Statistics in large galaxy redshift surveys

Lee Stothert

A Thesis presented for the degree of
Doctor of Philosophy

Institute for Computational Cosmology
Department of Physical Sciences
University of Durham
England

September 2018

Statistics in large galaxy redshift surveys

Lee Stothert

Abstract

This thesis focuses on modeling and measuring pairwise statistics in large galaxy redshift surveys. The first part focuses on two point correlation function measurements relevant to the Euclid and DESI BGS surveys. Two point measurements in these surveys will have small statistical errors, so understanding and correcting for systematic bias is particularly important. We use point processes to build catalogues with analytically known two point, and for the first time, 3-point correlation functions for use in validating the Euclid clustering pipeline. We build and summarise a two point correlation function code, `2PCF`, and show it successfully recovers the two point correlation function of a DESI BGS mock catalogue. The second part of this thesis focuses on work related to the PAU Survey (PAUS), a unique narrow band wide field imaging survey. We present a mock catalogue for PAUS based on a physical model of galaxy formation implemented in an N-body simulation, and use it to quantify the competitiveness of the narrow band imaging for measuring novel spectral features and galaxy clustering. The mock catalogue agrees well with observed number counts and redshift distributions. We show that galaxy clustering is recovered within statistical errors on two-halo scales but care must be taken on one halo scales as sample mixing can bias the result. We present a new method of detecting galaxy groups, Markov clustering (MCL), that detects groups using pairwise connections. We explain that the widely used friends-of-friends (FOF) algorithm is a subset of MCL. We show that in real space MCL produces a group catalogue with higher purity and completeness, and a more accurate cumulative multiplicity function, than the comparable FOF catalogue. MCL allows for probabilistic connections between galaxies, so is a promising approach for catalogues with mixed redshift precision such as PAUS, or future surveys such as 4MOST-WAVES.

Declaration

The work in this thesis is based on research carried out by Lee Stothert under the supervision of Dr Peder Norberg and Professor Carlton Baugh at the Institute for Computational Cosmology, the Department of Physics, Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification. Parts of this thesis are the author's contributions to published work.

- Some of the work in Chapter 2 was used by the Euclid Consortium Internal Documentation as part of the validation of the clustering processing functions.
- Section 3.5 of Chapter 3 reports on work presented in Smith et al. (2018).
- Chapter 4 is published in Stothert et al. (2018).

The figures in this thesis are created by the author unless stated otherwise in the figure caption.

Acknowledgements

I would first like to thank my family. Mum, Dad and Bob have been there to help at every stage of my studies, lending an ear, giving me advice or helping me move. I have always felt supported, and for that I am forever grateful.

I would like to thank Guinevere for patiently listening to my ramblings about my work. She has brightened every day, and I hope she will continue to put up with me and brighten many more.

I am heavily indebted to my two supervisors Peder and Carlton. Without their guidance this work would not have been possible. Their time and effort has been greatly appreciated, and watching and learning from them has made me a far better researcher.

I would also like to thank anyone at the ICC and in the PAUS or Euclid collaborations who has helped me along the way. I also thank STFC for sponsoring this work.

Contents

Chapter 1

Introduction

1.1 Theoretical models of cosmology

The field of cosmology is the study of the evolution of the Universe. This includes the beginning of the Universe, the formation and growth of structure, and the eventual fate of the universe. It is important to understand the rules and laws that govern this evolution. This section will provide a brief overview of theoretical cosmological models with a focus on the dominant model, ΛCDM.

The majority of cosmological models assume the cosmological principle. This states that on large enough scales (typically more than a few hundred Mpc[1]) the universe can be considered to be homogeneous (invariant with regards to position) and isotropic (invariant with regards to direction). This assumption allows for the temporal evolution of the universe on large scales to be described by a single scale factor $a(t)$, which increases (decreases) as the universe expands (contracts). The value of the scale factor at the present time t_0 is set to unity. Most models, including ΛCDM, are specific examples of the big bang cosmological model. In the big bang model the early Universe had a very small value of $a(t)$ and expanded over the age of the Universe to the size it is today, and is currently still expanding.

