), or the distance between the points and the angle between this vector and the line-of-sight ($\xi(r, \mu)$)). When only photo-zs are available the two-point angular correlation function (see Section 1.1.4.1) can be calculated in redshift bins. In this case, the BAO peak is at a different location for each redshift bin, shifting to smaller angles as the redshift increases(see Figure 1.10). As no distances must be inferred, this version of the 2PCF does not require a cosmology to be assumed, and is model independent. Few analyses have been performed using this method with real photometric data. One example of this is Carnero et al. (2012), but this method has also been applied to simulated future survey data (see Sánchez et al. 2011). Angular correlation functions have also been used with spectroscopic data in order to obtain model independent cosmological measurements (e.g. Alcaniz et al. 2016; Carvalho et al. 2016).

The BAO peak measured using galaxy clustering is not exactly equivalent to the value measured using the CMB due to non-linear effects such as: non-linear structure growth, redshift space distortions, shift in frequency and diffusion damping and galaxy bias. All these effects—although small on the large BAO scales—smooth and flatten the peak, making it more difficult to detect. There are also systematics in the measurement of the feature which could lead to errors: the accuracy of the 2-point estimator used, the size and appropriateness of the random set, covariance, fiducial cosmology and modelling of the 2-point statistic (Vargas-Magaña et al. 2018). Additionally, when photo-zs are used for BAO analyses, and redshift bins must be used, this results in projection effects which further flatten and shift

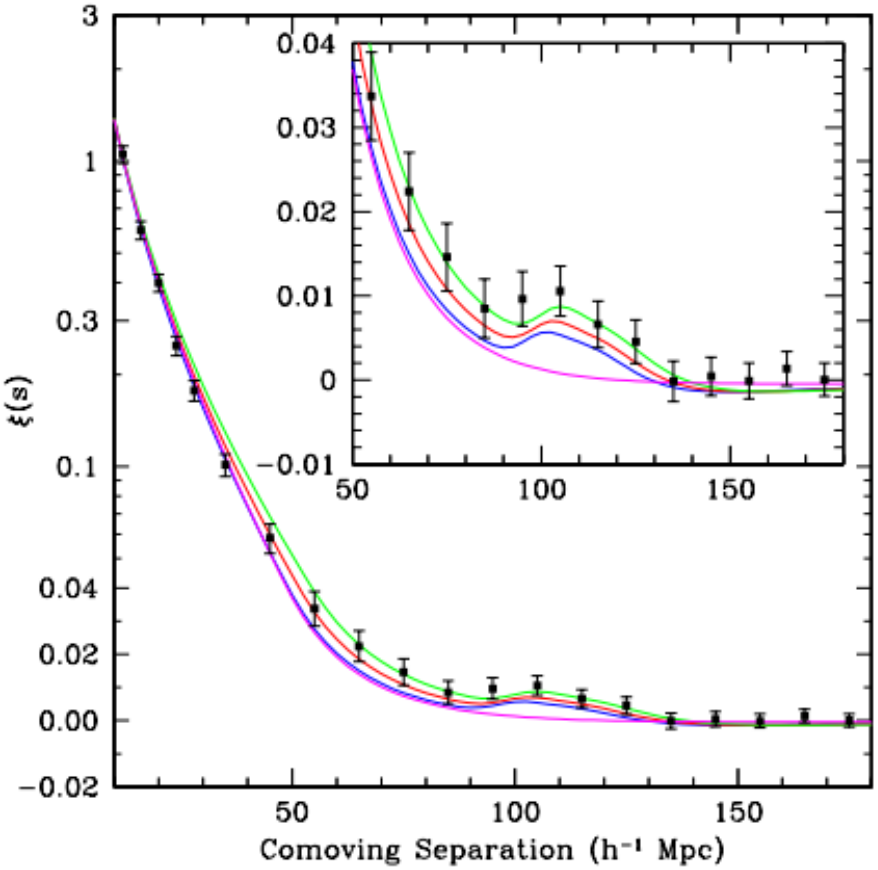



Figure 1.9: Two-point correlation function for 4 different cosmologies and showing the BAO peak. Figure taken from Eisenstein et al. (2005a).

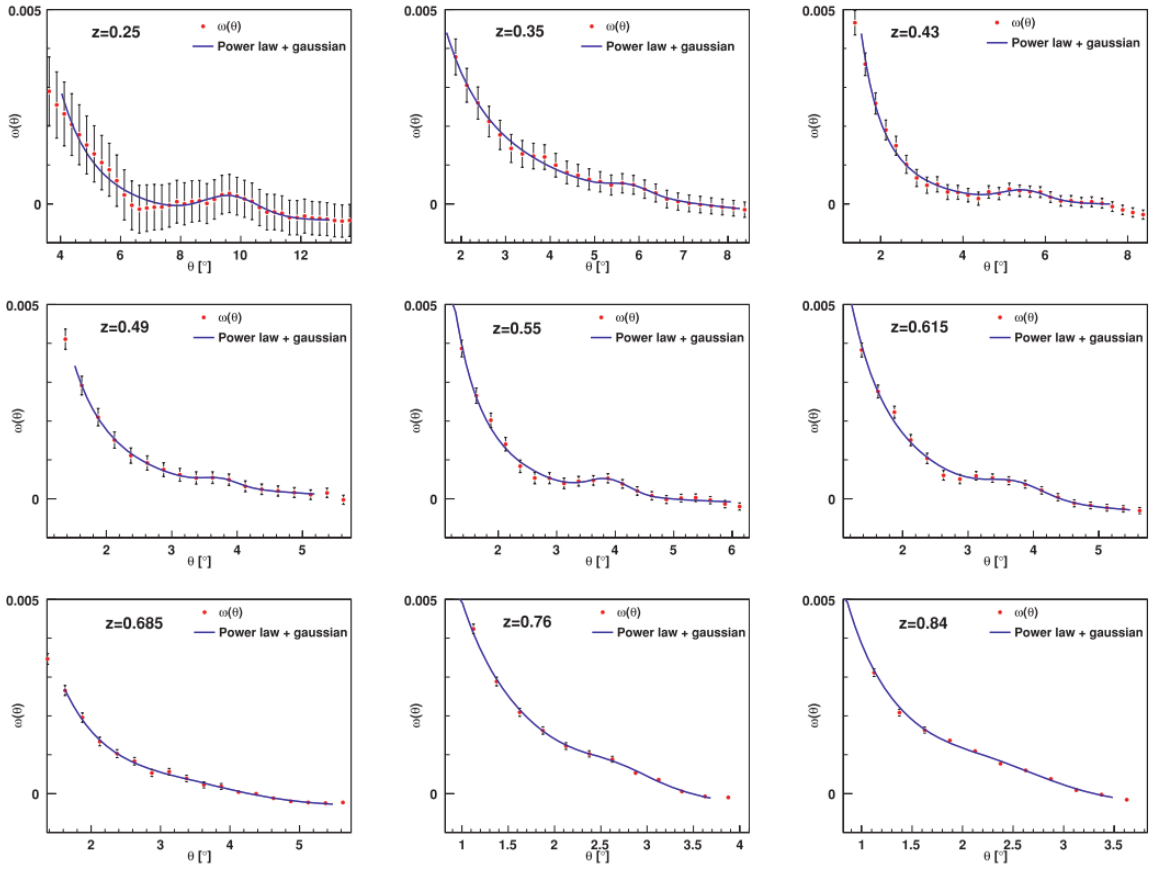



Figure 1.10: Two-point angular correlation functions in 9 redshift bins and showing the BAO peak for each case. Figure taken from Sánchez et al. (2011).

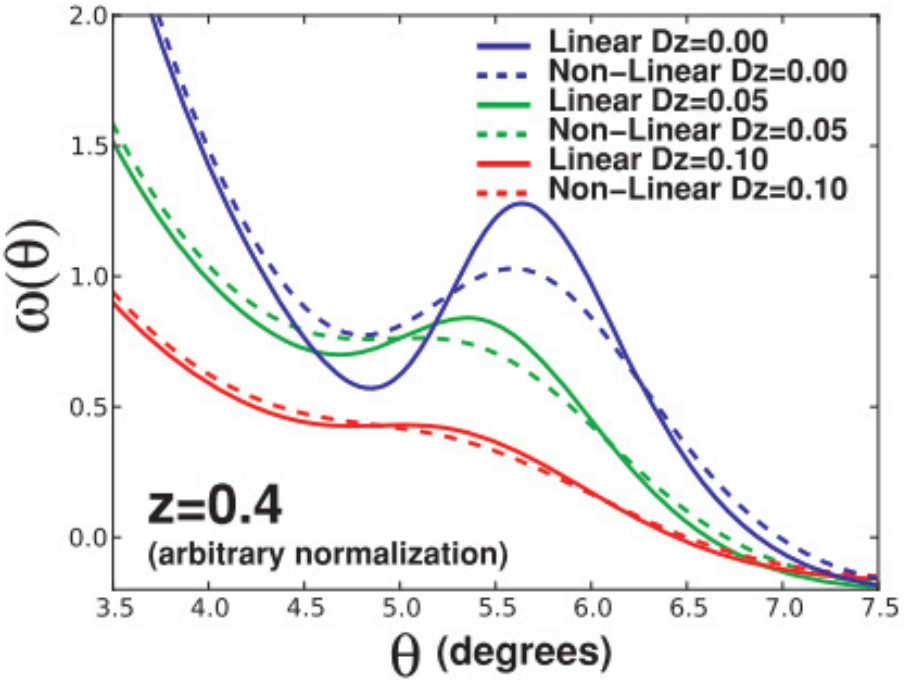



Figure 1.11: The effects of non-linearities and photo-z projection effects on the two-point angular correlation function and the position of the BAO peak. Figure taken from Sánchez et al. (2011).

the peak (see Figure 1.11). These projection effects are due to the assumption that galaxies of multiple redshifts are at the central redshift of the bin and their projected positions on this shell lead to a flattened and BAO peak that is shifted to lower redshifts.Sánchez et al. (2011) discuss these problems and suggest a method for reducing these effects.

Non-linear structure formation occurs when over-densities have amplitudes sufficiently large that perturbations on various scales become coupled. Initially on small scales, non-linear growth eventually causes large scale effects such as bulk flows which flatten the BAO peak. These effects are mapped in the displacement field of the Universe, and thus, if this field could be calculated and reversed, the effects of non-linear structure formation could be undone. This is the basis of the reconstruction process (Eisenstein et al. 2007; Padmanabhan et al. 2012) which uses the density field (from the positions of galaxies) to estimate the displacement field and move the particles backwards. This results in a more defined acoustic peak and RSDs can also be accounted for in this process. Unfortunately, since radial information is missing when only photo-zs are present, the 3-dimensional density field is not well known, and thus, reconstruction cannot be applied in this case.

#### 1.1.5.5 Gravitational Lensing

Light travelling from distant objects does not always follow a straight, undisturbed path to the observer: in many cases it will pass very near to an object (or many objects) located between the source and the observer and will be deflected by the gravitational field of this object. This effect is called gravitational lensing. This effect can be very strong if the light passes very near to or collides head on with a very massive object such as a massive galaxy or a galaxy cluster. This can lead to observations such as multiple images, Einstein rings and large, magnified arcs. It can also cause the background light to be magnified. Strong lensing has been used in a number of ways as a cosmological probe. First suggested by Refsdal (1964), the strong lensing of light from variable sources such as quasars that results in multiple images can be used to measure the Hubble constant via the measurement of the time delay between the observation of this variability in the different images. Another probe, suggested by Yamamoto et al. (2001) and Biesiada (2006), is based on the fact that the Einstein radius of a lens depends on the ratio of the angular diameter distances between the lens and the source and between the lens and the observer. Statistically significant measurements are performed using multiple lens systems. A similar analysis can be done using galaxy clusters that lens multiple background sources.

A much smaller, but significantly more common effect, is weak lensing. This occurs when the source light does not pass near enough to massive galaxies and instead of forming multiple images, arcs or rings, the image of the source is only slightly distorted by being slightly stretched (shear) and magnified. As these effects are very small, large numbers of galaxies must be averaged over to make measurements. One very important cosmological probe that uses weak lensing is cosmic shear. This effect is due to the fact that light from galaxies that are nearby will pass through similar cosmic structures on its way to an observer, and therefore the galaxy shapes will be correlated. The correlation function over different scales at different redshifts can then provide information on the growth of structure over time and the nature of dark energy. A major difficulty in measuring cosmic shear is the removal of the intrinsic alignments of galaxies based on the density field. Another useful probe is galaxy-galaxy lensing which is a measurement of the tangential shear of the background galaxies caused by the presence of the lens galaxies. This is calculated by cross-correlating the lensing galaxy positions and the shear of the source galaxies. This measurement relates the galaxy densities to matter densities and thus can be used to measure galaxy bias of a galaxy populations. Galaxy-galaxy lensing can also be combined with a measurement of the clustering of the lens galaxies in order to obtain the matter correlation function (Baldauf et al. 2010). Recent cosmological surveys such as the Kilo Degree Survey (KIDS; de Jong et al. 2013) and the Dark Energy Survey (DES; The Dark Energy

Survey Collaboration 2005)) have been conducting combined analyses of cosmic shear, galaxy-galaxy lensing and galaxy clustering to obtain measurements of both galaxy bias and cosmological parameters (Abbott et al. 2018a; van Uitert et al. 2018; Joudaki et al. 2018).

#### 1.1.5.6 HI or 21cm Line Observations and Intensity Mapping

HI is the neutral hydrogen atom which is present in large quantities throughout the interstellar and intergalactic media. This means that mapping the presence of HI would be a useful probe of matter and could provide a measurement of the matter spectrum. Fortunately, the spin-flip transition that takes place in HI atoms provides us with a method to detect its presence. This transition is due to the spontaneous flip of the atom's electron's spin orientation which results in a slightly lower energy state: when the spins are parallel, the system has slightly higher energy than when they are anti-parallel. The photon that is released when this transition occurs has a wavelength of 21cm leading to the alternative name '21cm line'. It is therefore detected using radio telescopes. This spontaneous flip occurs very rarely (once in 10 million years) but since HI atoms are so abundant in the interstellar and intergalactic media a signal is always present (Pritchard & Loeb 2012; Ewen & Purcell 1951). Radio frequency radiation is not obstructed by dust, making this measurement more robust and feasible at high redshifts. On the other hand, at low redshifts HI measurements can be complicated by the fact that the signal is very faint and the presence of radio frequency interference from man-made devices. At all redshifts these measurements are impacted by galactic foregrounds (Bernardi et al. 2009). Measurements of the very redshifted HI signal emitted during reionization are very useful for mapping the ionized regions or 'bubbles' being formed during this period and for testing comsological models (Scott & Rees 1990; Giri et al. 2018; Furlanetto et al. 2019). Lower redshift measurements can be used for clustering measurements and probes such as BAO and redshift space distortions using a technique called intensity mapping Bull et al. 2015; Santos 2016; Carucci 2018; Kovetz et al. 2019. This technique involves integrating the 21cm signal received over relatively large pixels of sky instead of resolving individual galaxies. These measurements are sufficient for these large scale structure measurements and combining the flux from a number of galaxies will address the problem of a low signal leading to measurements with larger signal to noise. Surveys that will make these measurements include the Canadian Hydrogen Intensity Mapping Experiment (CHIME; Newburgh et al. 2014) and the Square Kilometre Array (SKA; Square Kilometre Array Cosmology Science Working Group et al. 2018). Another way in which information is expected to be extracted from this signal is by measuring the weak lensing of the 21cm radiation due to foreground large scale structure

and the presence of 21cm radiation sources over a range of wavelengths will also allow measurements of the lensing signal for various source redshifts (Pourtsidou & Metcalf 2015; Romeo et al. 2018). Finally, another way in which HI line measurements are useful is as a method of obtaining spectroscopic redshifts of galaxies in large galaxy surveys (Maddox et al. 2013; Harrison et al. 2017)

#### 1.1.5.7 Galaxy Clusters

Large clusters of galaxies are present throughout the cosmic web and are located at the intersection of filaments. Due to their large size, their matter content is expected to be similar to the matter content of the Universe. The measurement made to take advantage of this fact is the gas fraction of the cluster:

$$f_{\text{gas}} = \frac{\text{baryon mass}}{\text{baryon mass} + \text{dark matter mass}} = \frac{\text{X-ray gas mass}}{\text{total cluster mass from weak lensing}}. \tag{1.1.5.6}$$

The measured gas fraction at various redshifts can be compared to simulations, and by assuming a fiducial cosmology, measurements of $\Omega_M$, $\Omega_\Lambda$ and the dark energy equation of state, $\omega$ can be made (see Mantz et al. 2014 for an example). The number counts of clusters as a function of mass and redshift are another useful probe because this function can be calculated theoretically and depends on the cosmological parameters $\Omega_M, \sigma_8$ and $\omega$ (see Campanelli et al. 2012 and Mantz et al. 2015 for examples).

#### 1.1.5.8 Primordial Non-Gaussianity in Large Scale structure

The initial density field present after inflation is expected to be a Gaussian random field. A Gaussian random density field is a random field such that the densities at any finite number of points form a multivariate Gaussian and these densities are multivariate Gaussian if any linear combination of them follows a univariate Gaussian distribution. More formally, a random field F(x) is said to be a Gaussian random field if a vector formed by the values of F at any combination of k points, $\{x_1, x_2, x_3, ..., x_k\}$ form a multivariate Gaussian. Now, a combination of variables, $Y_1, Y_2, Y_3, ...Y_m$, is said to be multivariate Gaussian if any linear combination of these variables $(a_1Y_1 + a_2Y_2 + a_3Y_3 + ... + a_mY_m)$ has a univariate Gaussian distribution. Primordial non-Gaussianity (PNG) describes deviations from a Gaussian random field in the initial density field present after inflation.

In reality, most theories of inflation are expected to lead to an initial density field that slightly deviates from a Gaussian random field. This is called primordial non-Gaussianity (PNG). In addition, of the models that fit with current observations there is significant

variation in the expected extent of non-Gaussianity present, thus, measuring this non-Gaussianity with observations will provide a valuable tool for discriminating between these models (see Bartolo et al. 2004 for a review). Simple slow-roll, single-field inflationary models predict almost Gaussian density fields with $f_{\rm NL} \ll 1$ ($\mathcal{O}(10^{-2})$) (Maldacena 2003; Creminelli & Zaldarriaga 2004) while multi-field models allow larger deviations from Gaussianity with $f_{\rm NL} \gtrsim 1$ (Lyth et al. 2003; Zaldarriaga 2004). Most inflationary models predict that the non-Gaussianity depends only on the local value of the potential, such PNG is said to be of the 'local type' and it is parameterized by the $f_{\rm NL}$ parameter which quantifies the deviation from a Gaussian random field $\phi$:

$$\Phi = \phi + f_{\rm NL}(\phi^2 - \langle\phi\rangle^2). \tag{1.1.5.7}$$

where $\Phi$ is the Bardeen gauge invariant potential.

The small fluctuations in the primordial density field are believed to seed the formation of all the large scale structure present in the Universe today as well as the Cosmic Microwave Background (CMB) temperature anisotropies that we observe. As a result, non-Gaussianity can be measured using either the CMB or the LSS. Local PNG has—to date—been measured most precisely using the bispectrum of the Cosmic Microwave Background (CMB) temperature anisotropy maps. The main downfalls of this method are that such measurements are limited by cosmic variance on large scales and by Silk damping on small scales (Alvarez et al. 2014; Ferraro & Smith 2015). Alternatively, LSS measurements can be used. The non-Gaussianity leads to an increase in the 3D power spectrum on large scales, which corresponds to an increase in the total halo bias on large scales (Dalal et al. 2008; Matarrese & Verde 2008; Carbone et al. 2008). One problem with this measurement is that general relativistic projection effects and non-linear evolution are expected to produce scale dependent effects on large scale structure measurements very similar to those of PNG with $f_{\rm NL}$ of order unity (Yoo 2010; Jeong et al. 2012; Bruni et al. 2012). Therefore, in order to start distinguishing between single-field and multi-field inflation, $\sigma(f_{\rm NL}) \sim 1$ is required, where $\sigma(f_{\rm NL})$ is the standard deviation in the measurement of $f_{\rm NL}$. This is therefore normally the target of analyses such as these (de Putter & Doré 2017; Moradinezhad Dizgah & Keating 2019). Large galaxy surveys (Slosar et al. 2008; De Bernardis et al. 2010; Xia et al. 2010; Ross et al. 2013; Leistedt et al. 2014; Castorina et al. 2019), combinations of galaxy surveys and CMB maps (Giannantonio & Percival 2014) and intensity mapping (Joudaki et al. 2011; Camera et al. 2013; Li & Ma 2017; Fonseca et al. 2018) have all been considered (or used) for measuring this non-Gaussianity effect. In addition, the multi-tracer method (Seljak 2009) that involves cross-correlating multiple populations of galaxies or intensity maps with different biases can reduce the effects of cosmic variance and improve constraints

(Ferramacho et al. 2014; Yamauchi et al. 2014; Alonso & Ferreira 2015; Witzemann et al. 2019; Ballardini et al. 2019).

A non-zero local primordial non-Gaussianity signal leads to peaks in the squeezed configuration of the matter bispectrum. The squeezed configuration is such that one scale is much smaller than the other two, i.e. $k_1 << k_2 \sim k_3$. As gravitational collapse and structure formation continue into the non-linear regime, and fluctuations on different scales become coupled, the power spectrum becomes affected by the mode-correlations caused by the non-Gaussianity and this leads to increases in power of the halo power spectrum on large scales. This corresponds to a scale dependent correction to the halo bias given by the expression below (Slosar et al. 2008; Dalal et al. 2008; Matarrese & Verde 2008; Carbone et al. 2008):

$$b_h(M,z) = b_L(M,z) + f_{\rm NL}\delta_c[b_L(M,z)-1]\frac{3\Omega_m H_0^2}{c^2k^2T(k)D(z)}, \tag{1.1.5.8}$$

where $b_h$ is the total halo bias, $b_L$ is the Gaussian linear bias, $f_{\rm NL}$ is the amplitude of the non-Gaussianity (defined in Equation 1.1.5.7), $\delta_c$ is the critical over-density for spherical collapse at redshift $z = 0$, $T(k)$ is the linear transfer function and $D(z)$ is the growth function. The primordial non-Gaussianity parameter $f_{\rm NL}$ is therefore normally determined from large scale structure by measuring the galaxy power spectrum or bispectrum (De Bernardis et al. 2010; Ross et al. 2013). The effect of a non-zero $f_{\rm NL}$ parameter on the 3D halo power spectrum is illustrated in Figure 1.12.

On the other hand, in situations where spectroscopic redshift data is not available (which will be the case with some current and future large scale surveys with volumes and depths that make it unfeasible to measure spectroscopic redshifts), and only the less accurate photometric redshifts or estimated redshift distributions are present, neither the 3D matter power spectrum nor the bispectrum can be reliably computed. Instead, a tomographic analysis with angular power spectra could provide a sufficient estimate.

## 1.2 Observations and Photometric Redshifts

Cosmological redshifts, defined in Section 1.1.1, are a vital tool for extra-galactic astronomers as they provide a measure of how quickly an object is moving away from the Earth due to the Hubble flow. As an object moves away from the observer the photons that travel lose energy on their journey and arrive with a longer wavelength (and correspondingly smaller frequency). The ratio of the wavelength of the photons that arrive to the wavelength of the photons that are emitted defines the redshift according to the following formula:

$$1 + z = \frac{\lambda_o}{\lambda_e} \tag{1.2.0.1}$$

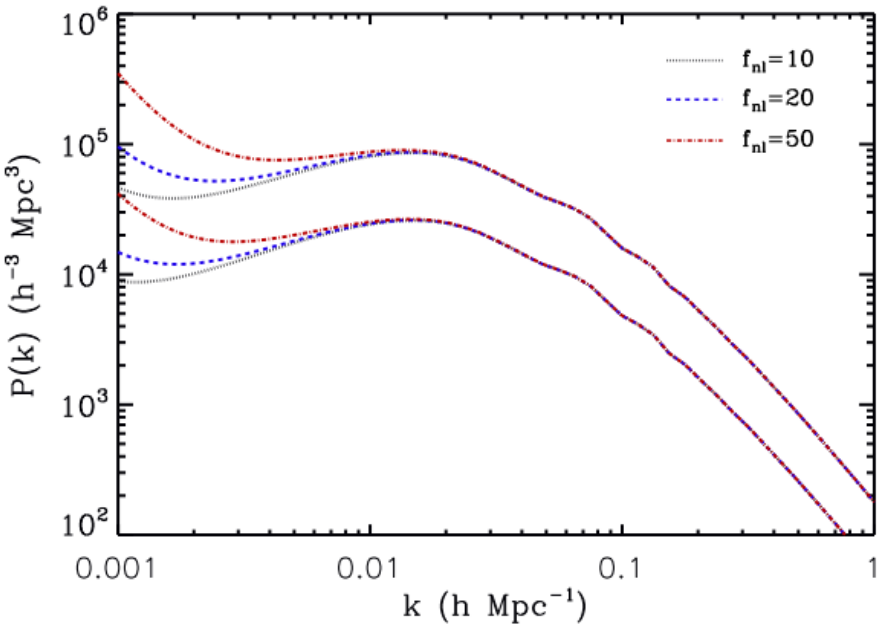



Figure 1.12: The theoretical 3D halo power spectrum at $z = 1$ for halos in the mass range $10^{11} - 10^{12} h^{-1} M_\odot$ for the lower curves and $10^{13} - 10^{14} h^{-1} M_\odot$ for the upper curves. Figure taken from Ferramacho et al. (2014).

where $\lambda_o$ is the observed redshift and $\lambda_e$ is the emitted redshift.

Galaxy redshifts can normally be measured using a spectrograph. The spectrograph produces the spectrum of the galaxy within which well known emission or absorption lines, or prominent features such as breaks (at known wavelengths), can be located and the observed wavelength can be extracted. The ratio of this observed wavelength to the emitted wavelength is then used to measure the galaxy's redshift. This is a straightforward process, with a high level of accuracy in most cases. One situation in which this is not the case is when similar lines or features exist in more than one part of the electromagnetic spectrum, allowing them to be mistaken for each other and resulting in significantly under or over estimated redshifts. Spectroscopic redshifts can be very expensive to obtain as photons must be collected over an extended period of time in order for a spectrum to be produced, thus requiring large exposure times for each pointing, and observing time is very limited and valuable for telescopes. This renders spectroscopic redshifts observationally infeasible to obtain for very large volumes of sky (large areas and deep images) for single surveys.

Photometric redshifts, on the other hand, are estimates of the redshift of a galaxy using the broad-band photometry of the galaxy, i.e. the flux from the galaxy that passes through filters with different wavelength ranges. These filters are placed in front of charge-coupled devices (CCDs) in the cameras of telescopes which receive the photons from galaxies.

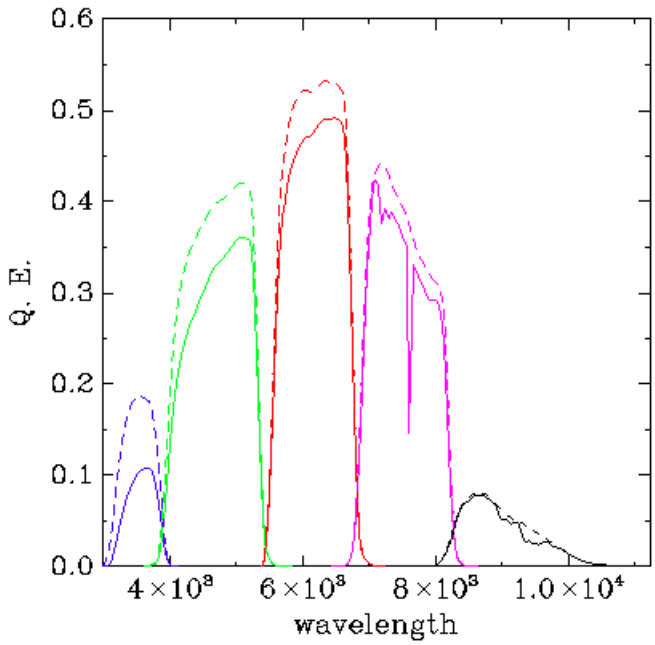



Figure 1.13: System response curves for the SDSS filters. The solid lines represent the combined response to the filter transmission, the quantum efficiency of the detectors, the optics of the telescope and the atmospheric conditions. The dashed lines do not include the effects of the atmosphere. Figure taken from the SDSS DR3 Camera webpage http://classic.sdss.org/dr3/instruments/imager/index.html

These filters normally have redshift ranges that span a section of the spectrum, depending on the type of photons the CCDs are designed to detect. For example, the Sloan Digital Sky Survey (SDSS) imaging camera detects photons in the optical range and has filters $u$, $g$, $r$, $i$ and $z$ (see Figure 1.13). Similarly, infrared telescopes will have filters with wavelength ranges in the infrared. Different telescopes have slightly different filter systems (with different transmission curves or different filter wavelengths) even if they cover the same large wavelength range. Using the flux transmitted in each filter, we can make a rough estimate of the shape of the galaxy's spectra, and use this to estimate the redshift of the galaxy.

Galaxy spectra are formed predominantly of the blackbody spectra of the galaxy's stars over a range of temperatures. This produces a relatively flat spectrum. Elements in the atmospheres of stars and hydrogen gas occupying the interstellar medium then cause emission/absorption lines and more prominent features such as breaks (sudden drops in the flux). One such feature is the 4000Å/Balmer break (see Figure 1.14) which is composed of two separate breaks that occur next to each other in the spectrum, appearing as one break. The Balmer limit (the shortest wavelength of the Balmer series) is actually at a wavelength of 3646Å and therefore the limit extends to wavelengths smaller than this, but before this

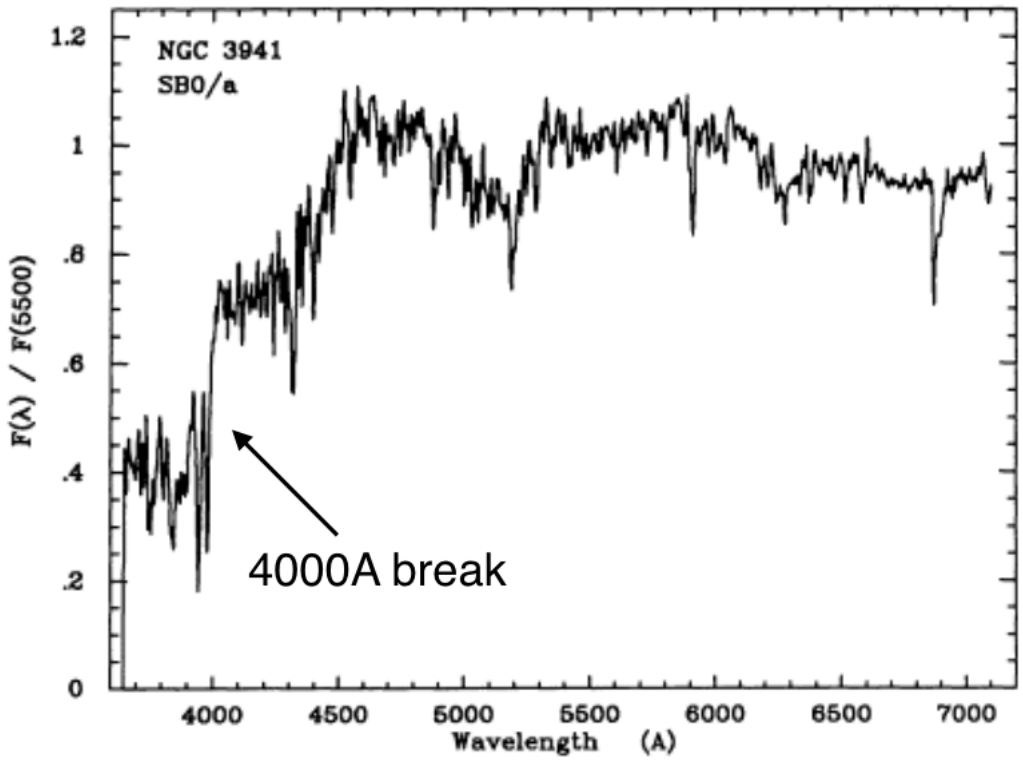



Figure 1.14: Figure showing the spectrum of an elliptical galaxy with the 4000Å break pointed out. Figure taken from Kennicutt (1992)

limit, at the 4000Å point, there is the sudden drop due to absorption by a number of metal elements in stellar atmospheres. These metal absorption lines do not appear as prominently if hot, blue stars are present. Therefore, the presence of the 4000Å break is dependent on the galaxy's star formation and age. This break is observed in the optical/near-IR portions of the spectrum for galaxies at $z < 1$ and therefore it is useful when estimating photometric redshifts as the flux would be decreased for filters at wavelengths shorter than the drop and the filter with the drop will have a decreased flux depending on where exactly along the filter's wavelength range the break occurred.

Another similar break is the Lyman break which begins at 912Å, the ionisation wavelength of neutral hydrogen gas, or the Lyman limit (of the Lyman series), and continues to higher energies (or smaller wavelengths). This break is seen in optical spectra if the galaxy has a redshift between 3 and 4, so this break is normally used for detecting these high-redshift galaxies, but this break can be observed for a larger range of redshifts if ultraviolet or infrared spectra are also present. This break is useful for measuring photometric redshifts for the same reasons as the 4000Å/Balmer break.

Filters of different wavelength ranges, from different telescopes can be combined to obtain improved spectral coverage, and allow more accurate photo-z estimation. The 2deg$^2$ COSMOS field is an example of this as it has coverage in 30 intermediate/narrow bands spanning the UV, visible, near-IR and mid-IR regions. As a result, very accurate photometric redshifts have been obtained for galaxies in this field (see Ilbert et al. 2009; Laigle et al. 2016).

The following section provides details of photometry and the photometric systems used. Details of the photometric-redshift estimation techniques currently employed are presented in Section 1.2.4.

### 1.2.1 Photometric Systems

Photometric redshifts are estimated using galaxy photometry. A photometric system is a set of filters with discrete bandwidths and known sensitivity to incoming radiation (based on the detectors used and the filter). An example of a photometric system and the transmission curves of the filters can be seen in Figure 1.13. Photometric systems are classified as wide-band or broadband ($\geq$300Å), intermediate band (100Å - 300Å) and narrow-band (<100Å). The choice of the width of passbands depends on the goals of the survey as narrower passbands lead to more detailed information about the spectral energy distributions of the observed galaxies while larger passbands increase the sensitivity of the telescope, allowing fainter galaxies to be detected. For each filter the incoming flux, $F_{\text{filter}}$, is measured by

convolving with the filter response function ($R_{\text{filter}}$, the function that describes what flux passes through the filter):

$$F_{\text{filter}} = \int_0^\infty f(\lambda) R_{\text{filter}}(\lambda) d\lambda. \tag{1.2.1.1}$$

The filter response function depends on a number of factors: the filter transmission, the quantum efficiency of the detectors, the optics of the telescope and the atmospheric conditions present at the observatory. This relationship can be expressed as:

$$R = R_{\text{theory}} T_{\text{CCD}} T_{\text{opt}} T_{\text{atm}} \tag{1.2.1.2}$$

where $R_{\text{theory}}$ is the theoretical response of the filter and the other terms are the transmission curves due to the CCD detectors, the telescope optics and the atmospheric conditions respectively.

### 1.2.2 Magnitude Systems

Magnitudes (which are used for historical reasons) are a conversion from flux onto a logarithmic scale. Apparent magnitude is a measurement of the brightness of an object as it appears from Earth and it is given by:

$$m = -2.5 \log \frac{F}{F_0} \tag{1.2.2.1}$$

where $F$ is the flux in a given filter and $F_0$ is a normalizing constant for that filter. This normalizing constant is set for each photometric system based on some reference star. One common system of this type is the Vega magnitude system, which uses the Vega star as a reference. The constants are chosen to be the value that makes the Vega star have magnitude 0.

AB magnitudes use a fixed normalizing constant flux of $F_0 = 3631$Jy for all frequencies, where Jy is the symbol for Jansky, a unit of flux density per unit frequency defined as $1\text{Jy} = 10^{-26}\text{WHz}^{-1}\text{m}^{-2} = 10^{-23}\text{erg s}^{-1}\text{Hz}^{-1}\text{cm}^{-2}$. The AB magnitude for a passband can be calculated by integrating the fluxes over the range of relevant frequencies:

$$m_{\text{AB, filter}} = -2.5 \log \frac{\int f_\nu d\nu}{\int F_0 d\nu}. \tag{1.2.2.2}$$

In practice, the numerator is simply the total flux detected in that filter. This magnitude system does not use a relative reference object, but a fixed value, for all colours. This makes it easier to compare magnitudes in different passbands. The magnitude system used in the SDSS survey, the asinh magnitudes (Lupton et al. 1999), is similar to the AB magnitude system, but accounts for the noise in the measurement. Thus, it performs similarly to AB

at high signal-to-noise but significantly better at low signal-to-noise as it can cope with negative fluxes due to noise. It is given by:

$$m_{\rm SDSS} = -\frac{2.5}{\ln(10)}\Big[\sinh^{-1}\Big(\frac{F}{2bF_0}\Big) + \ln b\Big]. \tag{1.2.2.3}$$

where $F_0$ is the same as for the AB system and b is a softening parameter chosen to be the 1-$\sigma$ sky noise in a point spread function aperature and in 1" seeing.

Bolometric magnitude is the magnitude based on the light measured at all wavelengths (bolometric luminosity, $L$) instead of the light passing through filters.

Another magnitude measurement, also called absolute magnitude, is the apparent magnitude a source would have if it were 10 pc away. Such a measurement is useful for comparing the intrinsic brightness of galaxies/stars. Absolute magnitude, $M$, is obtained from apparent magnitude, $m$, using the following equation:

$$M = m - 5\log_{10}\Big[\frac{D}{10\text{pc}}\Big], \tag{1.2.2.4}$$

where $D$ is the distance to the object. This relationship is complicated when broad band photometry is measured for objects at different wavelengths. This leads to the measurements representing different rest-frame wavelengths of the galaxies. The k-correction, $K_{QR}$, provides a correction to the conversion from apparent magnitude in bandpass $R$ to absolute magnitude in the emitted-frame bandpass:

$$m_R = M_Q + 5\log_{10}\Big[\frac{D_L}{10\text{pc}}\Big] + K_{QR}, \tag{1.2.2.5}$$

where $D_L$ is the luminosity distance which is necessary for redshifted objects as it accounts for the expansion and geometry of the Universe. This correction can be estimated using (Hogg et al. 2002; Oke & Sandage 1968):

$$K = -2.5\log\Big[\frac{\int_0^\infty f_\lambda(\frac{\lambda}{1+z})d\lambda}{\int_0^\infty f_\lambda(\lambda)R(\lambda)\lambda d\lambda}\Big], \tag{1.2.2.6}$$

where $R(\lambda)$ is the filter response and $f(\lambda)$ is the flux density. We see that the redshift of an object must be known to calculate its k-correction. The k-correction is not necessary if the light from known emission lines was measured as this is the same line in all spectra even if it was redshifted. Similarly, if we are using bolometric magnitudes this would not involve comparing different portions of the spectrum for different objects and therefore the k-correction would not be necessary.

### 1.2.3 Measuring the Flux from a Source

In addition to knowing how the magnitude of a galaxy will be measured, one must also decide on what area of the sky will be measured for each galaxy. The aim here is to include as much of the flux from the source as possible while simultaneously excluding as much of the background noise as possible. There are a number of different ways that have been developed to do this and the best choice depends on the science the photometry is being used for. The simplest method is to use a fixed-sized circular aperture for every object. This is called aperture photometry and is best for measuring very distant, unresolved galaxies. Isophotal magnitudes are another type of measurement. These involve measuring the flux from all the pixels that have a surface brightness greater than a particular threshold. This implies that different portions of galaxies are measured at different redshifts. Another method is to choose apertures with sizes that approximately enclose the total flux of the object. Kron magnitudes (Kron 1980) are an estimate of the total flux of an object and are obtained by estimating the half-light or effective radius of the galaxy, that is, the radius from the centre that surrounds an area from which half of the galaxy's light is emitted and measuring the flux in this aperture followed by multiplying this value by two. A similar measure is given by the Petrosian magnitude (Petrosian (1976)). This is the magnitude measured in a circular aperture with a radius of $N$ times $r_p$. SDSS uses an $N$ of 2. This Petrosian radius, $r_p$, is such that the Petrosian ratio, $R_p$, the ratio of the surface brightness in an annulus at $r_p$ to the mean surface brightness within $r_p$ is a given value (0.2 for SDSS).

Another useful type of magnitude is the model fit magnitude. This utilizes the flux obtained by fitting a model to the source light and integrating over the model for all radii. In SDSS de Vaucouleurs and exponential model fits are used. The de Vancouleurs profile is as follows:

$$I(r) = I_0 e^{-7.67 \frac{r}{r_e}^{1/4}} \tag{1.2.3.1}$$

and the exponential profile is:

$$I(r) = I_0 e^{-1.68 \frac{r}{r_e}} \tag{1.2.3.2}$$

where $r_e$ is the half-light radius. The de Vancouleurs profile provides a good fit to elliptical galaxies while the exponential profile is better for spiral galaxies. For each source, SDSS provides model magnitudes using both of these models, the model magnitude for the best fitting model for each source, determined from the r-band and applied to each of the other bands and a composite magnitude given by the linear combination of both models that best fits the source. Model magnitudes are normally used for galaxies while point spread function

(PSF) magnitudes are the analogous measure for stars. The PSF is often modelled as a two-dimensional Gaussian fit to the light from the star and sometimes a Moffat distribution is used instead.

Colours are another photometric measurement that can be made. The colour or colour index of an object is the difference between the magnitudes of two different filters. For example, in the SDSS photometric system, the u-g colour is simply the magnitude in the g-band subtracted from the magnitude in the u-band. Colours are measured such that the higher wavelength magnitude is subtracted from the lower wavelength one. Colours are useful for distinguishing galaxy/star types as they contain information on the temperature, age and composition of these objects. For example, galaxies with more positive colours are redder galaxies and tend to be larger, elliptical galaxies with little or no star formation while negative colours indicate bluer galaxies which are more likely to be younger, star-forming spiral galaxies. Magnitudes that are measured for each band through the same aperture allow more accurate measurements of colours as flux from the same area of the source is being compared.

#### 1.2.3.1 Galactic Extinction

An important correction that must be made when making photometric measurements is for galactic extinction. Galactic extinction, $A$, is the decrease in flux received due to the absorption or scattering of light from the source by galactic dust along the line of sight between the telescope and the source. Blue light is more strongly attenuated than red light and therefore extinction leads to an increase in the ratio of red photons to blue photons received by the observer and therefore the object appears to be 'reddened'. For this reason extinction is said to cause galactic or dust reddening. This reddening effect, $E(B-V)$, can be determined by measuring the dust content in the galaxy along different lines-of-sight. This measurement of dust maps has been made by Schlegel et al. (1998) who use these to estimate values of reddening for every line-of-sight in the sky. Other more recent work such as by Schlafly & Finkbeiner (2011) has provided re-calibrated conversions to reddening. Extinction of an object is the difference between the observed B-V colour and the intrinsic colour and this is equivalent to the difference between the extinction in the B band and that in the V band:

$$E_{B-V} = (B-V)_{\text{obs}} - (B-V)_{\text{int}} = A_B - A_V. \tag{1.2.3.3}$$

$A$ is also related to E(B-V) using an extinction curve, $R$, for the V band this relation is:

$$R_V = \frac{A_V}{E(B-V)}. \tag{1.2.3.4}$$

Reddening has a different effect on the light though each photometric filter, and therefore a different correction is necessary for each bandpass. The relationship above is adjusted with a different $R$ for each bandpass.

SDSS provide alongside the various magnitude measurements for each bandpass the galactic extinction measurements along with the corrected dereddened values of the magnitudes. This is important as most scientific work requires use of the dereddened magnitudes and colours.

### 1.2.4 Photometric Redshift Estimation

Large, deep redshift surveys are necessary for studying the large scale structure of the Universe and the evolution of dark energy (Seo & Eisenstein 2003; Hong et al. 2012), and a number of surveys have focused on achieving this goal [the Baryon Oscillation Spectroscopic Survey (BOSS) of the Sloan Digital Sky Survey III (SDSS-III; Dawson et al. 2013), the WiggleZ Dark Energy Survey (Blake et al. 2011b), the 2df Galaxy Redshift Survey (Colless et al. 2001), KIDS and DES and upcoming surveys such as *Euclid* (Laureijs et al. 2011) and LSST (LSST Science Collaboration et al. 2009) will provide unprecedented constraints on cosmological parameters. Although spectroscopic redshifts provide the most accurate redshift estimates, the process of obtaining spectroscopy is very time consuming, and is only feasible for nearby or bright galaxies, or very small areas containing faint galaxies (e.g. Alam et al. 2017; Lilly et al. 2009). Photometric redshifts (hereafter photo-zs) on the other hand, provide a more efficient method of obtaining redshifts to much greater depths than possible for spectroscopy (Connolly et al. 1995; Koo 1985; Blake et al. 2007; Oyaizu et al. 2008). Therefore, cosmological measurements that use large redshift samples will benefit from the use of accurate photo-zs. Mixed photometric and spectroscopic surveys such as SDSS also benefit from photo-z estimation as photometry is always deeper than spectroscopy and allows the most efficient use of the survey data (e.g. Almosallam et al. 2016a; Abdalla et al. 2011; Oyaizu et al. 2008; Li et al. 2007; Blake et al. 2007). In order to accomplish the science goals set, present and upcoming surveys have very stringent requirements for their photo-z errors. For example, *Euclid* requires root mean square error, $\sigma_{RMS} < 0.05(1+z)$, a catastrophic outlier fraction, i.e. the fraction of objects such that $|z-\bar{z}| > 0.15(1+z)$, $\eta_{0.15(1+z)} < 0.1$ and an error in the mean redshift of each tomographic bin, $e_{bin} < 0.002(1+z)$ (Laureijs et al. 2011). LSST requires $\sigma_{RMS} < 0.05(1+z)$, fraction of $3\sigma$ outliers, $\eta_{3\sigma} < 0.1$ and bias $(z-\bar{z})$, $b < 0.003(1+z)$ (LSST Science Collaboration et al. 2009). As a result, a significant amount of work is being done to increase the efficiency and accuracy of the process via the creation of new algorithms and optimization of existing

ones (e.g. Hildebrandt et al. 2010; Abdalla et al. 2011; Benítez et al. 2009; Brammer et al. 2008; Hogan et al. 2015; Almosallam et al. 2016a; Gomes et al. 2018).