Gravity is the dominant force on large scales in the Universe. Einstein's equations

[1] 1 pc is defined as the distance at which 1 astronomical unit (roughly the distance from the earth to the Sun) subtends an angle of 1 arcsecond ($1/3600^{\text{th}}$ of a degree) on the sky.

of general relativity provide accurate predictions for gravitational interaction in the local solar system. If the universe is homogeneous these equations must hold on similar scales in similar environments throughout the universe. A viable model of gravitational interaction on large scales must either be general relativity, or provide a mechanism to sufficiently recover these equations in environments similar to the solar system. The ΛCDM model assumes general relativity to be the correct model of gravitational interaction. Attempts at discovering other viable models of gravity defines the field of modified gravity (Koyama, 2016).

The CDM in ΛCDM stands for cold dark matter. This model assumes cold dark matter is the dominant mass contribution in the Universe. Cold dark matter is a fluid that interacts gravitationally, and only weakly through the other fundamental forces of nature. The prefix cold means that we are assuming that this fluid is non-relativistic, i.e. dark matter particles move at speeds significantly slower than the speed of light. Other models of dark matter can include stronger interactions (Tulin & Yu, 2018) or warm dark matter (Viel et al., 2013).

The Λ of ΛCDM represents the cosmological constant, which is associated with a vacuum energy (or dark energy) that attempts to explain the apparent accelerated expansion of the universe. If we model this dark energy as a perfect fluid, the density ρ of the fluid is related to the scale factor of the universe $a(t)$ by

$$\rho \propto a(t)^{-3(1+w)}\,, \tag{1.1.1}$$

where the value of w, the equation of state parameter, will vary depending on the physical nature of this fluid. A value of $w < -1/3$ will lead to an accelerated expansion. A true cosmological constant is a special case of this general dark energy model in which the vacuum pressure and density is invariant with changing scale factor of the universe ($w = -1$). The ΛCDM model assumes that dark energy acts as a cosmological constant. As the densities of matter (relativistic and non-relativistic) and spatial curvature fall as the universe expands, Λ will be the dominant contribution to the energy density at late times in an expanding universe.

While we assume the universe to be homogeneous on large scales, there is rich structure on small scales. This structure has all formed from small inhomogeneities in the early universe seeded by inflation (Linde, 2014). Gravity caused the overdense

regions of the universe to grow and eventually collapse into the first dark matter halos. Baryonic matter would collect at the centre of these halos due to radiative cooling. Conservation of angular momentum caused these cooling baryons to form into disks, and eventually the first stars and galaxies. From here these small structures began to grow and merge with each other to form larger structures, a process called hierarchical growth. During this process of structure growth, the complex physics of galaxy formation produces the wide variety of structure we can see in the universe. These processes include gas hydrodynamics, star formation and evolution, feedback from supernova and black holes and the dynamics of galaxy interactions and mergers. Many of these processes remain poorly understood, so galaxy surveys like the ones presented here, particularly the Physics of the Accelerating Universe Survey (PAUS), are needed to understand how galaxy properties are related to their host halos.

1.2 Observational cosmology

Observational cosmology aims to observe the real universe to place constraints on the theoretical models of cosmology and the astrophysics of galaxy formation.

1.2.1 Galaxy imaging

Galaxy imaging is typically done using band pass filters. A band pass filter only allows a certain wavelength range to pass. Figure 1.1 shows the filter response curves for the broad band filter set (u, g, r, i, z and Y) of the PAU Camera at the William Herschel Telescope in La Palma (Padilla et al., 2016). This is a very common filter set, overlapping with the near UV, optical and near infrared parts of the spectrum. These broad band filters are typically of the order of 1000Å in width.