#### 1.2.4.1 Photo-z Estimation: Template Fitting Methods

The method of using photometry to determine the redshift of galaxies was first developed in the 1960's by Baum (1962). This method involved using broad optical filters to collect the radiation from a galaxy followed by producing spectral energy distributions (SEDs). These SEDs were then compared to redshifted template SEDs of the same galaxy type in the rest frame—using the transmission curve of the filters—to find the best fit and the corresponding redshift. Modern SED template fitting requires a library of either observed or synthetic templates of typical galaxy SEDs with stellar populations of various ages and for different star-formation histories. The observed fluxes are then fitted to these templates, usually using a $\chi^2$ minimisation procedure, to find the set of templates that provide the closest match and the corresponding redshift. This is done by calculating the $\chi^2$ using the sum of the difference between the observed flux, $F_{obs}$, and the template flux at various redshifts, $F_{\text{temp}}(z)$, over all of the filters:

$$\chi^2_{\text{temp}}(z) = \sum_{\text{filters}} \left( \frac{F_{\text{obs}} - kF_{\text{temp}}(z)}{\sigma_{F_{\text{obs}}}} \right)^2, \tag{1.2.4.1}$$

where $k$ is a normalisation constant and $\sigma_{F_{obs}}$ is the error in the observed flux. The template and redshift that minimises this $\chi^2$ value are then chosen, providing a photo-z and the spectral type of the galaxy which can indicate other physical properties. This method works because SEDs can be distinguished based on the shape of the continuum as well as the presence and location of strong spectral properties such as the 4000Å break and strong emission lines [in the case of active galactic nuclei (AGN) and star forming galaxies (Bolzonella et al. 2000)]. Some commonly used examples of template-fitting codes are HYPERZ (Bolzonella et al. 2000), LE PHARE (Ilbert et al. 2006) and EAZY (Brammer et al. 2008). The set of template SEDs used for fitting is chosen based on a number of factors such as star formation rate (SFR), metallicity, initial mass function (IMF), interstellar reddening, flux decreases due to the Lyman alpha forest, and the limiting magnitude of each filter (e.g. Bolzonella et al. 2000). These templates could be either empirical, meaning that they are based on observed galaxies, or theoretical, in which case they are based on theoretical models of stellar population synthesis (the time evolution of the stellar populations in galaxies). One downfall of empirical template libraries is that most observations are of local galaxies which span a relatively limited volume of the parameter space available (parameters such as luminosity, morphology, metallicity etc) thus limiting the templates available for fitting

to observations. On the other hand, the theoretical templates can cover all of parameter space, but the assumptions made in the construction of the model might not be as accurate as observations. A study by Koo (1999) showed that the photo-zs obtained from theoretical and empirical template libraries that were available at the time differed by as much as $0.05$ at most redshifts.

The advantages of template fitting methods are that they allow easy extrapolation—allowing them to be used on very faint galaxies for which limited spectroscopy is available—and they also allow the determination of other physical properties of the galaxies, such as stellar mass and star-formation rate (e.g. Ilbert et al., 2015; Johnston et al., 2015). However, a major drawback is the possibility of template mismatch due to template set incompleteness. This is particularly important considering that the templates are normally based on local galaxies, and thus do not necessarily represent galaxies in the entire sample (e.g. Budavári et al., 2000; Abdalla et al., 2011). Despite this, a library with too many galaxy templates can also be disadvantageous as it can result in colour-redshift degeneracies (Benítez 2000). SED template fitting is sometimes combined with Bayesian techniques such that galaxies with known spec-z's and similar properties to the galaxies being observed are used as priors to calibrate the templates (Benítez 2000; Ilbert et al. 2006; Feldmann et al. 2006) and these methods often lead to improved results. Examples of such methods are: BPZ (Benítez 2000) and ZEBRA (Feldmann et al. 2006).

#### 1.2.4.2 Photo-z Estimation: Empirical and Machine Learning Methods

Empirical techniques for photo-z estimation were first developed in the 1990's (see Connolly et al. 1995; Wang et al. 1998) as an alternative to template based methods that were independent of any theoretical models about galaxy SEDs and did not require a large catalog of observed SEDs. These techniques involve using a sample of galaxies with spectroscopic redshifts and photometric data to develop an empirical relationship between magnitude and redshift for a particular passband. Connolly et al. (1995) studied the distribution of galaxies within magnitude space (Connolly et al. (1995) used a 4 passband system) and found that a simple quadratic equation could relate the magnitudes to the redshift. The coefficients were found using a least squares minimisation procedure. With this relationship, the intrinsic scatter of the photo-zs was found to be less than $\Delta z = 0.02$ for redshifts up to $z = 0.8$. Notably, Connolly et al. (1995) also noticed that the feature in the spectrum that was the main contributor to this quadratic relationship was the 4000Å break which shifted along the filters as redshift increased. This quadratic relationship was embraced and used a number of times (see Brunner et al. 1997; Subbarao et al. 1996; Sowards-Emmerd et al. 2000; Hsieh et al. 2005). Another empirical relationship was presented by Wang et al. (1998)

who found a linear relationship between redshift and three colours (U-B, B-V and V-I), but this equation was altered, with different coefficients based on the colours of the galaxy. The model created extended to redshifts $z \lesssim 4$ and produced results consistent with the template fitting methods available at the time (see Figure 1.15). Similarly, Richards et al. (2001) found a relationship between each colour and redshift for a sample of quasars. The redshifts of quasars were then estimated by performing a $\chi^2$ minimization over the sum of the differences in each of the observed colours and colour-redshift relation colours at various redshifts. This method was found to perform well despite the lack of the 4000Å break, suggesting that this was not a vital feature, but could improve accuracy.

In recent years, machine learning methods of photo-z estimation were developed. These algorithms develop complex models that fit the given data—making them superior to traditional empirical methods that are limited to simpler functions—(some examples are: ANNz (Collister & Lahav 2004) and ANNz2 (Sadeh et al. 2016), GAz (Hogan et al. 2015), TPZ (Carrasco Kind & Brunner 2013) and GPz (Almosallam et al. 2016a,b) which use artificial neural networks, genetic algorithms, random forests and Gaussian Processes, respectively).

Machine learning is a field of computer science that involves giving a computer the ability to learn without being explicitly programmed. 'Learn' here means determining a better model for performing a given task on the model. Such tasks include regression, classification and clustering. Machine learning methods can be classified as supervised, unsupervised and semi-supervised methods. Supervised models are given data with labels (for example the characteristics of a set of fruits and their names as the labels). This is called the training set and it is used to develop the model by finding model parameters based on the relationships between features and labels. Training is done by minimizing a loss function which is a measure of the difference between the output predicted by the model and the expected output (the label). As training is taking place, fitted models are run on a validation set, which is another set with labelled data, and it is used to optimize the relevant parameters/weights and prevent the model from fitting so exactly to the training data that it is unable to generalize to any new data (this is called overfitting). Following this the model is applied to a new set of labelled data (called the test set) and the performance of the model is determined using various metrics and used to assess the suitability of the model for a particular task. A suitable model can then be applied to the target set, the dataset with features of samples for which we need the predicted outputs. Supervised learning methods are normally used for regression and classification. Regression models describe the relationship between the independent variables, or features, $\mathbf{x}$ and the dependent one, $y$, the outputs, $y$ are continuous in this case. Classification models on the other hand describe

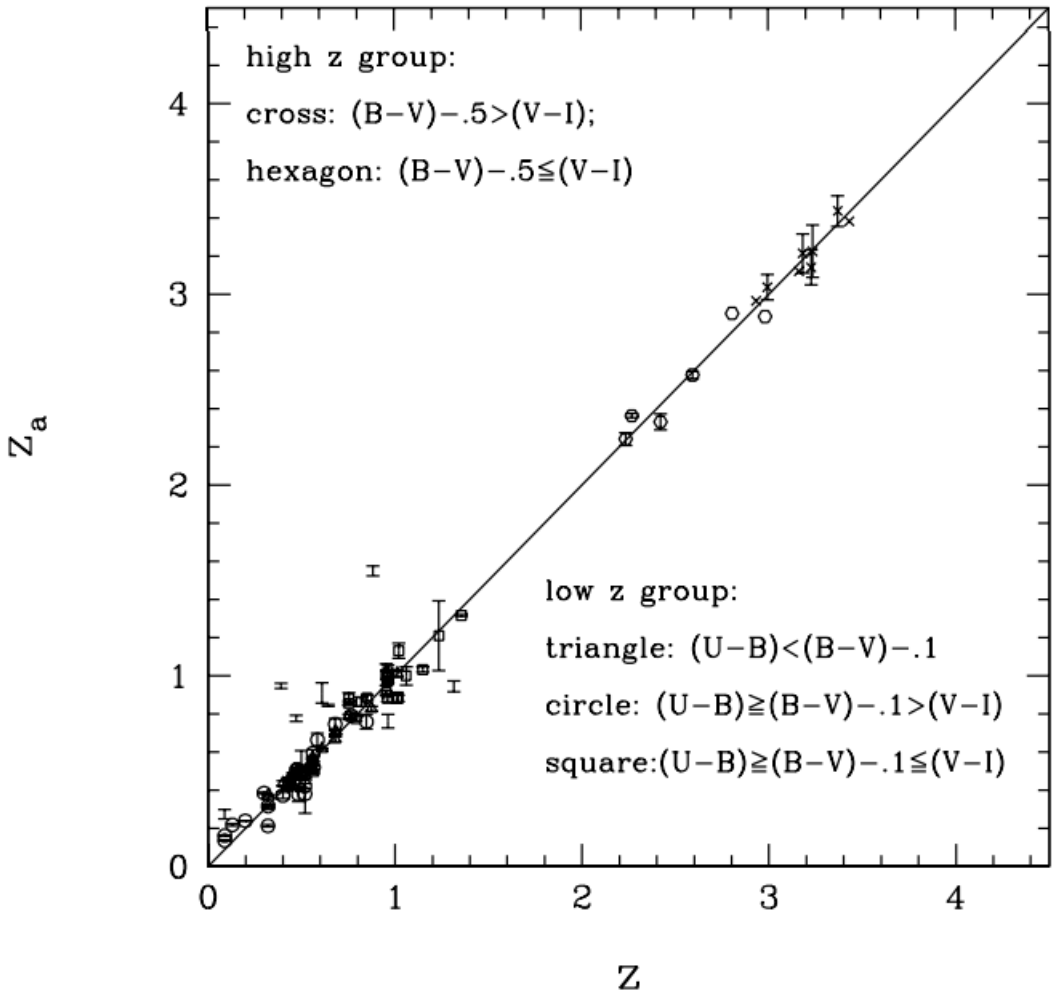



Figure 1.15: Figure showing the redshifts estimated by Wang et al. (1998) using a linear relationship between colour and redshift versus the spectroscopic redshift for 90 galaxies in the Hubble Deep Field. The different shapes represent the different colour restrictions that were assigned different coefficients in the linear relationship. Figure taken from Wang et al. (1998)

the relationship between the features and the likelihood of membership in each possible (discrete) class.

Unsupervised learning, on the other hand, has no training set and no need for labelled data. Instead of learning a relationship, these models find relationships between samples based on their features. Unsupervised learning is often used for tasks such as clustering in which case similar samples are grouped together based on their locations in the feature space and anomaly detection which involves determining if any samples are sufficiently different from most of the data. Finally, there are some techniques that utilise both supervised and unsupervised learning in different parts of the model. This is often used when there are large amounts of unlabelled data and small amounts of labelled data. An example of of a technique that uses both supervised and unsupervised learning is a face recognition algorithm which can detect the same face in multiple images in an unsupervised manner but needs to be provided with at least one labelled image to assign a name to that face.

Empirical photometric redshift estimation is a regression task as the model must map the flux in the various filters to the redshift. Supervised learning is therefore needed and spectroscopic redshift samples can be used as labels for the training, validation and test sets. After training is performed and the performance on the test set is satisfactory the resulting model is used for predicting the redshifts for a different set of galaxies given only their photometric data. A common problem with machine learning photometric redshift estimation is that the labelled data used for training, validation and testing, that comes from spectroscopic samples, can have a different colour and redshift distribution to the target set. This is because most spectroscopic samples are very shallow measurements of bright objects as it is much more difficult to obtain spectra from faint distant objects while the target set will contain galaxies that are much more distant. In addition, for objects in the range $1.4 < z < 3$ the most prominent spectral features (4000Å break and strong emission and absorption lines) are shifted into the near-IR portion of the spectrum, and therefore it is difficult to obtain spectroscopic redshifts from optical spectra. This region of redshift is called the 'redshift desert'. This lack of representativeness of the training set will lead to poorly determined redshifts for objects with colours that are not present in the training set. This problem has triggered significant work on obtaining spectra for galaxies at all relevant regions of colour-space (see Masters et al. 2015).

While the accuracy of photo-zs varies significantly depending on the method and the specific algorithm used, as well as the size and representativeness of the training set available, in general, these methods produce accurate redshift results coupled with acceptable estimates of uncertainty when a representative training set is available. Unlike SED fitting, these methods also do not require the nature of the observed galaxy to be explicitly known

or assumed (Connolly et al. 1995). On the other hand, the necessity of a representative spectroscopic sample is a major drawback. Thus, these methods normally outperform template-fitting methods when a spectroscopic sample that is large and representative is used, but perform poorly in comparison when such a sample is not available (such as at very faint magnitudes) (Oyaizu et al. 2008; Firth et al. 2003). Therefore, some combination of these methods depending on the science goal is likely to be the most accurate. New photo-z estimation techniques, mostly machine learning or hybrid techniques and improvements on existing ones, are being produced on a regular basis (e.g. Chapter 2, Duncan et al. 2018; Soo et al. 2018; Pasquet et al. 2019; Chong & Yang 2019; Leistedt et al. 2019).

#### 1.2.4.3 Photo-z Probability Distributions and Uncertainty Estimation

Modern photo-z codes, both template fitting and machine learning types, are expected to provide probability density functions (PDFs) describing the spread of the probability of the galaxy having a given redshift. These PDFs are useful for understanding the uncertainty in the photo-z point estimate as a greater spread implies greater uncertainty. In addition, in some cases the algorithm might find two or more likely redshifts (perhaps if multiple templates provide sufficient fits) and this is represented in the PDF as multiple peaks, i.e. a multi-modal distribution. In these ways an output PDF provides a significantly larger amount of information than a single point estimate and a point estimate can easily be obtained from the PDF. In addition, the PDF allows the user to choose a definition of the point estimate: for example it could be the mean, the mode or a Monte Carlo sample from the PDF. Different photo-z codes provide different types of PDFs. For example, the template fitting code BPZ (Benítez 2000) produces PDFs by summing the bayesian posterior probability of a galaxy having redshift $z$ over all templates. Figure 1.16 shows an example of this posterior probability for one galaxy and the corresponding likelihood and prior probability distributions used to create it. The PDFs produced by machine learning codes varies depending on the algorithm used. Gaussian process models provide Gaussian PDFs automatically and use a Bayesian approach to finding these PDFs. Most other algorithms do not naturally produce PDFs and creative methods for creating distributions similar to PDFs are implemented. For example, Ball et al. (2008) used a K nearest neighbours method which compares the photo-z to the spec-z's of the K nearest spectroscopic galaxies in magnitude space and determines the deviation of the photo-z from the spec-z's. The spread in the deviations, such as the $68^{th}$ percentile width, is then used to form the PDF. Sadeh et al. (2016) use the same approach for ANNz2. Polsterer et al. (2016) suggest a similar method in which if nearest neighbours are used to calculate the redshift, the nearest neighbours are fitted with a Gaussian mixture model to create detailed (possibly multimodal) distributions.

They also suggest using a random forest algorithm and fitting a Gaussian mixture model to the individual decision trees.

#### 1.2.4.4 Estimating Redshift Distributions using Reference Samples

A number of cosmological measurements require only the redshift distributions ($N(z)$) of galaxies in tomographic bins instead of the individual point estimates. There are a number of different methods currently utilized for obtaining $N(z)$ from individual galaxy estimates. A common method is to stack the PDFs allowing one galaxy to contribute to multiple bins with different magnitudes, then summing the probabilities in each bin. Other approaches are binning the expectation values of the individual redshift probability distributions, and similarly binning of Monte Carlo random draws from the PDFs. The Monte Carlo binning is expected to be less sensitive to the biases present in the estimation technique and approaches the results of the stacking method with infinite galaxies. Although to various degrees, these methods are all still susceptible to the biases inherent in the photo-z estimation code used. As an alternative to obtaining estimates for individual galaxies, some methods that obtain the $N(z)$ distributions directly, or obtain the redshift distributions of the galaxies binned using another method have been created. Some of these methods are cross-correlation techniques in which galaxies with unknown redshifts are cross-correlated with a reference sample with known redshifts (see Newman 2008 and Matthews & Newman 2010) or with HI intensity maps (Cunnington et al. 2019). Another common method of this type is the re-weighting of spec-z's or accurate photo-zs such as COSMOS 30-band photo-zs (Laigle et al. 2016) based on the magnitude space of the photometric sample (Lima et al. 2008).

## 1.3 Thesis Structure

The work presented in this thesis aims to investigate how improvements can be made to the cosmological measurements made using photometric redshifts. This is done in 3 chapters:

- Chapter 2 is an investigation into methods of improving the measurements obtained from the photometric redshift estimation algorithm GPz. I consider the use of additional features such as near-infrared filters and size measurements for creating the model as well as the use of a post-processing technique. I also compare the photo-z estimation performance with photometry of different quality. This chapter also details the application of a direct $N(z)$ estimation method that uses a re-weighting technique and the COSMOS 30-band reference sample to Hyper Suprime-Cam photometry.

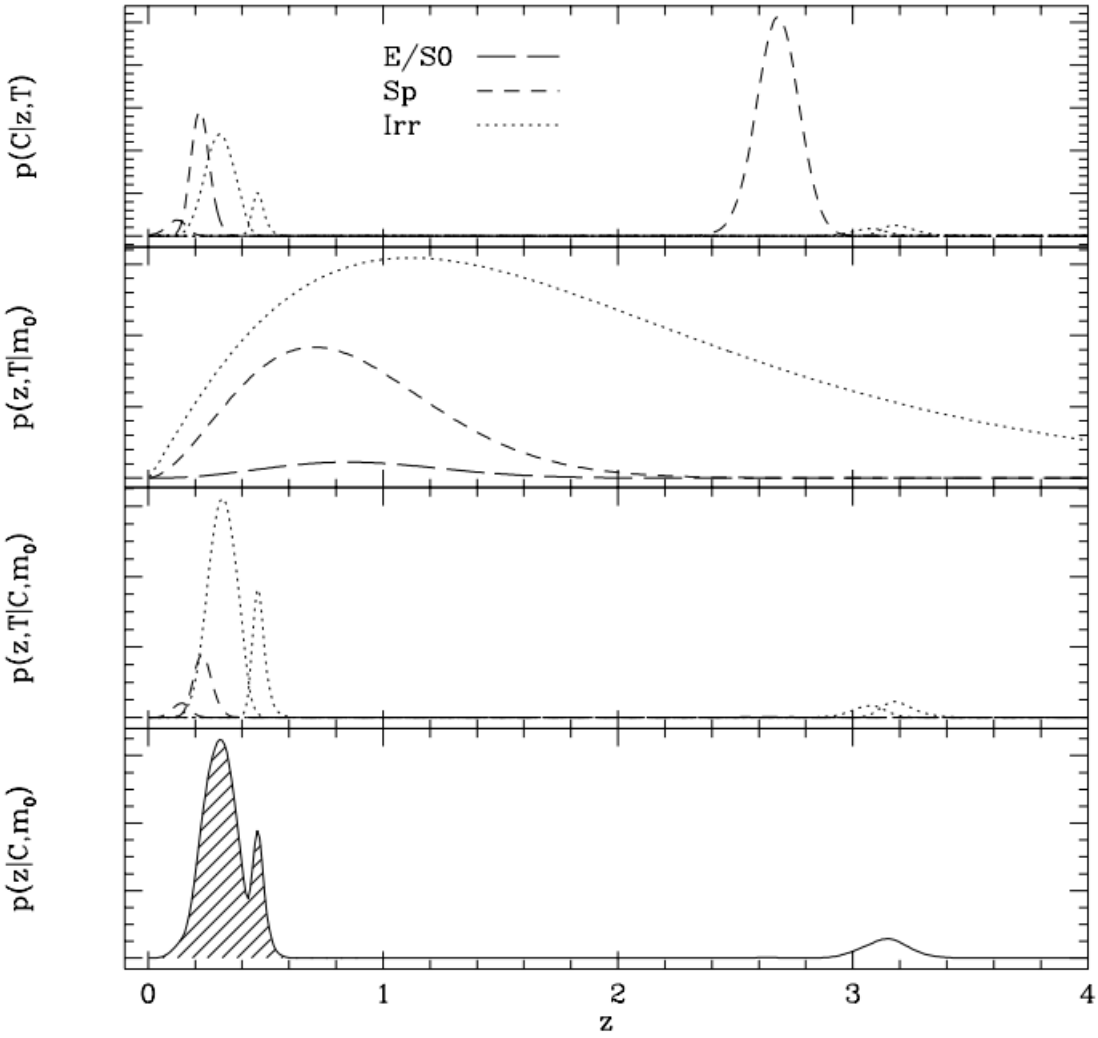



Figure 1.16: Figure showing the relevant probability distributions used by the BPZ algorithm for each galaxy. The top plot shows the likelihood functions for the templates used, this is the probability of measuring magnitudes/colours C given that the galaxy has the relevant SED (from the template) and redshift $z$. The second plot shows the prior probability that a galaxy with the SED of the relevant template will have redshift $z$. The next plot shows the posterior probability that the galaxy has redshift $z$ and SED given by the relevant template given that it has the observed magnitudes/colours, this is the product of the likelihoods and the priors. The final plot is the sum of this posterior probability over all templates. Figure taken from Benítez (2000)

.

- Chapter 3 provides the methods used and results of a clustering analysis and subsequent BAO peak detection of both luminous galaxies and all galaxy types using SDSS data and GPz for determining the redshifts. The aim of this analysis was to determine whether the accuracy and uncertainty estimates provided by GPz would allow cosmological measurements to be made using large samples of galaxies without colour cuts, thus significantly increasing sample sizes.

- Chapter 4 presents a Fisher forecast analysis of the constraints on the local primordial non-Gaussianity parameter, $f_{\rm NL}$ possible with the SKA radio continuum surveys. A multi-tracer analysis with multiple radio-selected galaxy populations that are differently biased tracers of dark matter was used. This is done using angular auto and cross correlation functions, assuming that only photometric redshifts are present. In addition, the use of $N(z)$'s from different simulations as well as halo bias functions from both simulations and observations were compared.

Finally, in Chapter 5, I summarise the methods used and results obtained in the three previous chapters and end with a discussion of future work that follows from or builds on the work presented in this thesis.

# Chapter 2

# Improving Photo-z Estimations

In this chapter I will discuss my work on improving redshift estimations. The first section, and the bulk of this chapter will be focused on the improvement of the results from the Gaussian Processes for photometric redshift estimation (GPz) algorithm. This is work that was published in the Monthly Notices of the Royal Astronomical Society under the title: 'Improving Photometric Redshift Estimation using GPz: size information, post processing and improved photometry' with the co-authors Matt J. Jarvis, Ibrahim A. Almosallam and Stephen J. Roberts.

The second section is about a magnitude-space re-weighting method for estimating the redshift distributions of galaxy populations without finding point estimates. I will discuss my application of this method to HSC data as part of a LSST project. This will be followed by a discussion of other applications of the method and its downfalls. This work was done predominantly alongside David Alonso, with input from other collaboration members.

## 2.1 Introduction

The next generation of large scale imaging surveys (such as those conducted with LSST and *Euclid*) will require accurate photometric redshifts in order to optimally extract cosmological information. GPz is a promising new method that has been proven to provide efficient, accurate photometric redshift estimations with reliable variance predictions. In the first part of this chapter, I investigate a number of methods for improving the photometric redshift (photo-z) estimations obtained using GPz (but which are also applicable to other photo-z estimation codes).

Some cosmological analyses need only the redshift distributions of the galaxies in each redshift bin or over the redshift range of interest. In the second part of this chapter I provide an overview of one method of obtaining these redshift distributions and I apply it to a photometric sample obtained from the HSC survey.

Part one of this chapter begins with a brief overview of the GPz algorithm and its advantages for photo-z estimation (Section 2.2.1). This is followed by a discussion on a number of approaches for improving the results obtained from GPz. One approach, presented in Section 2.2.2 involves the use of near-IR photometric filters and the angular size of galaxies as inputs for the training, validation and testing of the GPz model. Another approach is the use of a post-processing method that adjusts the positions of the probability distributions of the photo-zs—to minimize the deviation of the distributions obtained from those representative of the spectroscopic sample—based on their quantile-quantile plots, this is discussed in Section 2.2.4) and finally, the effect of photometric data with increased precision is discussed in Section 2.2.5. The second part of this chapter, section 2.3 will discuss the application of a magnitude-space re-weighting method of estimating redshift distributions to HSC galaxies. Conclusions are provided in Section 2.4

## 2.2 Improving results from GPz

### 2.2.1 Photometric Redshift Estimation using GPz

Gaussian Process (GP) regression (Rasmussen & Williams 2006) is a non-linear, Bayesian, non-parametric method of modelling distributions over functions. GP regression for photo-z estimation involves assuming that the input, $\boldsymbol{x}_i \in \mathbb{R}^d$ (the set of $d$ magnitudes for the $i$-th object and—in the case of GPz—the associated magnitude uncertainties) and output $y_i$ (the corresponding spec-z's) distributions are related such that:

$$y_i \sim \mathcal{N}\left(f(\boldsymbol{x}_i), \sigma^2\right), \tag{2.2.1.1}$$

assuming the following prior probability distribution over the function

$$f(\boldsymbol{x}) \sim \mathcal{N}\left(0, \boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{X}\right)\right), \tag{2.2.1.2}$$

where $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{n\times d}$ is the set of $n$ training samples and $\boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{X}\right) \in \mathbb{R}^{n\times n}$ is a covariance function such that the element in the $i$-th row and the $j$-th column is determined by a function of the pair of inputs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The covariance function is unbounded, i.e. it expands with the size of the training set, and it captures our prior knowledge that close-by inputs should be mapped to close-by outputs; e.g. the squared exponential kernel $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / \lambda\right)$ for $\lambda > 0$. From the likelihood in Equation 2.2.1.1 and the prior in Equation 2.2.1.2, one can obtain the predictive probability distribution, using Bayes' theorem, for an unseen test case $\boldsymbol{x}_*$ to be distributed as follows:

$$p(\boldsymbol{x}_*|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) = \mathcal{N}\left(\mu_*, \sigma_*^2\right) \tag{2.2.1.3}$$

where $\boldsymbol{y} = \{y_i\}_{i=1}^{n} \in \mathbb{R}^n$ is the set of $n$ outputs. Training the model then involves maximising the probability of obtaining the outputs $\boldsymbol{y}$ given the inputs $\boldsymbol{X}$, this is done by maximising the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \sigma^2)$ (using the training and validation sets) which allows the determination of the optimal hyper-parameters ($\lambda$ and $\sigma^2$).

The mean function is then given by,

$$\mu_* = \left(\boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{X}\right) + \boldsymbol{I}_n\sigma^2\right)^{-1} \boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{x}_*\right) \tag{2.2.1.4}$$

and the total variance, comprised of both the noise and model variance:

$$\sigma_*^2 = \nu_* + \sigma^2. \tag{2.2.1.5}$$

This process has a large computational cost, $O(n^3)$, but the sparse Gaussian Process introduced by Almosallam et al. (2016a) alleviates this problem by decreasing the number of kernel functions used ($m << n$) without significantly reducing the accuracy of the regression model. In order to accomplish this Almosallam et al. (2016a) allow each kernel function to have its own hyper-parameters in order to account for variable densities and patterns over the sample space, and the locations of these functions are optimised to represent the data distribution.

Almosallam et al. (2016a) also introduce cost-sensitive learning (CSL) methods, which allow the user to vary the weights and error functions of different regions of parameter space depending on the science goals the method is being used to achieve. One type of weighting that is utilised in the GPz code is the normalisation of the data points, these weights are defined as:

$$\omega_i = \left(\frac{1}{1+z_i}\right)^2, \tag{2.2.1.6}$$

where $\omega_i$ is the weight or error cost for sample $i$ and $z_i$ is the spec-z for sample $i$, thus giving lower redshift objects greater weight than higher redshift ones. In this analysis, the normalising weights and no weights cases were considered, with the application of these weights termed CSL method 'Normalised' and CSL method 'Normal' respectively.

The GPz algorithm was further modified to address the problem of heteroscedastic (non-uniform, input-dependent) uncertainties in photometric data. The predictive variance obtained from GP regression, Equation 2.2.1.5, consists of two components, the model variance $\nu_*$ and the noise variance $\sigma^2$. The model variance describes the confidence level for the model that is fit to the data. This decreases as the density of the data in the colour-redshift space of a given data point increases. On the other hand, the noise uncertainty describes the spread of the data points in a given region of colour-redshift space. It therefore depends on the factors such as precision of the data and number of relevant features used.

Noise uncertainty is normally assumed to be white Gaussian noise, but in this case, in order to account for heteroscedastic noise, Almosallam et al. (2016b) model this term as a function of the input $\sigma^2(\boldsymbol{x}_*)$[1] with its own hyper-parameters. This noise variance and the predictive mean function are then both learned over the optimisation process. Figure 2.1 provides an illustration of the model and noise variance.

This sparse Gaussian Process method used for estimating photo-zs was found to outperform other selected machine learning methods in terms of performance metrics, reliability of variance measurements and the length of time required for training (Almosallam et al. 2016a,b). The incorporation of CSL methods allows optimal weighting of sample space and the separation of the variance terms enables the selection of galaxy samples based on both data sparsity and photometric noise in order to provide the most appropriate photo-z sample for a given science goal.

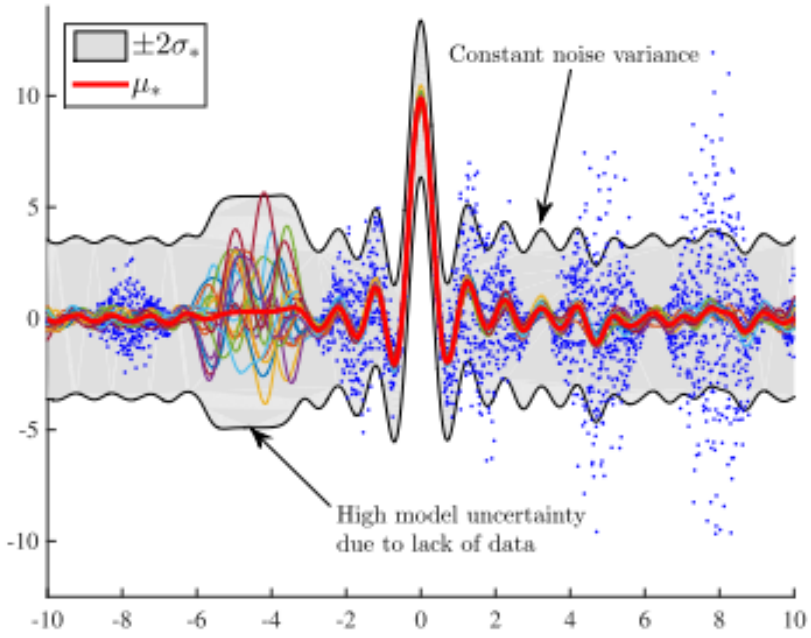
### 2.2.2 Additional features for learning

In this section the addition of commonly available features beyond optical colour/magnitude information and how it may help in improving the accuracy of photometric redshifts using GPz is investigated. Similar studies have been done by Tagliaferri et al. (2003) and Singal et al. (2011) which look at the effect of the addition of features such as galaxy morphology and size on neural network photo-z estimation methods. In addition, a comprehensive study of feature importance for photo-z estimation which included 85 derived or measured parameters such as magnitudes, colours, radii, morphology and ellipticity was presented by Hoyle et al. (2015).

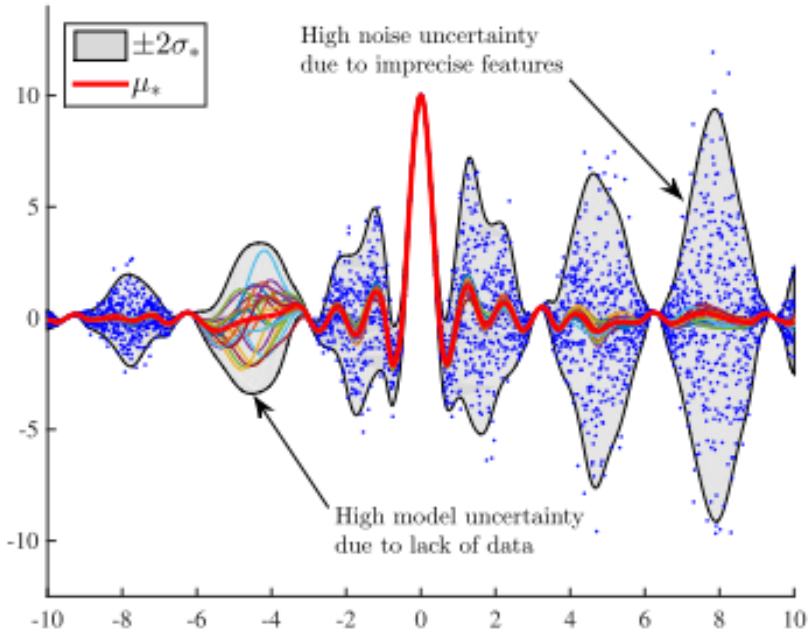
#### 2.2.2.1 Near-IR magnitudes

Large photometric redshift surveys often use photometric systems with 4-5 broad bands in the optical range (e.g. SDSS; Fukugita et al. 1996, DES; The Dark Energy Survey Collaboration 2005, PanStarrs; Chambers et al. 2016 and HSC; Aihara et al. 2018). Photometric redshift determination depends on the detection of continuum features in the SEDs of galaxies, and thus, localising these features is important; for this reason, the traditional broad band filter systems are not necessarily ideal for photo-z estimation (Benítez et al. 2009; Budavári et al. 2001). Study of the Jacobian matrix of fluxes as a function of the physical properties of galaxies has shown that a spectral feature is most noticeable when the feature is in the overlap of two filters (Budavári et al. 2001). One way of improving the probability of this occurrence is to use narrower, more numerous filters (Hickson et al. 1994), but this

[1]Almosallam et al. (2016b) in practice model the precision not the variance, i.e. $\beta(\boldsymbol{x}) = 1/\sigma^2(\boldsymbol{x})$, for numerical concerns but the variance notation is used here for simplicity.




(a) Full GP




(b) GPVL

Figure 2.1: Mean (red), two standard deviations from the mean (black envelopes) and sample functions (narrow lines of various colours) produced by two Gaussian process algorithms. A sinc generative function is modelled as it is a simple example of a case with a 1-D input. In regions with no visible (blue) data points they are all concentrated at the same position, behind the mean of the Gaussian process functions. The first model does not account for heteroscedastic noise so and the noise variance remains relatively constant. The second model accounts for this noise and the model variance is seen to increase when there is limited data while the noise variance increases when there is a spread in the features. Taken from Almosallam et al. (2016b).

would require many more exposures, making it unfeasible. Budavári et al. (2001) explored the possibility of using an additional broad filter formed by combining multiple intermediate width filters located within the original broad filters. This method aids in the location of continuum features within the broad bands and does not significantly increase the total exposure time required as only one additional filter is added. Another study conducted by Benítez et al. (2009) experimented with the number of filters used, the degree of overlap among these filters and constant versus logarithmically increasing filter width. Some conclusions of the study were that for small numbers of filters the colour-redshift degeneracy prevents accurate estimations, particularly for faint galaxies. However, including near-IR data improved the photometric redshift accuracy as it reduced colour-redshift degeneracies. The system that was found to produce the best redshift depth and precision contained nine filters, logarithmically increasing filter width and half-width overlaps.

In this study, the use of the five ugriz filters (Fukugita et al. 1996) that span the optical range is compared to the use of an additional four near-IR YJHK filters in the training of the GPz algorithm. As mentioned previously, near-IR photometry decreases the effect of the colour-redshift degeneracy as it provides data from an additional portion of the spectrum. Near-IR photometry will also aid in the determination of redshifts of galaxies in the 'redshift desert' ($1.2 < z < 1.8$) where there is a lack of emission lines at visible wavelengths and the Balmer break is redshifted into the near-IR part of the spectrum (Rudnick et al. 2001; Mobasher et al. 2004), but this is not explored in the present study.

#### 2.2.2.2 Angular size

Machine learning methods are expected to benefit from using additional features if they provide relevant additional information that aids in determining the relationship between input and output variables, thus providing better constraints on the resulting model. These additional features are not necessarily magnitudes/colours as machine learning methods can take input data of different forms. The classification of the galaxy morphology of SDSS objects for the Galaxy Zoo project provides one example of this: the inputs to machine learning algorithms were not limited to the de-reddened colours, but other features that were related to morphology such as axis ratio measurements and log likelihoods from de Vaucouleurs and exponential fits were also incorporated (Banerji et al. 2010; Gauci et al. 2010). By the same token, the inputs of machine learning algorithms for photo-z estimation are not limited to magnitudes or fluxes. Tagliaferri et al. (2003); Way (2011); Hoyle et al. (2015); Soo et al. (2018) provide examples of the improvements to photo-z accuracy made by including features such as morphology and size as input. Singal et al. (2011) on the other hand found no significant improvement in photo-z estimates when shape parameters

were added. In this analysis the effects of inputting the angular size of galaxies, a relatively simple measurement to make for most astronomical data sets was investigated.

Angular diameter distance, the ratio of the physical size of a body to the angular size we observe, is positively correlated to redshift for $z < 1$. This means that for objects of the same physical size, as redshift increases, the object will appear smaller, i.e. the angular size will decrease. In this experiment, this relationship is exploited by inputting the angular sizes of the major and minor axes of the galaxies as features for training. Since the data used here is in the range z<0.6 this effect is expected to have some effect. On the other hand, for redshifts above 1, the angular diameter distance flattens, to various degrees depending on the cosmology present. This is illustrated in Figure 1.2.

In addition to the evolution of angular diameter distance with redshift, and the resulting correlation between angular size and redshift, physical galaxy size is also related to redshift though galaxy evolution. The evolution of the halo mass function, measured using N-body simulations, indicates that the number of massive halos increased with decreasing redshift (Lukić et al. 2007). This suggests that there are larger numbers of more massive galaxies at lower redshifts. Furthermore, massive galaxies of the same stellar mass ($M > 10^{11} M_{\odot}$) were found to be denser, more compact objects at higher redshifts (Daddi et al. 2005; Trujillo et al. 2006), meaning that they have larger radii at lower redshifts. These relationships further motivate the use of size information in photo-z estimation.

One potential problem that arises when using any galaxy size or shape data in the estimation of the photometric redshifts of the galaxies is that the sizes become correlated with the photometric redshifts. This is a problem if the measurement being made is dependent on the galaxy size/shape—such measurements are cosmic shear and galaxy-galaxy lensing. In such cases it is better to exclude any size/shape data from the photo-z estimation method as this would otherwise result in degeneracies between shape measurements and the photo-zs which would lead to errors in the inferred cosmological parameters that are difficult to account for. One case in which biases can potentially arise is if some larger galaxies are assigned smaller redshifts because they are assumed to be closer then they are. This would in turn influence the size measurements in redshift bins.

#### 2.2.2.3 Experiment and Dataset

The main data set used in this analysis consists of the ugriz and YJHK photometry, angular semi-major and semi-minor axis measurements and spectroscopic redshifts from the GAMA DR2 database (Liske et al. 2015). The GAMA survey is a spectroscopic survey of 238,000 objects split into five survey regions covering a total area of 286 deg$^2$ with a limiting

Table 2.1: Equations defining the metrics used. Symbols $z_i$ and $\hat{z}_i$ are the spectroscopic and estimated photometric redshifts for source $i$, and $\sigma_i^2$ is the predictive variance.

| Metric | Equation |
|---|---|
| RMSE | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i-\hat{z}_i}{1+z_i}\right)^2}$ |
| BIAS | $\frac{1}{n}\sum_{i=1}^{n}\frac{z_i-\hat{z}_i}{1+z_i}$ |
| MLL | $\frac{1}{n}\sum_{i=1}^{n}-\frac{1}{2\sigma_i^2}\left(z_i-\hat{z}_i\right)^2-\frac{1}{2}\ln\left(\sigma_i^2\right)-\frac{1}{2}\ln\left(2\right)$ |
| $FR_{0.15}$ | $\frac{100}{n}\lvert i:\left\lvert\frac{z_i-\hat{z}_i}{1+z_i}\right\rvert<0.15\rvert$ |
| $FR_{0.05}$ | $\frac{100}{n}\lvert i:\left\lvert\frac{z_i-\hat{z}_i}{1+z_i}\right\rvert<0.05\rvert$ |

magnitude of $r < 19.8$ mag obtained using the AAOmega spectrograph on the Anglo-Australian Telescope. The second data release of GAMA contains the spectroscopic data along with photometric data and other additional information obtained from SDSS (which provided the ugriz magnitudes) and the UKIRT Infrared Deep Sky Survey Large Area Survey (UKIDSS-LAS) (which provided the YJHK magnitudes) for 72225 objects in three of the survey regions with total area 144 deg$^2$ with $r < 19$ in two regions and $r < 19.4$ in the third (Liske et al. 2015). A sample of 63937 galaxies was obtained after removing all object duplicates, all objects with missing relevant data and all objects with normalised redshift quality (NQ $\leq 3$). Next, the data set was randomly split into training, validation and testing sets with a ratio of 2:2:1 and these sets were maintained for all experiments performed.

The Gaussian Process with variable covariances (GP-VC) method was used with 100 basis functions and the modelling of heteroscedastic noise was included (Almosallam et al. 2016b). Experiments were done with the CSL methods normal and normalised (defined in Section 2.2.1), and for each of these, three datasets were used: the standard five ugriz magnitudes, the nine ugrizYJHK magnitudes, and the ugrizYJHK magnitudes together with angular size data. For each set of results, five metrics were evaluated: the normalised root mean squared error (RMSE), the normalised bias (BIAS), the mean log likelihood (MLL) and the fraction retained with outlier thresholds 0.15 ($FR_{0.15}$), and 0.05 ($FR_{0.05}$). These are defined in Table 2.1. It is also noted that the addition of the near-IR and angular size features did not significantly increase the training time necessary (this remained under 2 minutes).

#### 2.2.2.4 Results and Analysis

Figure 2.2 shows scatter plots of photometric redshift versus spectroscopic redshift resulting from running the GPz algorithm on the GAMA data (with SDSS/UKIDSS LAS photometry)

using both CSL methods and the three sets of inputs. Table 2.2 gives the corresponding performance measures and predictive variances. These values are calculated for $0 \leq z \leq 0.6$ (where $z$ is the spectroscopic redshift) because the small training set at higher redshifts renders the results unreliable. The straight line shown in Figures 2.2 is the line of $z = \hat{z}$ and thus represents perfect prediction. Consistent improvement is seen as the near-IR magnitudes and then size data are added for both the normal and normalised methods as the distribution becomes tighter and lines up more symmetrically along the straight line. Table 2.2 also shows that the normal and normalised methods result in very similar performance measures.

The noise variance term decreased consistently as the additional features were added for both methods. This is as expected as adding additional, relevant features decreases the spread of the data points in the multidimensional colour-redshift space (Almosallam et al. 2016b). It is clear that the model variance also generally decreases with additional filters and size data. The model variance depends on the confidence about the model, which improves with data density. As features are added, the dimensionality of the model increases, and the data becomes more sparse, but if this additional data improves the model then this can counteract the decrease in data density and model variance can decrease. The normalised method also had lower noise variance values than the normal method in all cases. The normalised CSL method causes the model to preferentially fit the lower redshift regions, producing a completely different fit to what is obtained from the normal method. If the spread of the data is smaller in the lower redshift range, and the higher redshift range is not as important, then the resulting noise variance of the entire model (at all redshifts) can be lower than it would be using the normal method, thus explaining this result. For real situations, in which no spectroscopic data is present, the variance terms may be the only method of determining the quality of the results obtained, thus this general decrease of the variance with additional features is important as it corresponds to improved performance measures.

Next, redshift bins of width 0.1 were defined and the five metrics and average variances for each redshift bin were calculated in order to understand the relationships between the CSL methods, features used and redshift range. Table 2.3 shows these metrics using ugriz features, and the same trends are observed when the additional features were added. It is clear that the results improve as the number of objects in the redshift bin increases: the 0.1-0.2 bin contained the largest number of objects and correspondingly produced the best results, the bin with fewest data points (0.5-0.6) produced the poorest results. This is because the GPz code minimises the total sum of squared errors, and therefore will preferentially fit the regions of sample space with higher densities of data points. The normalised method

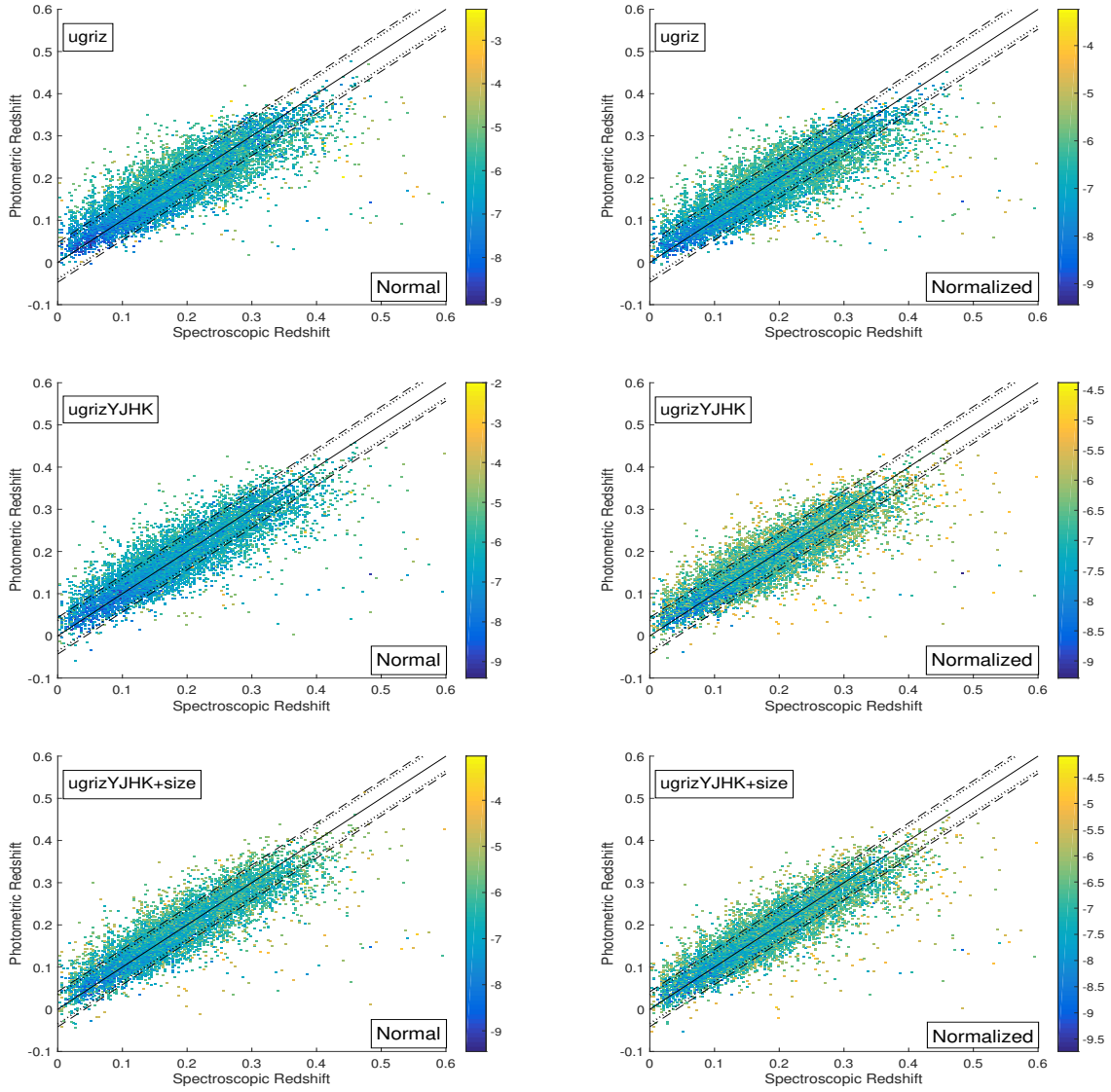



Figure 2.2: Photometric redshift versus spectroscopic redshift plots using the CSL methods normal and normalised and using ugriz, ugrizYJHK and ugrizYJHK filters with size data. The solid line is the photometric redshift = spectroscopic redshift line. The dashed and dotted lines represent the rms scatter, $\sigma_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - \hat{z}_i\right)^2}$ and the normalised rms scatter, $\sigma_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(z_i - \hat{z}_i\right)/\left(1+z_i\right)\right)^2}$, respectively, these are single values for each plot calculated using the whole redshift range $0 \leq z \leq 0.6$. The colour scale represents the predictive variance.

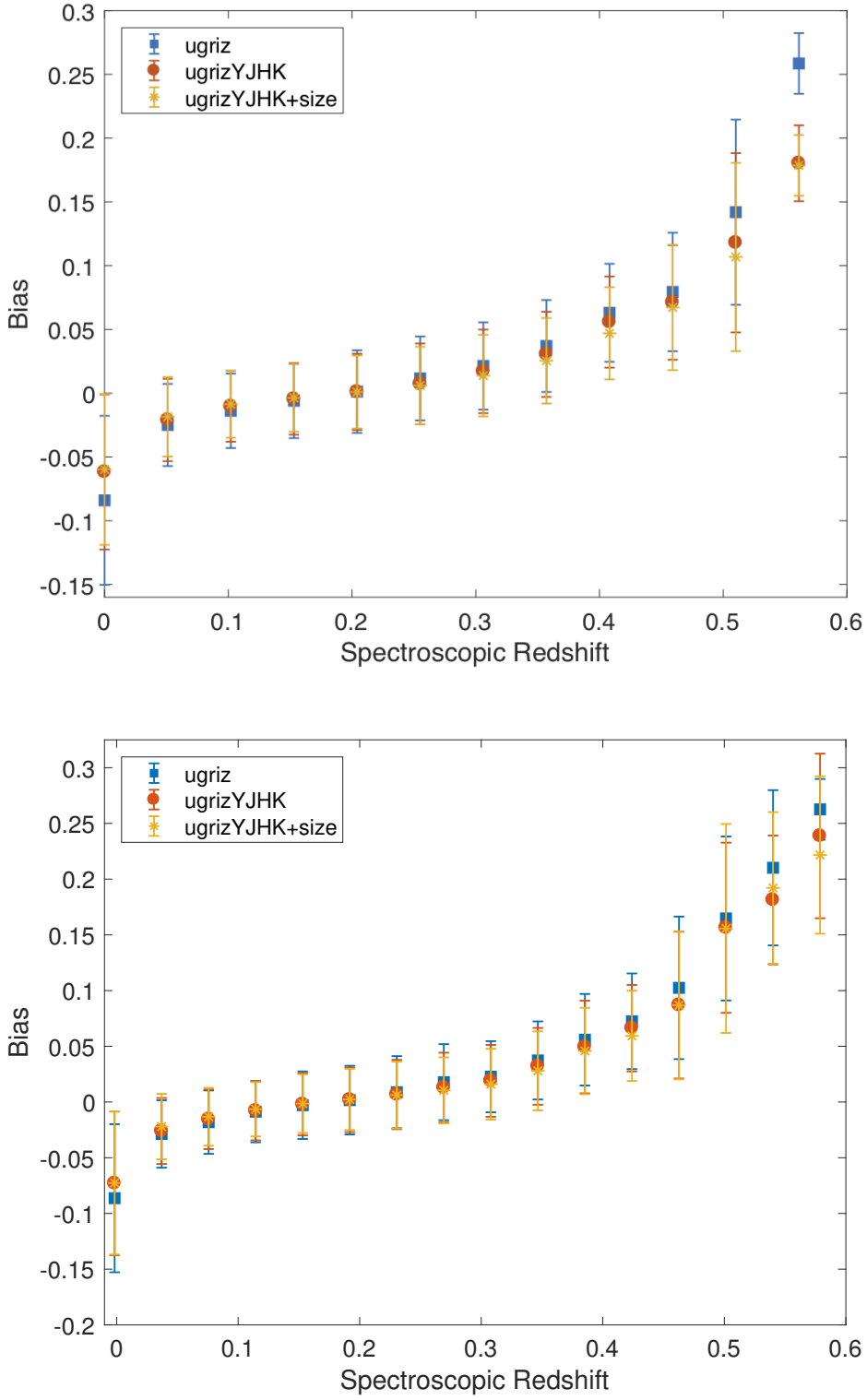



Figure 2.3: BIAS versus redshift plots using the CSL methods normal (upper figure) and normalised (lower figure) and using ugriz, ugrizYJHK and ugrizYJHK filters with size data.

Table 2.2: Summary performance measures and variances for the CSL methods normal and normalised with ugriz filters, ugrizYJHK filters and ugrizYJHK filters and size data. The number of training, validation and testing objects are: 25574, 25575 and 12788 respectively. The best metrics and variances are highlighted.

| CSL Method | Filters | RMSE | BIAS | MLL | $FR_{0.15}$ | $FR_{0.05}$ | Variance | Model Var | Noise Var |
|---|---|---|---|---|---|---|---|---|---|
| Normal | ugriz | 0.0393 | -0.00228 | 1.79 | 99.35 | 85.08 | 0.0023 | 7.0E-06 | 0.0023 |
| | ugrizYJHK | 0.0360 | -0.00185 | 1.84 | 99.54 | 87.92 | 0.0018 | 6.5E-06 | 0.0018 |
| | ugrizYJHK+size | 0.0347 | -0.00177 | **1.91** | 99.50 | 89.61 | 0.0018 | 7.7E-06 | 0.0018 |
| Normalised | ugriz | 0.0387 | 0.00069 | 1.73 | 99.38 | 85.09 | 0.0014 | 6.3E-06 | 0.0014 |
| | ugrizYJHK | 0.0357 | **0.00031** | 1.80 | 99.53 | 87.77 | 0.0013 | 6.9E-06 | 0.0013 |
| | ugrizYJHK+size | **0.0340** | 0.00048 | 1.85 | **99.55** | **89.87** | **0.0011** | **6.2E-06** | **0.0011** |

Table 2.3: Table showing performance measures and variances by redshift bin for the CSL methods normal and normalised with ugriz filters. The best metrics and variances are highlighted.

| | | | | | | Normal | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Redshift Bin | $N_{train}$ | $N_{valid}$ | $N_{test}$ | RMSE | BIAS | MLL | $FR_{0.15}$ | $FR_{0.05}$ | Variance | Model Var | Noise Var |
| 0-0.1 | 4592 | 4621 | 2297 | 0.0510 | -0.0298 | 1.86 | 98.04 | 78.49 | 0.0020 | 9.2E-06 | 0.0020 |
| 0.1-0.2 | 11613 | 11442 | 5888 | 0.0309 | -0.0063 | 2.02 | 99.92 | 89.93 | 0.0019 | 5.6E-06 | 0.0019 |
| 0.2-0.3 | 6809 | 6973 | 3366 | 0.0345 | 0.0095 | 1.73 | 99.88 | 86.54 | 0.0028 | 7.1E-06 | 0.0028 |
| 0.3-0.4 | 2117 | 2105 | 1012 | 0.0471 | 0.0313 | 1.25 | 99.31 | 74.41 | 0.0030 | 9.2E-06 | 0.0030 |
| 0.4-0.5 | 259 | 244 | 134 | 0.0954 | 0.0766 | -2.20 | 91.04 | 37.31 | 0.0071 | 1.6E-05 | 0.0071 |
| 0.5-0.6 | 27 | 28 | 14 | 0.2012 | 0.1879 | -10.22 | 35.71 | 0.00 | 0.0115 | 2.3E-05 | 0.0115 |
| | | | | | | Normalised | | | | | |
| 0-0.1 | 4592 | 4621 | 2297 | 0.0460 | -0.0263 | 1.86 | 98.65 | 81.15 | **0.0013** | 7.5E-06 | 0.0013 |
| 0.1-0.2 | 11613 | 11442 | 5888 | **0.0299** | **-0.0036** | **2.05** | **99.93** | **90.64** | **0.0013** | **5.1E-06** | **0.0012** |
| 0.2-0.3 | 6809 | 6973 | 3366 | 0.0357 | 0.0123 | 1.68 | 99.79 | 84.28 | 0.0017 | 6.4E-06 | 0.0017 |
| 0.3-0.4 | 2117 | 2105 | 1012 | 0.0493 | 0.0341 | 0.89 | 98.91 | 72.43 | 0.0018 | 8.6E-06 | 0.0018 |
| 0.4-0.5 | 259 | 244 | 134 | 0.1039 | 0.0858 | -4.89 | 87.31 | 33.58 | 0.0036 | 1.6E-05 | 0.0036 |
| 0.5-0.6 | 27 | 28 | 14 | 0.2194 | 0.2067 | -16.24 | 35.71 | 0.00 | 0.0054 | 2.2E-05 | 0.0054 |

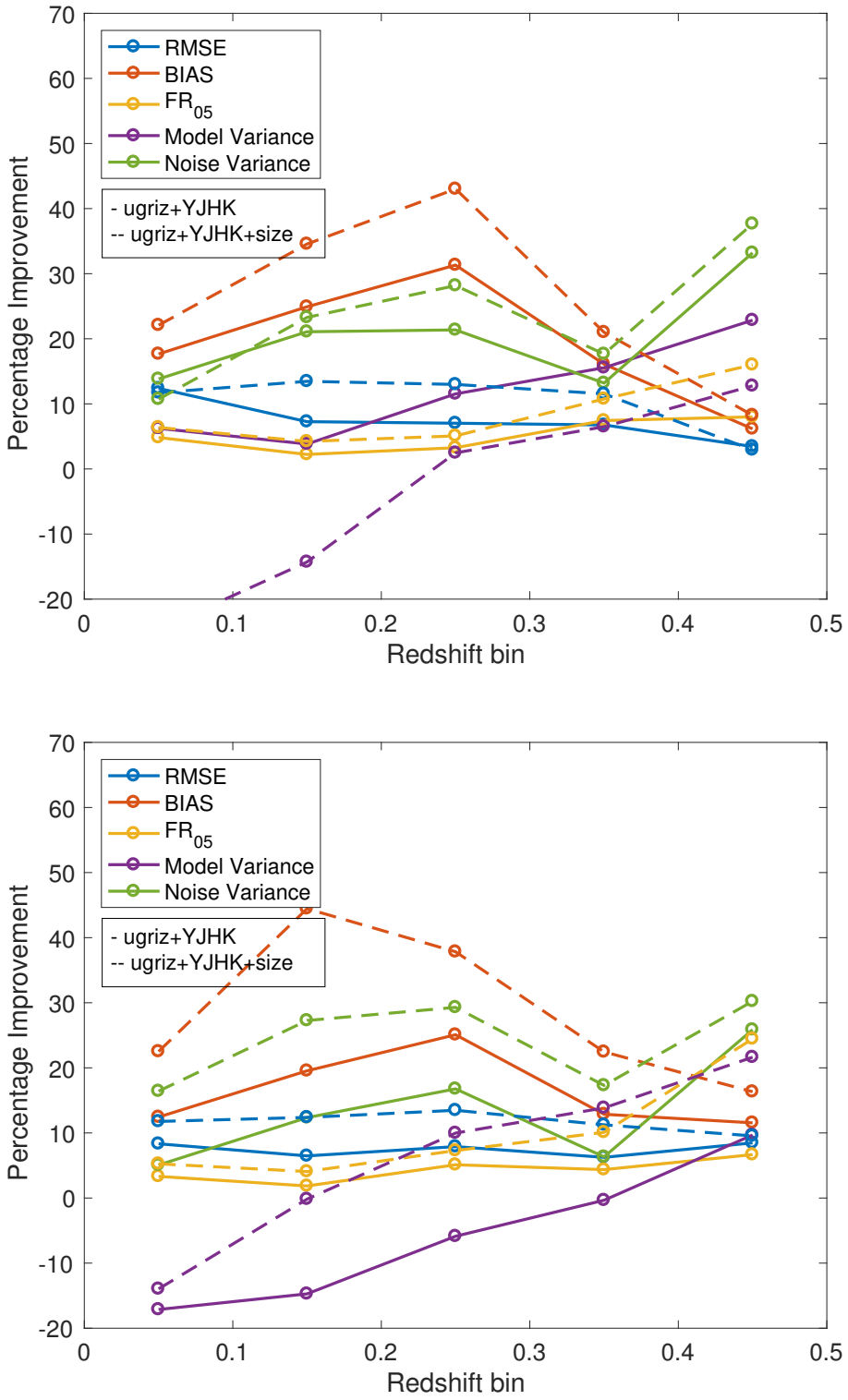



Figure 2.4: Percentage improvements of performance measures and variances by redshift bin due to the use of ugrizYJHK filters and size data using the normal (upper figure) and normalised (lower figure) methods. The solid lines represent the improvements due to adding the near-IR features and the dashed lines represent the improvements due to adding both the near-IR and angular size features. Note that percentage improvements across the metrics are not equivalent as some metrics are much larger than others, therefore small changes in the smaller metrics can lead to large percentage improvements when compared to larger metrics.

performed better than the normal method at lower redshifts ($0 < z < 0.2$), while the normal method performed better in the higher redshift regions ($0.2 < z < 0.6$). This is expected since the normalised method gives more weight to the lower redshift objects than the higher redshift ones in the training of the model. This effect of the normalising weights implies that this method would not be appropriate for science goals which require accurate photometric redshifts of higher redshift galaxies for which data is scarce. On the other hand, if a sample is expected to be mostly at low redshifts, and the accuracy of the few high redshift objects is not important then the normalised method will provide some added accuracy in the low redshift regime. When analysing the variance estimates by redshift bin it was found that in all redshift bins, the normalised method again resulted in noise variance values that were lower than for the normal method. The model variance decreases with increasing number of objects in each redshift bin—this is expected behaviour for the model uncertainty as it decreases as the data density increases. On the other hand, the noise variance increases with increasing redshift, this is because although the higher redshift regions contain less training data, the photometry is likely to be less accurate as these galaxies are fainter on average, leading to a larger spread of the estimated photo-zs.

Figure 2.4 shows the percentage improvements resulting from adding the near-IR followed by the angular size features. Improvements are clearly seen across all metrics in all redshift bins and apart from one case involving the model variance, the angular size features clearly provide significant added improvements compared to the near-IR features alone. The RMSE and $\text{FR}_{0.05}$ metrics both undergo smaller improvements in regions with higher data densities, where the original estimates were more accurate, while they increase more significantly ($\text{FR}_{0.05}$ in particular) in the lower density regions. This is not clear from the plot because the values of these metrics vary significantly with redshift. Regions with larger data densities have smaller metric values and therefore very small improvements can lead to sizeable percentage improvements, thus, a relatively constant percentage improvement with redshift implies larger improvements in the less dense redshift bins and smaller ones in the more dense bins. The BIAS appears to undergo significant improvements over all redshift bins, but because the original values were particularly small (see Table 2.3), very small changes led to large percentage improvements. Figure 2.3 shows the BIAS as a function of redshift, and here, a similar trend to the other metrics is observed: more significant improvements occur in regions of lower number densities. Improvements of $\text{FR}_{0.15}$ (not shown) were negligible, while those of $\text{FR}_{0.05}$ were more significant in all redshift bins, this implies that the addition of these features does not have a significant influence on the worst outliers, but decreases the scatter of objects with smaller initial deviations.

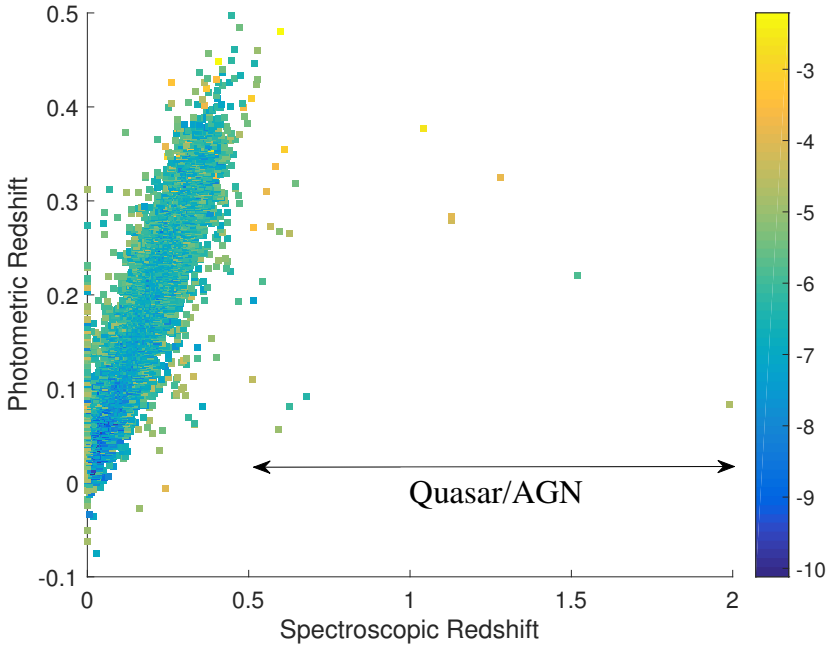



Figure 2.5: Plot showing the location of catastrophic outliers.

### 2.2.3 Outliers

When the entire redshift range for which data is present ($z < 2.1$) was studied, objects with the highest fractional errors $\left(\frac{|z-\hat{z}|}{1+z} > 0.15\right)$ were identified and it was found that most of these objects were either quasars, narrow-line AGN, or had noisy spectra that made it difficult to determine a spectroscopic redshift although in the catalogues provided they were listed as high confidence redshifts. In addition, although this set contained objects with a range of redshifts, all the outliers with high spectroscopic redshifts ($z > 0.55$) were contained in this group, the positions of these outliers are shown in Figure 2.5.

The reason why the GPz algorithm was unable to correctly predict the redshift of these quasars and AGN is because too few of these were present in the training data to allow the algorithm to make realistic estimates. The analysis in the previous section highlights that the number of objects available for training in each bin is an important factor in obtaining accurate estimates, thus if a large sample of quasars and other AGN was present, it is expected that photo-z estimation of these objects would be greatly improved. For the objects with noisy spectra, the spectroscopic redshifts may have been incorrectly determined (as our constraint of NQ $\leq 3$ will not result in 100 per cent accuracy for the spectroscopic redshifts), in which case the photometric redshift estimate may be more accurate than the spectroscopic redshift.

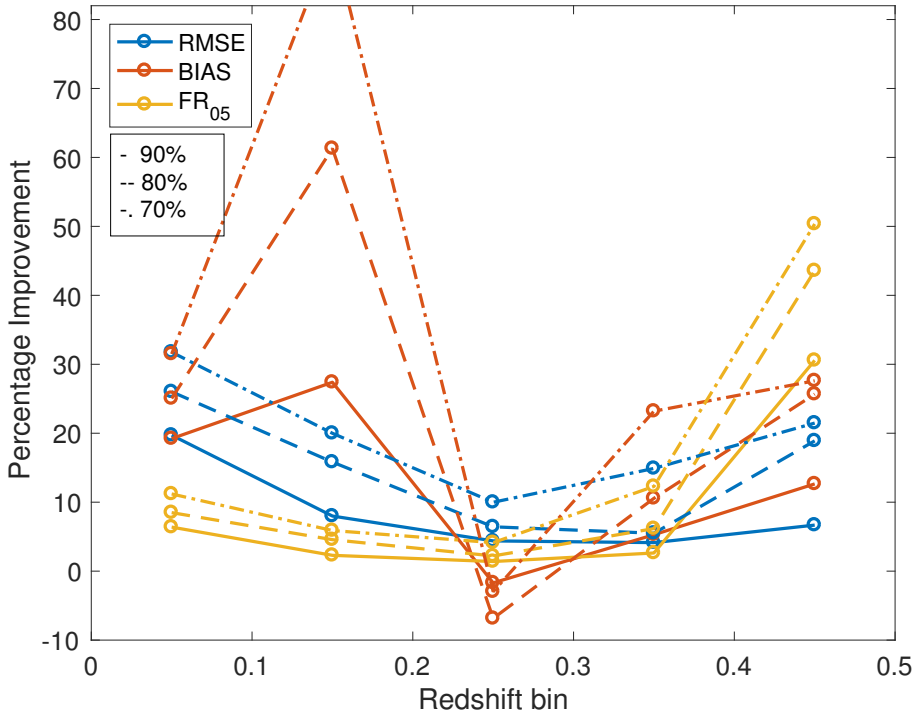



Figure 2.6: Percentage improvements of performance measures by redshift bin when 90, 80 and 70 per cent of the testing data with the lowest uncertainties were used. CSL method normal using ugrizYJHK filters and size data was used.

In Figure 2.6 the improvement in metric performance as the sample is reduced according to the variance prediction is illustrated. As higher variance estimates are removed, our results improve greatly: cutting the values with higher uncertainties and using 90, 80 and 70 per cent of the test data with the lowest variances shows consistent improvement in all metrics. The percentage improvements in the bias are very large in the redshift bin 0.1-0.2 because the initial bias in this bin is very small due to the high data density (see Table 2.3) and therefore small changes in the bias can result in large percentage improvements. In the redshift bin 0.2-0.3 the bias is also very small as this is the second most dense region. If a few more of the high uncertainty redshifts were biased in one direction then this could result in a slightly larger bias causing a negative percentage improvement but since the bias is so small this change is negligible. In general, larger percentage improvements are obtained in the redshift bins with lower number densities.

This analysis was done using ugrizYJHK filters and size data, but the same trend is observed using only ugriz and ugrizYJHK filters. As this removal of data was based solely on the variance values, this method can also be used for real surveys to obtain appropriate samples—based on the specifications for data density and variance necessary for the specific science case.

### 2.2.4 Optimising the Probability Density Functions

It has become clear that single point estimates of the photometric redshifts are insufficient for many scientific applications, and the full probability density function (PDF) is preferred. However, it is extremely difficult to obtain reliable PDFs from both template fitting (due to non-representative templates) and empirical methods (where for example absence of data is traditionally difficult to quantify). Some methods employ post-processing to give their estimated PDFs the correct statistical properties (see Bordoloi et al. 2012; Polsterer et al. 2016), GPz overcomes this by introducing an additional noise term to alleviate some of these issues.

In this section the accuracy of the probability density functions (PDFs) of the photometric redshifts using Quantile-Quantile (Q-Q) plots is investigated (as in Wittman et al. 2016). These Q-Q plots are used to provide appropriate alterations to the PDFs with the aim of further optimising the redshift estimates.

#### 2.2.4.1 Quantile-Quantile Plots

The first step in obtaining Q-Q plots is calculating the percentiles of the spectroscopic redshift values relative to the PDFs of the photometric redshifts. The PDF of the photo-z

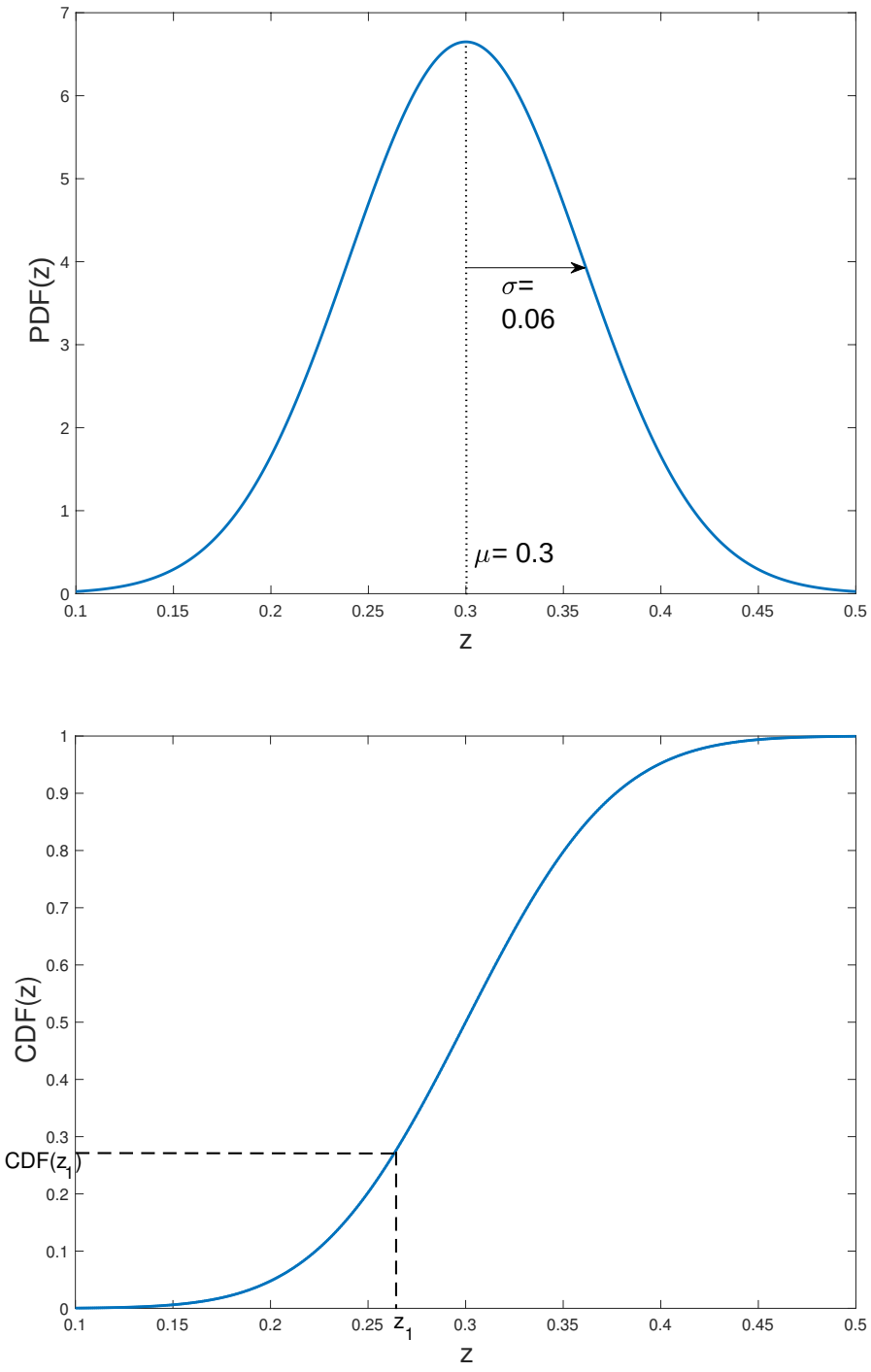
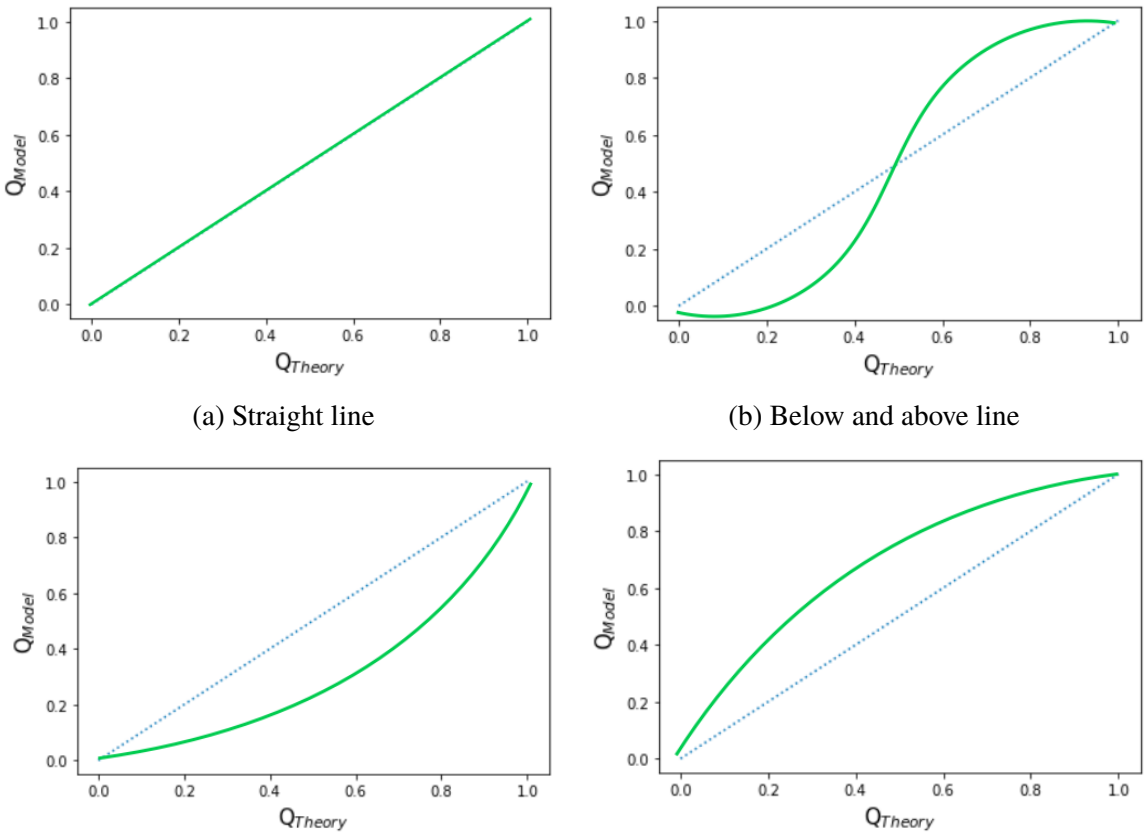



Figure 2.7: The top figure is an example of a PDF of a photometric redshift estimation obtained from GPz. The bottom figure is the corresponding CDF for this estimation.




(a) Straight line

(b) Below and above line

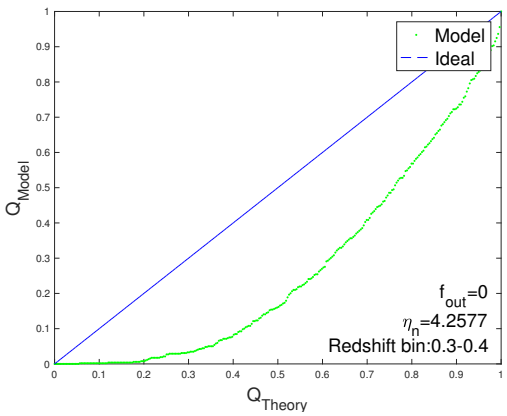
(c) Under line

(d) Over line

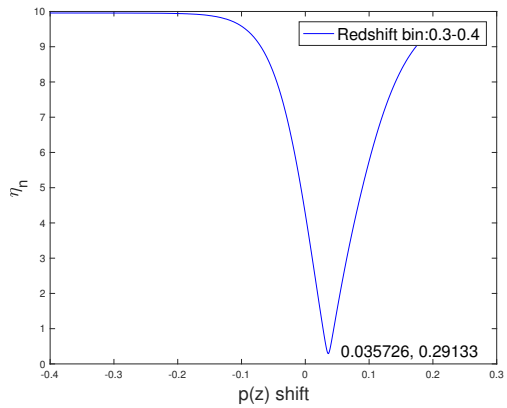
Figure 2.8: Figure showing examples of different forms of a Q-Q plot. (a) Straight line Q-Q plot. This implies that the measured quantiles are the same as the theoretical quantiles. (b) Q-Q plot is both below and above the theoretical line. (c) Q-Q plot is consistently below the theoretical line. (d) Q-Q plot is consistently above the theoretical line.

for a given object obtained using the GPz algorithm is Gaussian with mean equal to the photo-z estimate and variance given by the sum of the model and noise variances. An example of such a PDF is given in Figure 2.7 (top plot). The percentile of a given redshift ($z_1$) is given by the value of the cumulative distribution function (CDF) at that redshift [CDF($z = z_1$)]. This is illustrated in the bottom plot of Figure 2.7. In this way, the percentiles of every spec-z relative to its photo-z PDF can be calculated.

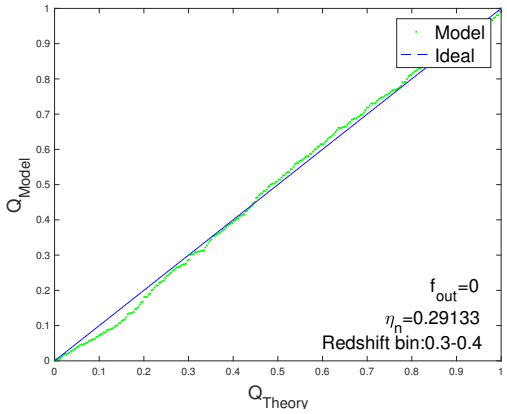
Next, the percentiles were used to determine the quantiles: the quantile at a value $x$ is defined as the fraction of percentiles that are below the fraction $x$ (for example: the quantile at 0.2 is given by the fraction of objects with percentiles less than 0.2). Theoretically, for perfect sampling of a distribution, for all fractions $x$ ($0 < x < 1$), quantile($x$) = $x$ as 20 per cent of values are expected to have percentiles less than or equal to 0.2 and so on. A plot of the calculated quantiles versus the theoretical quantiles is the Q-Q plot and examples are




(a) Before shift




(b) $\eta_n$ vs $P(z)$




(c) After shift

Figure 2.9: (a) Q-Q plot in the redshift bin 0.3-0.4 with the CSL method normal using ugriz filters before applying shifts, (b) the corresponding $\eta_n$ vs $P(z)$ and (c) the Q-Q plot after the $p(z)$ shift was applied.

shown in Figures 2.9a and 2.9c. For the case of finding the calculated quantiles from this photometric data, we count the number of spectroscopic redshifts with percentiles (based on their corresponding PDFs) less than the fraction $x$ and divide this by the total number of spectroscopic redshifts.

A Q-Q plot with a straight line (see Figure 2.8a) indicates that the photometric redshift PDFs (and thus the means and variances) are appropriately representing the spectroscopic redshift distribution, i.e. the spectroscopic redshift values are representative of a random sampling of each of the photometric redshift PDFs. On the other hand, a Q-Q plot that is initially below the straight line and later above it implies that there were too few spec-z's with percentiles less than the fractions below 0.5 and there were too many more than the higher fractions. This in turn indicates that the standard deviation of the PDFs was likely too small, i.e. the distributions were too peaked and narrow. The inverse situation would

imply that the standard deviations were too large.

A Q-Q plot that is consistently below the 'ideal' line (Figure 2.8c) has less spec-z's than expected with percentiles less than each fraction. This means that the PDFs estimated by GPz had means that were generally too low (causing the spec-z values to have higher percentiles), so the photo-zs are being underestimated. Conversely, a Q-Q plot that is consistently above the 'ideal' line (Figure 2.8d) has more spec-z's than expected with percentiles less than each fraction, implying that the PDF estimates had means that were too high. This suggests that the photo-zs are being overestimated. These insights can then be used to our advantage if we then shift the photo-z estimates in the appropriate direction, for example, if the redshifts were underestimated, one could shift them to larger values. This is the technique applied in this section.

#### 2.2.4.2 Applying Shifts to Redshift Bins

Deviations from the straight 'ideal' line discussed above indicate deviation from ideal photometric PDFs and this can be quantified using the Euclidean distance ($\eta_n$). $\eta_n$ versus the shift in the PDF required for the measurement of this $\eta_n$ (e.g. Figure 2.9b) can then be determined. This is done by applying multiple positive and negative shifts to the means of the PDFs, finding the respective percentiles of the spectroscopic redshifts, followed by the corresponding quantiles, and then finding $\eta_n$ of the Q-Q plots. The $p(z)$ shift that minimises the $\eta_n$ can then taken as the shift to be applied to the photometric redshifts.

In this analysis, one half of the test data was used to produce $\eta_n$ versus $p(z)$ for each redshift bin in order to find the optimal $p(z)$ shift (Figure 2.9). These $p(z)$ shifts were then applied to all the best fit photometric redshift values in the respective photometric redshift bins of the second half of the test data. Half of the test data was used instead of the entire set and the shifts were applied in photo-z instead of spec-z bins in order to illustrate the utility of this method when spectroscopic data is present for only a subset of the data. All spectroscopic redshifts with percentiles equal to 0 or 1 were not used for producing the Q-Q plots, as the corresponding photo-zs—defining the relevant PDFs—are considered to be outliers. The fraction of sources not within the PDF (fraction of outliers; $f_{\rm out}$) gives the number of these outliers divided by the number of objects in the relevant redshift bin. In this analysis, the $f_{\rm out}$ after applying the $p(z)$ shifts was 0 in all bins considered (0-0.5) apart from the 0.4-0.5 bin in which the $f_{\rm out}$ did not surpass $0.05$ which corresponds to $3$ objects.

The Q-Q plots obtained after applying the photo-z shifts in all redshift bins were near to straight lines, with very low $\eta_n$values. This means that the GPz algorithm produced appropriate variance estimates with a slight BIAS on the mean values. The optimal shifts found were very small in redshift bins with large numbers of data points, meaning that

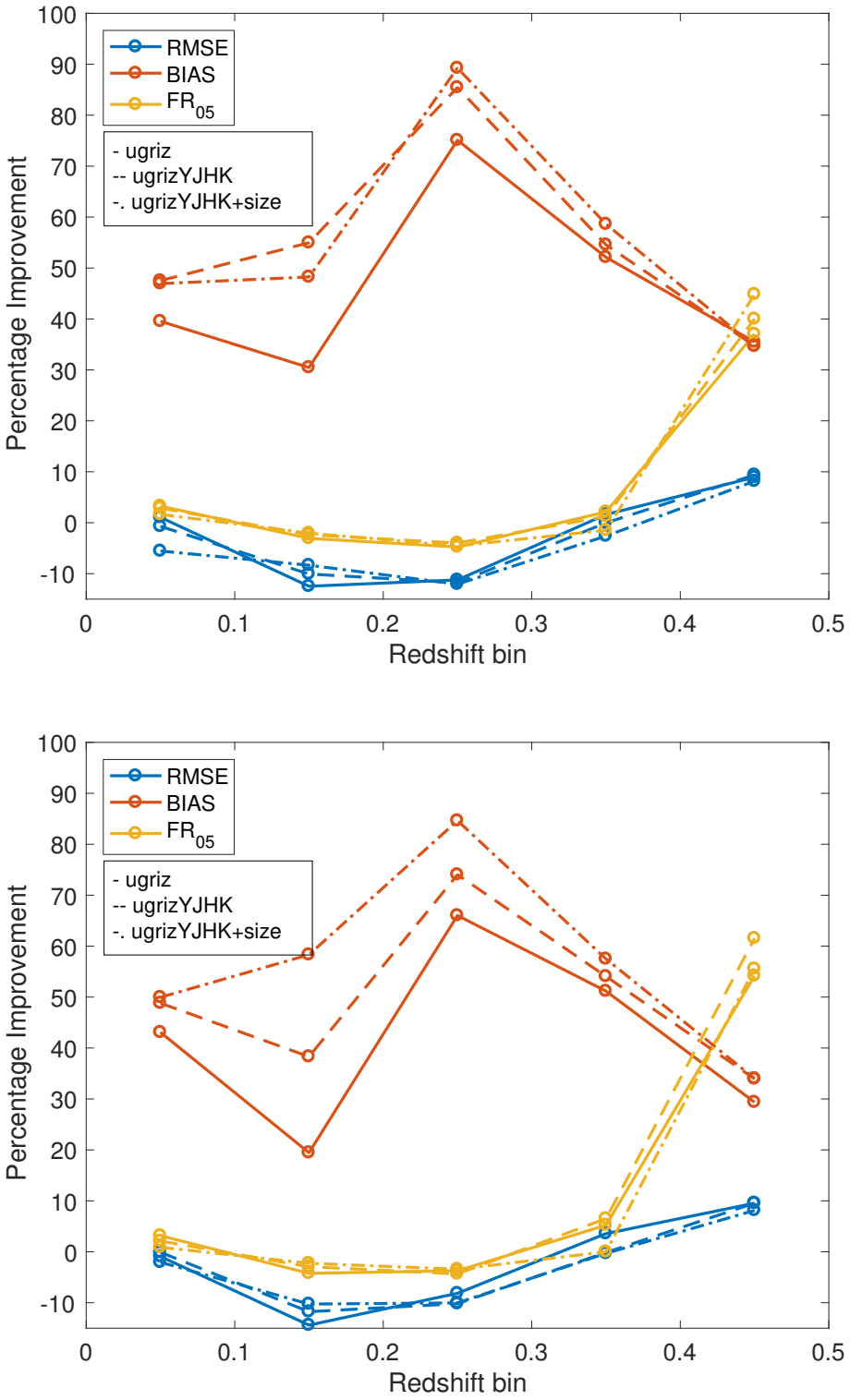



Figure 2.10: Percentage improvements of performance measures by redshift bin due to shifting the means of the photo-z PDFs compared to before the means were shifted using the same training, validation and testing objects and using the normal (top) and normalised (bottom) methods.The solid lines represent using the ugriz features, the dashed line represents using the ugrizYJHK features and the dash-dotted line represents using the ugrizYJHK and angular size features.

the mean values produced in these bins were accurate. On the other hand, the shifts were significant in the redshift bins with lower densities (see Figure 2.9), indicating that some BIAS was present in these bins but that this post-processing method was able to adjust the positions of the PDFs such that they were more representative of the spectroscopic redshifts. Curves like the one shown in Figure 2.9a which represent a lack of objects with percentiles below the given fraction for all quantiles, correspond to the PDFs generally being biased to low redshifts and thus a shift to higher redshift is suggested by the $\eta_n$ vs $p(z)$ plot. Q-Q plots that indicated that the PDFs were biased to low redshift were obtained for the higher redshift bins ($z > 0.2$) while the opposite was obtained for the lower redshift bins ($z < 0.1$). This can be explained by the fact that the best fit mean function will be found where the data density is highest, which is around the redshift of $z \sim 0.2$. In some redshift bins (e.g. $z > 0.5$) the number of galaxies was too small for this analysis to be carried out.

Figure 2.10 gives the percentage improvements of the performance metrics by redshift bin due to shifting the PDFs (the variances are not included as they are not affected by the shifts). For all the redshift bins in which this method was applied ($z < 0.5$), for all configurations of input variables and for both CSL methods there is significant improvement in the BIAS metric, while the other metrics only worsen slightly in some redshift bins. The BIAS shows the most significant improvements because this method of using the Q-Q plots to shift the PDFs specifically targets the BIAS. The RMSE and $\mathrm{FR}_{0.05}$ metrics both show improvements in redshift bins with lower number densities, while improvements are minimal in the highest density bins. This is because the original model was such that it fit the more dense regions better than the less dense ones, resulting in biases on both sides of the central dense region. These clear improvements demonstrate the efficiency of this post-processing method and further improvements are expected if smaller redshift bins are used.

### 2.2.5 Effects of Improved Photometry

In this section the improvement in the photometric redshifts with deeper imaging data is investigated and quantified.

The Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) Data Release 1(Aihara et al. 2018) provides photometry in grizy filters with magnitude errors that are an order of magnitude smaller than the SDSS/UKIDSS LAS photometry. The galaxies used for the previous analyses were cross-matched with the HSC galaxies and the corresponding photometry was combined with the GAMA spectroscopy. After eliminating spec-z's with NQ<3 and removing any objects with missing SDSS/UKIDSS LAS or HSC photometry

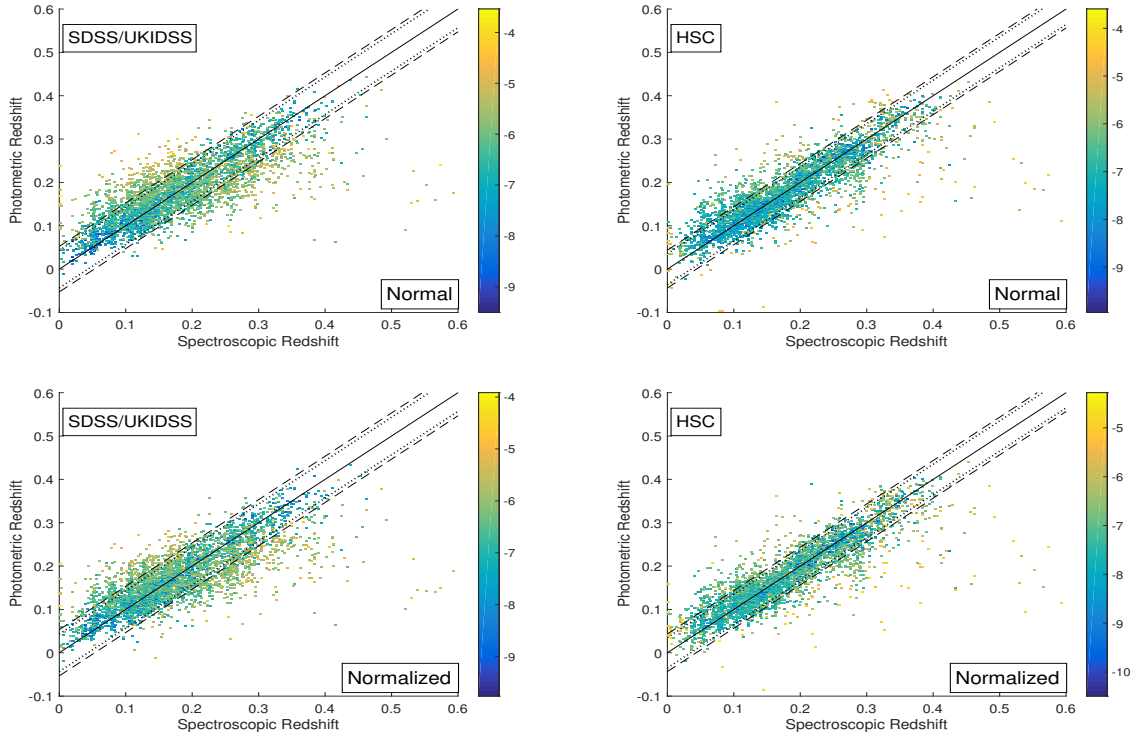



Figure 2.11: Photometric redshift versus spectroscopic redshift plots showing the performance of the GPz code using SDSS/UKIDSS LAS photometry (left) and HSC photometry (right) with the CSL methods normal and normalised (in the first and second rows respectively). The colour scale represents the variance of data points in that area of the plot and the straight line is the $z = \hat{z}$ line.