Different filters can correlate with different properties of galaxies. For example, redder filters may correlate more with the stellar mass of a galaxy, while bluer filters may correlate with the population of young stars and therefore the star formation rate of a galaxy.

Brightness in a particular filter is typically quoted using the magnitude system,

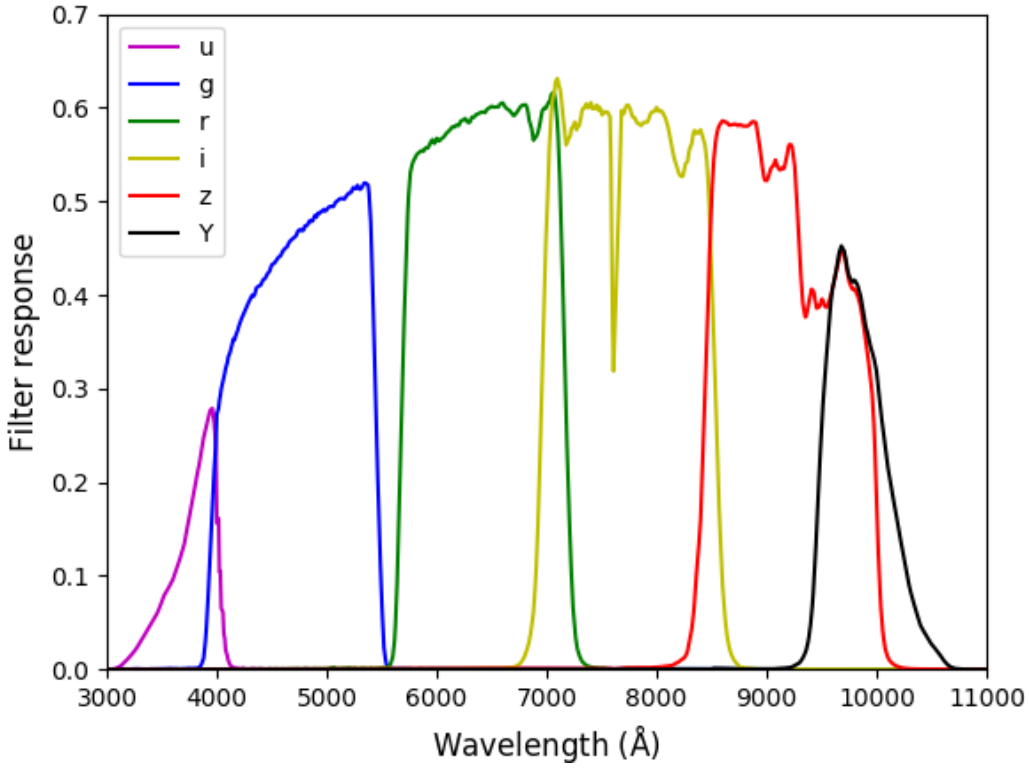

Figure 1.1: The PAUCam broad band (u, g, r, i, z, Y) filter responses as a function of wavelength. The filter response is defined as the fraction of energy at that wavelength that reaches the CCDs. The filter response also includes telescope optics and simulated atmospheric transmission.

which is logarithmic in flux relative to a reference object. This work will use the AB magnitude system, which defines the apparent magnitude, $m_{\rm AB}$, relative to the flux f_ν integrated over a filter with quantum efficiency $q(\nu)$, as

$$m_{\rm AB} = -2.5 \log_{10} \left(\frac{\int f_\nu q(\nu) \mathrm{d}\nu}{\int 3631\mathrm{Jy}\, q(\nu) \mathrm{d}\nu} \right). \tag{1.2.2}$$

Here the reference object has a spectral flux density of 3631Jy independent of λ, where 1 Jy $= 10^{-26}$ W Hz^{-1} m^{-2}. Different communities will define the flux to use in equation 1.2.2 in different ways. Often, only the flux lying within a certain angular radius of the centre of an object is measured. In this work we always use the total flux as we are dealing with simulations where this quantity is easily known.