Table 2.4: Table showing summary performance measures and variance for the HSC photometry with grizy filters and the SDSS/UKIDSS LAS photometry with all three configurations of features. The number of training, validation and testing objects are: 8047, 8048 and 4024 respectively. The best metrics and variances are highlighted.

| CSL Method | Survey | Features | RMSE | BIAS | MLL | $FR_{0.15}$ | $FR_{0.05}$ | Variance | Model Var | Noise Var |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | SDSS/ | grizY | 0.0432 | -0.0013 | 1.70 | 99.18 | 81.25 | 0.0025 | 2.3E-05 | 0.0025 |
| | UKIDSS LAS | ugrizYJHK | 0.0352 | -0.0014 | 1.87 | **99.60** | 88.22 | 0.0018 | 2.1E-05 | 0.0018 |
| | | ugrizYJHK+size | 0.0336 | -0.0012 | 1.92 | 99.53 | 89.64 | 0.0016 | 2.1E-05 | 0.0016 |
| | HSC | grizy | 0.0357 | -0.0008 | **1.94** | 99.33 | 89.14 | 0.0015 | 1.9E-05 | 0.0015 |
| Normalised | SDSS/ | grizY | 0.0432 | 0.0021 | 1.64 | 99.25 | 80.67 | 0.0018 | 2.3E-05 | 0.0018 |
| | UKIDSS LAS | ugrizYJHK | 0.0355 | 0.0010 | 1.83 | 99.58 | 88.17 | 0.0014 | 2.2E-05 | 0.0013 |
| | | ugrizYJHK+size | **0.0334** | **0.0007** | 1.87 | 99.55 | **89.86** | **0.0011** | 1.9E-05 | 0.0011 |
| | HSC | grizY | 0.0349 | 0.0011 | 1.85 | 99.43 | 89.64 | **0.0011** | **1.8E-05** | **0.0010** |

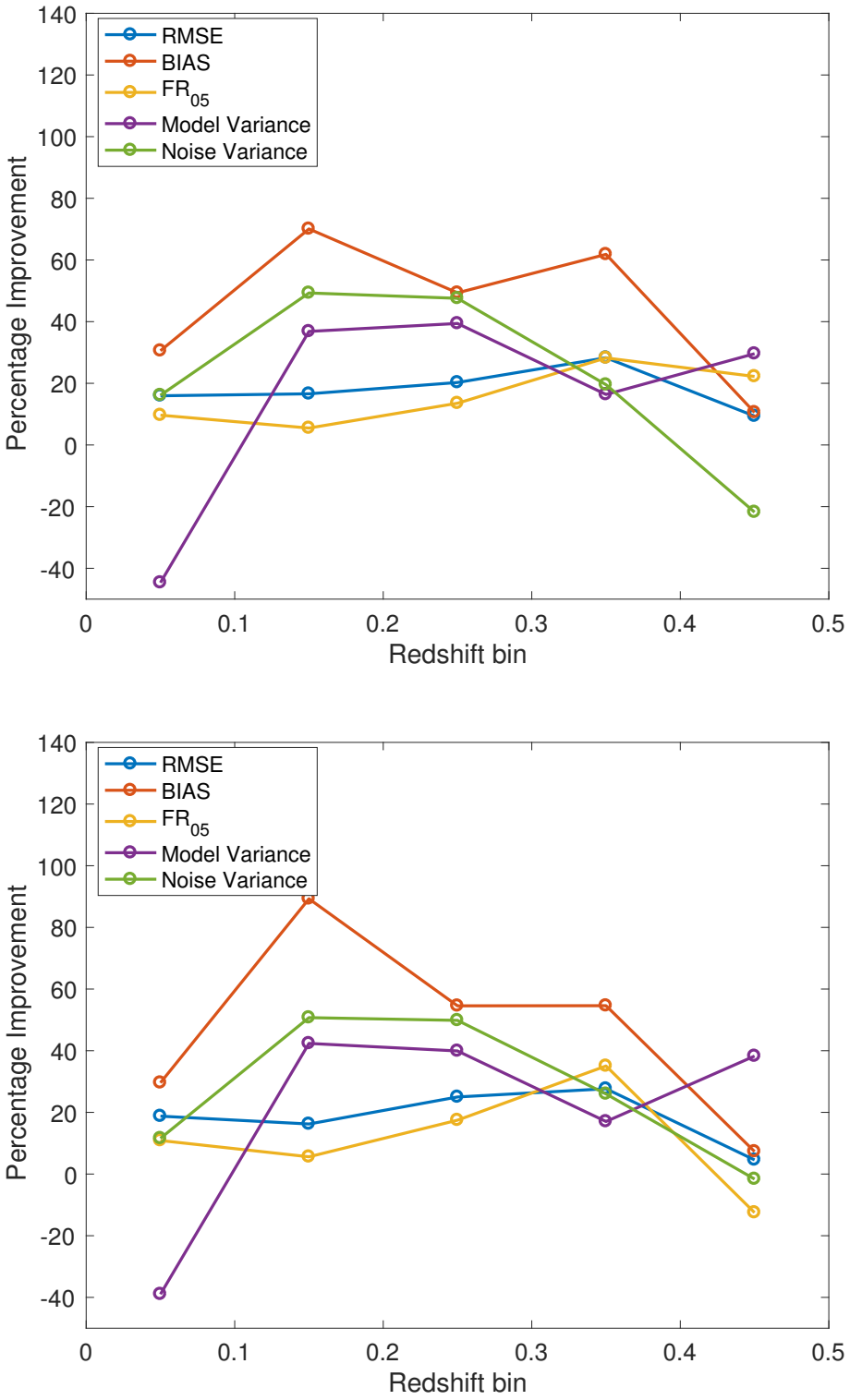



Figure 2.12: Percentage improvements of performance measures and variances by redshift bin due to the use of HSC grizy filters compared to SDSS/UKIDSS LAS grizY filters using the same training, validation and testing objects and using the normal (left) and normalised (right) methods.

only 20253 GAMA objects were matched to HSC objects. This was because only limited portions of the GAMA fields were covered in the HSC-SSP survey (see Aihara et al. 2018).

The GPz algorithm with 100 basis functions and modelling of heteroscedastic noise was then used to estimate the photometric redshifts from both the HSC grizy photometry and the SDSS/UKIDSS LAS grizY photometry for an identical set of galaxies. It should be noted that the Y system response curve from the UKIDSS LAS photometry is of a notably different shape to the y system response curve from the HSC photometry. Photometric redshift versus spectroscopic redshift plots are shown in Figure 2.11, the percentage improvements of the performance measures and variances by redshift bin are shown in Figure 2.12, and metrics are given in Table 2.4. It is evident from these figures that the HSC photometry produces a tighter distribution than the SDSS/UKIDSS LAS photometry, and the metrics show significant improvements when the HSC data is used. In addition to improved metrics, the model and noise variances are also improved. As previously discussed, improved precision of the input data should decrease the noise variance as the spread of the data should decrease. The model variance also improves as the algorithm is much more confident about the model fit with the more precise data.

Next, all available filters as well as size data were added to the estimates obtained using the GPz algorithm from the GAMA SDSS/UKIDSS LAS dataset. The results are summarised in Table 2.4. Although the HSC data clearly outperforms the SDSS/UKIDSS LAS data when only five filters are used, its performance is similar to the SDSS/UKIDSS LAS data when near-IR data are added and it is outperformed in most metrics by a small margin when size data is also added.

Considering the limited size of the data set, the improvements in the metrics provided by the improved quality of the photometry are very significant and thus, further improvements in photometry over large survey regions—as will be provided by future surveys such as the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009) and *Euclid* (Laureijs et al. 2011)—will have significant impacts on the ability of the GPz algorithm to accurately predict the photometric redshifts of galaxies.

## 2.3 Estimating redshift distributions using COSMOS 30-band photometry

Some cosmological analyses such as weak lensing and the measurement of primordial non-Gaussianity require only a redshift distribution ($N(z)$, the number of galaxies with given redshifts over a range of redshift) of the galaxies in the sample instead of requiring the redshift or probability distribution function (PDF) of each galaxy individually. As

previously mentioned this can be obtained by using point estimate algorithms (such as template fitting or machine learning algorithms) and simply binning the galaxies by their expectation values, or binning by Monte Carlo random draws from the PDFs. An alternative method is to stack the PDFs allowing one galaxy to contribute to multiple bins with different weights, then summing the probabilities in each bin. On the other hand, there are methods of determining $N(z)$ without calculating the point estimates. These methods do not use an algorithm to estimate the photo-z values, but instead compare the galaxy sample to a sample with known redshifts (a reference sample) to infer the redshift distribution. This makes these methods less susceptible to algorithmic biases but sensitive to biases in the reference sample. One such method is the cross-correlation method in which galaxies with unknown redshifts are cross-correlated with a reference sample with known redshifts and another is a re-weighting technique that involves the re-weighting of the reference sample in magnitude space such that the magnitude distribution matches that of the photometric sample of interest. Surveys such as the Dark Energy Survey (DES, Hoyle et al. 2018), and the Kilo Degree Survey (KiDS, Hildebrandt et al. 2017a) have used both these methods to validate their redshift distributions. I apply this final method to a clustering analysis of Hyper Suprime Cam (HSC) data being done by the LSST LSS working group. I re-weight COSMOS 30-band photo-zs based on the magnitude space of the HSC sample following the method presented by Lima et al. (2008) which will be expanded upon in this section.

This re-weighting method (sometimes called the weighted direct estimation or DIR approach) was first presented by Lima et al. (2008). They developed a nearest neighbour method of re-weighting a limited sample of spectroscopic (or very accurate photometric) galaxies based on the magnitudes of the larger photometric sample in order to obtain redshift distributions. The principle governing this method is that two galaxy sets with identical distributions of magnitudes/colour will have identical $N(z)$ distributions. Thus, assuming the relevant requirements are satisfied, one can re-weight the spectroscopic/accurate photometric galaxies using the ratios of the densities of the photometric galaxies in magnitude space, resulting in a spectroscopic sample with the same magnitude distribution as the photometric sample. This can then easily be binned and used to produce a redshift distribution that is independent of the photo-zs of the main catalogue. The only requirements of this method are that the spectroscopic sample spans the entire magnitude-space volume of the photo-z catalogue (in order to allow the $N(z)$ to be representative) and that the magnitude space dimensionality is high enough to allow colour/magnitude and redshift to be uniquely matched. Lima et al. (2008) argue that this method can be superior to using the photo-zs as it is independent of the biases that can arise in the photo-z estimation process and they found that this method was able to outperform ANNz photo-zs. This process has been used

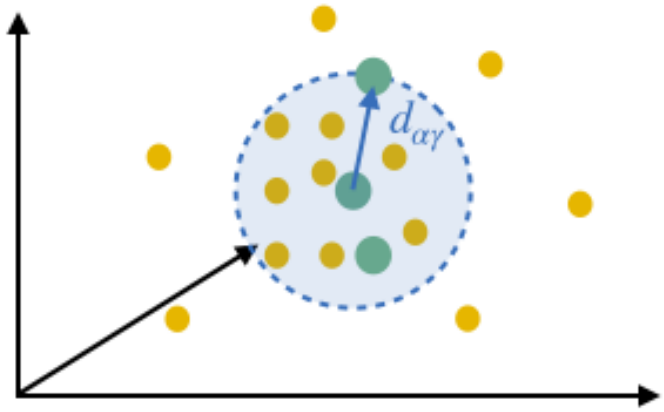



Figure 2.13: Illustration of the method of counting the number of photometric galaxies within a radius around each reference sample galaxy. The yellow points represent the photometric sample galaxies and the larger green points represent the reference sample. The axes represent three dimensional magnitude space, but in reality, these analyses are carried out in higher dimensional space.

in the KiDS (Hildebrandt et al. 2017a) and DES Year 1 (Hoyle et al. 2018) analyses, for the purpose of $N(z)$ calibration.

As part of an LSST Large Scale Structure working group project I was tasked with re-weighting the COSMOS 30-band photometric galaxies based on the magnitudes of the HSC galaxies in order to obtain redshift distributions. This project aims to do a clustering analysis of the HSC data and this re-weighting method will be used as a validation technique, to be compared with the other $N(z)$ measurements obtained using binning of the photo-z estimates and the stacking of the photo-z PDFs. The reason why COSMOS 30-band photometry is used instead of spectroscopic data is because current spectroscopic samples overlapping with the HSC data are not as deep as the HSC data, and therefore a portion of magnitude space would be excluded. While COSMOS photometry is not as accurate as spectroscopy, it is considered to be accurate enough to use for this purpose. The photometric redshifts of photometry with 30 bands is expected to be very accurate as most of the structure of the spectrum is resolved and any important features should be detected. Put differently, it is unlikely for there to be colour-redshift degeneracies. The steps taken in this analysis are listed below:

- The first step was to cross-match the COSMOS galaxies with the HSC photometric galaxies (in the COSMOS field) in order to find the HSC magnitudes associated with the COSMOS redshifts. Quality cuts were first applied to the COSMOS galaxy sample in order to ensure only robust photo-zs and this was performed using the same

quality cuts used by the HSC team when producing a training set (see Tanaka et al. 2018). Next, since the HSC sample in the COSMOS field was much shallower and smaller than the COSMOS sample, each of the HSC galaxies was matched to the closest COSMOS galaxy if it fell within a distance of 1 arcsec.

- Next, for each matched galaxy, $\alpha$, with a COSMOS redshift and HSC photometry, a fixed number, $N_{nei}$, (chosen to be 20) of nearest neighbours (of other matched galaxies) in multidimensional magnitude space was found and the distance to the furthest neighbour, $d_{\alpha\gamma}$, was calculated. $N_{nei}$ was chosen such that it was small enough to reflect the changes in density across magnitude space, while also being large enough to avoid very large shot noise errors.

- The number of HSC neighbours (these can be HSC galaxies in the COSMOS field or another field, because we are looking for neighbours in magnitude space, so physical position is not important) within this radius $d_{\alpha\gamma}$ around galaxy $\alpha$ was then counted and labelled $N_P(\vec{m}_\alpha)$. This step is illustrated in the diagram in Figure 2.13.

- The weight of galaxy $\alpha$ was then computed using the ratio of the numbers of neighbours in the HSC and COSMOS samples.

$$W_\alpha = \frac{1}{N_{p,tot}} \frac{N_p(\vec{m}_\alpha)}{N_T(\vec{m}_\alpha)}$$

  This was then repeated for each galaxy. The nearest neighbour algorithm of the python scikit-learn library was used for these computations.

- These weights were then summed in redshift bins to obtain the re-weighted $N(z)$ estimate. The resulting distribution can be seen in Figure 2.14.

$$N_{P,est}(z_i) = N_{P,tot} \sum_{\alpha=1}^{N_T(z_i)} W_\alpha.$$

- Finally, since this is a tomographic analysis, the $N(z)$ for the galaxies placed into each tomographic bin (based on the best performing point estimate) is required. The HSC galaxy sample was binned into the 4 broad tomographic bins to be used in the clustering analysis based on the best photo-zs estimated by the leading photo-z estimation technique (the Extended Photometric Redshift (Ephor) method using the PSF-matched aperture photometry, see Tanaka et al. 2018). All the COSMOS-HSC matched galaxies in each redshift bin were extracted along with their corresponding weights, which were then summed in smaller bins, allowing the production of Figure 2.15.

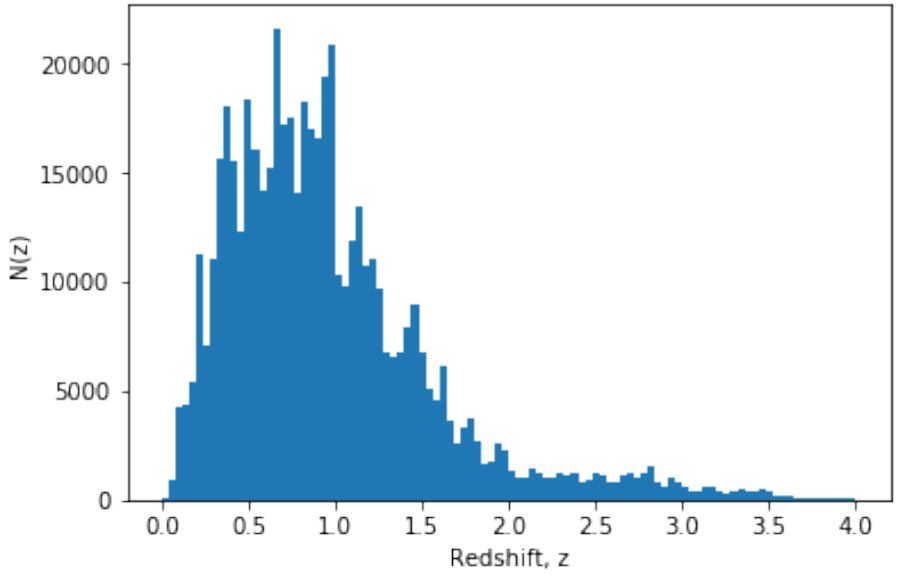



Figure 2.14: $N(z)$ distribution obtained after re-weighting a sample of COSMOS galaxies based on the magnitude distribution of the HSC COSMOS field galaxies.

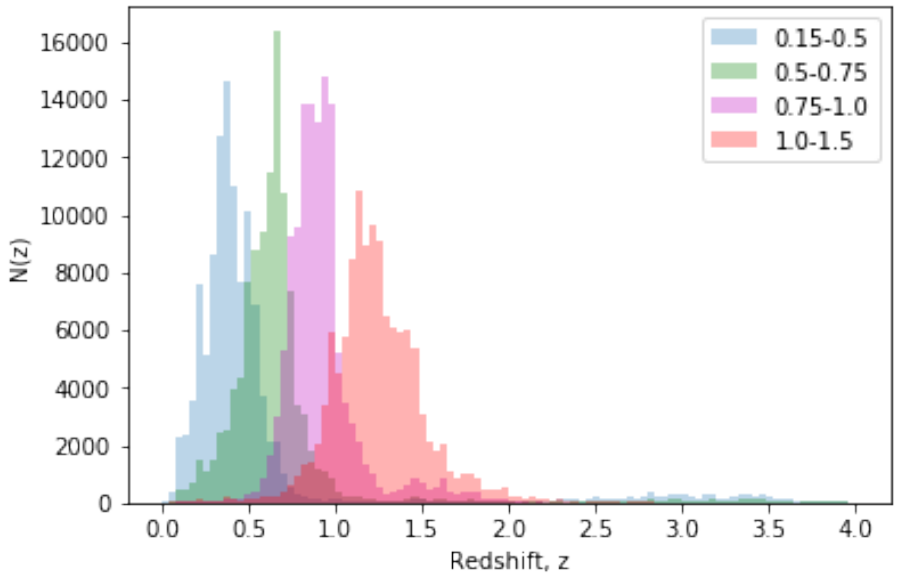



Figure 2.15: $N(z)$ distributions by redshift bin based on the photo-z estimates of the Ephor ab method found by the HSC collaboration Tanaka et al. (2018).

These redshift distributions were found to be very similar to the result of stacking PDFs as well as binning the Monte Carlo estimates of the photo-zs using the Ephor ab estimation technique. The shapes of these distributions (Figure 2.15) show that the binning with the best Ephor ab estimates was relatively correct, with the majority of the galaxies falling within the expected limits in each bin. These $N(z)$ distributions found with this re-weighting technique were used in the later parts of the clustering analysis. One peculiar feature was the small but noticeable bump in galaxy number at redshifts between 3 and 3.5 for the first (0.15 - 0.5) redshift bin (the blue histogram in Figure 2.15). This appears to be a feature in the HSC data and not a problem with the method.

Since this analysis, the HSC collaboration have gone on to conduct a similar analysis on the HSC data, using the COSMOS 30-band photometry in order to get redshift distributions for a cosmic shear analysis of the HSC Year 1 data (Hikage et al. 2019). The technique used in this analysis differed from the one used here in the manner in which galaxies are grouped. In the method used here a nearest neighbour approach is utilized but Hikage et al. (2019) used self organising maps (SOMs) which similarly clusters galaxies with similar magnitudes together.

Despite the usefulness of this method for producing redshift distributions when point estimates are not available or trusted, there are still a number of drawbacks. The application of this method is limited by the depth and overall filled magnitude-space of the reference sample used. If there are gaps in this space then very distant reference galaxies will be up-weighted, leading to biases. In addition, if the reference sample has biases or outliers, it will lead to biases in the estimated photometric redshifts. The reference sample must also overlap with the photometric survey of interest as they need to be cross-matched in order for both samples to have the same magnitude system. Finally, if the photometric system being used is limited, ie. there are very few filters being used, then there may not be unique mappings between magnitude and redshift, i.e. colour-redshift degeneracies will occur.

In addition, if the COSMOS 30-band photometry or any of the spectroscopic redshift samples used as reference samples had notable errors then this could result in inaccuracies in the photo-zs estimated with this method. Laigle et al. (2016) found a catastrophic outlier fraction of 6% at fainter magnitudes with a residual bias of about 2%. This suggests that it is possible that the COSMOS 30-band sample could lead to inaccurate and possibly biased results.

This magnitude-space weighting technique has been used to obtain redshift distributions in two of the largest cosmic shear cosmological surveys: DES and KiDs. The original DES Year 1 analysis (Hoyle et al. 2018), utilized the COSMOS 30-band photometry as a reference sample to shift the results of stacking the pdfs of the photo-z estimates obtained using the

Bayesian photometric redshift code bpz (Benítez 2000). On the other hand, the original KiDS450 analysis (Hildebrandt et al. 2017a) also used bpz, and binned the point estimates but used the redshift distributions obtained using the magnitude re-weighting method with a reference sample consisting of overlapping spectroscopic data instead of using them to shift the stacked $N(z)$.

A combined KiDS and VIKING analysis was performed by Hildebrandt et al. (2020) who compared the effect of COSMOS 30-band data and a combination of spectroscopic reference data on the calibration of the redshift distributions. As the combination of KiDS and VIKING leads to photometry in 9 bands: ugriZYJHK, compared to the ugri of the KiDS-only analysis and the grizY of the DES analysis, this analysis is expected to lead to more accurate photometric redshifts and redshift distributions, with less colour-redshift degeneracies contaminating the result. It was found that the COSMOS data produced a systematic shift in the redshifts to lower redshifts, (leading the $\sigma_8$ measurement nearer to the original DES Year 1 result). A clustering redshift method was also applied, and found to agree with the spectroscopic redshift calibration method. Due to the questionable accuracy of the COSMOS photometric redshifts Hildebrandt et al. (2020) concluded that the spectroscopic data, and the corresponding redshift distributions were more reliable.

In order to conduct a joint analysis of the cosmology from cosmic shear using KiDS and VIKING (VISTA Kilo-Degree Infrared Galaxy Survey) data (KV450) and DES Year1 Joudaki et al. (2019) attempted to standardize the redshift distribution calibration methods of both surveys. Therefore, instead of using the COSMOS 30-band data for the DES data, they used a spectroscopic reference sample similar to that used in the KV450 analysis, containing spectroscopic data from a combination of surveys, with the deepest survey providing reference data for the deepest photometric galaxies. This recalibration resulted in a systematic shift in the resulting redshift distributions of the DES galaxies to higher redshifts (the same effect that was observed by Hildebrandt et al. (2020)). This led to a decrease in the $\sigma_8$ value inferred and therefore a decrease in the tension between the values of $\sigma_8$ obtained by DES and KV450.

The cases above both suggest that the COSMOS 30-band photometric sample induces an incorrect systematic bias on the redshift distributions that are calibrated with it, and this leads to a bias on the cosmological parameters inferred from these high precision, low cosmic variance surveys. This systematic shift is to lower redshifts and is likely due to the misclassification of high redshift galaxies as low redshift ones (the presence of which is indicated by the catastrophic outlier fraction) as well as the overall bias of 2% towards lower redshifts. Therefore, the current COSMOS 30-band photometric sample is not the ideal sample for use in cases in which an alternative sample, such as a spectroscopic sample

is available. For the case of the HSC photometry, there are no sufficiently deep spectroscopic reference samples available and therefore a recalibration analysis using spectroscopic redshifts instead of COSMOS is not yet possible for HSC.

## 2.4 Conclusions

This chapter was split into two main sections. In the first and more substantial part of the chapter methods of obtaining improved photometric redshift estimations from the GPz machine learning algorithm were investigated. Spectroscopy from the Galaxy and Mass Assembly Data Release 2 with a limiting magnitude of $r < 19.4$, along with corresponding Sloan Digital Sky Survey visible ($ugriz$) photometry and UKIRT Infrared Deep Sky Survey Large Area Survey near-IR ($YJHK$) photometry were used. These methods were introducing near-IR magnitudes and angular size features, post-processing the results by shifting the photo-z estimates based on their Q-Q plots and utilising photometry with higher precision. It was found that the inclusion of near-IR (YJHK) filters and angular size data in the training, validation and testing of photometric redshift estimation resulted in significantly improved accuracy, by $\sim 15-20$ per cent and thus, when available, these data should be utilised. The process of shifting the probability distributions of the estimated redshifts by minimising the $\eta_n$ value has proven to substantially improve the BIAS of the estimated photometric redshifts (by $\sim 40$ per cent). Therefore, when a suitable spectroscopic sample is available, this method could be applied to supply additional accuracy to the predictions from GPz and other methods. Finally, it was illustrated that improvements in the accuracy of the photometry improved the accuracy of the photometric redshifts, to a very similar extent as adding the near-IR and angular size data, and therefore, work should continue to be done to improve the quality of the photometric data obtained.

It is worth mentioning that galaxies predominantly at $z < 0.5$ were targeted in this study, where one might expect the size information to have more of an influence. On the other hand, the near-infrared filters are expected to add a comparatively larger amount of information at $z > 1$, where the 4000Å break moves out of the visible wavelength filters and into the near-infrared.

The second part of this chapter illustrated a method of estimating the redshift distributions of the galaxies put into different redshift bins via another photo-z estimation technique. The results of this were used in a clustering project of HSC data conducted by the LSST LSS working group. The method used entailed re-weighting a reference sample of galaxies with well known redshifts (the COSMOS 30-band sample) to represent the magnitude space

distribution of the photometric galaxies in each bin. This was done successfully and the results are illustrated in Figure 2.15.

It is possible to combine the methods of both the GPz point estimation and the magnitude re-weighting technique. Weights for the training galaxies can be determined based on the density of galaxies in the magnitude space distribution of the training dataset versus the target dataset. Via the cost sensitive learning mechanism, these weights will then cause training galaxies that are in magnitude-space regions with limited training galaxies but a more significant presence of the target galaxies to be up-weighted and have a greater impact on the training of the model (down-weighting will take place for training galaxies in the opposite environment). This effectively allows the GPz model to train on a sample that is representative of the target dataset. This should lead to increased accuracy and should reduce overfitting to the training set. This is particularly useful if the training set has a very different magnitude-space distribution to the target set, which is likely the case with spectroscopic versus photometric samples due to the limited depths of spectroscopic surveys. This is currently being implemented and will be discussed in further detail in Chapter 5.

# Chapter 3

# Measuring the BAO using SDSS galaxies and GPz Photometric Redshifts

## 3.1 Introduction

Baryon acoustic oscillations are a useful cosmological probe that can allow us to measure the Hubble parameter, $H(z)$, and the angular diameter distance $D_A(z)$, as functions of redshift. This can lead to measurements of the dark energy equation of state, $w_0$ and $w_a$. The nature of baryon acoustic oscillations (BAO) was discussed in Section 1.1.5.4 of Chapter 1. When accurate spectroscopic redshifts are not available, the BAO peak can be measured using the angular correlation function (or angular power spectrum) in two dimensions and the angular diameter distance can be calculated.

The clustering signal is greatest for luminous red galaxies (described in Chapter 1) due to their high bias and their prominent spectral features such as the 4000 Å break that make photometric redshifts easier to estimate and correspondingly more accurate. As a result, such analyses have been carried out using luminous red galaxy or luminous galaxy samples in the past. Here, I perform a similar analysis, using SDSS luminous galaxies and GPz photo-zs and I also use the entire galaxy set (red and blue galaxies) as GPz has reliable uncertainty estimates that allow the cutting of the sample to exclude galaxies with untrustworthy redshifts. There are also a number of systematics to consider when doing such and these can be addressed by adjusting the mask used or weighting the galaxies. These effects will be discussed and adjustments will be implemented.

The structure of this chapter is as follows: in Section 3.2 I present the steps taken in pre-processing of the data and dealing with systematics. The photometric redshift estimation procedures used and their performance is discussed in Section 3.3. In Section 3.4 I discuss the code used to create the angular correlation function and also provide a discussion on the creation of the random catalog and the measurement of error bars. The technique used

Table 3.1: Table showing the data obtained for each galaxy from the SDSS DR14 database.

| ID | Position | Optical Photometry | Photometric Uncertainties | WISE Infrared Photometry | WISE Photometric Uncertainties | Shape Measure-ments | Image Quality and Accuracy |
|---|---|---|---|---|---|---|---|
| objID | ra | $dered_u$ | $modelMagErr_u$ | $w1_{mag}$ | $w1_{magerr}$ | $petroRad_i$ | $sky_i$ |
| | dec | $dered_g$ | $modelMagErr_g$ | $w2_{mag}$ | $w2_{magerr}$ | $expAB_i$ | $extinction_i$ |
| | | $dered_r$ | $modelMagErr_r$ | | | | $psffwhm_r$ |
| | | $dered_i$ | $modelMagErr_i$ | | | | |
| | | $dered_z$ | $modelMagErr_z$ | | | | |

for locating the BAO peak in the angular correlation function is presented in Section 3.5. Finally, in Section 3.6 I present the angular correlation functions for the entire galaxy set and the luminous galaxy set and the corresponding BAO peak locations. This is followed by a conclusion in Section 3.7.

## 3.2 Data Cleaning and Systematics

### 3.2.1 SDSS Data

The data used for this analysis was obtained from the SDSS DR14 database (Abolfathi et al. 2018) using the Catalog Archive Server (CAS) jobs interface. A query was performed on the *Galaxy* view joined with the *WISEForcedTarget* table. The *Galaxy* view contains resolved primary survey objects that are classified as galaxies and the *WISEForcedTarget* table provides forced photometry data from the Wide-Field Infrared Survey Explorer (WISE; Wright et al., 2010). Table 3.1 shows the data that was extracted for each galaxy in the *Galaxy* view to obtain the sample of all the galaxies. For the luminous galaxy (LG from here on) sample, the same data was extracted but only for those galaxies that satisfied the following constraints:

- 17.5 < $cModelMag_i$ - $extinction_i$ < 19.9
- $dered_r$ - $dered_i$ < 2
- $dered_r$ - $dered_i$ - ($dered_g$-$dered_r$)/8.0 > 0.55
- $fiber2Mag_i$ - $extinction_i$ < 21.7
- $cModelMag_i$ - $extinction_i$ < 19.86 +1.6*(($dered_r$ - $dered_i$ - ($dered_g$-$dered_r$)/8.0) - 0.8)
- ((($psfMag_z$-$extinction_z$) - $dered_z$) > 9.125 - (0.46*$dered_z$) AND (($psfMag_i$ - $extinction_i$) - $dered_i$) > 0.2 + 0.2*(20-$dered_i$)) OR ((($cmodelMag_r$ - $extinction_r$) < 13.6 + (0.7*($dered_g$

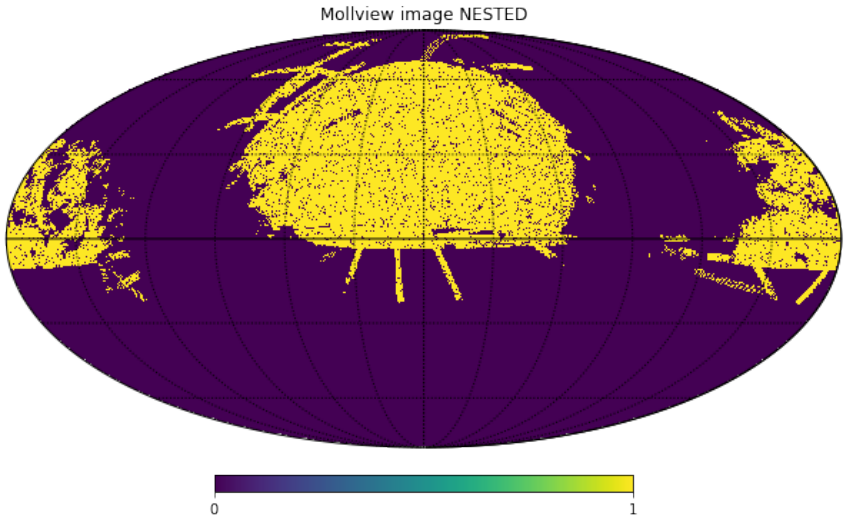



Figure 3.1: SDSS imaging mask. The yellow regions represent areas with imaging data present.

- $\text{dered}_r$) + 1.2*($\text{dered}_r$ - $\text{dered}_i$ - 0.18))/0.3) AND (ABS($\text{dered}_r$ - $\text{dered}_i$ - ($\text{dered}_g$ - $\text{dered}_r$)/4.0 - 0.18) < 0.2) AND (16<($\text{cmodelMag}_r$ - $\text{extinction}_r$)<19.6)))

These constraints used for extracting LGs are the same ones used by Ross et al. (2011). The photometry, their errors and the size measurements are used for estimating photometric redshifts. The image quality and accuracy data are used for making cuts to remove systematic errors. The magnitudes used for these measurements are the inverse hyperbolic sine or asinh magnitudes used by SDSS (refer to Chapter 1). The asinh function is such that for brighter objects the magnitude follows a scale similar to a log scale but for fainter objects it is approximately linear, and objects that are not detected are assigned a faint value. This can lead to incorrect photo-z estimates, but if the uncertainties in these magnitude measurements are large then they will not have significant effects on the photo-z results.

### 3.2.2 Masking

Masks based on imaging area overlap are necessary to ensure that the galaxies and randoms cover the same sky area (as the randoms will be assigned random positions within the mask). This masking of the galaxy samples was done using the mask used by Ross et al. (2011) and Ho et al. (2012). This mask is a HEALPix mask (Górski et al. 2005) with $N_{side}$ = 1024, meaning that the sky is broken into 12,582,912 pixels of equal area. Each pixel is assigned a weight based on the extent of its overlap with the imaging footprint: a weight of 1 means that the entire pixel is within the imaging footprint, a weight of 0 means that the entire pixel is outside of this footprint and a weight that is between 0 and 1 means that part

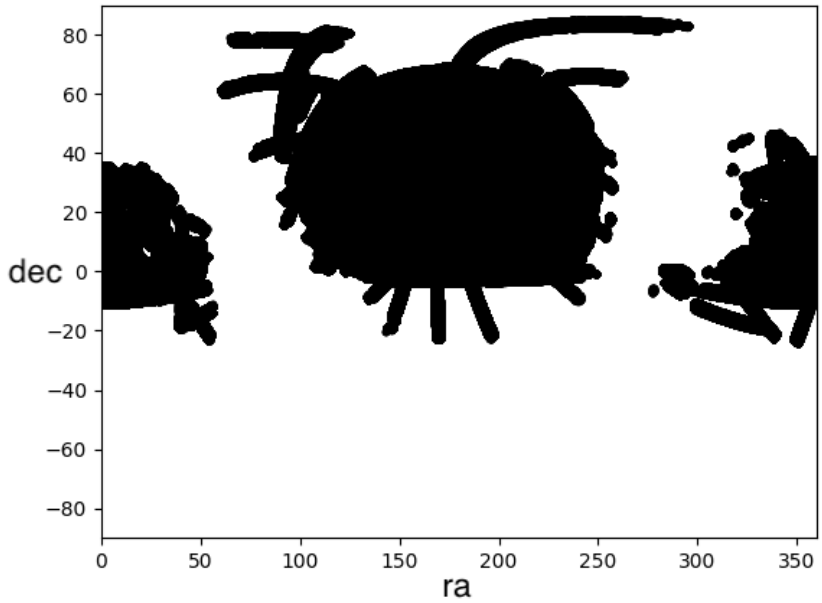



Figure 3.2: A sky map of the positions of the randoms created using the SDSS imaging mask.

of the pixel is within the imaging area. A visualization of this mask is provided in figure 3.1. Figure 3.2 is a plot of the positions of the random catalog (needed for calculating the angular correlation function, as detailed in Chapter 1) created using this mask and we see that it follows the same pattern as the mask.

#### 3.2.2.1 Seeing and Extinction Cuts

This imaging mask was also made in such a way that it excludes regions of the sky in which atmospheric seeing and extinction are very large. Seeing is the distortion in the light detected due to the atmosphere, this results in blurring and leads to a spreading of the light received. The effect of atmospheric seeing can therefore be measured using the point spread function full width half maximum measurement. The point spread function (PSF) is a two-dimensional function of flux received versus position. It therefore describes the (two-dimensional) spread of the light received from a point source, and is due to both the telescope optics (diffraction of the aperture) and the atmospheric seeing. In SDSS, the seeing is measured using the full width half maximum (FWHM) of the one-dimensional Gaussian that has the same noise effective area as the two-dimensional Gaussian fit to the

measured PSF. The one-dimensional FWHM is given by:

$$PSFFWHM = \sqrt{\frac{8\ln 2}{4\pi}}\sqrt{n_{eff}}, \tag{3.2.2.1}$$

where $n_{eff}$, the effective number of pixels is defined as the inverse of the sum of squares of the normalised two-dimensional Gaussian fit to the PSF, over all pixels. The PSF differs slightly for each photometric band-pass, therefore there are measurements of this PSF FWHM for each band. For this analysis the r-band measurement (psffwhm$_r$) is used and HEALPix pixels with average psffwhm$_r > 2''$ were excluded from the mask.

Extinction is the decrease in flux received due to the presence of galactic dust in the line of sight between the telescope and the source and was discussed in Chapter 1. SDSS provide galactic extinction measurements in the various band-passes and these can be converted into galactic dust reddening estimates from the Schlegel et al. (1998) dust maps by multiplying by a constant (these constants can be found at https://irsa.ipac.caltech.edu/applications/DUST/). Although the effects of extinction can be mitigated by subtracting the reddening in each band from the magnitude measured (i.e. using the dered magnitudes) in regions of significant reddening some objects will not be detected, introducing a bias in any clustering measurements made. The mask used in this analysis is such that HEALPix pixels with E(B-V)$> 0.08$ were excluded.

### 3.2.3 Colour Space Distributions and Incompleteness Cuts

Before proceeding with the training of GPz in order to estimate the photo-zs, I evaluated the colour space distribution of the training set and compared it to that of the target set. This has two purposes, the distribution of points on colour-colour plots will provide some insight into whether the data is reasonable as any unexpected features would suggest errors in the dataset. In addition, comparing the distributions of the training and target sets can provide some indication of how well a model developed using the training set would be able to make photo-z predictions for data in the target set.

In Figure 3.3 I show various colour-colour plots of the SDSS training galaxies. They show a number of strange lines that do not appear to be natural and are instead due to issues in the catalogue.

As the volume of space that we observe increases as we look deeper into space, toward higher redshifts (and fainter magnitudes), the number of galaxies present also increases. When the numbers start to decrease with increasing (fainter) magnitude this means that all the galaxies at these magnitudes are not being detected and the galaxy sample is incomplete. Figure 3.4 shows the histogram of galaxies detected with different i-band magnitudes. We

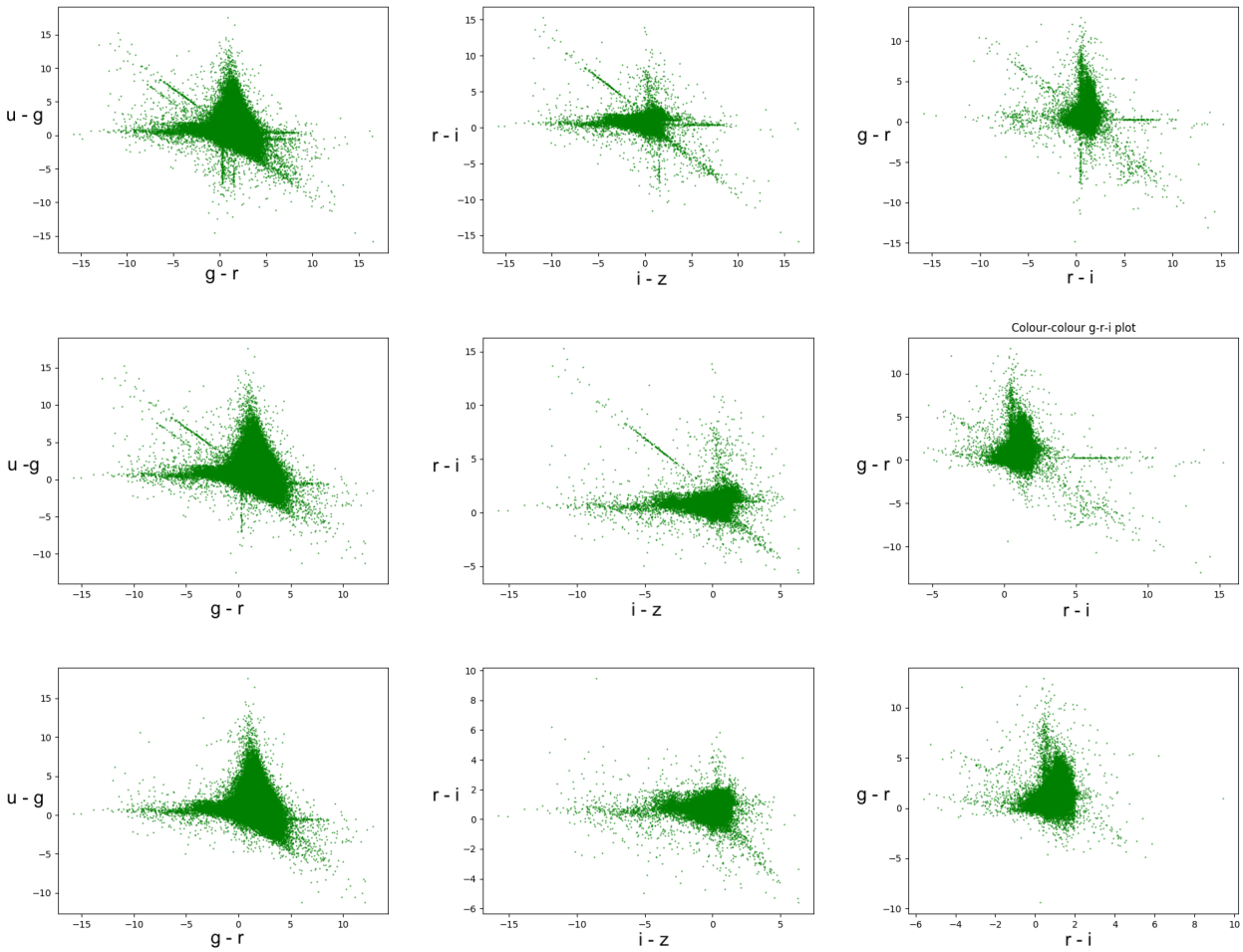



Figure 3.3: Colour-Colour plots of the training data. The top row uses the training data before any magnitude cuts, the second row uses the training data with the i-band magnitude cut and the last row uses the training data with both the i-band and r-band magnitude cuts. The three columns are plots of u-g versus g-r, r-i versus i-z and g-r versus r-i respectively. The magnitude ranges are very large because of the nature of the asinh magnitudes used.

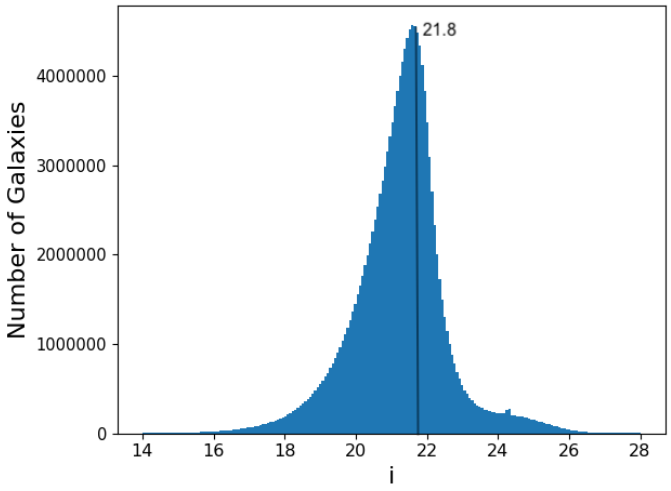



Figure 3.4: Histogram of i-band magnitudes for the entire galaxy set. The black line represents the cut made at the turnover point.

see that the numbers increase steadily until a turnover at a particular point (found to be about 21.8 mag). This is where the catalogue begins to become incomplete, as there are always more faint galaxies than bright. Thus, if we exclude the galaxies beyond this turnover we should have a complete galaxy sample that includes all the galaxies over the whole survey region used at the magnitudes included. A similar histogram was produced for the r-band magnitudes and the corresponding turnover was found to be 22.2 mag. The sample was then adjusted such that all galaxies with i-band magnitude fainter than 21.8 mag or r-band magnitude fainter than 22.2 mag were removed and the updated colour-colour plots are shown in the second and third rows of Figure 3.3 respectively. It is clear that these cuts remove the strange linear features appearing in the plots.

The corresponding colour-colour plots for the large galaxy sample target galaxies are given in Figure 3.5. It is clear that the target galaxies span a much larger region of colour space than the training galaxies. This suggests that photometric redshift estimation might be challenging as there are areas of colour space included in the target set that are not present in the training set to provide a relationship between colour and redshift. I will return to this in Section 3.3.

### 3.2.4 Area Around Bright Stars

The presence of bright stars within the survey mask can lead to systematic effects on the light received from galaxies. If a bright star is near to the galaxy being observed, the light from the star will contribute to the light from the galaxy. In addition, if the star light is much brighter than the galaxy and depending on the positioning of the galaxy and star the galaxy

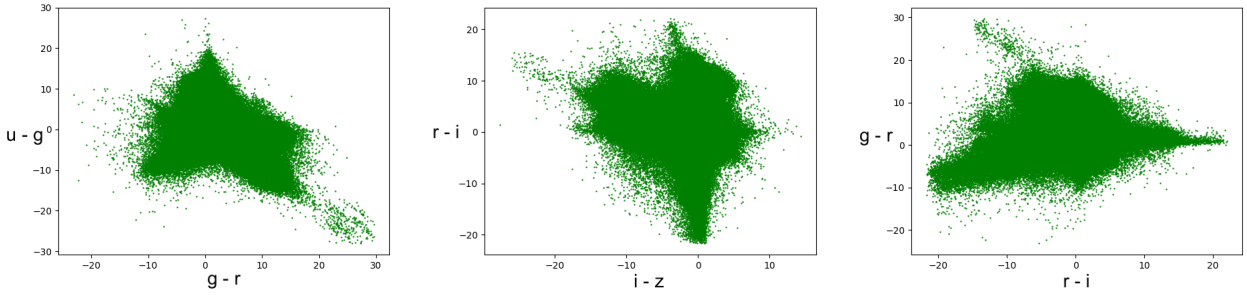



Figure 3.5: Colour-Colour plots of the target data. The three plots, left to right are of u-g versus g-r, r-i versus i-z and g-r versus r-i respectively. The magnitude ranges are very large because of the nature of the asinh magnitudes used.

might not be detected at all. To address this problem one can simply remove all galaxies that are within a fixed radius of each of the bright stars in the mask. In this analysis I located the bright stars by extracting from the Star view (objects identified as stars) of SDSS DR14 those with magnitudes: $i_{mod} < 19.9$, this returned a large sample of 93646864 stars. Galaxies within 9.48 arcsec around each of these stars were then removed. This radius was found to be optimal by Ross et al. (2011). This process was done using the Starlink Tables Infrastructure Library Tool Set (STILTS) tskymatch2 tool which matched galaxies to the bright stars with an error of 9.48 arcsec and returned all galaxies that were not matched. In order to account for these artificially removed galaxies, randoms within these regions were also removed in the same way.

### 3.2.5 Sky Background Weights

The sky background can influence the number of galaxies observed at different positions of the sky. Again following Ross et al. (2011), one way of removing this effect is by calculating the average number density of galaxies observed in regions with different sky backgrounds and weighting galaxies in these regions based on this density. Specifically, the ratio of the average number density of galaxies in bins of sky background to the average number density of the total area is calculated and the inverse of this is used as the weight.

The distributions of sky background for the LGs and all galaxy types are shown in Figure 3.6 (after some extreme outliers were removed). Galaxies with the largest sky background (> 3 standard deviations above the mean, 16.66 for the sample of all galaxy types) were further removed. The remaining galaxies were then binned into 30 sky background bins and the ratio discussed above was determined for each bin. These results are plotted in Figure 3.7. The inverse of this value was found for each bin and this was multiplied by the mask weight for each galaxy in the corresponding bin.

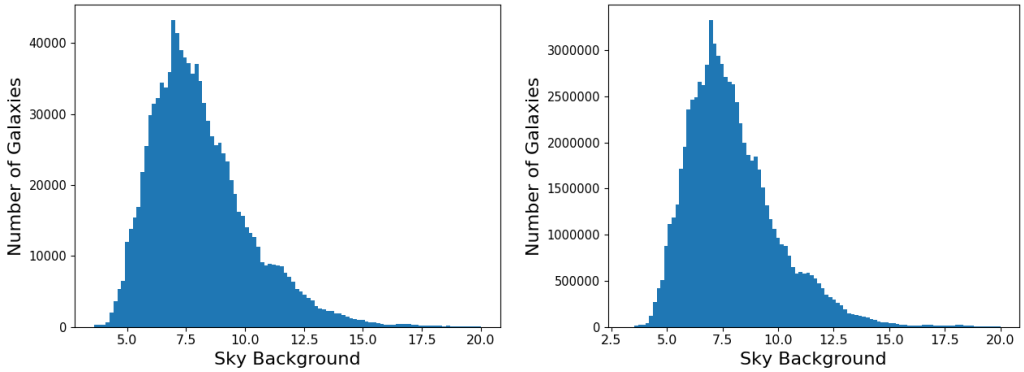



Figure 3.6: Histogram of the sky background in the i band for the LGs (left) and all galaxy types (right). The units of sky background are nanomaggies/arcsec$^2$.

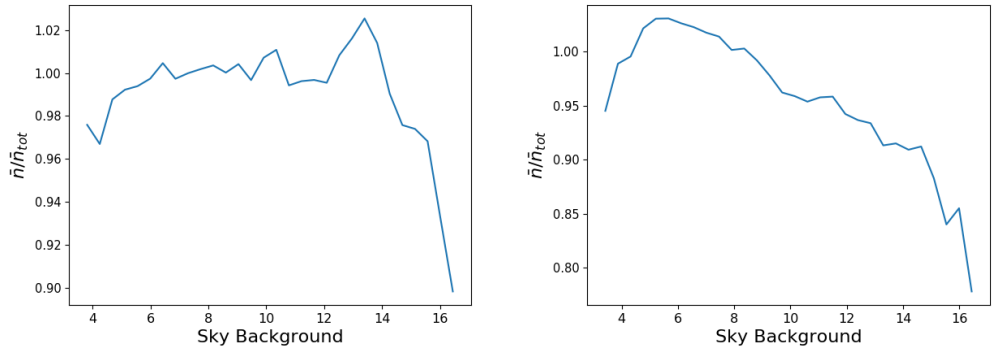



Figure 3.7: Plot of the ratio of the average galaxy number density in regions with different sky backgrounds and the average number density over the entire area versus sky background for the LGs (left) and all galaxy types (right). The units of sky background are nanomaggies/arcsec$^2$.

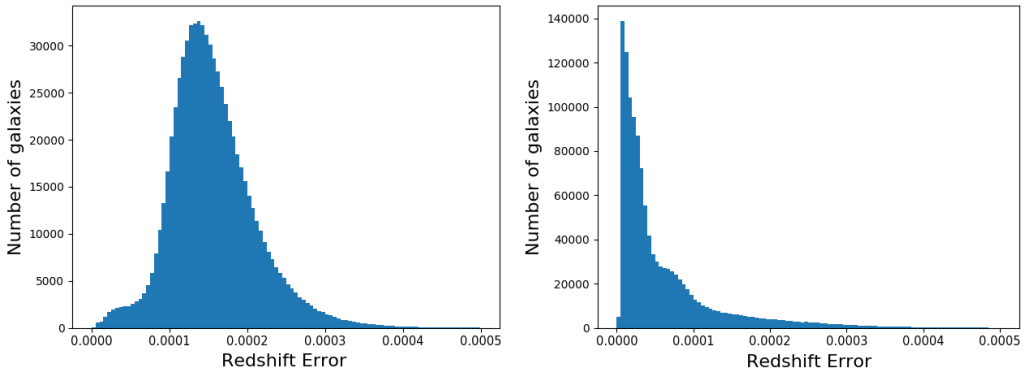



Figure 3.8: Histogram of the spectroscopic redshift errors of the training galaxies for the LG sample (left) and the non-LG sample (right).

## 3.3 Photo-z Estimation

The GPz algorithm was used to estimate the photo-zs of the galaxies. The training set was obtained by selecting the columns in Table 3.1 along with the spectroscopic redshifts and measures of the redshift quality from the SDSS DR14 Galaxy table joined with the WISEForcedTarget, SpecPhoto and SpecObj tables. Only galaxies with spectra with no warnings (zWarning = 0) were selected. The testing set was adjusted such that the redshift assigned to BOSS LGs is the $z_{noqso}$ value instead of the $z$ value as this is expected to be the better value. In addition to this, only galaxies with a spectroscopic redshift error below a threshold, chosen to be $0.0005$ were used in the training set in order to prevent redshifts that might be erroneous from entering the training. Histograms of the spectroscopic redshift errors for the LGs and galaxies that are not LGs (non-LGs from here on) are presented in Figure 3.8. These show that the errors for the LGs peak at higher values than those for the non-LGs and this is likely due to the fact that most of the LGs will not have emission lines, making it more difficult to measure the redshift. Despite this, in both cases the vast majority of the galaxies have spectroscopic redshift errors below the $0.0005$ cut applied.

The training task was separated into LGs and non-LGs as the LG galaxies have different spectra and colour distributions to the other galaxies. In addition, some galaxies had complete WISE data while others (a significant number) had some missing values (galaxies with any other relevant data that was missing were discarded). In order to not artificially cut the sample by removing all galaxies with missing WISE data, I further split the training and target samples into two sets: those with all WISE data and those with missing WISE data. The set with missing WISE data was trained without any WISE data. In summary, both the training and the target data were split into 4 sets:

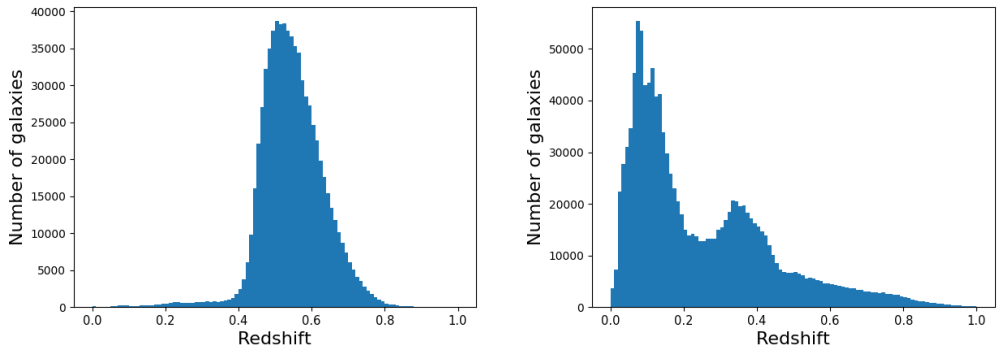



Figure 3.9: Spectroscopic redshift distributions of the training galaxies for the LG sample (left) and the non-LG sample (right).

- LG and all WISE data
- LG and missing WISE data
- non-LG and all WISE data
- non-LG and missing WISE data

For the training data, the two missing WISE data sets were replaced with the entire LG and non-LG sets without the WISE photometry included, as having more galaxies for training helps to optimise the GPz algorithm when estimating photo-zs.

The training data used were the optical filters and their uncertainties, the WISE filters and their uncertainties (where present) and the shape measurement data (see Table 3.1). The spectroscopic redshift distributions of the training data are given in Figure 3.9. It is notable that the vast majority of the LG training galaxies are in the range $0.4 < z < 0.7$ while the other galaxies have a distribution that spans a range of $0 < z \lesssim 0.9$ with a peak at $z = 0.1$. This suggests that the results of the non-LG training will be better at lower redshifts and might be biased at higher redshifts where data density is low.

Following this, GPz models were trained on all of the training samples using a train: validate: test ratio of 2:2:1. The resulting photometric versus spectroscopic redshift plots for each of the four training cases are given in Figures 3.10 and 3.11 and the RMSE performance metrics are provided in Table 3.2. The training that included the WISE data led to better performance on the testing set for both the LGs and the non-LGs. This is expected as the WISE data adds measurements from a different part of the spectrum and should therefore improve the fitting procedure (refer to Chapter 2). It should also be noted that the majority of the galaxies have WISE data, and therefore this training performance

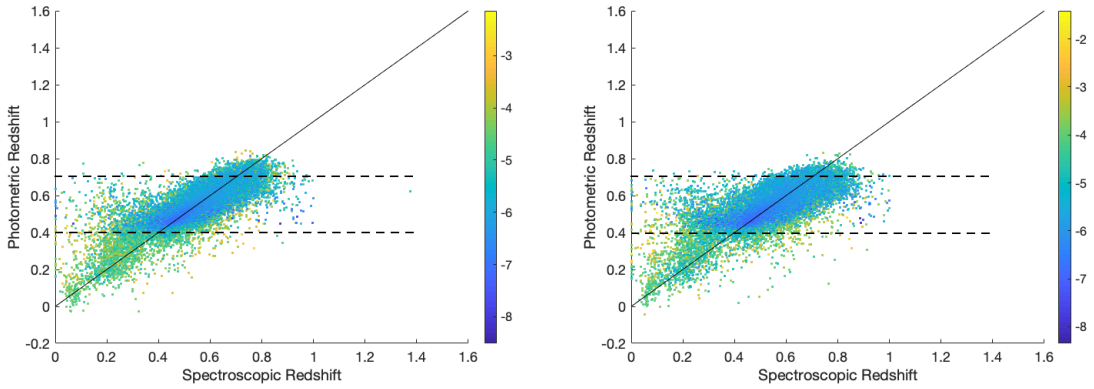



Figure 3.10: Photo-z versus spectroscopic redshift plots for LG testing galaxies. The plot on the left was obtained using the WISE data ($w1_{mag}$ and $w2_{mag}$) in the training while this data was excluded when creating the plot on the right. The straight lines enclose the region of interest for this clustering analysis ($0.4 < z < 0.7$).

is the more relevant one. Table 3.2 shows that the performance on the full testing sets with the non-LGs is better than for those that are LGs, this is predominently because the LG training, validation and testing sets had a relatively small number of galaxies outside of the range $0.4 < z < 0.7$, therefore training outside of these regions was poor. The plots in Figure 3.10 show a much smaller spread of galaxies within the $0.4 < z < 0.7$ compared to outside of it. It is also observed that many lower redshift galaxies were assigned higher redshifts while higher redshift galaxies were assigned lower redshifts, this was due to the higher density of training data in the central region and is an effect that was also observed in Chapter 2. The photometric versus spectroscopic redshift plots show that for the redshift range of interest ($0.4 < z < 0.7$) the estimates are relatively good, with no clear biases such as in the lower and higher redshift regions. Table 3.3 provides the RMSE performance metrics for those galaxies assigned photometric redshifts in the range $0.4 < \bar{z} < 0.7$, i.e. the galaxies that will be included in this analysis. In this case it is clear that the photometric redshifts for the LGs are significantly more accurate than those for the non-LGs and these metrics provide a better indication of the accuracy of the redshift estimates relevant for this analysis. In addition, it is clear that many outlying points have higher uncertainty than those near the central line (green or yellow points compared to the deeper blue nearer the centre). This suggests that galaxies with higher uncertainty measurements are more likely to be galaxies with poor estimates. This was discussed in Chapter 2. On the other hand, there are still a number of significantly outlying points that are deep blue, this could be because these galaxies are in fact quasars or AGN or because they have noisy spectra leading to an incorrectly assigned spectroscopic redshift (see Chapter 1).

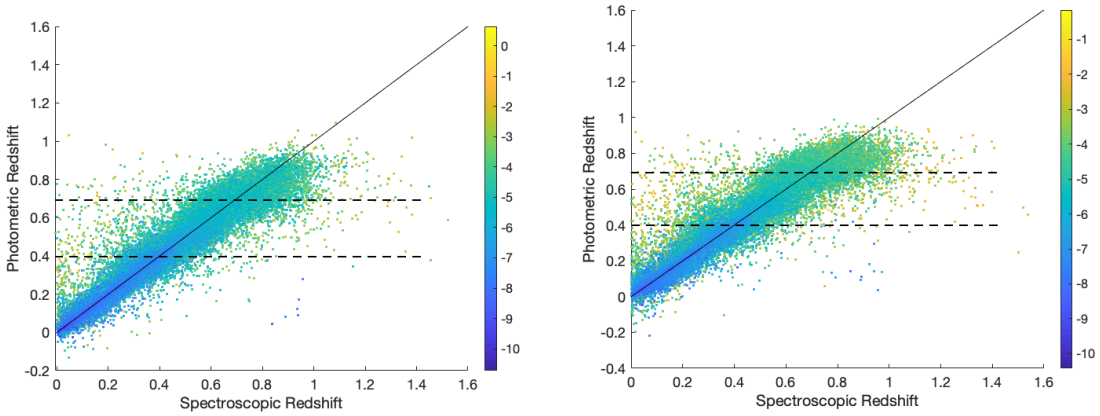



Figure 3.11: Same as above but for the non-LG testing galaxies.

Table 3.2: Table showing the RMSE performance metric for each of the four training cases.

| | LG | non-LG |
|---|---|---|
| WISE data | 0.046 | 0.037 |
| No WISE data | 0.052 | 0.043 |

Table 3.3: Table showing the RMSE performance metric for each of the four training cases within the relevant bounds of $0.4 < \bar{z} < 0.7$.

| | LG | non-LG |
|---|---|---|
| WISE data | 0.044 | 0.064 |
| No WISE data | 0.050 | 0.076 |

Table 3.4: Table showing the RMSE performance metric for each of the four training cases within the relevant bounds of $0.4 < \bar{z} < 0.7$ and with the cut uncertainty$< 0.05$.

| | LG | non-LG |
|---|---|---|
| WISE data | 0.044 | 0.059 |
| No WISE data | 0.050 | 0.068 |

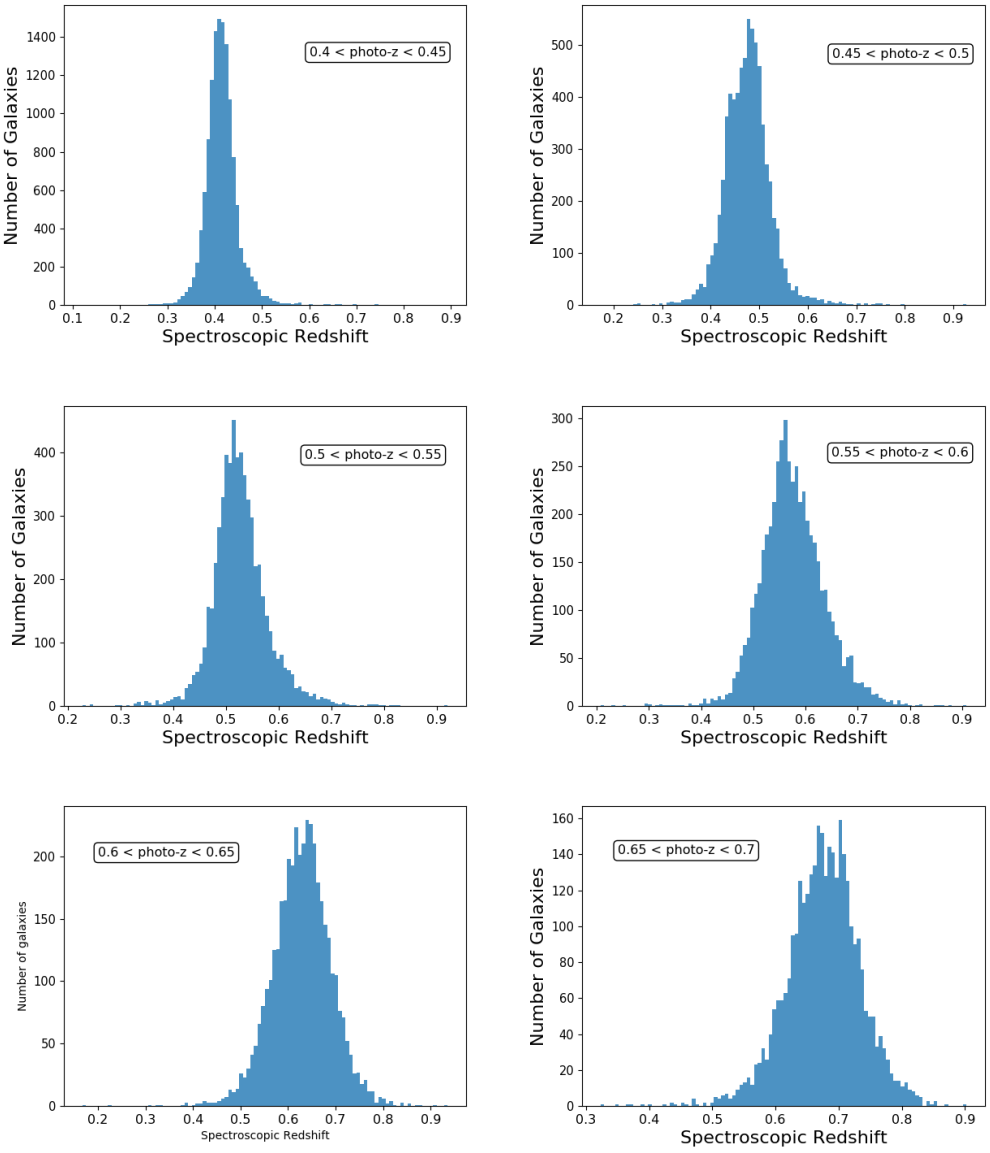



Figure 3.12: Spectroscopic redshift distributions of galaxies binned into each of the $\Delta z = 0.5$ bins based on their photometric redshift estimates.

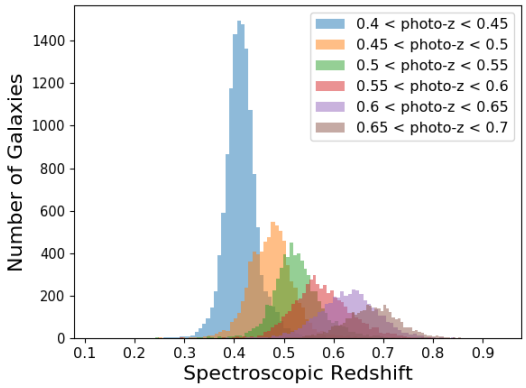



Figure 3.13: Spectroscopic redshift distributions of galaxies binned into bins of width $\Delta z = 0.5$ based on their photometric redshift estimates, overlaid on the same axes

In order to visualise the effects that the inaccuracies in the predictions have on the redshift distributions of galaxies binned by photo-z we plot the spectroscopic redshift distributions for a series of photometric bins using the testing data (for the non-LG sample). This is presented in Figures 3.12 and 3.13. These resulting distributions are approximately Gaussian and centred at the correct spectroscopic redshift. We note that the analogous distributions for a target sample is likely to have increased spread if the target sample has a different distribution in colour/magnitude space to the training sample (the test set gives the best possible performance as it is drawn from the same spectroscopic sample as the training and validation sets).

Next, the galaxies with large photometric redshift errors are removed from the LG and non-LG samples. Figure 3.14 shows the distribution of uncertainty values for the two target samples. The LG sample has much smaller uncertainties than the sample of other galaxies so a reasonable cut-off was chosen based on the uncertainties for the non-LG sample. This cut was chosen so as to retain the majority of the sample while still removing a significant portion of high-uncertainty estimates. Thus, a value of $0.05$ was chosen as the cut off and applied to both samples. The RMSE metric was recalculated for the testing set with this cut applied and in the range $0.4 < \bar{z} < 0.7$. This is provided in Table 3.4. This cut clearly does not make a difference to the LGs but does reduce RMSE for the non-LG galaxies as expected.

The photo-z estimation performance on the LG sample can be compared to that obtained in other analyses. Ross et al. (2011) used the ANNz algorithm to estimate photo-zs for a LG sample with the same selection cuts as those used in this paper and found an RMSE value of 0.0585 which is significantly higher than the values of 0.046 (with WISE) and 0.052 (without WISE) found in this analysis.

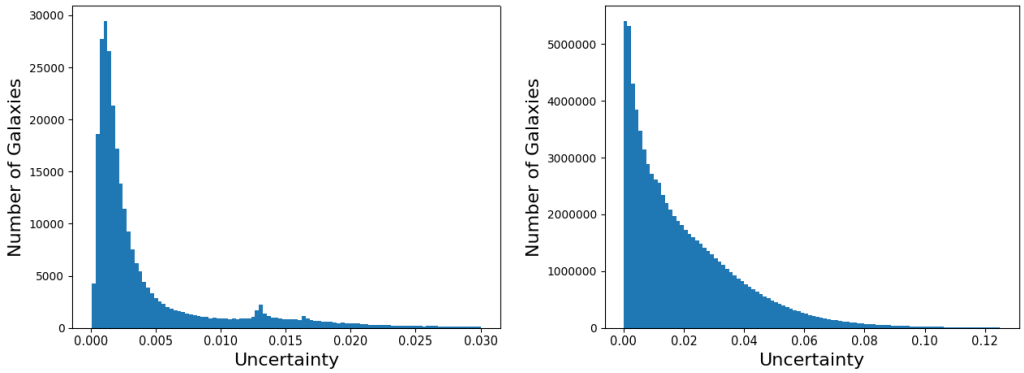



Figure 3.14: Histogram of photometric redshift uncertainty of the target data for the sample of LGs (left) and non-LGs (right). The cut off used for this analysis is an uncertainty of 0.05.

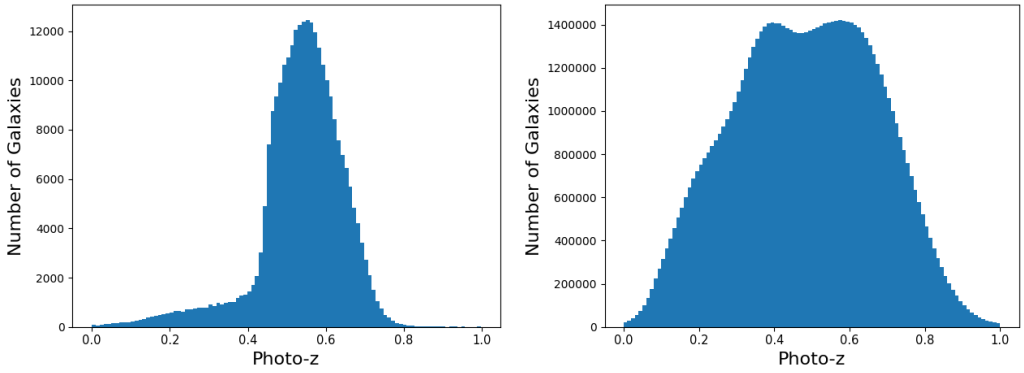



Figure 3.15: Photometric redshift histogram obtained after applying the trained GPz model to the LG target sample (left) and the non-LG sample (right).

The resulting photometric redshift distributions obtained after applying the trained GPz models to the target sets are provided in Figure 3.15.

## 3.4 Angular Correlation Functions

### 3.4.1 CUTE code for finding correlation functions

Correlation Utilities and Two-point Estimates (CUTE; Alonso, 2012) is a correlation function code that allows the calculation of various types of auto and cross-correlation functions. These include: the angular correlation function, the radial correlation function, the monopole correlation function, the 3-dimensional correlation function and the 'full' correlation function. CUTE utilizes the Landy-Szalay estimator to do these calculations. This

estimator was introduced in Chapter 1 but is also provided below:

$$\xi(s) = \frac{DD(s) - 2DR(s) + RR(s)}{RR(s)}, \tag{3.4.1.1}$$

where DD(s), RR(s) and DR(s) are the normalized numbers of galaxy-galaxy, random-random and galaxy-random pairs with a comoving (or angular) separation $s$. The creation of the random catalog is discussed in Section 3.4.2.

Correlation is a very time consuming process, the brute force approach requires correlation of every pair of galaxies/randoms, this is an O($n^2$) algorithm. In order to increase efficiency, CUTE uses a nearest neighbour searching algorithm to avoid unnecessarily correlating many objects that are separated by distances/angles that are larger than the scales of interest. This works by splitting the occupied space into cells and only correlating the objects in each cell with objects in a fixed number of cells surrounding the first cell such that the maximum distance/angle is included in these surrounding cells. In addition to this, for the angular and radial correlation functions CUTE provides the option of using pixels instead of individual galaxies to calculate the correlation function. CUTE also allows the straightforward inclusion of weights to this correlation calculation. The weights for each galaxy and random must be provided as the fourth column of the input files (the first three containing the ra, dec, and redshift columns).

For this analysis I utilize CUTE to find the angular correlation (for the LG and combined datasets) using logarithmic binning and a scale of up to 20 degrees. I use the pixelization option with 2048 pixels and I include the weights discussed in Section 3.2. For the entire galaxy sample, the use of the pixelization option reduced the computing time from > 2 days to $\sim$3 hours.

### 3.4.2 Random files

As discussed in Chapter 1, in order to measure the angular correlation function, a catalogue of randoms must be produced. These randoms must be a spatially random distribution of 'galaxies' with the same sky coverage and redshift distribution as the galaxy sample and are necessary for comparison with the galaxy sample when computing the correlation function (which measures the excess probability of two galaxies being separated by a given distance/angle compared to if they were randomly distributed). Randoms are produced by randomly generating ra and dec positions on a sphere within the relevant survey mask and then matching them to a randomly chosen redshift from the masked galaxy sample. This automatically replicates the dN/dz of the galaxies in the randoms. Random samples of approximately fifty times the size of the galaxy samples were used for this analysis. Such large numbers should sufficiently reduce the effect of the shot noise in the random sample.

### 3.4.3 Error Bars

The simplest form of error bars that could be used for correlation function measurements is Poisson errors in the galaxy galaxy pair counts. The randoms can be ignored if the number of randoms is much larger than the number of galaxies. These Poisson errors are defined as:

$$\Delta\omega = \frac{1+\omega(\theta)}{\sqrt{DD}}, \tag{3.4.3.1}$$

where $\omega$ is the angular correlation function and $DD$ represents the number of galaxy-galaxy pairs with separation $\theta$. One drawback of these error bars is that they ignore any correlation between DD bins, which is large when the number of galaxies is large, this leads to underestimated errors.

Alternatively, significantly more informative error bars can be obtained by using either bootstrap re-sampling or jack-knife re-sampling methods. Bootstrap re-sampling is the method of sampling galaxies (or sub-volumes of the survey volume) with replacement followed by calculating the correlation function for each of these samples (e.g. Ling et al. 1986). The covariance function (which includes the variances for each bin) can then be calculated using the equation below:

$$C_{boot}(x_i, x_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i^k - \bar{x_i})(x_j^k - \bar{x_j}), \tag{3.4.3.2}$$

where $N$ is the number of bootstrap samples drawn (this number is unlimited), x represents the correlation function and $i$ and $j$ represent two (angular) bins of the correlation function.

Jack-knife errors are found by splitting the sky area within the mask into a number of equally-sized blocks followed by calculating the correlation function multiple times, with one of these blocks excluded each time (see Weinberg et al. 2004 for an example). As one block is removed for each sample, there are therefore $N$ samples where $N$ is the number of blocks. The covariance function is then calculated in a similar manner to the bootstrap covariance function:

$$C_{jack}(x_i, x_j) = \frac{N-1}{N}\sum_{k=1}^{N}(x_i^k - \bar{x_i})(x_j^k - \bar{x_j}). \tag{3.4.3.3}$$

The coefficient is different in order to account for the dependence between jack-knife samples (samples only differ by two blocks). Both of these sampling methods are limited by the size of the observed dataset as measurements on scales larger than the survey volume cannot be made. In addition Norberg et al. (2009) found that these methods overestimate the variance on a range of scales (all scales for the bootstrap method and small scales (<2-3 $h^{-1}$Mpc) for the jack-knife method)

Another method, one that does not utilize the observed galaxy sample directly is the use of mock catalogs produced using $N$-body simulations. These simulations are produced by assuming a cosmology and using a Monte Carlo sampling method to sample mock galaxies using the same redshift distribution and number densities in each bin as the observed sample with the aim of reproducing the statistical distribution of the observed sample. The covariance matrix is then calculated from these mocks using the same equation used for the bootstrap samples (Equation 3.4.3.2). See Hildebrandt et al. (2017b) and Zhu et al. (2018) for recent examples. This is a very accurate method of producing covariance matrices of correlation functions and has become the standard approach for modern galaxy surveys, but it is very time consuming and computationally expensive as large simulation boxes with large numbers of particles, populated with galaxies, are necessary for large survey analyses. Large numbers of simulations are also necessary as cosmological parameter uncertainty scales with the ratio $\frac{n_{bins}}{n_{simulations}}$ where $n_{bins}$ is the number of angular bins used in estimating the correlation function (Dodelson & Schneider 2013). In addition, this method does not necessarily take into account all the systematics present in the observed data while the re-sampling methods account for this automatically. For an in-depth comparison of the methods discussed here see Norberg et al. (2009).

In this analysis a jack-knife re-sampling technique is adopted. As the relevant scales probed in this analysis (those that are in the vicinity of the BAO) are large, this method does not suffer from the inaccuracies discussed by Norberg et al. (2009). In addition, Crocce et al. (2011b) conduct a similar analysis to the one presented here and found that the jack-knife was compatible with the analytical covariance matrix.

In this scenario, due to the irregularity of the imaging mask used, obtaining equally sized jack-knife regions within the mask would be a very challenging task. Instead of doing this I chose to use equal sky areas which accounted for different masked areas (the full sky was split into 100 equal areas) and then I accounted for the differences in the regions by introducing a weight for each jack-knife region based on the fraction of the masked area included within that region. In addition, as I required error bars from the best estimate of the correlation function: the one computed using the entire area, $\omega^{best}$, I chose to use the deviations from this value when calculating the uncertainties. The adjusted equation is as follows:

$$C_{jack}(x_i, x_j) = \frac{N-1}{N} \frac{n}{\sum_k wt_k} \sum_{k=1}^{N} wt_k(\omega_i^k - \omega_i^{best})(\omega_j^k - \omega_j^{best}), \tag{3.4.3.4}$$

where $wt_k$ is the weight for jack-knife region $k$ and $\omega^k$ are the angular correlation functions measured using jack-knife region $k$. These jack-knife errors are the error bars seen on the correlation function plots presented in this chapter unless otherwise stated.

## 3.5 Extracting the BAO Feature

After a correlation function is measured, the position of the BAO peak must be extracted. There are a number of methods that have been proposed for extracting the BAO feature from an angular correlation function. These methods generally involve fitting the correlation function to a template using a method such as Markov Chain Monte Carlo or Maximum Likelihood and obtaining cosmological parameters from the parameters of the fit. The fitting procedure presented by Sánchez et al. (2011) entails modelling the correlation function as a power law and the BAO peak as a Gaussian:

$$\omega(\theta) = A + B\theta^{\gamma} + Ce^{-(\theta-\theta_{FIT})^2/2\sigma^2}. \tag{3.5.0.1}$$

A fit to this function using the free parameters shown above is performed, and the $\theta_{FIT}$ parameter (the mean of the Gaussian) gives the BAO scale. This parameter is then corrected for projection effects (introduced in Chapter 1) which are due to the non-negligible width of the redshift bins used, by using the following equation:

$$\theta_{BAO} = \alpha\theta_{FIT}, \tag{3.5.0.2}$$

where $\alpha$ depends on the mean redshift and redshift bin width. This correction due to the projection effects is much greater at low redshifts and for wider redshift bins. The correction factor $\alpha$ can be determined using a fiducial cosmology along with a theoretical calculation of the angular correlation function with the relevant redshifts for both a single redshift or redshift bins so that the difference in the two peak positions can be calculated (Sánchez et al. 2011; Carvalho et al. 2016). Sánchez et al. (2011) calculate alpha for a range of mean redshifts and bin widths as well as a range of cosmologies and found that $\alpha$ varied very little with cosmology (see Figure 3.16). This method is not affected by redshift space distortions or different cosmologies and it is a very common choice in the literature for obtaining BAO measurements using two point angular correlation functions (e.g.Carnero et al. 2012; Alcaniz et al. 2016; Carvalho et al. 2016; Carvalho et al. 2017).

A similar, but alternate method was presented by Chan et al. (2018) and was applied to the Dark Energy Survey BAO analysis. The template used to fit the measured angular correlation function is given below:

$$T(\theta) = B\omega(\alpha\theta) + A_0 + \frac{A_1}{\theta} + \frac{A_2}{\theta^2}, \tag{3.5.0.3}$$

where $T(\theta)$ is the template (intended to match the measured angular correlation function) and $\omega(\alpha\theta)$ is the theoretical angular correlation function obtained using a fiducial cosmology. In contrast to the previous method, this template method is cosmology dependent as

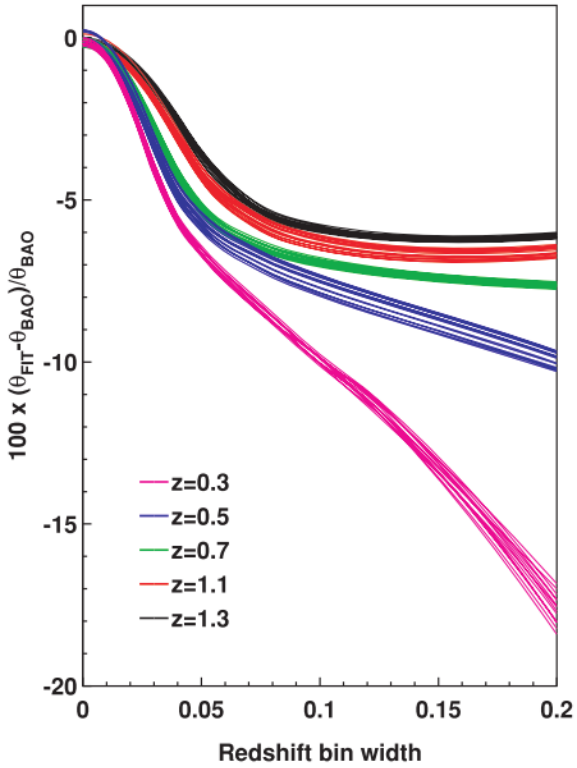



Figure 3.16: The shift in the angular BAO peak for a series of mean redshifts, redshift bin widths and cosmologies. The shift is observed to increase with increasing redshift bin width and decreasing mean redshift. The effect of the projection effect becomes smaller with increasing redshift because the redshift bins become physically very small, even with larger redshift bin widths. The cosmology, represented by the different lines within one colour, has a relatively small effect on the shift in comparison to the other parameters. Figure taken from Sánchez et al. (2011).

we are directly comparing the measured function to a specific cosmology. The parameter $\alpha$ represents the shift between the measured BAO peak and the fiducial BAO peak (versus the true BAO peak as with the previous method). A fit is then performed between the measured correlation function and the template and this requires a covariance matrix which can be the theoretical covariance or the mock covariance if a large sample of mocks are available.

After the fit is performed using an appropriate method, and $\alpha$ is obtained and applied, the following cosmological information can be extracted:

$$\frac{D_A}{r_s} = \alpha \frac{D_A^{fid}}{r_s^{fid}}, \tag{3.5.0.4}$$

where 'fid' represents the fiducial cosmology, $D_A$ is the angular diameter distance and $r_s$ is the sound horizon. This is the case because:

$$\theta_{BAO} = \alpha \theta_{fid}, \tag{3.5.0.5}$$

and

$$\theta_{BAO} = \frac{r_s}{D_A}. \tag{3.5.0.6}$$

The model in Equations 3.5.0.1 and 3.5.0.2 is such that $\alpha$ produces a shift in the BAO peak position while the remaining correlation function remains the same. It therefore corrects for the projection effect that shifts the peak along the correlation function. On the other hand, the model provided in 3.5.0.3 is such that the shift of the BAO peak, $\alpha$ compared to a fiducial position shifts the entire correlation function which is not necessarily ideal for accounting for projection effects.

In this analysis we use the cosmology independent model suggested by Sánchez et al. (2011) and the measured correlation functions are fitted with the sum of a power law and a Gaussian. This fitting is performed using the curve fit algorithm from the SciPy library (Jones et al. 2001). This uses a non-linear least squares algorithm and provides best fit parameters and the corresponding covariance matrix. Initial best guesses can be supplied to assist with finding the local minimum and parameter bounds can be set as priors. This algorithm also allows the user to input 1-$\sigma$ errors on the data and this is incorporated into the fitting by minimising the following:

$$\chi^2 = \sum_{n=1}^{N} \left( \frac{y_n - f(x_n)}{\sigma} \right)^2, \tag{3.5.0.7}$$

where $x_n$ is the independent variable (the angular scale in the case of angular correlation functions), $y_n$ is the dependent variable (the magnitude of the correlation), $f$ is the function of a specified form that takes the independent variable as well as fit parameters and outputs

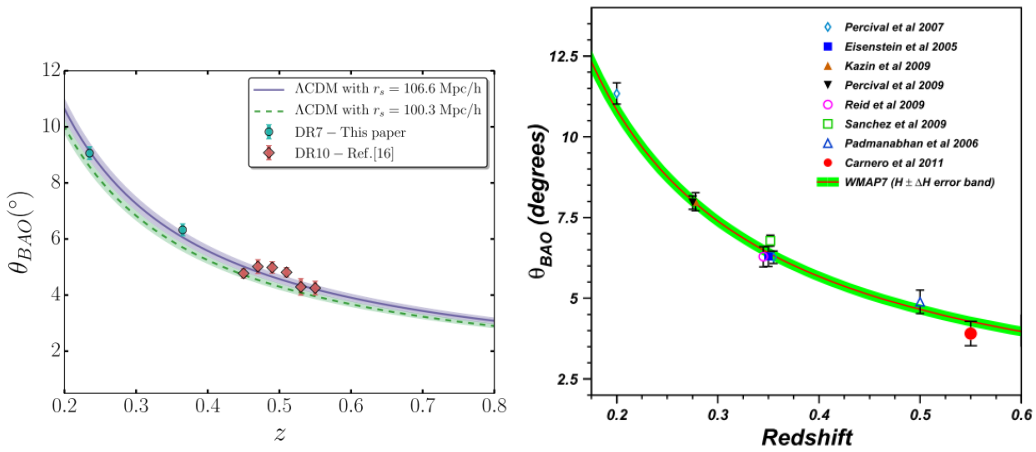



Figure 3.17: The angular BAO scale as a function of redshift. *Left:* The green and red data points represent measured results from Alcaniz et al. (2016) and Carvalho et al. (2016) respectively and the solid lines represent the ΛCDM prediction using the WMAP and Planck acoustic scale values (Figure taken from Alcaniz et al. (2016)). *Right:* the solid line is the ΛCDM using the 7-year WMAP results and the points are results from a number of measurements which are listed in the top right corner (Figure taken from Carnero et al. (2012)).

an estimate of the dependent variable. $n$ is the total number of data points available for fitting. For this analysis the function $f$ used has the form of the sum of a power law and Gaussian and the $\sigma$ errors used are those obtained from a jack-knife resampling using 100 regions.

## 3.6 Results

The results presented in this section can be compared to the results expected for the ΛCDM model and CMB measurements, as well as other similar measurements. Figure 3.17 provides an overview of past measurements and Figure 3.18 provides an example of an analysis within the redshift range used here.

### 3.6.1 Correlation Function of LGs

I chose to calculate correlation functions both for the entire redshift range of interest ($0.4 < z < 0.7$) as well as for smaller redshift bins of size $\Delta z = 0.1$. One expects the result with the entire redshift range to be a smoother function with smaller errors as the sample size is larger than for the individual bins. On the other hand, as this covers a wider range of redshift, the BAO peak is expected to be more shifted and flattened due to projection

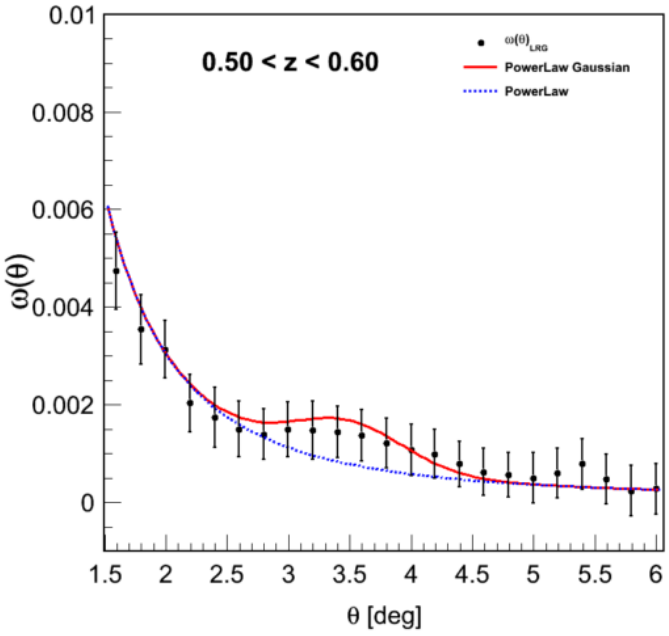



Figure 3.18: Measured angular correlation function and fit using a power law and Gaussian obtained using SDSS luminous red galaxies by Carnero et al. (2012). The best fit mean of the Gaussian was found to be $3.48° \pm 0.19°$.

effects.

Figure 3.19 shows the correlation functions obtained when the sky background weights and removal of galaxies around bright stars were applied separately and together. This allows for a comparison of the quality of the correlation functions in the different situations. The clustering on large scales, apart from at the acoustic scale, is expected to be very small, therefore, the correlation functions with a smaller signal in this plot are expected to be more accurate and the extra power is likely due to systematic effects. This means that removing the galaxies around bright stars has a notable positive effect on the correlation function in comparison to the effect of the sky background weights. The combination of the two corrections leads to a correlation function that is marginally better than that produced with only the star correction. Thus, we chose to use both corrections for all other correlation functions produced.

The LG sample in the redshift range $0.4 < z < 0.7$ with incompleteness cuts, bright stars removed and sky background weights contained 902802 galaxies and the resulting correlation function is shown in Figure 3.20. The correlation functions for the individual redshift bins are given in Figure 3.21. The result for the whole range has a noticeable peak between $\theta = 0.3$ and $\theta = 0.35$, which, if corrected for the large projection effects present, roughly corresponds to the expected value for a $\Lambda$CDM cosmology of $\sim 3.97° - 4.22°$ for

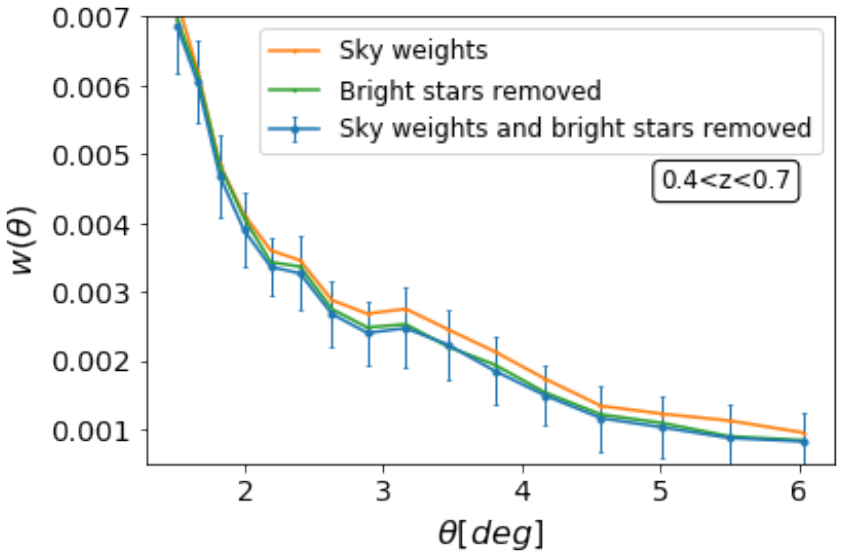



Figure 3.19: Angular correlation functions for the DR14 LG sample with various cuts or weights imposed. Lines connecting the data points were inserted and error bars were removed for clarity. All correlation functions have the incompleteness cuts. The orange line is the correlation function obtained using weights based on density for different sky backgrounds and the green line is the correlation function obtained after galaxies and randoms within a radius of 9.48 arcsec around bright stars were removed. The blue line is the combination of both of the above cuts/weights and the blue error bars are jack-knife error bars for this correlation function.