1.2.2 Redshift

One of the main challenges in observational cosmology is measuring the distances to objects in the sky. Without estimates of distance, intrinsic properties such as the brightness and size of objects are far more difficult to infer. Distance measurements also allow us to build a three dimensional picture of structure in the universe.

The main tool used to determine how far away distant (beyond the scale where the local gravitation field make a contribution) objects are is the cosmological redshift of their light. The wavelength of light propagating through space is stretched by the expansion of the Universe such that it is received redder than when it was emitted. The amount of redshift is related to the scale factor of the Universe at the time of emission and observation. The redshift of an object is defined as

$$1 + z \equiv \frac{\lambda_o}{\lambda_e} = \frac{a(t_o)}{a(t_e)}, \tag{1.2.3}$$

where λ is the wavelength of light, $a(t)$ the scale factor of the universe and the subscripts e and o signify the quantity at the time of emission and observation respectively. If we can measure the redshift of an object, we can infer the scale factor at the time when the light was emitted. For a given cosmological model this can then be used to infer a distance to the object.

1.2.3 Redshift-distance relations

The expanding nature of the Universe leads to multiple definitions of distance. It is useful to define a measure of distance that is independent of how the universe has expanded since the light was emitted and when it was received. For this we define the comoving distance D_C as

$$D_C = \int_{t_e}^{t} \mathrm{d}t' \frac{c}{a(t')} \,. \tag{1.2.4}$$

This value may still change due to the local movements of an object, but will not change as the scale factor of the universe changes. We would like to express this equation in terms of an object's redshift. In order to do this we need to understand how the scale factor $a(t)$ varies with time, for which we need a cosmological model. The Friedmann equation is a solution to the equations of general relativity in the case of a universe described solely by the scale factor $a(t)$ and is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2(\Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{k,0}a^{-2} + \Omega_{\Lambda,0}) \,, \tag{1.2.5}$$

where H_0 is the Hubble constant, defined as the value of $\dot{a}/a$ evaluated at the present day. The values of $\Omega_{r,0}$, $\Omega_{m,0}$, $\Omega_{k,0}$ and $\Omega_{\Lambda,0}$ are the present day densities of radiation, matter, curvature and the cosmological constant in units of the critical density[2]. The sum of these densities must equal one. The Friedmann equation can be used to express equation (1.2.4) as

$$D_C = D_H \int_{a(t_e)}^{a(t_0)} \frac{\mathrm{d}a'}{\sqrt{\Omega_{r,0} + \Omega_{m,0}a' + \Omega_{k,0}a'^2 + \Omega_{\Lambda,0}a'^4}} \,, \tag{1.2.6}$$

where D_H is the Hubble distance defined as c/H_0. Using the definition that $a = 1/(1+z)$ gives

$$D_C = D_H \int_o^z \frac{\mathrm{d}z'}{\sqrt{\Omega_{r,0}(1+z')^4 + \Omega_{m,0}(1+z')^3 + \Omega_{k,0}(1+z')^2 + \Omega_{\Lambda,0}}} \equiv D_H \int_o^z \frac{\mathrm{d}z'}{E(z')} \,. \tag{1.2.7}$$

The comoving distance is inversely proportional to the value of the Hubble constant H_0. In order to produce results that are independent of the value of the Hubble

[2]The critical density, ρ_c, is given by $\rho_c = 3H_0^2/8\pi G$.

constant distances are often quoted in h^{-1}Mpc where $h = H_0/(100\text{kms}^{-1}\text{Mpc}^{-1})$ ($\sim$ 0.67 (Planck Collaboration et al., 2018)).