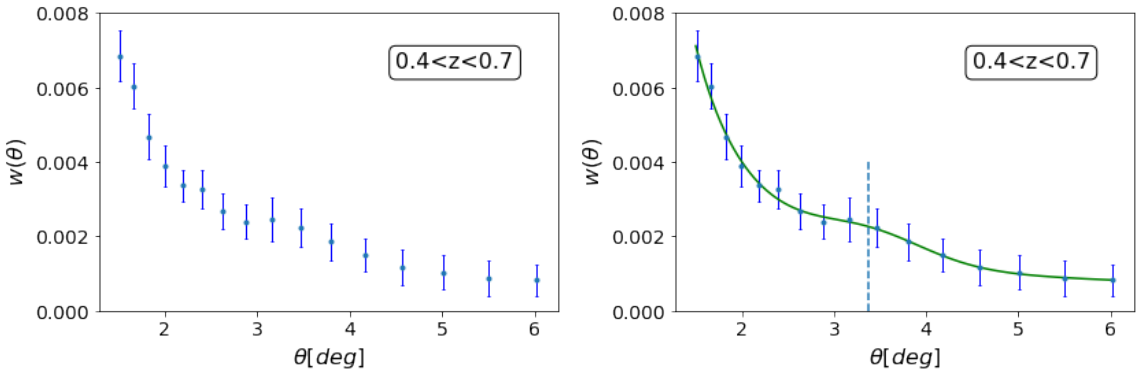



Figure 3.20: Angular correlation functions for the DR14 LG sample with incompleteness cuts, removal of the area around bright stars and sky background weights. The figures on the right show the fit to the sum of a power law and a Gaussian and the dashed line is at the mean of the Gaussian. Error bars are from jack-knife resampling using 100 regions.

a bin with a median redshift of $z = 0.55$ (see Figure 3.17 and Table 3.7). The correlation functions in the $\Delta z = 0.1$ bins also showed peaks at reasonable scales, with the peak scale decreasing with increasing redshift as expected. The jack-knife error bars were larger for the redshift bins than for the entire redshift range, making the results from these bins less robust. Smaller redshift bins, with $\Delta z = 0.05$, were also experimented with and were found to result in very noisy correlation functions with little or no noticeable signal.

Next, the technique described in Section 3.5 was used to fit the correlation functions and locate the BAO peaks. Figures 3.20 and 3.21 show the angular correlation functions along with the fit to the sum of a power law and Gaussian. The parameters for the Gaussian parts of this fit are provided in Table 3.5. The power law parameters are well constrained. The position of the BAO peak was found to be $3.37 \pm 0.13°$ for the entire redshift range ($0.4 < z < 0.7$). The mean redshift of this sample is $z = 0.55$ and therefore this result can be compared to the result found by Carnero et al. (2012) who find the BAO peak using the angular correlation function and galaxies with redshifts in the range $0.5 < z < 0.6$ (see Figure 3.18). The scale of the peak measured was $3.48 \pm 0.19°$ which is similar to this result, but it is notable that our result, with a larger redshift bin width and therefore a greater expected projection effect, is shifted to a lower redshift. This is the expected direction of the shift as galaxies with a wider range of redshift are assumed to be at the mean redshift of the bin and therefore the corresponding projected positions of galaxies separated by the BAO scale have smaller angular separations, shifting the peak to smaller angular scales (Sánchez et al. 2011).

The result obtained from the redshift bin $0.5 < z < 0.6$ is found to be $3.84 \pm 0.11°$. This result can be more directly compared to the result of Carnero et al. (2012), it is

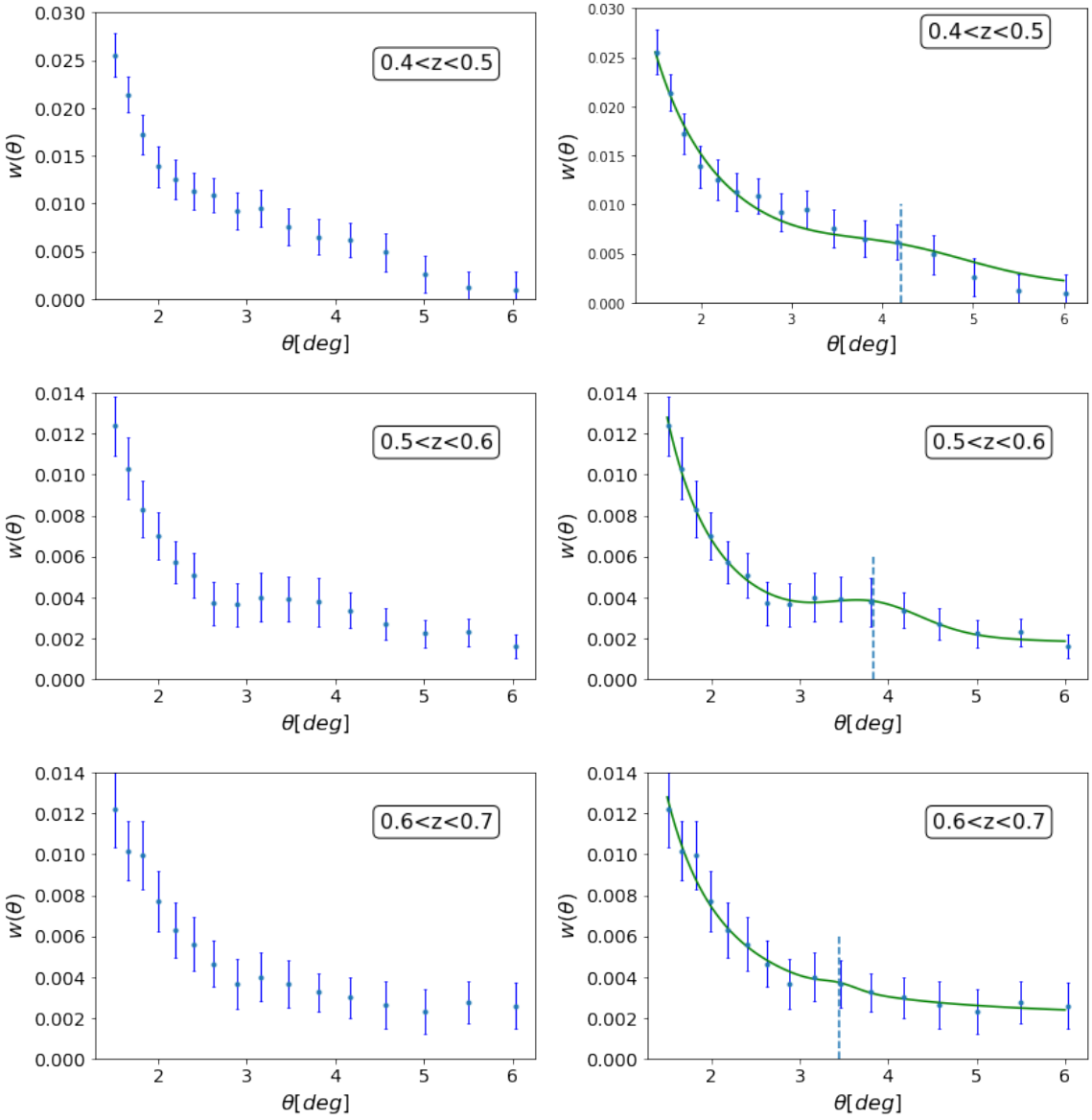



Figure 3.21: Angular correlation functions for the LG sample in bins of width $\Delta z = 0.1$. The figures on the right show the fit to the sum of a power law and a Gaussian and the dashed line is at the mean of the Gaussian. Error bars are from jack-knife resampling using 100 regions.

Table 3.5: Table showing the best fit parameters for the Gaussian part of the fit of the sum of a power law and Gaussian to the correlation functions for the LG sample. The uncertainties are 1 standard deviation errors on the fits and do not include the effect of systematic errors.

| LG Correlation Function | C | $\theta_{fit}$ | $\sigma$ | $N_{gal}$ |
|---|---|---|---|---|
| $0.4 < z < 0.7$ | 6.88 $\pm$ 1.34e-4 | 3.37 $\pm$ 0.13 | 6.25 $\pm$ 1.67e-1 | 902802 |
| $0.4 < z < 0.5$ | 2.30 $\pm$ 1.45e-3 | 4.11 $\pm$ 0.44 | 7.98 $\pm$ 6.50e-1 | 244069 |
| $0.5 < z < 0.6$ | 1.32 $\pm$ 0.24e-3 | 3.84 $\pm$ 0.11 | 5.53 $\pm$ 1.29e-1 | 448128 |
| $0.6 < z < 0.7$ | 2.55 $\pm$ 3.98e-4 | 3.45 $\pm$ 0.43 | 2.04 $\pm$ 3.25e-1 | 210605 |

larger than the peak measured by Carnero et al. (2012) (3.48$\pm$0.19°) but smaller than the corresponding result after it was corrected for projection effects (3.90$\pm$0.38°). The projection effect correction, $\alpha$, calculated and used by Carnero et al. (2012) for a mean redshift of $z = 0.55$ and a photo-z bin width of $\Delta z = 0.1$ is 1.12. This measured BAO peak can also be compared to the analysis by Carvalho et al. (2016) which uses spectroscopic redshifts. Carvalho et al. (2016) finds a peak position of $4.25°$ for this mean redshift. This result is similar to the result assuming a $\Lambda$CDM cosmology and using the WMAP9 acoustic scale and is more similar to the result measured in this analysis after it is corrected for projection effects.

The results for the other redshift bins on the other hand did not have peaks that were as clear or easily fitted with the parametric model. For example, the correlation function for the bin $0.4 < z < 0.5$ had a spurious peak around $\theta = 3$ making the fitting procedure more difficult and the result less accurate. Despite this, the best fit models provide BAO peak positions at 4.11$\pm$0.44° and 3.45$\pm$ 0.43° for the redshift bins $0.4 < z < 0.5$ and $0.6 < z < 0.7$ respectively. The result for a mean redshift of 4.5, i.e. $4.11°$ is smaller than the corresponding result by Carvalho et al. (2016), $4.77°$, but after projection effect corrections it should be comparable to this result and the predicted results assuming a $\Lambda$CDM cosmology, $4.72°$ - $5.03°$ (see Table 3.7). The result in the $0.6 < z < 0.7$ bin on the other hand is very similar to the Planck $\Lambda$CDM and therefore after the correction for projection effects it will increase and become nearer to the WMAP predicted result.

In Figure 3.20 an unexpected peak can be observed at $2° < \theta < 2.5°$ and a similar peak at $3° < \theta < 3.5°$ is observed in the correlation function for the $0.4 < z < 0.5$ redshift bin in Figure 3.21. It is normal for nuisance peaks to be present in the angular correlation function and they are due to the insufficient removal of systematic effects and the presence of noise (Alcaniz et al. 2016). The true bump is normally identified by comparing the positions of the peaks to the position expected from a cosmological model (see Carnero et al. 2012), but this makes the result model dependent. A method for identifying the true peak was

presented by Alcaniz et al. (2016) and entailed changing the angular coordinates of the galaxies by small and random amounts and observing which peak remains, as only the true BAO peak is expected to survive this change. Another method, suggested by Carvalho et al. (2016), involves finding the peak that remains after changing the angular bin size used for calculating the correlation function as well as shifting the angular coordinates slightly, as in Alcaniz et al. (2016). In this chapter we do not implement a similar model independent approach but such an analysis can be implemented in future work.

### 3.6.2 Correlation Function with the Entire Galaxy Sample

The data used in this section was obtained using the methods discussed in Section 3.2.1. After photo-zs were obtained for both the LGs and the non-LGs (see Section 3.3) these samples were combined and the cuts and weights discussed in Section 3.2 were applied to the entire galaxy sample, resulting in a sample size of 35848084 galaxies in the range $0.4 < z < 0.7$. Randoms were created as before and the correlation function was then measured using CUTE, jack-knife resampling was again performed in order to obtain the error bars. No peak was visible in the $0.4 < z < 0.7$ correlation function obtained. Correlation functions for $\Delta z = 0.1$ redshift bins were then measured and were also found to have no noticeable peaks.

Due to the lack of a detectable peak in this correlation function, I decided to impose a magnitude cut at $i < 21$ to determine whether the peak would be detectable in this sample of brighter galaxies of all types. Brighter galaxies are more likely to lead to detectable BAO peaks for two reasons: their photo-zs should be more accurate than for fainter galaxies and they are more likely to be more biased tracers of the underlying dark matter. This cut results a sample of 20887127 galaxies within the range $0.4 < z < 0.7$ after all the relevant cuts were made. The resulting correlation function for the entire redshift range was very similar to the result without the magnitude cut and no noticeable peak was present. On the other hand, when redshift bins of width $\Delta z = 0.1$ were used a peak was visible in the $0.5 < z < 0.6$ bin and a lower significance peak was present in the $0.6 < z < 0.7$ bin. These correlation functions and the corresponding fits to the sum of a power law and Gaussian are presented in Figure 3.22 and Table 3.6. These correlation functions had smaller error bars than those obtained for the LG sample as the number of galaxies used was significantly higher. The Gaussian peaks obtained: $4.09 \pm 0.16°$ and $3.44 \pm 0.13°$ are consistent with the results obtained for the LG sample and have similar uncertainties (see Table 3.7) but are a little larger, and therefore nearer to the theoretical expected results for $\Lambda$CDM from Planck and WMAP. On the other hand, the correction for projection effects will increase

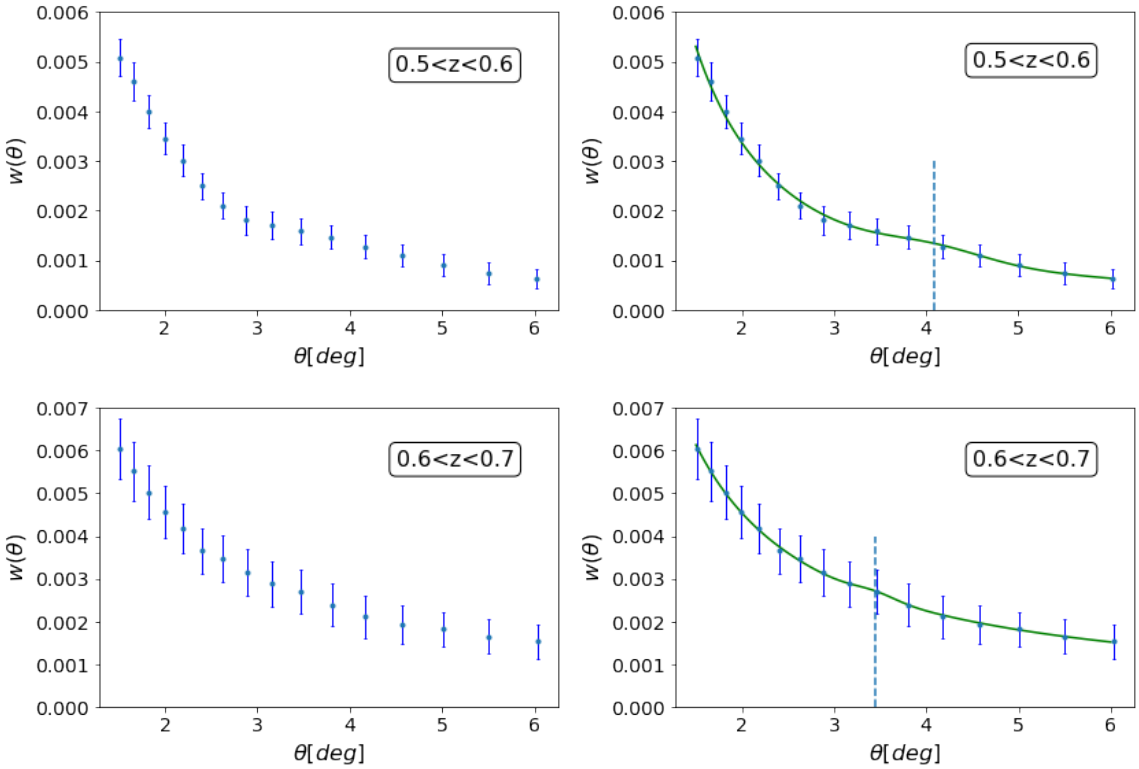



Figure 3.22: Angular correlation functions for the entire galaxy sample with a magnitude cut of $i < 21$ in bins of width $\Delta z = 0.1$. The figures on the right show the fit to the sum of a power law and a Gaussian.

Table 3.6: Table showing the best fit parameters for the Gaussian part of the fit of the sum of a power law and Gaussian to the correlation functions using all galaxy types with an $i < 21$ magnitude cut. The uncertainties are 1 standard deviation errors on the fits and do not include the effect of systematic errors.

| Correlation Function $i < 21$ | C | $\theta_{fit}$ | $\sigma$ | $N_{gal}$ |
|---|---|---|---|---|
| $0.5 < z < 0.6$ | 2.25 $\pm$ 0.59e-4 | 4.09 $\pm$ 0.16 | 5.59 $\pm$ 1.99e-1 | 6730945 |
| $0.6 < z < 0.7$ | 1.14 $\pm$ 0.49e-4 | 3.44 $\pm$ 0.13 | 2.29 $\pm$ 1.10e-1 | 5342497 |

these values, making them larger than predicted, and nearer to the WMAP predicted result than the Planck results.

BAO peaks obtained for both samples are compared to expected results from CMB measurements and assuming a $\Lambda$CDM cosmology. Table 3.7 provides these results. The values from the Planck data are obtained assuming the best-fit parameters for the $\Lambda$CDM model and the comoving physical acoustic scale values: 100.09$\pm$1.4 and 99.08$\pm$0.81 from the Planck 2015 and Planck 2018 results (Planck Collaboration et al. 2016a, 2018) and 106.61$\pm$3.47 from the WMAP9 results (Hinshaw et al. 2013). It should be noted that the errors on the measured peaks do not account for systematic effects and therefore should not be directly compared to those the predictions.

The uncertainty in the measurements of the BAO peak position would ideally be compared to these obtained for the equivalent measurement using spectroscopic redshifts in order to evaluate the trade-offs between accurate redshifts and larger galaxy numbers. This cannot be done correctly in this analysis as the reported uncertainties are simply uncertainties in the fit, systematic uncertainties were not accounted for. Carnero et al. (2012) perform such a comparison and include three dimensional spectroscopic analyses by converting the result to an angular scale given the fiducial cosmology used. The result of this comparison is the second plot in Figure 3.17. The spectroscopic measurement uncertainties are clearly smaller than the result by Carnero et al. (2012), 3.90$\pm$0.38. The uncertainty in the fits for the luminous galaxies with mean redshifts 0.45 and 0.65 are larger than this and therefore these measurements are less accurate compared to the spectroscopic measurements. On the other hand, the uncertainties for the luminous galaxy sample in the redshift bin 0.55 and the all galaxy samples had small uncertainty values which would rival some spectroscopic measurements. For example Alcaniz et al. 2016 find the measurements 9.06$\pm$0.23 at $z = 0.235$ and 6.33$\pm$0.22 at $z = 0.365$ while Carvalho et al. 2016 have uncertainties ranging from 0.17 and 0.30 for a range of redshift intervals with means spanning 0.45-0.55. However, as the systematic uncertainties were not quantified little can be concluded from

Table 3.7: Table showing the ΛCDM predictions for the BAO peak position with the acoustic scale at Planck and WMAP values, as well as best fit BAO peak positions before correcting for projection effects for the correlation functions using the LG sample and all galaxy types with an $i < 21$ magnitude cut. The uncertainties are 1 standard deviation errors on the fits and 68% confidence intervals for the ΛCDM predictions.

| Average redshift | ΛCDM, Planck2018 | ΛCDM, Planck2015 | ΛCDM, WMAP9 | LG sample | All galaxies $i < 21$ |
|---|---|---|---|---|---|
| $\bar{z} = 0.45$ | $4.72 \pm 0.04$ | $4.76 \pm 0.07$ | $5.03 \pm 0.16$ | $4.11 \pm 0.44$ | |
| $\bar{z} = 0.55$ | $3.97 \pm 0.03$ | $4.00 \pm 0.05$ | $4.22 \pm 0.14$ | $3.84 \pm 0.11$ | $4.09 \pm 0.16$ |
| $\bar{z} = 0.65$ | $3.46 \pm 0.03$ | $3.48 \pm 0.05$ | $3.67 \pm 0.12$ | $3.45 \pm 0.43$ | $3.44 \pm 0.13$ |

this comparison.

## 3.7 Conclusions

In this chapter I presented a clustering analysis of the SDSS DR14 galaxies using photometric redshifts from the GPz algorithm. Two analyses were performed: one including all primary objects identified as galaxies and another with only luminous galaxies (using the CMASS selection criteria). Magnitude cuts were first applied to both samples based on the magnitude at which the numbers start to decrease, galaxies with magnitudes fainter than this limit were removed in order to made the samples complete. Following this, an imaging area mask was selected and applied to both samples. This mask removed regions that were not imaged as well as areas with poor seeing and large amounts of galactic extinction. The additional systematic errors due to the sky background and the presence of bright stars were then removed. This was done by weighting the galaxies based on their sky background in order to remove the effects of this background on the galaxy density and by removing all galaxies near to bright stars from both the galaxy samples and the random samples. Angular correlation functions were measured using the CUTE code for both the luminous galaxies and all galaxy types and error bars were estimated using jack-knife resampling. These correlation functions with different adjustments for systematic effects were compared and the removal of galaxies near to bright stars was found to have a notable, positive effect on the clustering measurement.

Potential BAO peaks were detected in the results from the luminous galaxies and a model with a power law (for the continuum) and Gaussian (for the BAO peak) was then fit to these correlation functions and the best fit positions for the BAO peaks were found. This peak was located at 3.37±0.13 for a redshift range $0.4 < z < 0.7$ and the results for the redshift bins were $4.11 \pm 0.44°$, $3.84 \pm 0.11°$ and $3.45 \pm 0.43°$ in order of increasing redshift.

These positions will shift to larger angles if corrections for projection effects are applied. These projection effects can be calculated theoretically for the relevant mean redshifts and redshift bin sizes and this will be carried out in future work. Thus, the peaks measured are consistent with the predicted results using $\Lambda$CDM and previous results of $\sim$4.72°-5.03°, 3.97°-4.22° and 3.46°-3.67° (see figure 3.17 and Table 3.7).

The BAO peak was not detectable with the entire galaxy set, so a magnitude cut of $i < 21$ was applied to determine whether the clustering of this brighter sample, with more accurate photo-zs would lead to a detectable peak. The correlation functions measured in the redshift bins $0.5 < z < 0.6$ and $0.6 < z < 0.7$ had peaks at 4.09 $\pm$ 0.16° and 3.44 $\pm$ 0.13° respectively. These positions are similar to the results obtained with the LGs and are consistent with the expected results. The aim of this analysis was to determine whether the peak would be detectable with a large sample of galaxies of all types and using photometric redshifts and this was proven to be the case. This is very important for future cosmological measurements as large photometric surveys are conducted and spectroscopic redshifts are not available. In addition, the removal of a dependence on the samples such as luminous galaxies for clustering measurements such as these will be important for extracting information at high redshifts where luminous galaxies are rare. This will also be important for cases in which angular clustering measurements for galaxies of different types are required, such a measurement is that of the local primordial non-Gaussianity signal using the multi-tracer method and different galaxy populations (discussed in detail in Chapter 4).

# Chapter 4

# Forecasting Primordial Non-Gaussianity Constraints using the Multi-tracer Method and Radio-Selected Galaxies

The work in this chapter was carried out with the help of Matt Jarvis, Stefano Camera, Catherine Hale and José Fonseca and has been published in the Monthly Notices of the Royal Astronomical Society under the title: 'Non-Gaussianity Constraints using Future Radio Continuum Surveys and the Multi-Tracer Technique'.

## 4.1 Introduction

The local value of primordial non-Gaussianity, $f_{\rm NL}$, defined in Equation 1.1.5.7 describes the deviation of the initial density field from a Gaussian random field. Simple slow-roll, single-field inflationary models predict almost Gaussian density fields with $f_{\rm NL} \ll 1$ ($\mathcal{O}(10^{-2})$) (Maldacena 2003; Creminelli & Zaldarriaga 2004) while multi-field models allow larger deviations from Gaussianity with $f_{\rm NL} \gtrsim 1$ (Lyth et al. 2003; Zaldarriaga 2004). The current tightest constraint on the local PNG parameter is $f_{\rm NL} = 0.8 \pm 5$ which was obtained using Planck 2015 CMB data (Planck Collaboration et al. 2016b). Planck 2018 data (Planck Collaboration et al. 2019) produced the slightly less constrained result $f_{\rm NL} = 0.9 \pm 5.1$. This recent result was less constrained due to the use of more realistic polarization simulations with higher noise levels for estimating the errors.

As discussed in Chapter 1, local PNG is normally measured using the bispectrum of the CMB temperature anisotropy maps (Komatsu et al. 2003; Planck Collaboration et al. 2016b, 2019), but this measurement is limited on both small and large scales. Large scale structure measurements are a viable alternative to the CMB bispectrum as the non-Gaussianity also leaves an imprint on the large scale structure (LSS) by increasing the 3D power spectrum

on large scales. See Chapter 1 for more details on this measurement and the precision required. Large galaxy surveys (both optical and radio) such as SDSS and the National Radio Astronomy Observatory (NRAO) Very Large Array (VLA) Sky Survey (NVSS; Condon et al., 1998) have been used to measure the clustering of dark matter and constrain the effects of PNG (Slosar et al. 2008; De Bernardis et al. 2010; Xia et al. 2010; Ross et al. 2013). One approach to this LSS measurement is to use quasar populations as these are highly biased dark matter tracers and cover large volumes of space. Leistedt et al. (2014) and Castorina et al. (2019) study the clustering of the SDSS quasar sample and provide the tightest constraints with LSS methods to date. Combinations of galaxy surveys and CMB maps have also been considered (Giannantonio & Percival 2014).

As this PNG effect is most prominent on large scales, very large surveys are used, but these are still limited by cosmic variance on the larger scales and forecasts of the tightest constraints possible with single tracer surveys have found that $\sigma(f_{\rm NL}) \sim 1$ is not possible. The multi-tracer method, first proposed by Seljak (2009) significantly reduces the effect of cosmic variance by cross-correlating multiple dark matter tracers, with different biases located in the same sky volume. These populations have the same underlying density distributions and thus the ratio of their biases will cancel the effects of cosmic variance resulting in a measurement that is only limited by shot noise. This multi-tracer technique, applied to the large upcoming galaxy surveys with multiple differently biased galaxy populations is expected to rival the current CMB constraints (see Ferramacho et al. 2014; Yamauchi et al. 2014; Alonso & Ferreira 2015).

An alternative avenue for probing the effects of PNG on the LSS is intensity mapping and future surveys of this type seem promising due to the combination of their large sky coverage and the potential to push to high redshifts (Joudaki et al. 2011; Camera et al. 2013; Li & Ma 2017; Fonseca et al. 2018). The multi-tracer method has also been suggested for use on HI intensity mapping surveys. For example, Fonseca et al. (2018) forecast that performing such an analysis on the HI survey from the SKA and the H$\alpha$ survey from SPHEREx (Doré et al. 2014) could lead to constraints on $f_{\rm NL}$ on the order of $\sigma(f_{\rm NL}) \sim 1$. Applying the multi-tracer method to combinations of intensity mapping and photometric galaxy surveys is also possible and was first suggested by Alonso & Ferreira (2015) and Fonseca et al. (2015). More recently the combination of intensity mapping from SKA and galaxy surveys from LSST was suggested and investigated by Witzemann et al. (2019) and intensity mapping from MeerKAT and galaxy surveys from DES was suggested by Fonseca et al. (2017) and forecasts have found that these analyses should provide constraints more than two times better than Planck (Fonseca et al. 2017). Recently a forecast analysis has been performed on the use of the multi-tracer method for combining HI intensity mapping

with both photometric galaxy surveys and CMB lensing and the constraints of $\sigma(f_{\rm NL}) \sim 1$ were obtained (Ballardini et al. 2019).

Further research on improving the methods currently being used for measuring this $f_{\rm NL}$ signal is also being done. For example, work is being done on improving the measurement of the galaxy bispectrum and accounting for systematics (Tellarini et al. 2016; Welling et al. 2016; Mueller et al. 2019; Karagiannis et al. 2018; Uhlemann et al. 2018). Another example is a recently proposed method which involves the creation of a sample with zero bias, allowing the effect of PNG on the clustering to be measured more easily (Castorina et al. 2018). Similarly, studies on the optimal galaxy survey for providing $f_{\rm NL}$ constraints of $\sigma(f_{\rm NL}) = 1$ are also being done (de Putter & Doré 2017).

Radio surveys are appealing for use with multi-tracer techniques for a number of reasons: radio surveys cover very large areas of sky, which is particularly advantageous since the increased bias signal is detected on large scales, and is also useful for reducing shot noise. Just as importantly, radio surveys contain a range of distinct radio galaxy populations that have significant spread in their biases (Hale et al. 2018). Other advantages of radio surveys are that they allow the observation of galaxies to very high redshifts and without dust obscuration (Jarvis et al. 2015), which has been shown to provide advantages in determining cosmological parameters (Camera et al. 2012). On the other hand, one disadvantage of using radio surveys is that it is very difficult to get precise redshifts for the individual sources: either multi-wavelength data or HI 21-cm line measurements are necessary (Jarvis et al. 2015).

The Square Kilometre Array (SKA, Carilli & Rawlings 2004; Dewdney et al. 2009) will be the world's largest radio telescope with thousands of dishes and antennas distributed over South Africa and Australia. It will consist of a mid-frequency instrument (SKA-MID) and a low-frequency instrument (SKA-LOW) and will span a frequency range of 50 MHz to 15 GHz with unprecedented angular resolution and sensitivity. It will be constructed in two main phases, with the first composed of $10\%$ of the intended eventual collecting area. The SKA's radio continuum survey will cover $2\pi$ steradians while its spectroscopic H1 survey will detect galaxies up to redshift $\sim 2$. HI intensity mapping surveys that cover redshifts up to 5 may also be possible with the SKA (Santos et al. 2015; Quinn et al. 2015).

In this chapter I update the work by Ferramacho et al. (2014) who perform a tomographic multi-tracer analysis driven by multiple radio galaxy populations with simulated masses and then use this to forecast the PNG constraints possible with the SKA Phase 1. I performed a similar analysis, using both the galaxy masses estimated by the $S^3$ simulation (Wilman et al. 2008) and those measured by Hale et al. (2018) who use Very Large Array (VLA) COSMOS field observations, to estimate the average bias values and corresponding halo

masses of the radio galaxy populations. In addition, redshift distributions predicted by both $S^3$ and the more recent Tiered Radio Extra-galactic Continuum Simulation (T-RECS) by Bonaldi et al. (2019) are used and compared. In section 4.2 I present an overview of the theoretical framework used for this analysis, followed by a discussion of the choice of radio galaxy populations in section 4.3. The methods used in the Fisher analysis are described in section 4.4 and the results and analysis are presented in section 4.5. Finally, conclusions are presented in section 4.6. The fiducial cosmology used has $H_0 = 67.74$, $\Omega_\Lambda = 0.6911$, $\Omega_{CDM} = 0.26$, $\Omega_b = 0.05$, $A_s = 2.142$ x $10^{-9}$ and $f_{\rm NL} = 0$ and was taken from the Planck 2015 results (Planck Collaboration et al. 2016b).

## 4.2 Multi-tracer Analysis for Measuring Primordial Non-Gaussianity in Large Scale Structure

The effects of primordial non-Gaussianity on the large scale structure were presented and discussed in the introduction of this thesis. Here I will discuss how a multi-tracer analysis is performed.

As previously discussed, the main effect of non-Gaussianity on the large scale structure is to increase the halo power spectrum on large scales. The primordial non-Gaussianity parameter $f_{\rm NL}$ is normally determined from large scale structure by measuring the galaxy power spectrum or bispectrum (De Bernardis et al. 2010; Ross et al. 2013). On the other hand, in situations where spectroscopic redshift data is not available (which will be the case with some current and future large scale surveys with volumes and depths that make it unfeasible to measure spectroscopic redshifts), and only the less accurate photometric redshifts or estimated redshift distributions are present, neither the 3D matter power spectrum nor the bispectrum can be reliably computed. Instead, a tomographic analysis with angular power spectra can be used. The photo-z estimates are used to place the galaxies into redshift bins and then the angular power spectrum, which is the projection of the line-of-sight halos (within a particular redshift bin), into the 2D plane can be computed. Although the angular power spectra do not provide as much information as the full 3D analysis, it is a sufficient estimate. Some analyses that have used this method are: Slosar et al. (2008) and Xia et al. (2010).

For the case of a multi-tracer analysis using these angular power spectra, the statistical information is given by the auto and cross correlation power spectra that are split into multipoles (see Huterer et al. 2001):

$$C_l^{i,j} = \frac{2}{\pi} \int_{k_{min}}^{k_{max}} k^2 P_\delta(k) W_l^i(k) W_l^j(k) dk. \tag{4.2.0.1}$$

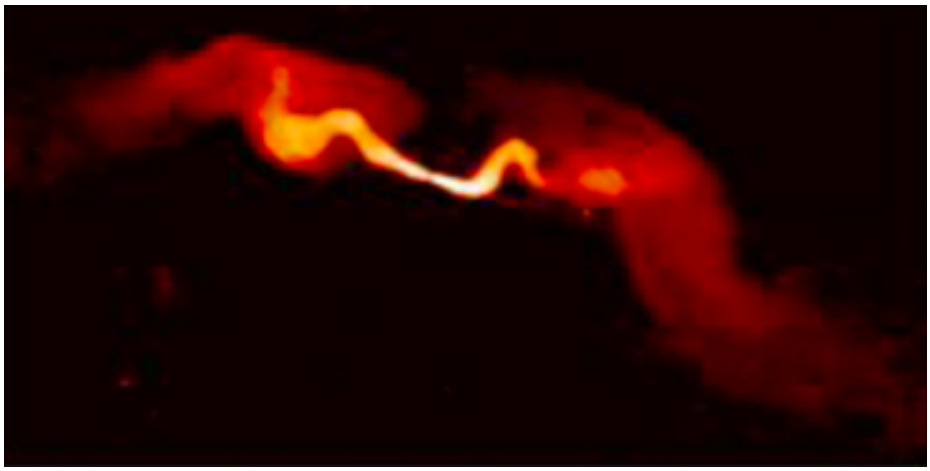


Figure 4.1: Image of a galaxy (3C31) that is classified as a FRI. Image taken from the National Radio Astronomy Observatory.

$P_\delta(k)$ is the dark matter power spectrum at redshift 0. $W_l^i$ are window functions that incorporate the redshift distribution ($dn/dz$) and bias $b$ of the tracer $i$ as well as the angular geometry of multipole $l$ via a spherical Bessel function $j_l(kr)$ where $r$ is the comoving radial distance to redshift $z$. The form of these window functions is given below:

$$W_l^i = \int \frac{dn}{dz}^i D(z) b_h^i(z) j_l(kr) dz. \tag{4.2.0.2}$$

In this analysis I compute these auto and cross power spectra for a variety of tracers, defined by radio galaxy type and redshift bin, using estimated bias functions and redshift distributions given the constraints of the SKA telescope. Following this, I perform a Fisher analysis of this multi-tracer method to determine the constraints on the $f_{\rm NL}$ parameter that will be possible with the SKA. The relatively large width of the redshift bins used in this analysis allows us to neglect the effects of redshift space distortions.

## 4.3 Radio-Selected Galaxy populations

### 4.3.1 Radio AGN Classification

Radio observations allow us to detect a wide range of galaxies that can be classed on the broadest terms as star forming (star forming galaxies or SFGs) and active galactic nuclei (AGN). Both the SFGs and AGN are detected in the radio due to their synchrotron radiation. This radiation comes predominantly from the acceleration of relativistic electrons in supernova remnants for the SFGs and from jet emission for the AGN. The SFGs can be distinguished from the AGN via the shape of their radio spectral energy distributions,

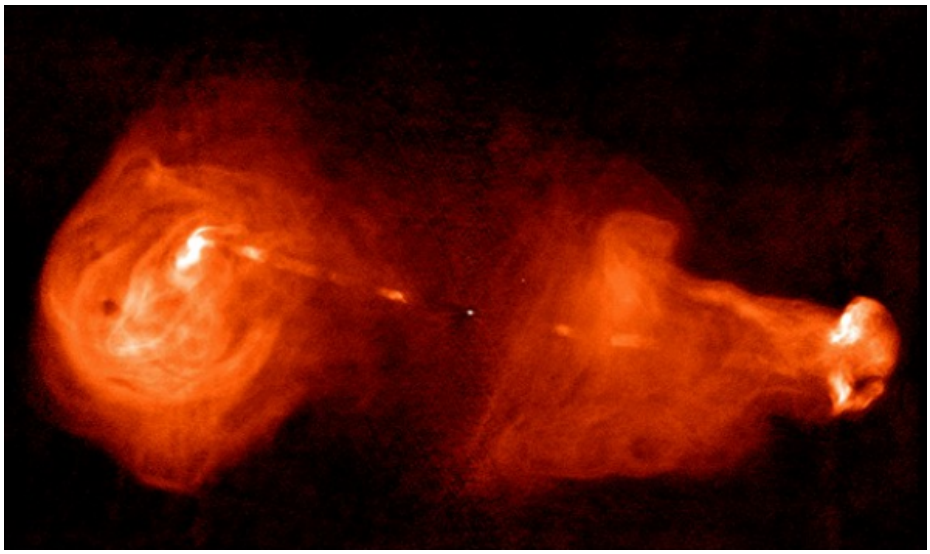


Figure 4.2: Image of a galaxy (3C353) that is classified as an FRII. Image taken from the National Radio Astronomy Observatory.

as well as based on their morphologies. Multi-wavelength data, in particular optical and infrared is also very useful for this task.

The AGN can be split into different categories, and these differ based on the classification scheme used. The two most common ways of classifying radio AGN are by morphology and by accretion mode. The morphological method or Fanaroff-Riley classification (Fanaroff & Riley 1974) splits AGN based on the ratio of the distance between their regions of greatest surface brightness and the total length of the galaxy. Galaxies with their brightest regions near their centres are called FRIs (see Figure 4.1 for an example) while those with bright regions at the edges of extended lobes are labelled FRIIs (see Figure 4.2 for an example). This is a relatively straightforward classification that can be applied when only radio images, with sufficient resolution, are present. Despite this, not all AGN will fall neatly into one of these categories, making them difficult to classify.

The second method, the accretion mode classification distinguishes AGN based on their radiative efficiency. Radiatively efficient AGN are those with a radiative accretion mode, meaning that they have an accretion disk and follow the orientation-based AGN unification model (Urry & Padovani 1995). They also have luminosities of $L \gtrsim 0.01 L_{\rm Edd}$ where $L_{\rm Edd}$ is the Eddington luminosity and, in the classification scheme of Smolčić et al. (2017a) they are called high luminosity AGN (HLAGN). When these AGN are radio-loud and extended they are also called High Excitation Radio Galaxies (HERGs), but QSOs are also included in this HLAGN category. HERGs have strong high excitation emission lines in their spectra, hence the term 'high excitation', these are narrow emission lines and examples of such spectra are

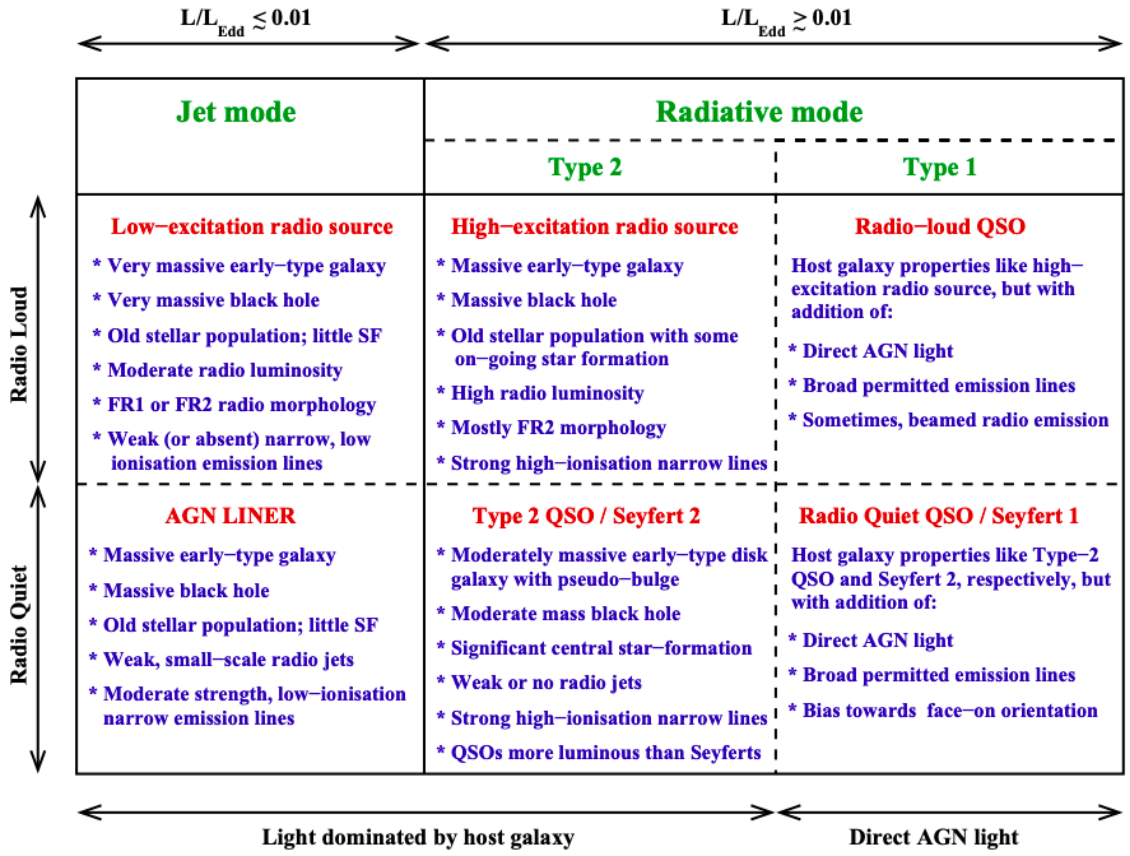



Figure 4.3: Table showing the MLAGN/HLAGN classification of AGN and the galaxy types that occupy the categories. Taken from Heckman & Best (2014)

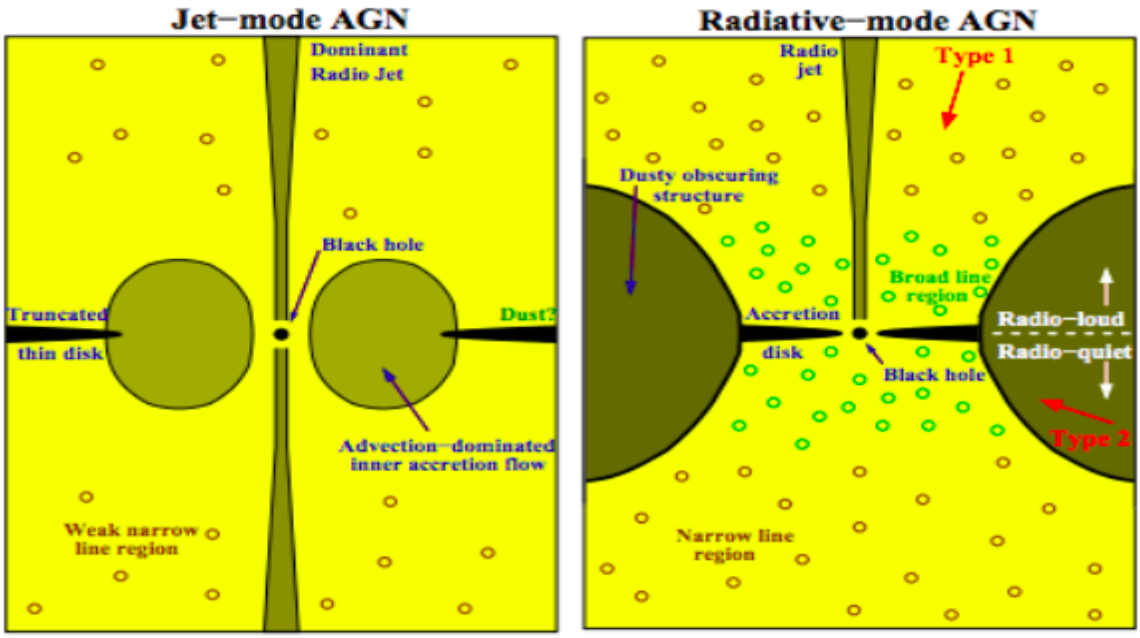



Figure 4.4: Diagrams illustrating the physical components present in jet mode (left) and radiative mode (right) AGN. Taken from Heckman & Best (2014).

shown in Figure 4.5. On the other hand, the radiatively inefficient AGN accrete in the jet mode, they have no accretion disk and instead their accretion flow is advection dominated. These have luminosities $L \lesssim 0.01 L_{\rm Edd}$ and are called Moderate luminosity AGN (MLAGN) in the Smolčić et al. (2017a) classification scheme. The radio-loud MLAGN are called Low Excitation Radio Galaxies (LERGs) and radio quiet AGN LINERS are the other members of this group. The spectra of these LERGs have few emission lines and the only ones that may be present are weak, narrow, low ionisation emission lines. Examples of these spectra can be seen in Figure 4.6. The physical components of these two types of AGN are illustrated in Figure 4.4. The table in Figure 4.3, taken from Heckman & Best (2014), provides a succinct description of this classification and the galaxy types that fall into these categories along with the connection to the Fanaroff-Riley classification. See Section 2.1 of Heckman & Best (2014) for a full description of the classification. Unlike the FRI/FRII classification, this method requires a wealth of multi-wavelength data which will not always be present.

### 4.3.2 Populations for Multi-Tracer Analysis

There have been a number of attempts to analyse radio galaxies by population and determine their average dark matter halo masses and biases (Wilman et al. 2008; Lindsay et al. 2014; Magliocchetti et al. 2017; Hale et al. 2018; Bonaldi et al. 2019). Wilman et al. (2008) performed a semi-empirical simulation of the extra-galactic radio-continuum sky covering

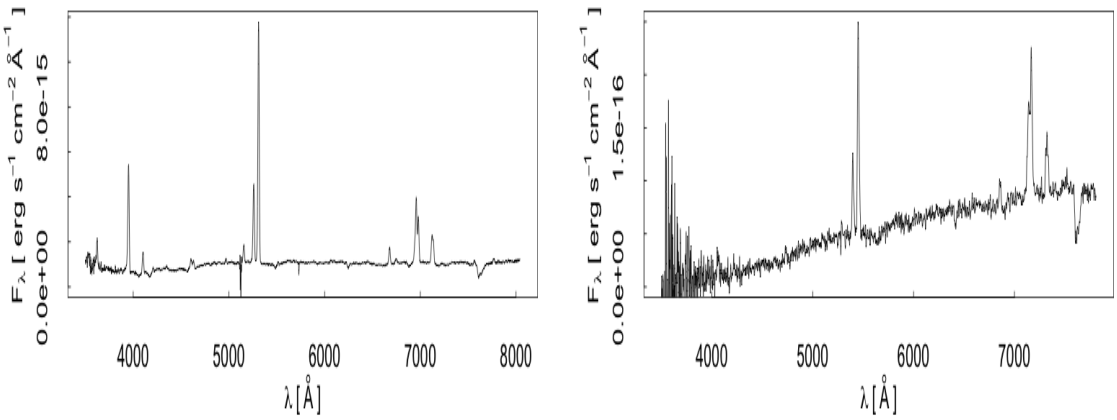



Figure 4.5: Spectra of two AGN classified as HERGs. The left spectrum is galaxy 3C033 and the right spectrum is galaxy 3C105. The spectra were taken from the NASA/IPAC Extragalactic Database (NED).

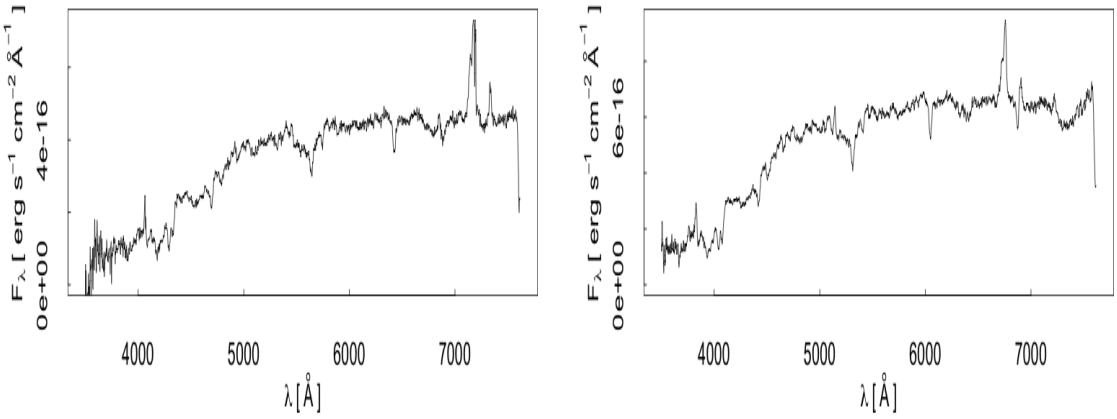



Figure 4.6: Spectra of two AGN classified as LERGs. The left spectrum is galaxy 3C388 and the right spectrum is galaxy 3C442. The spectra were taken from the NASA/IPAC Extragalactic Database (NED).

an area of 400 deg$^2$ with a flux density limit of 10 nJy over a range of frequencies. This simulation was created as part of the SKA Design Study Simulated Skies ($S^3$). Five distinct galaxy populations were identified: star forming galaxies (SFGs), star bursts (SBs), Fanaroff-Riley Class I (FRIs), Fanaroff-Riley Class II (FRIIs) and radio quiet quasars (RQQs) and average halo masses and biases were computed for each population. The very low flux limits of $S^3$ make it easily scalable to the proposed 5 $\mu$Jy flux density limit at 1.4 GHz of SKA Phase 1 and the five distinct populations of radio galaxies make this set-up appealing for forecasting the constraints on $f_{\rm NL}$ from the SKA using the multi-tracer method. This was done by Ferramacho et al. (2014) who used the redshift distributions and the mean masses of these five populations to compute auto and cross correlation angular power spectra and perform a Fisher analysis.

The Tiered Radio Extra-galactic Continuum Simulation (T-RECS; Bonaldi et al., 2019) is a more recent simulation created for the SKA with a sky area of 25 deg$^2$ and a 150 MHz - 20 GHz frequency range. Bonaldi et al. (2019) do not provide independent estimates for halo masses of galaxy populations but do compare clustering measurements for the T-RECS sources to Hale et al. (2018), and find that the observed biases for the SFGs are higher than predicted while for the active galactic nuclei (AGN) there is reasonable agreement.

In contrast to the simulated data of $S^3$ and T-RECS, Hale et al. (2018) explored the clustering of different radio-selected galaxy populations taken from 3 GHz VLA imaging over the $\sim$2 deg$^2$ COSMOS field (see Smolčić et al. 2017b) using a 5.5 $\sigma$ cut which corresponds to a mean flux density limit of 22 $\mu$Jy at 1.4 GHz. Using the angular two-point correlation functions of the radio galaxies, they estimated average bias values for the median redshifts of the samples and corresponding halo masses for these galaxy populations. See Figure 10 of Hale et al. (2018) for a comparison of the measured biases and the $S^3$ simulated bias trends as a function of redshift. The populations considered were: high and low radio luminosity SFGs, high and low radio luminosity AGN and high and moderate luminosity AGN (HLAGN and MLAGN). The radio galaxies were classified into these categories by Smolčić et al. (2017a) who gathered a wealth of multiwavelength (optical, near-UV near-infrared, mid-infrared, and X-ray) data to perform the classifications.

In order to conduct this multi-tracer analysis, distinct populations with different biases must be identified. The biases of SFGs have been found to be much lower than those of AGN (Hale et al. 2018) and this is expected because AGN are hosted by more massive galaxies than the blue SFG population. Despite this, the clustering of SFGs varies with star formation rate, galaxies with high star formation rates tend to have higher stellar masses (consider the star formation main sequence Noeske et al. (2007); Whitaker et al. (2012); Johnston et al. (2015)) which is correlated with halo mass. I chose not to distinguish between low and high

luminosity SFGs or SFG and star burst galaxies (galaxies with very high star formation rates) as these distinctions are difficult to make with only radio continuum data. The physically motivated choice of classification for the AGN is the accretion mode classification because MLAGN and HLAGN are expected to reside in different environments and evolve differently over time due to their different accretion properties. Hale et al. (2018) found that MLAGN are significantly more biased than HLAGN and therefore inhabit more massive halos. This might be because the hot gas in the most massive halos is not easily accreted onto AGN, leading to more massive halos hosting less efficient AGN. This result supports the findings of Hardcastle 2004; Tasse et al. 2008; Janssen et al. 2012; Ramos Almeida et al. 2013; Gendre et al. 2013. Unfortunately, the wealth of multi-wavelength data required for classifications of MLAGN/HLAGN to be made will not be present for large volumes of sky in the early stages of the SKA. On the other hand, it will be possible to identify the FRI and FRII galaxies from radio images,as long as the angular resolution is better than 1 arcsec (Wilman et al. 2008; Ferramacho et al. 2014). Therefore, since radio-loud MLAGN are normally associated with FRI galaxies while HLAGN are associated with FRII, for the purposes of this forecast analysis, I assume that these populations are the same although this is not strictly true. This assumption allows us to use the biases obtained from $S^3$ and Hale et al. (2018) (see section 4.4.1) and the number distributions found for FRI/FRII from the $S^3$ and T-RECS simulations (see section 4.4.2). Therefore, in this analysis, with the aim of predicting constraints that are realistic, I use only the populations: SFGs, FRI/MLAGN and FRII/HLAGN. It is important to keep in mind the caveat that FRIs and FRIIs are similar to, but not the same as HLAGN and MLAGN respectively. For instance, the HLAGN/MLAGN distinction of the VLA-COSMOS 3 GHz survey (Smolčić et al. 2017a) includes the radio quiet quasars which are included as a distinct population in Wilman et al. (2008).

An important point to note here is that when using these radio galaxy populations for a clustering analysis such as this one, angular masks will need to be applied in a similar manner to optical analyses (see Chapter 3). Such a mask should remove the effects of large galactic sources such as supernova remnants as well as very large extended radio sources such as FRII galaxies.

## 4.4 Fisher Analysis for SKA Forecasts

### 4.4.1 Halo Bias

Bias values for the halos of each galaxy population and for each redshift bin were required to perform the multi-tracer analysis (see Equations 4.2.0.1 and 4.2.0.2). In order to find these bias distributions, a mean mass $M_{cent}$ was adopted for each population, based on the

observed/simulated estimates and it was assumed that the galaxy halos spanned a Gaussian distribution of masses, $f$, with a mean $M_{cent}$ and a standard deviation of $0.2M_{cent}$. The linear halo bias distribution over a Gaussian distribution of masses, $b^i(z)$ was then calculated using the equation below:

$$b^i(z) = \int f^i(M, M^i_{cent}) b^i_L(M, z) dM, \tag{4.4.1.1}$$

where $i$ represents the galaxy population and $b^i_L(M, z)$ is the linear bias as a function of halo mass and redshfit. Ferramacho et al. (2014) found that this simplistic approach was sufficient to describe the mass distributions and that changing this form did not notably change the constraints obtained. In order to calculate the Gaussian linear bias function above, Equation 4.4.1.2 was used along with the estimated average halo masses (obtained from Hale et al. (2018) and $S^3$) and the mass variance $\sigma$ found using the Halo Mass Function (the number density of halos as a function of mass) calculator from Murray et al. (2013). The linear bias is given by:

$$b_L(M, z) = 1 + \frac{q\nu - 1}{\delta_c(0)} + \frac{1}{\delta_c(0)} \frac{2p}{1 + (q\nu)^p}, \tag{4.4.1.2}$$

where $\nu = \delta^2_c(0)/\sigma^2(M, z)$, $\delta_c(0)$ is the critical over-density for spherical collapse at redshift $z = 0$ and $\sigma(M, z)$ is the mass variance, the root mean square fluctuation in spheres which, on average, contain mass $M$ at the initial time. The accepted values of $p$ and $q$ are 0.3 and 0.75 respectively (Sheth & Tormen 1999). The angular power spectra code utilized in this analysis (Fonseca et al. 2015) required a single bias for each redshift bin and therefore I extracted the average bias in each bin from the models found and used these values.

Two sets of mean halo masses were used to produce bias functions. One set of mass values were from the radio observations of Hale et al. (2018) and the other was from $S^3$. The bias functions obtained for both the Hale et al. (2018) and the $S^3$ masses can be seen in Figures 4.7 and 4.8 respectively. The bias values found by Hale et al. (2018) for the MLAGN (FRI) were higher than those from $S^3$, values for the SFGs were slightly lower and values for the HLAGN (FRII) were significantly lower than those predicted by $S^3$. The bias values predicted by $S^3$ were produced by assuming one constant halo mass for each population and using the formalism of Mo & White (1996). These constant halo masses were chosen such that they are consistent with the clustering measurements made by Overzier et al. (2003) using the NVSS and the VLA FIRST (Faint Images of the Radio Sky at Twenty-cm) radio surveys. On the other hand, Hale et al. (2018) measure the clustering of various radio populations using 3 GHz VLA imaging over the COSMOS field which has significantly lower flux limits than the NVSS and FIRST measurements used by Overzier et al. (2003).

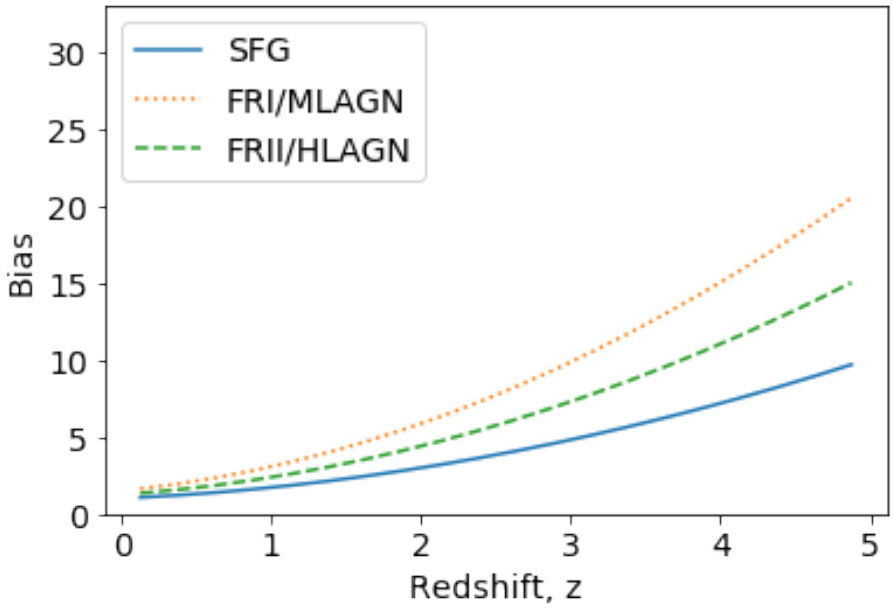



Figure 4.7: Bias redshift evolution for the different galaxy populations assuming a Gaussian distribution of masses for each population, using the central masses determined by Hale et al. (2018).

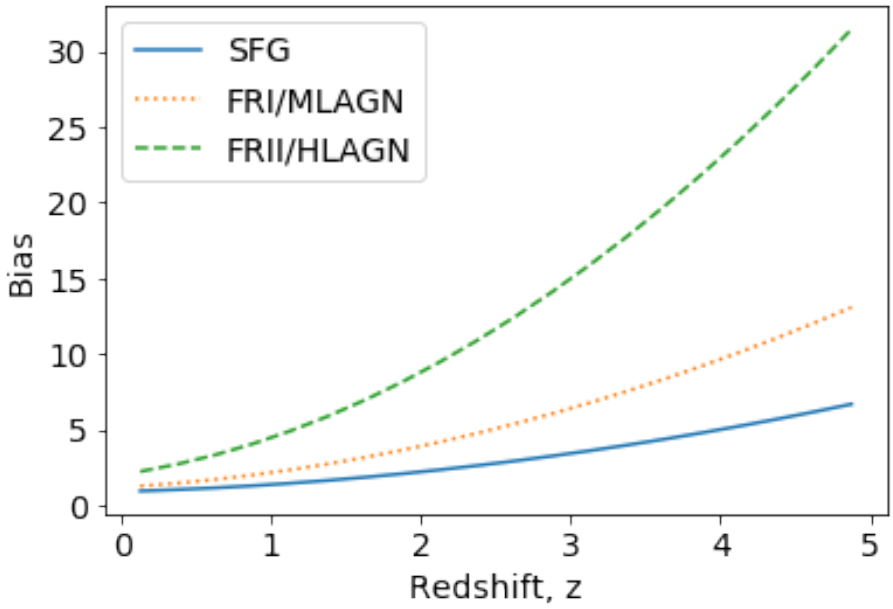



Figure 4.8: Same as figure 4.7 using the $S^3$ distribution (Wilman et al. 2008).

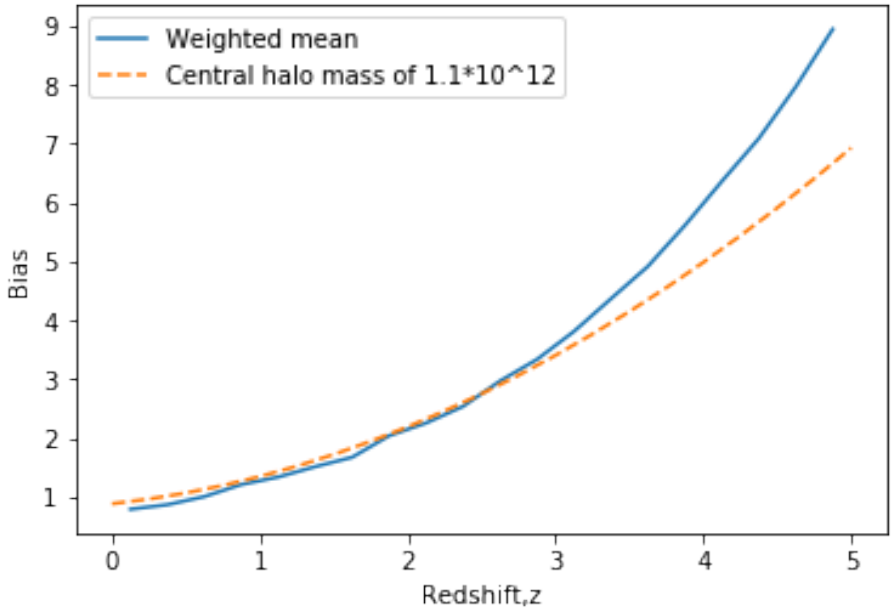



Figure 4.9: Bias redshift evolution for the combined SFG and SB galaxy populations from the SKADS simulation. The solid line is the result obtained from taking a weighted mean of the biases estimated for the individual populations and the dashed line is the best fit to this line using a central halo mass of 1.1x$10^{12}$.

An appropriate halo mass was then estimated from the bias values measured. The lower flux limit allows fainter sources to be detected at each redshift and this is suggested as the reason for the differences in the bias functions (Hale et al. 2018).

Hale et al. (2018) identified one population of star forming galaxies which consisted of both normal star forming galaxies (SFGs) and star-burst galaxies (SBs), while the $S^3$ simulation separated these populations. In this analysis I did not wish to distinguish between star forming and star-burst galaxies as they are not expected to be distinguishable with SKA data and therefore, I combined these galaxy populations of the $S^3$ simulation and used these combined biases (and redshift distributions, which will be discussed in the following section). I found the bias for the combined SFG and SB set using a weighted mean based on the numbers of each population in each redshift bin:

$$b_{comb}(z) = \frac{\sum_i n^i(z) b^i(z)}{\sum_i n^i(z)}. \tag{4.4.1.3}$$

This produced the bias function shown in Figure 4.9 (the solid line). Following this I found the central halo mass that best fit this bias distribution (see the dashed line of Figure 4.9) and used this halo mass to conduct the analysis. The fit shown in Figure 4.9 is not perfect, but is the best that could be obtained using a single central halo mass and a Gaussian

distribution around these masses. In addition, it is important to note that the vast majority of the relevant galaxies fall within the lower redshift range ($< 3$) and therefore this will be the part of the bias function that has the greatest impact. In addition, in the Fisher analysis conducted in this chapter the case in which accurate photometric redshifts are not available for the majority of the population beyond a redshift of 2 was considered, and in this case, the analysis will not in any way be affected by the mismatch in this bias function fitting.

In the case in which I use the $S^3$ biases and the T-RECS number distributions, I do not have differentiated SB and SFG distributions from T-RECS and therefore, I assumed that the ratio of SFG to SB remains the same as in $S^3$, and the same bias function was used. The best fit halo mass was found to be $1.1 \times 10^{12} M_{\odot} h^{-1}$ and the corresponding bias function is plotted in Figure 4.8. It was noticed that the combined bias of the SFG and SB of the $S^3$ simulation was similar in magnitude to the corresponding bias obtained by Hale et al. (2018).

When exploring the constraints obtained with different populations, I also considered the case in which it is not possible to differentiate the MLAGN/FRI from the HLAGN/FRII and therefore the two populations were combined and the combined bias was calculated with the method shown above.

The bias functions presented above are extrapolated to $z = 5$ while the bias measurements obtained from observations such as Hale et al. (2018) do not extend upwards of $z \sim 2$, Wilman et al. (2008) suggest a cut on the bias function at $z = 1.5$ for AGN and $z = 3$ for SFGs with a flat bias after this in order to prevent the bias from becoming unrealistically large, but there is little observational evidence for such an abrupt cut. Indeed, recent work on measuring the bias of Lyman-break galaxies at high redshifts suggest that the bias could be as high as $b \sim 8 - 11$ at $z \sim 6$ (Hatfield et al., 2018). This is therefore in the range of the assumed bias for our FRI/MLAGN and SFG samples, which are the dominant radio populations at these redshifts for the survey depth considered. Furthermore, one might expect that the host galaxies of the FRI/MLAGN population are more massive than the Lyman-break galaxies at similar redshifts, thus a higher bias may be expected. However, due to the uncertainty present at redshifts above $z \sim 2$ analyses performed with redshifts below this are more trustworthy. In this analysis I provide results for the cases in which redshift is restricted to $z < 2$ and to $z < 5$ and the constraints are compared.

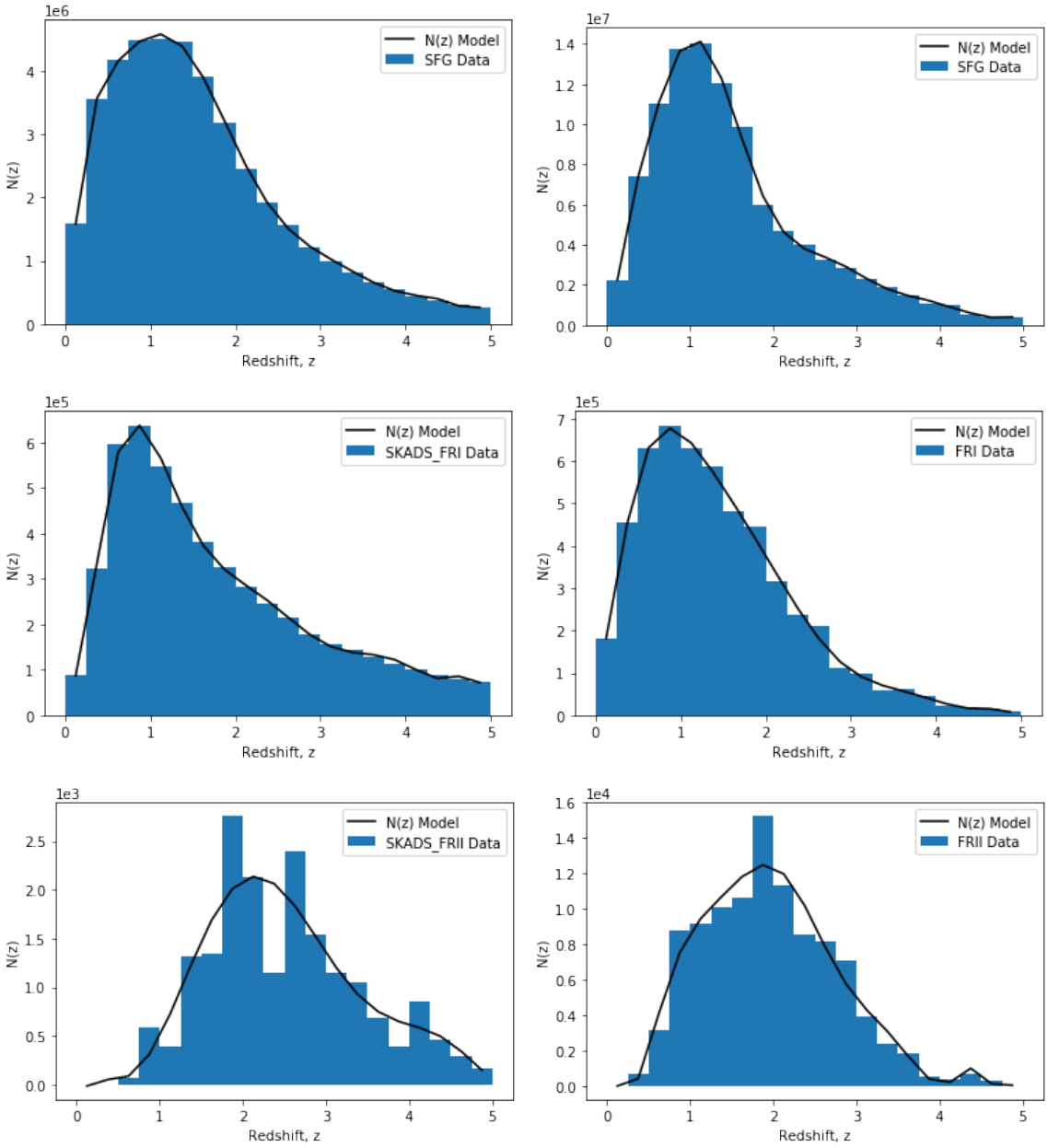



Figure 4.10: Functional fits to the histograms of the counts of galaxies across the redshift range 0 - 5 for the $S^3$ (left) and T-RECS (right) simulations and the three galaxy populations.

### 4.4.2 Redshift Distributions

#### 4.4.2.1 Fitting Functional Forms to the Redshift Distributions

Another requirement for this analysis was redshift distributions of the galaxy populations over the redshift range of interest. Due to a lack of observed radio data at the flux limits expected for SKA (the VLA COSMOS data from Smolčić et al. 2017b is limited to 22 $\mu$Jy at 1.4 GHz), the redshift distributions obtained from SKA simulations, namely, the $S^3$ and T-RECS simulations were used. The MLAGN/FRI, HLAGN/FRII and SFG galaxy populations were extracted, restricted to a flux limit of 5 $\mu$Jy (a conservative flux limit for SKA Phase 1), a redshift $z < 5$ and then rescaled to 1 steradian of sky. The SFG population for the $S^3$ simulation was obtained by combining the SFG and SB galaxy populations (T-RECS has one population including both of these galaxy types). The redshifts of the galaxies were then plotted in histograms with appropriate binning schemes and functional forms were fitted. The functional forms fitted to the different populations were chosen via trial and error. The functions that I found provided the best fits were polynomials with varying orders. All of the number distributions were fit to a function of the form:

$$N(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + a_5 z^5 + a_6 z^6 + a_7 z^7 + a_8 z^8 + a_9 z^9 + a_{10} z^{10} + a_{11} z^{11} + a_{12} z^{12}, \quad (4.4.2.1)$$

where the coefficients of some of the higher powers of $z$ are 0 for some of the populations. As an example, the function for the SKADS SFG population is a polynomial of degree 9 with the form:

$$N(z) = -944191 + 27158293z - 66307122z^2 + 93602874z^3 - 78050179z^4 + 38973905z^5 - 11806548z^6 + 2128776z^7 - 210297z^8 + 8768z^9. \quad (4.4.2.2)$$

The general functional form, Equation 4.4.2.1, was coded into the angular power spectrum code and the relevant coefficients were passed in the appropriate sections. The fits found can be seen in Figure 4.10.

The resulting redshift distributions can be seen in Figures 4.11 and 4.12 and comparisons of the FRI and FRII distributions are shown in Figure 4.13. Clearly, the T-RECS simulation predicts significantly higher numbers of SFGs than $S^3$ and these numbers seem to be more reasonable, as observations (Smolčić et al. 2017b) have suggested that $S^3$ underestimates SFGs (see also Bonaldi et al. 2016). A larger number of FRII/HLAGN was predicted by T-RECS while similar numbers of FRI/MLAGN were predicted by both simulations. The AGN populations found in both simulations were developed in very different manners. $S^3$

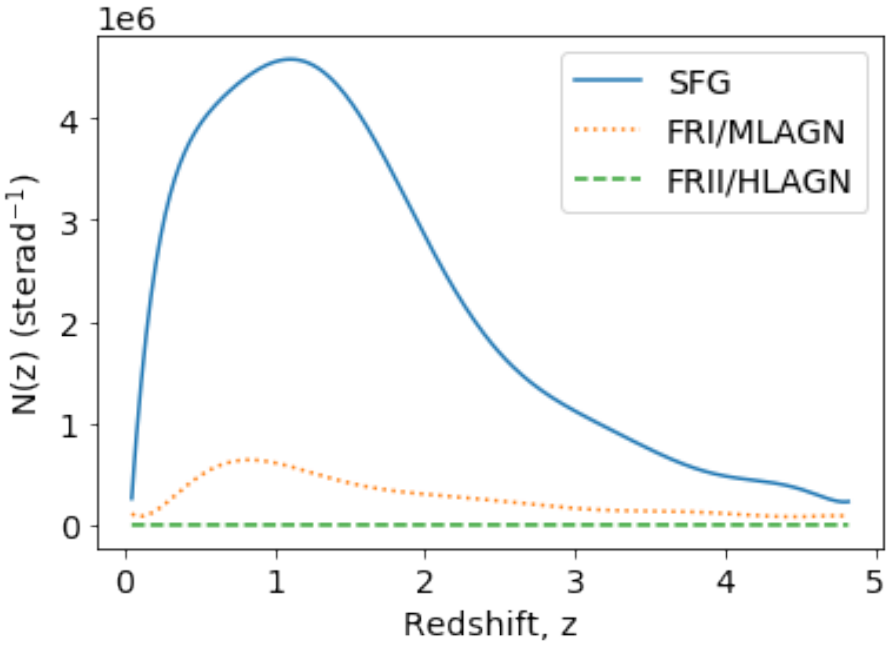



Figure 4.11: Redshift distributions for the SFG, FRI and FRII populations using the $S^3$ simulations.

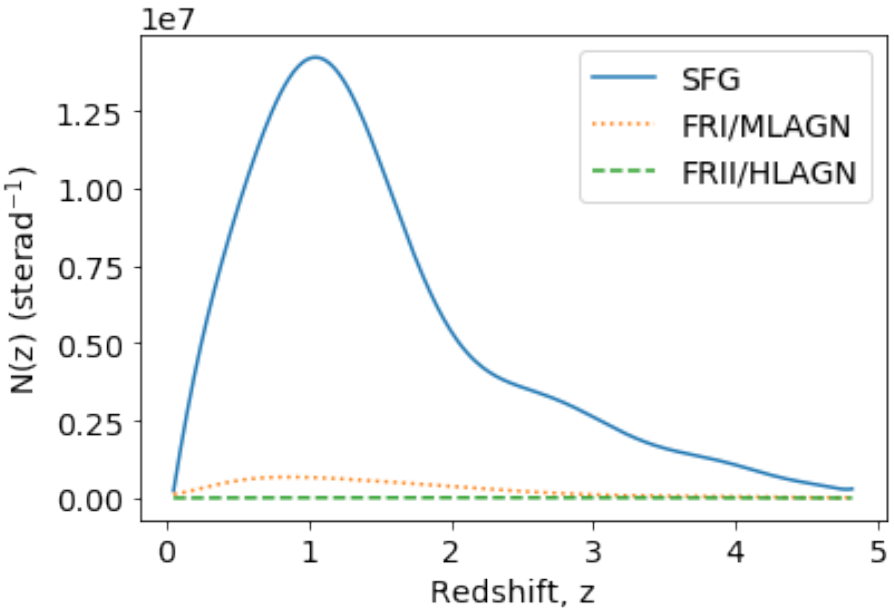



Figure 4.12: Redshift distributions for the SFG, FRI and FRII populations using the T-RECS simulations.

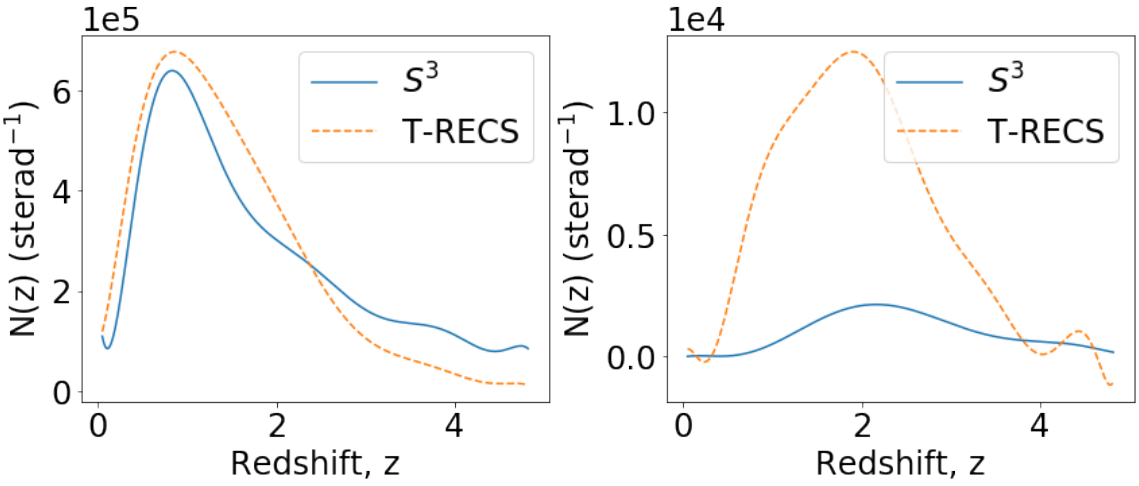



Figure 4.13: Comparison of the redshift distributions for the FRI/MLAGN (left) and FRII/HLAGN (right) populations using the $S^3$ and T-RECS simulations.

use a simple radio luminosity function evolution model from Willott et al. (2001) which provides functional forms for high luminousity sources and low luminosity sources. All of the high luminosity sources are assumed to be FRIIs and all of the low luminosity sources are assumed to be FRIs, although this is not necessarily the case. T-RECS on the other hand model the entire steep spectrum population and then define FRIs and FRIIs based on size and radio luminosity. Despite the major differences discussed above, both simulations follow similar general trends, with the population size decreasing significantly in the order of SFGs, MLAGN/FRI and HLAGN /FRII.

As previously stated, in order to compare the constraints with different possible scenarios the case in which differentiation of the MLAGN/FRI from the HLAGN/FRII is not possible is considered. Therefore the two populations were also combined and the total N(z) obtained, as with the SFG and SB populations.

#### 4.4.2.2 Obtaining Redshift Distributions in Practice

Unfortunately, since redshifts cannot be obtained from radio continuum surveys as the synchrotron radiation is featureless, it might be difficult to obtain detailed redshift distributions for large numbers of radio galaxies over large areas of sky. The emergence of large optical/near-IR surveys that will have accurate photo-zs and overlapping sky coverage will mitigate this problem. On the other hand, some smaller, deep fields such as those used by the MIGHTEE survey (Jarvis et al. 2016) will soon have all the radio continuum, redshift and multi-wavelength data necessary for such an analysis, albeit on much smaller scales.

These surveys can be used to inform the key observations necessary for the large scale $f_{\rm NL}$ studies.

The accuracy of the photometric redshifts of the survey used for cross-matching must also be sufficiently high to produce these redshift distributions, although as only distributions are necessary, precise point estimates are less important. A number of methods have been proposed for estimating redshift distributions and significant improvements on accuracy and error estimation have been made over the last few years (Almosallam et al. 2016b; Gomes et al. 2018; Soo et al. 2018; Sánchez & Bernstein 2019 and Chapter 2). In addition, apart from standard template fitting, machine learning or hybrid methods that combine machine learning and template fitting (see Duncan et al. 2018 for an example), there are now techniques for determining the redshift distributions directly (without the point estimates). Two such methods are clustering redshifts (Newman 2008; Matthews & Newman 2010; Ménard et al. 2013; Schmidt et al. 2013) and magnitude space re-weighting (Lima et al. 2008), both of which require a spectroscopic reference set. Most of these methods still require cross-matching the radio galaxies to a optical/near-IR survey, but the clustering redshift method does not require this as it involves the spatial cross-correlation between the galaxies with unknown redshifts and a distribution with known redshifts. This method has been shown to produce reasonable redshift distributions for radio galaxy surveys (Ménard et al. 2013). The accuracy of the $N(z)$ distribution will become increasingly important as smaller binning schemes are chosen for analyses such as these. The redshift bin sizes of $\Delta z = 0.25$ and $\Delta z = 1$ were used in this analysis.

An alternative option is to use the spectroscopic HI survey of the SKA but this is limited to lower redshifts $\sim 2$ and would only detect a limited sample of the star forming galaxies. Thus, it could provide a more reliable redshift distribution for the SFGs but would contribute little to the AGN number distributions.

### 4.4.3 Fisher Analysis

Fisher matrix analyses (Tegmark et al. 1997) allow the determination of the minimum errors on parameters that are obtainable using a particular method and experiment. This is possible because the precision to which a parameter can be determined depends on the dependence of the model on the parameter (the greater the model dependence, the greater the constraining ability) and on the covariance of the observed data (the greater the independence of the observed data, the more information can be extracted). In order to obtain the Fisher matrix for multiple tracers, in different redshift bins, I use the following expression, provided by

Ferramacho et al. (2014):

$$F_{\alpha\beta} = \sum_{l=l_{min}}^{l_{max}} \frac{2l+1}{2} f_{sky} Tr\Big(\frac{\partial \mathbf{C}(l)}{\partial \theta_\alpha}(\mathbf{\Gamma}_l)^{-1}\frac{\partial \mathbf{C}(l)}{\partial \theta_\beta}(\mathbf{\Gamma}_l)^{-1}\Big), \tag{4.4.3.1}$$

where $\mathbf{C}(l) = [C_l^{ij}]$ (defined in Equation 4.2.0.1) contains all the auto and cross correlation angular power spectra for all tracers in different redshift bins and $\theta_\alpha$ and $\theta_\beta$ are any two model parameters. $f_{sky}$ is the fraction of the sky expected to be covered by the radio survey, for this analysis I use a survey size of $2\pi$ steradian as the SKA is expected to cover at least half of the sky and therefore a value of $f_{sky} = 0.5$ is used. $\mathbf{\Gamma}$ is the covariance matrix and if it is assumed that the cosmic variance and shot noise are the dominant contributors to the noise (the experimental systematics are sub-dominant) then it is given by:

$$\Gamma_l^{ij} = C_l^{ij} + \delta^{ij} N^{ii}, \tag{4.4.3.2}$$

where $N^{ii}$ is the noise power spectrum for a given redshift bin and galaxy population $i$ and is given by $1/n^i$ where $n^i$ is the number of sources per steradian. These values of $n^i$ were obtained using the scaled redshift distributions discussed above.

The bias and number distributions were then used to calculate the auto and cross correlation power spectra using Equations 4.2.0.1 and 4.2.0.2. For this computation a modification of CAMB sources (Challinor & Lewis 2011) presented by Fonseca et al. (2015) was used. This code allows the user to input the redshift distributions and biases of the various tracers. In the most ideal case the redshift distribution are split into top-hat redshift bins of width 0.25 (a conservative size as the photo-z estimates/spectroscopic redshifts are expected to be much more precise than this, e.g. Chapter 2) leading to a total of 20 windows for each of the three galaxy populations. The fiducial parameters were then input and power spectra were obtained with $l_{min} = 2$ and $l_{max} = 200$ (as in Ferramacho et al. 2014). The cosmological parameters that are most degenerate with the non-Gaussianity signal are dark matter density $\Omega_{CDM}$, the Hubble parameter $H_0$ and the amplitude of the primordial fluctuations $A_s$ and therefore these are included in the Fisher analysis. I use the parameters obtained by Planck 2015 (Planck Collaboration et al. 2016b) with CMB temperature and polarization power spectra as well as lensing and external (Baryon Acoustic Oscillations (BAO), supernova Joint Light-curve Analysis and $H_0$) measurements. The other cosmological parameters were used in producing the angular power spectra. The parameters used for this analysis (the cosmological parameters and the estimated central masses of the galaxy populations) and their fiducial values are given in Table 4.1.

The next step was to obtain the derivatives of these power spectra with respect to each parameter. Some of the derivatives were computed analytically while others were

Table 4.1: Fiducial values chosen for the parameters in this model. 'Hale biases' and '$S^3$ biases' are the bias values derived from the measurements made by Hale et al. (2018) and from the $S^3$ simulation respectively. The masses are in units of $M_{\odot}h^{-1}$.

| Parameter | Fiducial Value | |
|---|---|---|
| $\Omega_{\rm CDM}$ | 0.26 | |
| $H_0$ | 67.74 | |
| $ln(10^{10}A_s)$ | 3.064 | |
| $f_{\rm NL}$ | 0 | |
| | Hale Halo Masses | $S^3$ Halo Masses |
| $M_{\rm cent}^{\rm SFG}$ | $4 \times 10^{12}$ | $1.1 \times 10^{12}$ |
| $M_{\rm cent}^{\rm MLAGN}$ | $3.5 \times 10^{13}$ | $1 \times 10^{13}$ |
| $M_{\rm cent}^{\rm HLAGN}$ | $1.5 \times 10^{13}$ | $1 \times 10^{14}$ |
| $M_{\rm cent}^{\rm AGN}$ | $3.46 \times 10^{13}$ | $1.01 \times 10^{14}$ |

computed numerically. In order to compute the numerical derivatives, increments of each of the relevant parameters were input while keeping the remaining parameters fixed and a 5-point stencil method was applied. This is a numerical method that estimates the derivative of a function (in this case the angular power spectrum) given the values of the function evaluated at the relevant value (x) plus or minus an extra increment (h and 2h). The formula for this method is given below:

$$f'(x) \approx \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}. \tag{4.4.3.3}$$

The covariance matrix, $\Gamma$, was then computed using the $C_l$ results with the fiducial cosmology according to Equation 4.4.3.2. The Fisher matrix was then computed using Equation 4.4.3.1 with $f_{sky}$ = 0.5. The constraint on the $f_{\rm NL}$ parameter was obtained by inverting the Fisher matrix and extracting the $f_{\rm NL} \times f_{\rm NL}$ component.

## 4.5 SKA Forecasts - Results and Analysis

In order to understand the relationships between the constraints that can be obtained with different sized redshift bins (corresponding to photometric redshift quality) and different redshift ranges (which will depend on the performance of the SKA as well as the availability of multi-wavelength data with photometric redshifts to cross match to the radio data) a number of cases were considered:

- Redshift bins of width 0.25 and a range of $0 < z < 5$

- Redshift bins of width 1 and a range of $0 < z < 5$
- Redshift bins of width 0.25 and a range of $0 < z < 2$

The first of these cases represents the most optimistic view, implying that there will be complete and accurate photometric redshift coverage to match to the radio galaxies up to a redshift of $z \sim 5$. This is unlikely to be the case due to obstacles such as dust obscuration at high redshifts which could introduce artificial selection effects (with brighter objects more likely to be detected) and impact our measurements of $f_{\rm NL}$. In addition, there are limited spectroscopic redshifts at high redshifts, making the training of machine learning photometric redshift estimation methods as well as the redshift distribution estimation methods more difficult and less accurate. For these reasons the case in which photometric redshifts are available up to $z \sim 5$ but the photometric redshift accuracy is much lower was also evaluated. Much larger bins of width 1 were used to represent this decrease in redshift accuracy. Finally, the much more realistic case of obtaining galaxies with photometric redshifts only up to a redshift of 2 was considered. Below this redshift it is more likely that multi-wavelength data, and correspondingly photometric redshifts will be available for cross matching to most of the radio sources. In addition, in this redshift range it is also possible to obtain some redshifts spectroscopically from HI line measurements. Photometric redshift accuracy is also expected to be higher in this redshift range, making the bin width of 0.25 a conservative choice. In addition to obtaining better redshifts below $z \sim 2$, the bias functions measured/simulated are also more accurate below these redshifts. As a result, constraints obtained for this case of restricted redshift are much more realistic than other cases.

For each of these cases the constraints on $f_{\rm NL}$ with the different combinations of either the $S^3$ or Hale et al. (2018) biases and the $S^3$ or T-RECS $N(z)$ distributions were evaluated in order to obtain a range of constraints depending on the range of possibilities for the radio source populations. In addition, the result of having 3 independent distributions: SFG+SB, MLAGN/FRI and HLAGN/FRII was compared to the case in which MLAGN/FRI and HLAGN/FRII are not distinguished and there are 2 populations: SFG+SB and AGN. The morphology of radio galaxies is not a discrete distribution of FRI and FRII objects, but a continuous distribution of morphologies, with many objects that are difficult to classify as either an FRI or a FRII. This distinction also becomes increasingly difficult as redshift increases and the images obtained resolve less structure. Therefore, for the most conservative case it is assumed that this distinction is not possible. Moreover, the number density of the HLAGN in both redshift distributions was relatively low in comparison to the other populations and therefore I decided to compare the constraints obtained when the SFG+SB

Table 4.2: Forecasts on $f_{\rm NL}$ 1-$\sigma$ errors obtained using a multi-tracer Fisher analysis with redshift bins of size $\Delta z = 0.25$ and a range of $0 < z < 5$.

| | $S^3$ $N(z)$ | T-RECS $N(z)$ |
|---|---|---|
| $S^3$ biases | 1.90 | 1.88 |
| Hale biases | 1.54 | 1.56 |

Table 4.3: Forecasts on $f_{\rm NL}$ 1-$\sigma$ errors obtained using a multi-tracer Fisher analysis with redshift bins of size $\Delta z = 1$ and a range of $0 < z < 5$.

| | | SFG+SB MLAGN HLAGN | SFG+SB MLAGN | SFG+SB AGN |
|---|---|---|---|---|
| $S^3$ biases | $S^3$ $N(z)$ | 2.23 | 2.27 | 2.26 |
| | T-RECS $N(z)$ | 2.35 | 2.47 | 2.46 |
| Hale biases | $S^3$ $N(z)$ | 1.83 | 1.84 | 1.84 |
| | T-RECS $N(z)$ | 1.98 | 1.99 | 1.98 |

and MLAGN populations are used to when the SFG+SB and AGN (where the HLAGN is included in the AGN) populations are used.

The 1-$\sigma$ constraints obtained for the three cases: small redshift bins with a large redshift range, large redshift bins with a large redshift range and small redshift bins with a small redshift range are presented in Tables 4.2, 4.3 and 4.4 respectively. It was found that for all cases (of redshift bin size and range as well as combinations of galaxy types) the biases from Hale et al. (2018) led to significantly tighter constraints than those from $S^3$, while the different redshift distributions of the two simulations led to less notable differences. This means that if the biases calculated by Hale et al. (2018) are nearer to the truth, tighter constraints than previously expected (Ferramacho et al. 2014) will be possible with this method. The Hale et al. (2018) biases may have caused this effect because the separation in the biases of the FRI and SFGs is much greater than for $S^3$ (compare Figures 4.7 and 4.8) and the numbers of SFG and FRI are much greater than the number of FRII present.

Table 4.4: Forecasts on $f_{\rm NL}$ 1-$\sigma$ errors obtained using a multi-tracer Fisher analysis with redshift bins of size $\Delta z = 0.25$ and a range of $0 < z < 2$.

| | | SFG+SB MLAGN HLAGN | SFG+SB AGN |
|---|---|---|---|
| $S^3$ biases | $S^3$ $N(z)$ | 6.53 | 6.58 |
| | T-RECS $N(z)$ | 5.39 | 5.82 |
| Hale biases | $S^3$ $N(z)$ | 4.61 | 4.63 |
| | T-RECS $N(z)$ | 4.07 | 4.08 |

Therefore, although the difference between the bias functions of the populations is greater for $S^3$, with the FRII galaxies having very high biases, the very limited number of FRIIs and larger numbers of FRIs and SFGs lead to the Hale et al. (2018) biases producing tighter constraints. I note that ideally, the $N(z)$ distribution parameters would be marginalized over in this analysis. However, the differences in the constraints caused by the $N(z)$ distributions is subdominant to the differences due to the biases using the $S^3$ simulation and Hale et al. (2018). Therefore, the fact that the number distributions have substantially different distributions (see Figures 4.11 and 4.12) implies that small variations in the number distributions will not result in notable differences in constraints.

The choice of simulation for determining the redshift distributions had varied influences on the resulting constraints across redshift bin size and range. For the case with $z < 2$ (Table 4.4), the T-RECS number distributions produced significantly tighter constraints while for the case of larger redshift bins but a range of $0 < z < 5$ (Table 4.3) the $S^3$ number distributions resulted in tighter constraints. This pattern is likely because the numbers of MLAGN in the T-RECS simulation tapers off to much lower numbers beyond $z = 2$. This causes the MLAGN in the $S^3$ simulation to contribute more to the $0 < z < 5$ measurement than those in the T-RECS simulation, while the MLAGN in the T-RECS simulation contributes more to the $z < 2$ measurement than those in the $S^3$ simulation. On the other hand, in the most ideal case with small redshift bins and a large redshift range the differences induced by the different redshift distributions are insignificant.

Excluding the HLAGN (see Table 4.3) and joining the MLAGN and HLAGN (Tables 4.3 and 4.4) both reduced the constraints (increased the 1-$\sigma$ errors) by small but notable amounts with the combined MLAGN and HLAGN performing marginally better than removing the HLAGN entirely. This is expected as the HLAGN population is very small and therefore does not contribute much to the constraints when added to the MLAGN. The addition of the HLAGN (combined with the MLAGN to form 'AGN') does increase the number of galaxies to a small extent, which decreases the shot noise, while also contributing a greater ratio of higher redshift galaxies, these effects lead to the marginal improvement observed. Removing the HLAGN as an independent tracer was found to produce a greater reduction in constraining power for the combination of the $S^3$ biases and the T-RECS $N(z)$ distributions than for the other cases. This is likely due to the very large bias values assigned to the HLAGN by the $S^3$ simulation coupled with the larger numbers of HLAGN predicted by the T-RECS simulation. Another important observation is that the increases in errors imposed by the removal of the HLAGN as an independent tracer are much less significant than the variation between the bias and number distribution configurations. This is clearly observed in Tables 4.3 and 4.4 and highlights that the SFG and MLAGN populations are

contributing the majority of the constraining power. This implies that if image resolution or multi-wavelength data is not sufficient for distinguishing between MLAGN/FRI and HLAGN/FRII then this is not a major problem as it will not significantly decrease the possible constraints.

When comparing the cases with different redshift bin widths it was found that increasing the bin size increased the 1-$\sigma$ errors on $f_{\rm NL}$ on all the bias and $N(z)$ configurations by a factor less than 1.5 (compare Table 4.2 to Table 4.3) while reducing the redshift range to $z < 2$ had a much more significant impact with a factor nearer to 3 (see Table 4.2 and Table 4.4). This implies that while having accurate photometric redshift estimates that allow smaller bins is important it is not as vital to adding constraining power as having deeper surveys with galaxies and cross-matched photometric redshifts going to higher redshifts. The reduced redshift led to more significant reductions in constraining power because the volume of the survey is decreased, thus reducing the number of large scale measurements contributing to the angular power spectra. Despite this, as previously discussed, it is not expected that these data will be available for $z \gtrsim 2$ in the near future, and the bias functions have not been measured at higher redshifts. Therefore, the results in Table 4.4 are taken as the expected limits of constraints with the SKA Phase 1 continuum surveys. Note here that our choice of $f_{sky} = 0.5$ was conservative for the SKA and it is possible that greater coverage will be obtained, as the errors scale with $\sqrt{f_{sky}}$ this will lead to tighter constraints. Also note that these fisher matrix constraints are the minimum possible errors and that their calculation assumes that any systematic errors inherent in the experiment have a much smaller impact than the cosmic variance or shot noise contributions.

It should be noted that this was a conservative analysis, using only the three or two galaxy populations that are most likely to be correctly distinguished. If the multi-wavelength data available becomes sufficient to distinguish star burst galaxies from the other star forming galaxies and to similarly differentiate the radio quiet quasars then additional tracers can be utilized and the results will be further improved, as evaluated in Ferramacho et al. (2014).

## 4.6 Conclusions

In this chapter I presented an analysis of the constraints on primordial non-Gaussianity that could be obtained using the SKA radio continuum surveys. This chapter serves as an update to the analysis undertaken by Ferramacho et al. (2014). I investigated the possibility of using multiple radio galaxy populations, with different halo mass and bias properties to perform a multi-tracer analysis, thus reducing the effect of cosmic variance and increasing the constraining power that can be obtained. The galaxy populations considered were

star forming galaxies, moderate luminosity AGN (or FRIs) and high luminosity AGN (or FRIIs). The MLAGN and HLAGN were considered to be similar enough to the FRI and FRII populations that the classifications were used interchangeably based on the simulation/observations utilized. In contrast to Ferramacho et al. (2014), star forming galaxies and star burst galaxies are not distinguished as this may not be observationally feasible at the required level of precision.

I perform a Fisher matrix analysis following the formalism of Ferramacho et al. (2014) which utilizes the auto- and cross-correlation angular power spectra of the multi-tracers. These power spectra require the redshift and bias distributions of the radio galaxy populations. I use the redshift distributions of radio galaxy populations given by two SKA simulations: the $S^3$ simulation and the more recent T-RECS simulation and they were re-scaled to the appropriate area and flux limit. For the bias distributions I used the average halo masses obtained from $S^3$ and from Hale et al. (2018) who use VLA COSMOS-field observations. Gaussian distributions around these masses were then assumed and biases were calculated using an analytical expression for the linear bias. I considered all of the combinations of the bias and redshift distributions as well as different redshift ranges (dependent on the depth of the continuum survey and the availability of visible/near-IR data with reliable photometric redshifts that can be cross-matched to the radio sources) and redshift bin sizes (dependent on photometric redshift accuracy). In addition to treating the SFG, MLAGN/FRI and HLAGN/FRII as independent populations I also considered the combined MLAGN and HLAGN population to determine the effect on the constraints if the MLAGN and HLAGN cannot be differentiated. I take the case of small redshift bins (width = 0.25) and a small redshift range ($0 < z < 2$) as the most realistic case for SKA observations and this provides a range in the 1-$\sigma$ errors of 4.07 to 6.53 if the AGN populations are differentiated and 4.08 to 6.58 if they are not. The similarity between these is due to the fact that the high-luminosity AGN are much rarer than the dominant AGN population at lower luminosities. The results with the observed biases (Hale et al. 2018) and the more recent simulations (T-RECS) provide the lower bounds of these ranges: 4.07 and 4.08 with 3 populations and 2 populations respectively. These constraints will surpass the existing tightest constraints ($\sim 5$ obtained with Planck 2015 and 2018 data; Planck Collaboration et al., 2016b, 2019) but will also provide an independent precise measure of $f_{\rm NL}$, obtained from the large scale structure instead of the CMB bispectrum. These results also indicate that if redshift information is available for the galaxy populations to higher redshifts, constraints will improve significantly, and with a redshift range of $0 < z < 5$ the $1\sigma$ error on $f_{\rm NL}$ will fall between 1.5 and 2, representing unprecedented precision on the measurement of $f_{\rm NL}$ and approaching the target of 1. On the other hand, some limitations of this analysis are that

the masses of the galaxy populations might have non-Gaussian distributions or different standard deviations than the ones assumed in this analysis. In addition, the bias evolution above redshift $z \sim 2$ is currently very uncertain. Moreover, the redshift distributions used are based on simulations, which are constrained by observations, but they still might not be very accurate, particularly at the higher redshifts where the lever-arm for the constraints on $f_{\rm NL}$ are strongest.

It is clear that in order to use wide-area radio continuum surveys for constraining the influence of non-Gaussianity on the large-scale structure, then robust measurements of the distribution of biases for the different populations, along with their redshift distributions are needed. To some extent, this is underway, as I have used the observed constraints from the VLA-COSMOS survey in order to estimate the bias of the variety of radio source populations. However, VLA-COSMOS covers a relatively small area, and does not reach the depth of the planned "all-sky" SKA continuum surveys. Therefore, the imminent MeerKAT International Giga-Hertz Tiered Extragalactic Exploration (MIGHTEE; Jarvis et al., 2016) survey will certainly provide much better information on the bias (through a clustering analysis covering many square degrees, as opposed to a single $\sim 2$ deg field, and to a depth similar to that of the all-sky SKA survey. A key element of MIGHTEE is it also covers areas of the extragalactic sky with excellent multi-wavelength data, thus more accurate redshift distributions will also be available based on the photometric redshifts in these fields (similar to VLA-COSMOS now).

The combination of these multi-wavelength data, coupled with deep radio data may also allow the characterisation of the sources into the populations discussed here using morphological information. Such an effort could potentially supply the necessary training sample for a variety of machine learning algorithms currently being tested on radio data (e.g Lukic et al., 2018; Glaser et al., 2019), and then applied to wider field data where the data extent and quality is not comparable to in these deep fields.

As previously highlighted, multi-wavelength data for the relevant radio galaxies will be vital for analyses such as these. At redshifts $z \sim 5$ detection limits coupled with seeing and transparency effects could lead to patchy multi-wavelength detections. This might be avoidable if additional work such as doing follow up observations or measuring the flux based on the radio position is done.

Thus, although constraining $f_{\rm NL}$ requires large cosmological volumes, information from the deep narrow surveys will be crucial in planning the final strategy.

Furthermore, joint analyses with other multi-tracer probes (other than radio continuum) such as HI intensity mapping or low redshift galaxy surveys would likely lead to further improvements on $f_{\rm NL}$ constraints.

# Chapter 5

# Conclusions and Future Work

Photometric redshifts are a vital tool for a number of astronomical measurements and this thesis has explored methods of improving the estimation of photometric redshifts and some ways in which these improved redshifts can be used for making cosmological measurements. In this final chapter I provide a short summary of this work followed by a discussion of future projects or improvements that would build on or follow logically from this work.

## 5.1 Summary

### 5.1.1 Chapter 2: Improving Photo-z Estimations

In Chapter 2 the GPz algorithm for photometric redshift estimation was introduced. This was followed by an analysis of some methods of improving the results obtained using this algorithm. One method investigated was introducing additional near-IR and angular size features. This analysis was performed using spectroscopy from the Galaxy and Mass Assembly Data Release 2 along with corresponding Sloan Digital Sky Survey visible ($ugriz$) photometry and angular size measurements and UKIRT Infrared Deep Sky Survey Large Area Survey near-IR ($YJHK$) photometry. These additions were found to improve results by $\sim 15-20$ per cent and thus it was concluded that these data should be included when available. One exception to this is that size data should be excluded when the measurement being made depends on galaxy size/shape such as gravitational lensing. As weak lensing relies on size/shape measurements, if these are used to help with photo-zs it may result in unaccounted for degeneracies between shape measurements and the photo-zs. This would bias the inferred parameters in ways that are difficult to account for.

Next, a post-processing method of shifting the binned photometric redshifts by small amounts based on Q-Q plots was applied. These shifts were determined such that they minimized the deviation of the Q-Q plot from an ideal Q-Q plot which corresponds to the

case of the photometric redshift PDFs appropriately representing the spectroscopic redshifts. This method was found to substantially improve the bias of the estimated photometric redshifts (by $\sim 40$ per cent). Following this, the results of GPz using SDSS photometry was compared with those obtained using HSC photometry and it was found that the results were significantly improved by more accurate photometry, implying that work should continue to be done to improve the quality of the photometric data available.

The next section of Chapter 2 presented an application of a method of determining redshift distributions $N(z)$ using a reference sample of galaxies with known redshifts. The reference galaxies are re-weighted in order to represent the magnitude space of the galaxies with unknown redshifts. This was performed using HSC photometric data and the COSMOS 30-band sample was used as the reference sample. This direct determination of redshift distributions is an alternative to using point estimates and obtaining the redshift distributions from these. This method can also be used to quantify the spread in redshift of the galaxies placed into a redshift-bin using a point estimate method.

### 5.1.2 Chapter 3: Measuring the BAO using SDSS galaxies and GPz Photometric Redshifts

A galaxy clustering analysis using SDSS data and photo-zs from GPz was presented in Chapter 3. This analysis was carried out using two galaxy samples: one which included all galaxy types and another with only luminous galaxies, defined by the selection criteria used for the BOSS CMASS galaxy sample. A number of modifications were made to these samples in order to remove possible systematic errors. In particular, cuts on the magnitude were applied in order to make the sample complete to a particular magnitude, weights based on the galaxy density of regions with different sky backgrounds were calculated and applied and all galaxies and randoms within a given radius from any bright star (defined as stars with $i_{mod} < 19.9$) were removed. The removal of galaxies near to bright stars was found to cause a notable improvement to the correlation function.

The GPz algorithm was trained using SDSS data with spectroscopic redshifts. The training set was split based on whether or not the galaxy was a luminous galaxy as well as whether or not there were WISE mid-infrared measurements for the galaxy. GPz was trained with the 5 optical SDSS magnitudes, the two WISE magnitudes, where available, the magnitude uncertainties and some morphological data. The resulting root mean squared errors for the photo-z estimates were 0.044 and 0.050 for the luminous galaxies and 0.064 and 0.076 for the other galaxies in the redshift range of interest (0.4 < z < 0.7). The first metrics correspond to the result when WISE data was present and the second ones correspond to the result without WISE data and the majority of the target galaxies have

WISE data. The result for the luminous galaxies can be compared to the value of 0.0585 obtained by Ross et al. (2011) using ANNZ.

Angular correlation functions were measured using the CUTE algorithm which utilizes the Landy-Szalay estimator along with a nearest neighbour searching algorithm and a pixelization option to increase efficiency. Randoms fifty times the size of the datasets were made by sampling randomly from the imaging mask and randomly assigning redshifts from the real galaxies to the randoms in order to produce the same redshift distribution. A jack-knife resampling method was used to measure the covariances and determine the error bars for the angular correlation functions. This was done using a weighing method due to the irregularity of the imaging mask used in this analysis.

The measured angular correlation functions which had a visible peak were then fitted with the sum of a power law and a Gaussian following the method suggested by Sánchez et al. (2011) and the angular scale of the BAO peak was determined. The angular correlation function obtained with the luminous galaxy sample in the redshift range $0.4 < z < 0.7$ had a clear peak with a best fit Gaussian mean at $3.37 \pm 0.13°$. This result will be shifted to a larger angle when corrected for projection effects and is consistent with existing measurements and expected results based on fiducial cosmologies ($\sim$3.97°-4.22°, see Alcaniz et al. (2016) and Carnero et al. (2012) for figures summarising past results). Correlation function were also found for this galaxy sample in redshift bins of width $\Delta z = 0.1$. The Gaussian means of the resulting peaks for the redshift bins in order of increasing redshift were $4.11 \pm 0.44°$, $3.84 \pm 0.11°$ and $3.45 \pm 0.43°$ which followed a pattern of decreasing angular scale as expected. The peaks in the first and third redshift bins were not very robust and were difficult to fit (refer to the large errors associated with these fits) but the $0.5 < z < 0.6$ peak was clear and provided a result that was also consistent with expected results for a mean redshift of 0.55. The BAO peak positions in the other redshift bins were also found to be reasonable when compared to predicted results assuming a $\Lambda$CDM cosmology.

The sample of galaxies of all types on the other hand did not produce a visible BAO peak. As a result, I applied a magnitude cut of $i < 21$ to the sample to determine whether the peak would be present for brighter galaxies with better photo-zs. The correlation function for the redshift range $0.4 < z < 0.7$ did not show a visible peak but the redshift bins $0.5 < z < 0.6$ and $0.6 < z < 0.7$ resulted in faint peaks at the expected scales. The positions of these peaks were $4.09 \pm 0.16°$ and $3.44 \pm 0.13°$ respectively. These are also consistent with the expected results assuming a $\Lambda$CDM cosmology.

This result implies that the photo-zs from GPz are sufficiently accurate to make cosmological measurements such as the BAO without being restricted to specific galaxy types. This will be very useful as current and future photometric surveys will detect very large

numbers of galaxies over large volumes without spectra. At higher redshifts in particular luminous galaxies that normally have well defined 4000Å break and are normally used for clustering analyses will be rarer as most galaxies will have some star formation in progress. As a result, clustering measurements such as measuring the BAO will need to be done with other galaxy types as well.

### 5.1.3 Chapter 4: Forecasting Primordial Non-Gaussianity Constraints using the Multi-tracer Method and Radio-Selected Galaxies

Chapter 4 contained a forecast analysis of the constraints on primordial non-Gaussianity that could be obtained using the SKA radio continuum surveys and the multi-tracer method. Three populations that are very different tracers of dark matter were used: SFGs, MLAGN/FRI and HLAGN/FRII. The fisher forecast analysis for the multi-tracer method assuming that only photometric redshifts are available involves the auto and cross correlation angular power spectra of the tracers in different redshift bins. These power spectra require the redshift and bias distributions of the radio galaxy populations. The halo biases were calculated using both simulations from $S^3$ and observations of the VLA COSMOS-field from Hale et al. (2018). The effects of these bias functions on the constraints were later compared. The redshift distributions were obtained using two simulations: $S^3$ and T-RECS and their effects on the constraints were similarly compared.

Constraints were calculated for all of the combinations of the bias and redshift distributions. Different redshift ranges (which would depend on the depth of the continuum survey as well as the availability of visible/near-IR data with reliable photometric redshifts that can be cross-matched to the radio sources) and redshift bin sizes (dependent on photometric redshift accuracy) were also compared. The situation in which the MLAGN and HLAGN cannot be clearly differentiated was also considered by combining them and treating them as one distribution.

The most realistic case, that of small redshift bins (width = 0.25) and a redshift range ($0 < z < 2$), provides a range in the 1-$\sigma$ errors of 4.07 to 6.53 if the AGN populations are differentiated and 4.08 to 6.58 if they are not. This result highlights the fact that the additional information provided by the differentiation of the AGN population is minimal and this is explained by the low number densities of the HLAGN population. The combination of the observed biases (Hale et al. 2018) and the more recent simulations (T-RECS) leads to the lower bounds of these ranges: 4.07 and 4.08 with 3 populations and 2 populations respectively. These constraints would surpass the existing tightest constraints from Planck ($\sim 5$). If the redshift range available was greater more large scales would be measured and if this range extends to $z \sim 5$ the forecasted $1\sigma$ error will reduce significantly to between

1.5 and 2 representing unprecedented precision on the measurement of $f_{\rm NL}$. Any of these measurements will be useful as they will provide an independent measure of $f_{\rm NL}$ obtained from the LSS instead of the CMB.

The limitations involved in this analysis were related to the determination of the bias redshift evolution and the redshift distributions. In particular, the bias functions were extrapolated to higher redshifts than the range for which observations have been made, making higher redshift forecasts less reliable. In addition, the halo masses of the galaxy populations may not have Gaussian distributions or standard deviations similar to what was assumed to estimate the bias functions given the mean halo masses. The redshift distributions used were also based on simulations which might not be very accurate at all redshifts, particularly at $z > 3$ where observational constraints are poorer.

## 5.2 Future Work

### 5.2.1 Improvements to GPz

One disadvantage of the GPz algorithm and most machine learning algorithms in general is the inability to extrapolate them to regions without training data and obtain accurate results. Specifically, in regions of feature space (colours, morphological features, etc.) with limited training data the predictions are often biased. In both Chapters 2 and 3 it was observed that good fits were obtained for redshift ranges that had a high data density in the training set while lower/higher redshift estimates tended to be over/under estimated, biasing these results. One way of ameliorating this problem is to shift the photo-z estimates in these biased bins based on the Q-Q plots as presented in Chapter 2, but this approach is limited to adjusting the results obtained for the feature space included in the training/testing set as spectroscopic redshifts are necessary for this calibration. It is possible that the target set will have a different redshift/colour distribution to the training set and therefore this correction will be insufficient in some regions. An improved and more thorough method for dealing with this problem is to up-weight the training galaxies in regions of feature space that have low density in the training set if they have a high density in the target set. This can be thought of as a combination of the methods of both the GPz point estimation and the magnitude re-weighting technique. The weights are calculated using the magnitude reweighting technique discussed in Chapter 2 such that the training galaxies that are surrounded by larger numbers of target galaxies (in feature space) are assigned larger weights. These weights will then have a greater impact on the training of the algorithm via the cost sensitive learning mechanism of GPz, effectively allowing the GPz model to train on a sample that is representative of the target dataset.

Assuming the above implementation leads to significantly improved photo-z estimates for many target galaxy samples, it is still possible that a portion of the target galaxy feature space will lie in a region in which there are very few or no training galaxies. In this case it will not be possible to obtain accurate estimates for these regions using GPz, or any other machine learning method. In such a situation, incorporating a template fitting method is a viable solution. As discussed in Chapter 1 template fitting methods can be extrapolated well and are therefore ideal for such situations assuming that the template library used is sufficient. Duncan et al. (2018) explore the combination of GPz and template fitting methods using a hierarchical Bayesian method. This was applied to an AGN population and found to reduce scatter by up to a factor of $\sim 4$ compared to a template-only method, proving that this is a viable method for improving estimates. Such a combination should be considered for obtaining the photo-zs for the present and upcoming photometric surveys.

For the analyses presented in this thesis, photo-zs were obtained using GPz with magnitude uncertainties included as additional features. The algorithm has recently been updated to allow these uncertainties to be accepted as input noise values instead of as additional features. This additional input noise then contributes (along with the model variance and noise variance) to the uncertainty in the photo-z estimate. This is a better way of utilizing this uncertainty information as they are not expected to influence the training in the same way as the main features.

Another modification to the GPz algorithm currently being implemented is the creation of more complex PDFs instead of the simple Gaussian PDFs normally produced. More detailed PDFs will provide additional information about the redshift estimate, such as multi-modality and interesting distribution shapes. One way in which this can be done is by training the algorithm multiple times and combining the resulting PDFs.

### 5.2.2 Clustering Measurements using GPz

One straightforward way in which this analysis could be improved would be to improve the photo-zs used. This could be done by implementing some of the changes to the GPz algorithm discussed above in Section 5.2.1. In particular, the weighting based on colour space could have significant effects on the accuracy obtained by improving the estimates in regions with little training data. Using the photometry uncertainty as an input noise instead of as additional features could also provide improvements. Similarly, the post-processing method of applying shifts to the photo-zs in redshift bins based on their Q-Q plots presented in Chapter 2 could be implemented to reduce the bias in less populated regions of feature space. This would be particularly relevant for the LG sample which was concentrated in the range $0.4 < z < 0.7$ and resulted in biased estimates above and below this region.

Another way in which this analysis could be improved is by searching for quasars and stars that were not correctly identified in the SDSS catalogs and therefore remained as contaminating objects in the galaxy sample used for this analysis. Ross et al. (2011) attempt to separate the quasars and stars by using the ANNz algorithm as a classifier and this can be implemented in a similar way using 5.2.1.

Improvements on the way in which the BAO peak is measured from the correlation functions can also be made. One improvement would be to use a model independent method for detecting the true BAO peak instead of choosing the peak that is near to the scale expected based on a cosmological model. Some methods for doing this were discussed in Chapter 3 and included making small random adjustments to the angular positions of the galaxies and using different angular bin sizes to calculate the correlation function and identifying the peak that is robust to these changes. An alternative method for measuring the correlation function that could provide smoother, more accurate results is the kernel-based density estimator method used by Hatfield et al. (2016).

Finally, modifications to this analysis that will improve the final BAO peak position measurement include making corrections for projection effects and implementing a more thorough investigation into the errors involved in this analysis to obtain a comprehensive uncertainty for the measured peaks.

### 5.2.3 Forecasting Primordial Non-Gaussianity Constraints

One adjustment to the analysis performed in Chapter 4 would be to account for the uncertainty in the redshift distributions and bias functions. This could be done by marginalizing over the means of the N(z) and bias values in the redshift bins. As these are two very important measurements for the determination of the $f_{\rm NL}$ parameter and therefore this marginalization will result in more robust forecasts. Such marginalization should also be applied when measuring $f_{\rm NL}$ (instead of forecasting constraints) when the SKA data is available and the redshift distributions and bias functions can be measured, as there will still be some error in their determination and it will be important to account for it.

Another analysis that would be an interesting extension of the work presented in this thesis would be to utilize the red and blue SDSS galaxy populations along with the MLAGN and HLAGN radio galaxies observed with the SKA as the galaxy populations to be used with the multi-tracer method for measuring the local primordial non-Gaussianity parameter, $f_{\rm NL}$. The redshifts of these galaxy populations would be calculated with GPz and used to determine the redshift distributions. These populations are useful for the multi-tracer analysis as they each have different biases: red galaxies are known to be much more biased than blue galaxies, and the powerful radio galaxies observed with the SKA are expected

to be more biased than the red galaxies from SDSS. In addition, as discussed in Chapter 4 the MLAGN population is expected to be more biased than the HLAGN population (based on the observations of Hale et al. 2018). Such an analysis would require accurate photo-z estimates for the galaxies that are not luminous galaxies and GPz, along with the weighting based on target galaxy magnitude space technique discussed above will likely provide this.

# Bibliography


Abbott T. M. C., et al., 2018a, Phys. Rev. D, 98, 043526

Abbott T. M. C., et al., 2018b, MNRAS, 480, 3879

Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891

Abolfathi B., et al., 2018, ApJS, 235, 42

Aihara H., et al., 2018, PASJ, 70, S8

Alam S., et al., 2017, MNRAS, 470, 2617

Alcaniz J. S., Carvalho G. C., Bernui A., Carvalho J. C., Benetti M., 2016, arXiv, arXiv:1611.08458

Alcock C., Paczynski B., 1979, Nature, 281, 358

Alcock C., et al., 2001, ApJS, 136, 439

Almosallam I. A., Lindsay S. N., Jarvis M. J., Roberts S. J., 2016a, MNRAS, 455, 2387

Almosallam I. A., Jarvis M. J., Roberts S. J., 2016b, MNRAS, 462, 726

Alonso D., 2012, arXiv, arXiv:1210.1833

Alonso D., Ferreira P. G., 2015, Phys. Rev. D, 92, 063525

Alvarez M., et al., 2014, arXiv, arXiv:1412.4671

Baldauf T., Smith R. E., Seljak U., Mandelbaum R., 2010, Phys. Rev. D, 81, 063531

Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12

Ballardini M., Matthewson W. L., Maartens R., 2019, MNRAS, 489, 1950

Banerji M., et al., 2010, MNRAS, 406, 342

Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, Phys. Rep., 402, 103

Baum W. A., 1962, in McVittie G. C., ed., IAU Symposium Vol. 15, Problems of Extra-Galactic Research. p. 390

Baxter E. J., Rozo E., 2013, ApJ, 779, 62

Benítez N., 2000, ApJ, 536, 571

Benítez N., et al., 2009, ApJ, 692, L5

Bennett C. L., et al., 2003, ApJ, 583, 1

Bennett C. L., et al., 2013, ApJS, 208, 20

Bernardi G., et al., 2009, A&A, 500, 965

Biesiada M., 2006, Phys. Rev. D, 73

Blake C., Collister A., Bridle S., Lahav O., 2007, MNRAS, 374, 1527

Blake C., et al., 2011a, MNRAS, 415, 2876

Blake C., et al., 2011b, MNRAS, 415, 2876

Bolzonella M., Miralles J. M., Pelló R., 2000, AA, 363, 476

Bonaldi A., Harrison I., Camera S., Brown M. L., 2016, MNRAS, 463, 3686

Bonaldi A., Bonato M., Galluzzi V., Harrison I., Massardi M., Kay S., De Zotti G., Brown M. L., 2019, MNRAS, 482, 2

Bordoloi R., et al., 2012, MNRAS, 421, 1671

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Bruni M., Crittenden R., Koyama K., Maartens R., Pitrou C., Wands D., 2012, Phys. Rev. D, 85, 041301

Brunner R. J., Connolly A. J., Szalay A. e. S., Bershady M. A., 1997, ApJ, 482, L21

Budavári T., Szalay A. S., Connolly A. J., Csabai I., Dickinson M., 2000, AJ, 120, 1588

Budavári T., Szalay A. S., Csabai I., Connolly A. J., Tsvetanov Z., 2001, AJ, 121, 3266

Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21

Camera S., Santos M. G., Bacon D. J., Jarvis M. J., McAlpine K., Norris R. P., Raccanelli A., Röttgering H., 2012, MNRAS, 427, 2079

Camera S., Santos M. G., Ferreira P. G., Ferramacho L., 2013, Phys. Rev. Lett, 111, 171302

Campanelli L., Fogli G. L., Kahniashvili T., Marrone A., Ratra B., 2012, European Physical Journal C, 72, 2218

Carbone C., Verde L., Matarrese S., 2008, ApJ, 684, L1

Carilli C., Rawlings S., 2004, New Astronomy Reviews, 48, 979

Carnero A., Sánchez E., Crocce M., Cabré A., Gaztañaga E., 2012, MNRAS, 419, 1689

Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483

Carucci I. P., 2018, Journal of Physics Conference Series, 956, 012003

Carvalho G. C., Bernui A., Benetti M., Carvalho J. C., Alcaniz J. S., 2016, Phys. Rev. D, 93, 023530

Carvalho G. C., Bernui A., Benetti M., Carvalho J. C., de Carvalho E., Alcaniz J. S., 2017, arXiv, arXiv:1709.00271

Castorina E., Feng Y., Seljak U., Villaescusa-Navarro F., 2018, Phys. Rev. Lett, 121, 101301

Castorina E., et al., 2019, JCAP, 2019, 010

Challinor A., Lewis A., 2011, Phys. Rev. D, 84, 043516

Chambers K. C., et al., 2016, arXiv, arXiv:1612.05560

Chan K. C., et al., 2018, MNRAS, 480, 3031

Chong K., Yang A., 2019, arXiv, arXiv:1901.07544

Colless M., et al., 2001, MNRAS, 328, 1039

Collister A. A., Lahav O., 2004, PASP, 116, 345

Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693

Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655

Creminelli P., Zaldarriaga M., 2004, J. Cosmology Astropart. Phys., 2004, 006

Crocce M., Cabré A., Gaztañaga E., 2011a, MNRAS, 414, 329

Crocce M., Gaztañaga E., Cabré A., Carnero A., Sánchez E., 2011b, MNRAS, 417, 2577

Cunnington S., Harrison I., Pourtsidou A., Bacon D., 2019, MNRAS, 482, 3341

Daddi E., et al., 2005, ApJ, 626, 680

Dalal N., Doré O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514

Dawson K. S., et al., 2013, AJ, 145, 10

De Bernardis F., Serra P., Cooray A., Melchiorri A., 2010, Phys. Rev. D, 82, 083511

Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, IEEE Proceedings, 97, 1482

Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537

Doré O., et al., 2014, arXiv, arXiv:1412.4872

Duffy L. D., van Bibber K., 2009, New Journal of Physics, 11, 105008

Duncan K. J., Jarvis M. J., Brown M. J. I., Röttgering H. J. A., 2018, MNRAS, 477, 5177

Einstein A., 1915, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, Berlin, pp 844–847

Eisenstein D. J., et al., 2005a, ApJ, 633, 560

Eisenstein D. J., et al., 2005b, ApJ, 633, 560

Eisenstein D. J., Seo H.-J., White M., 2007, ApJ, 664, 660

Ewen H. I., Purcell E. M., 1951, Nature, 168, 356

Fabian A. C., 1999, Proceedings of the National Academy of Science, 96, 4749

Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P

Feldmann R., et al., 2006, MNRAS, 372, 565

Ferramacho L. D., Santos M. G., Jarvis M. J., Camera S., 2014, MNRAS, 442, 2511

Ferraro S., Smith K. M., 2015, Phys. Rev. D, 91, 043506

Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195

Fonseca J., Camera S., Santos M. G., Maartens R., 2015, ApJ, 812, L22

Fonseca J., Maartens R., Santos M. G., 2017, MNRAS, 466, 2780

Fonseca J., Maartens R., Santos M. G., 2018, MNRAS, 479, 3490

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748

Furlanetto S., et al., 2019, arXiv, arXiv:1903.06204

Gauci A., Zarb Adami K., Abela J., 2010, arXiv, arXiv:1005.0390

Gendre M. A., Best P. N., Wall J. V., Ker L. M., 2013, MNRAS, 430, 3086

Giannantonio T., Percival W. J., 2014, MNRAS, 441, L16

Giri S. K., Mellema G., Dixon K. L., Iliev I. T., 2018, MNRAS, 473, 2949

Glaser N., Wong O. I., Schawinski K., Zhang C., 2019, MNRAS, 487, 4190

Gomes Z., Jarvis M. J., Almosallam I. A., Roberts S. J., 2018, MNRAS, 475, 331

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759

Griest K., 1991, ApJ, 366, 412

Hale C. L., Jarvis M. J., Delvecchio I., Hatfield P. W., Novak M., Smolčić V., Zamorani G., 2018, MNRAS, 474, 4133

Hamilton A. J. S., 1998, in Hamilton D., ed., Astrophysics and Space Science Library Vol. 231, The Evolving Universe. p. 185

Hardcastle M. J., 2004, A&A, 414, 927

Harrison I., Lochner M., Brown M. L., 2017, arXiv, arXiv:1704.08278

Hatfield P. W., Lindsay S. N., Jarvis M. J., Häußler B., Vaccari M., Verma A., 2016, MNRAS, 459, 2618

Hatfield P. W., Bowler R. A. A., Jarvis M. J., Hale C. L., 2018, MNRAS, 477, 3760

Hawkins E., et al., 2003, MNRAS, 346, 78

Heckman T. M., Best P. N., 2014, ARA&A, 52, 589

Hewett P. C., 1982, MNRAS, 201, 867

Hickson P., Gibson B. K., Callaghan K. A. S., 1994, MNRAS, 267, 911

Hikage C., et al., 2019, PASJ, 71, 43

Hildebrandt H., et al., 2010, A&A, 523, A31

Hildebrandt H., et al., 2017a, MNRAS, 465, 1454

Hildebrandt H., et al., 2017b, MNRAS, 465, 1454

Hildebrandt H., et al., 2020, A&A, 633, A69

Hinshaw G., et al., 2013, ApJS, 208, 19

Ho S., et al., 2012, ApJ, 761, 14

Hogan R., Fairbairn M., Seeburn N., 2015, MNRAS, 449, 2040

Hogg D. W., Baldry I. K., Blanton M. R., Eisenstein D. J., 2002, arXiv, astro-ph/0210394

Hong T., Han J. L., Wen Z. L., Sun L., Zhan H., 2012, ApJ, 749, 81

Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275

Hoyle B., et al., 2018, MNRAS, 478, 592

Hsieh B. C., Yee H. K. C., Lin H., Gladders M. D., 2005, ApJS, 158, 161

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

Huterer D., Knox L., Nichol R. C., 2001, ApJ, 555, 547

Ilbert O., et al., 2006, A&A, 457, 841

Ilbert O., et al., 2009, ApJ, 690, 1236

Ilbert O., et al., 2015, A&A, 579, A2

Janssen R. M. J., Röttgering H. J. A., Best P. N., Brinchmann J., 2012, A&A, 541, A62

Jarvis M., Bacon D., Blake C., Brown M., Lindsay S., Raccanelli A., Santos M., Schwarz D. J., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), p. 18

Jarvis M., et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 6

Jeong D., Schmidt F., Hirata C. M., 2012, Phys. Rev. D, 85, 023504

Johnston R., Vaccari M., Jarvis M., Smith M., Giovannoli E., Häußler B., Prescott M., 2015, MNRAS, 453, 2540

Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, `http://www.scipy.org/`

Joudaki S., Doré O., Ferramacho L., Kaplinghat M., Santos M. G., 2011, Phys. Rev. Lett, 107, 131304

Joudaki S., et al., 2018, MNRAS, 474, 4894

Joudaki S., et al., 2019, arXiv, arXiv:1906.09262

Jungman G., Kamionkowski M., Griest K., 1996, Phys. Rep., 267, 195

Kaiser N., 1987, MNRAS, 227, 1

Karagiannis D., Lazanu A., Liguori M., Raccanelli A., Bartolo N., Verde L., 2018, MNRAS, 478, 1341

Kazin E. A., et al., 2013, MNRAS, 435, 64

Kennicutt Jr. R. C., 1992, ApJS, 79, 255

Kerscher M., Szapudi I., Szalay A. S., 2000, ApJ, 535, L13

Komatsu E., et al., 2003, ApJS, 148, 119

Komatsu E., et al., 2014, Progress of Theoretical and Experimental Physics, 2014, 06B102

Koo D. C., 1985, AJ, 90, 418

Koo D. C., 1999, in Weymann R., Storrie-Lombardi L., Sawicki M., Brunner R., eds, Astronomical Society of the Pacific Conference Series Vol. 191, Photometric Redshifts and the Detection of High Redshift Galaxies. p. 3

Kovetz E., et al., 2019, BAAS, 51, 101

Kron R. G., 1980, ApJS, 43, 305

LSST Science Collaboration et al., 2009, arXiv, arXiv:1903.11083

Laigle C., et al., 2016, ApJS, 224, 24

Landy S. D., Szalay A. S., 1993, ApJ, 412, 64

Laureijs R., et al., 2011, arXiv, arXiv:1110.3193

Leistedt B., Peiris H. V., Roth N., 2014, Phys. Rev. Lett., 113, 221301

Leistedt B., Hogg D. W., Wechsler R. H., DeRose J., 2019, ApJ, 881, 80

Li Y.-C., Ma Y.-Z., 2017, Phys. Rev. D, 96, 063525

Li L.-L., Zhang Y.-X., Zhao Y.-H., Yang D.-W., 2007, Chinese J. Astron. Astrophys., 7, 448

Li X.-D., Park C., Forero-Romero J. E., Kim J., 2014, ApJ, 796, 137

Lilly S. J., et al., 2009, ApJS, 184, 218

Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118

Linder E. V., 2003, Phys. Rev. Lett., 90, 091301

Lindsay S. N., et al., 2014, MNRAS, 440, 1527

Ling E. N., Frenk C. S., Barrow J. D., 1986, MNRAS, 223, 21P

Liske J., et al., 2015, MNRAS, 452, 2087

López-Corredoira M., 2014, The Astrophysical Journal, 781, 96

Lukić Z., Heitmann K., Habib S., Bashinsky S., Ricker P. M., 2007, ApJ, 671, 1160

Lukic V., Brüggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246

Lupton R. H., Gunn J. E., Szalay A. S., 1999, AJ, 118, 1406

Lyth D. H., Ungarelli C., Wands D., 2003, Phys. Rev. D, 67, 023503

Maddox N., Hess K. M., Blyth S. L., Jarvis M. J., 2013, MNRAS, 433, 2613

Magliocchetti M., Popesso P., Brusa M., Salvato M., Laigle C., McCracken H. J., Ilbert O., 2017, MNRAS, 464, 3271

Maldacena J., 2003, Journal of High Energy Physics, 2003, 013

Mantz A. B., Allen S. W., Morris R. G., Rapetti D. A., Applegate D. E., Kelly P. L., von der Linden A., Schmidt R. W., 2014, MNRAS, 440, 2077

Mantz A. B., et al., 2015, MNRAS, 446, 2205

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018, ApJ, 856, 2

Masters D., et al., 2015, ApJ, 813, 53

Matarrese S., Verde L., 2008, ApJ, 677, L77

Matthews D. J., Newman J. A., 2010, ApJ, 721, 456

McLure R. J., Kukula M. J., Dunlop J. S., Baum S. A., O'Dea C. P., Hughes D. H., 1999, MNRAS, 308, 377

Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, arXiv, arXiv:1303.4722

Miralda-Escudé J., 2003, Science, 300, 1904

Mo H. J., White S. D. M., 1996, MNRAS, 282, 347

Mobasher B., et al., 2004, ApJ, 600, L167

Moradinezhad Dizgah A., Keating G. K., 2019, ApJ, 872, 126

Mueller E.-M., Percival W. J., Ruggeri R., 2019, MNRAS, 485, 4160

Murray S. G., Power C., Robotham A. S. G., 2013, Astronomy and Computing, 3, 23

Newburgh L. B., et al., 2014, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 9145, 91454V

Newman J. A., 2008, ApJ, 684, 88

Nock K., Percival W. J., Ross A. J., 2010, MNRAS, 407, 520

Noeske K. G., et al., 2007, ApJ, 660, L43

Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009, MNRAS, 396, 19

Nusser A., 2005, MNRAS, 364, 743

Oke J. B., Sandage A., 1968, ApJ, 154, 21

Okumura T., Matsubara T., Eisenstein D. J., Kayo I., Hikage C., Szalay A. S., Schneider D. P., 2008, ApJ, 676, 889

Ota K., et al., 2017, ApJ, 844, 85

Overzier R. A., Röttgering H. J. A., Rengelink R. B., Wilman R. J., 2003, A&A, 405, 53

Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., Sheldon E. S., 2008, ApJ, 674, 768

Padmanabhan N., White M., 2008, Phys. Rev. D, 77, 123540

Padmanabhan N., et al., 2007, MNRAS, 378, 852

Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26

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

Peebles P. J. E., Groth E. J., 1975, ApJ, 196, 1

Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19

Penzias A. A., Wilson R. W., 1965, The Astrophysical Journal, 142, 419

Percival W. J., et al., 2004, MNRAS, 353, 1201

Perlmutter S., et al., 1999, ApJ, 517, 565

Petrosian V., 1976, ApJ, 209, L1

Planck Collaboration et al., 2014, A&A, 571, A1

Planck Collaboration et al., 2016a, A&A, 594, A13

Planck Collaboration et al., 2016b, A&A, 594, A17

Planck Collaboration et al., 2018, arXiv, arXiv:1807.06209

Planck Collaboration et al., 2019, arXiv, arXiv:1905.05697

Polsterer K. L., D'Isanto A., Gieseke F., 2016, arXiv, arXiv:1608.08016

Pourtsidou A., Metcalf R. B., 2015, MNRAS, 448, 2368

Primack J. R., Seckel D., Sadoulet B., 1988, Annual Review of Nuclear and Particle Science, 38, 751

Pritchard J. R., Loeb A., 2012, Reports on Progress in Physics, 75, 086901

Quinn P., Axelrod T., Bird I., Dodson R., Szalay A., Wicenec A., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 147

Ramos Almeida C., Bessiere P. S., Tadhunter C. N., Inskip K. J., Morganti R., Dicken D., González-Serrano J. I., Holt J., 2013, MNRAS, 436, 997

Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. The MIT Press

Refsdal S., 1964, MNRAS, 128, 307

Reid B. A., et al., 2012, MNRAS, 426, 2719

Richards G. T., et al., 2001, AJ, 122, 1151

Riess A. G., et al., 1998, AJ, 116, 1009

Romeo A., Metcalf R. B., Pourtsidou A., 2018, MNRAS, 474, 1787

Ross A. J., et al., 2011, MNRAS, 417, 1350

Ross A. J., et al., 2013, MNRAS, 428, 1116

Rudnick G., et al., 2001, AJ, 122, 2205

Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502

Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801

Sánchez E., et al., 2011, MNRAS, 411, 277

Santos M. G., 2016, in Proceedings, 51st Rencontres de Moriond, Cosmology session: La Thuile, Italy, March 19-26, 2016. ARISF, pp 307–314

Santos M., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 19

Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307

Scott D., Rees M. J., 1990, MNRAS, 247, 510

Seljak U., 2009, Phys. Rev. Lett, 102, 021302

Seo H.-J., Eisenstein D. J., 2003, ApJ, 598, 720

Seymour N., et al., 2007, ApJS, 171, 353

Sheth R. K., Tormen G., 1999, MNRAS, 308, 119

Shields G. A., 1999, PASP, 111, 661

Silk J., 1968, ApJ, 151, 459

Singal J., Shmakova M., Gerke B., Griffith R. L., Lotz J., 2011, PASP, 123, 615

Slosar A., Hirata C., Seljak U., Ho S., Padmanabhan N., 2008, JCAP, 8, 031

Smolčić V., et al., 2017a, A&A, 602, A2

Smolčić V., et al., 2017b, A&A, 602, A6

Soo J. Y. H., et al., 2018, MNRAS, 475, 3613

Sowards-Emmerd D., Smith J. A., McKay T. A., Sheldon E., Tucker D. L., Castander F. J., 2000, AJ, 119, 2598

Square Kilometre Array Cosmology Science Working Group et al., 2018, arXiv, arXiv:1811.02743

Subbarao M. U., Connolly A. J., Szalay A. S., Koo D. C., 1996, AJ, 112, 929

Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Lecture Notes in Computer Science, 2859, 226

Tanaka M., et al., 2018, PASJ, 70, S9

Tasse C., Best P. N., Röttgering H., Le Borgne D., 2008, A&A, 490, 893

Tauber J. A., et al., 2010, A&A, 520, A1

Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22

Tellarini M., Ross A. J., Tasinato G., Wands D., 2016, JCAP, 6, 014

The Dark Energy Survey Collaboration 2005, arXiv, astro-ph/0510346

Thomas S. A., Abdalla F. B., Lahav O., 2011, MNRAS, 412, 1669

Trujillo I., et al., 2006, MNRAS, 373, L36

Uhlemann C., Pajer E., Pichon C., Nishimichi T., Codis S., Bernardeau F., 2018, MNRAS, 474, 2853

Urry C. M., Padovani P., 1995, PASP, 107, 803

Vargas-Magaña M., et al., 2013, A&A, 554, A131

Vargas-Magaña M., et al., 2018, MNRAS, 477, 1153

Wang Y., Bahcall N., Turner E. L., 1998, AJ, 116, 2081

Way M. J., 2011, ApJ, 734, L9

Weinberg D. H., Davé R., Katz N., Hernquist L., 2004, ApJ, 601, 1

Welling Y., van der Woude D., Pajer E., 2016, JCAP, 8, 044

Whitaker K. E., van Dokkum P. G., Brammer G., Franx M., 2012, ApJ, 754, L29

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Willott C. J., Rawlings S., Blundell K. M., Lacy M., Eales S. A., 2001, MNRAS, 322, 536

Wilman R. J., et al., 2008, MNRAS, 388, 1335

Wittman D., Bhaskar R., Tobin R., 2016, MNRAS, 457, 4005

Witzemann A., Alonso D., Fonseca J., Santos M. G., 2019, MNRAS, 485, 5519

Wright E. L., et al., 2010, AJ, 140, 1868

Xia J.-Q., Viel M., Baccigalupi C., De Zotti G., Matarrese S., Verde L., 2010, ApJ, 717, L17

Yamamoto K., Kadoya Y., Murata T., Futamase T., 2001, Progress of Theoretical Physics, 106, 917

Yamauchi D., Takahashi K., Oguri M., 2014, Phys. Rev. D, 90, 083520

Yoo J., 2010, Phys. Rev. D, 82, 083508

Zaldarriaga M., 2004, Phys. Rev. D, 69, 043508

Zhu F., et al., 2018, MNRAS, 480, 1096

de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Experimental Astronomy, 35, 25

de Putter R., Doré O., 2017, Phys. Rev. D, 95, 123513

van Uitert E., et al., 2018, MNRAS, 476, 4662