In Euclidean space the energy density of isotropically emitted radiation ρ_r follows the inverse square distance law

$$\rho_r \propto \frac{1}{D^2}\,, \tag{1.2.8}$$

for euclidean distance D. We would like to be able to use the same law in an expanding universe for luminosity calculations so we define the luminosity distance D_L as the distance for which this law will hold. The flux received by an observer goes as a factor of $(1+z)^{-2}$, so in order for the inverse square law to hold (in a flat universe where $\Omega_{k,0} = 0$) the luminosity distance D_L is related to the radial comoving distance D_C by

$$D_L = (1+z)D_C\,. \tag{1.2.9}$$

In an intrinsically curved spacetime ($\Omega_{k,0} \neq 0$) the relationship is slightly more complex, being

$$D_L(z) = \begin{cases} \frac{(1+z)D_H}{\sqrt{\Omega_{k,0}}} \sinh\left(\frac{\sqrt{\Omega_{k,0}}D_C(z)}{D_H}\right) & \text{for } \Omega_{k,0} > 0 \\ (1+z)D_C(z) & \text{for } \Omega_{k,0} = 0 \\ \frac{(1+z)D_H}{\sqrt{|\Omega_{k,0}|}} \sin\left(\frac{\sqrt{|\Omega_{k,0}|}D_C(z)}{D_H}\right) & \text{for } \Omega_{k,0} < 0\,. \end{cases} \tag{1.2.10}$$

1.2.4 Redshift space

The local (peculiar) velocity, v_{pec}, of an object along the line of sight to the observer also makes a contribution to the redshift in addition to that from the expansion of the Universe. This redshift due to the peculiar velocity, z_{pec}, can be given by

$$z_{\text{pec}} = \frac{v_{\text{pec}}}{c}\,, \tag{1.2.11}$$

provided $v_{\text{pec}} \ll c$. The observed redshift, z_{obs}, is given in terms of z_{pec} and the redshift due to the expansion of the universe, z_{H}, by

$$1 + z_{\text{obs}} = (1 + z_{\text{pec}})(1 + z_{\text{H}})\,. \tag{1.2.12}$$

So if a distance is inferred from a measured redshift, the true position of the object isn't recovered, rather, the measurement is in redshift space, which includes the

contribution of the peculiar velocity of the object. A very large galaxy velocity of 300 kms^{-1} gives rise to a peculiar redshift of ~ 0.01. This contribution is subdominant to the cosmological redshift measured in a typical galaxy redshift survey but still acts to smear galaxy positions along the line of sight. This can bee seen later on in the introduction in Figure 1.4 or in Chapter 4 in Figure 4.6.

1.2.5 Measuring redshift

In order to measure the redshift of an object we need to be able to identify known features in its spectrum so we can infer how far they have been reddened compared to their rest-frame wavelength. These measurements are typically made in two ways, spectroscopically or photometrically.

Spectroscopic redshift measurements make use of high resolution spectra to identify specific features in the spectrum of an object, such as emission or absorption lines. Typically, objects will be identified in an imaging survey then spectra will be taken for these objects using a fibre fed spectrograph. An example of a redshifted galaxy spectrum is shown in Figure 1.2 which shows an SDSS spectrum (Smee et al., 2013) and the identified emission and absorption features. Looking at one line, Hα, which is emitted at 6563Å, it is found in this spectrum at $\sim$ 7450Å, giving $z = 0.135$.

Photometric redshift measurements use multiple flux measurements from imaging bands to infer the most likely redshift for an object. The spectral resolution of imaging bands is typically far lower than it is for spectrography, so the the precision of the redshift measurement is usually lower. The narrower the bands, the larger the number of bands, and the greater the wavelength range they cover, the better the typical precision of the redshift measurement. Often, a redshift probability distribution is calculated rather than just the most likely redshift. The main method for inferring photometric redshifts is template fitting. This involves finding the best fit linear combination of templates and a redshift for a representative set of rest frame template spectra. These template spectra can be real data from spectroscopic surveys or taken from models.

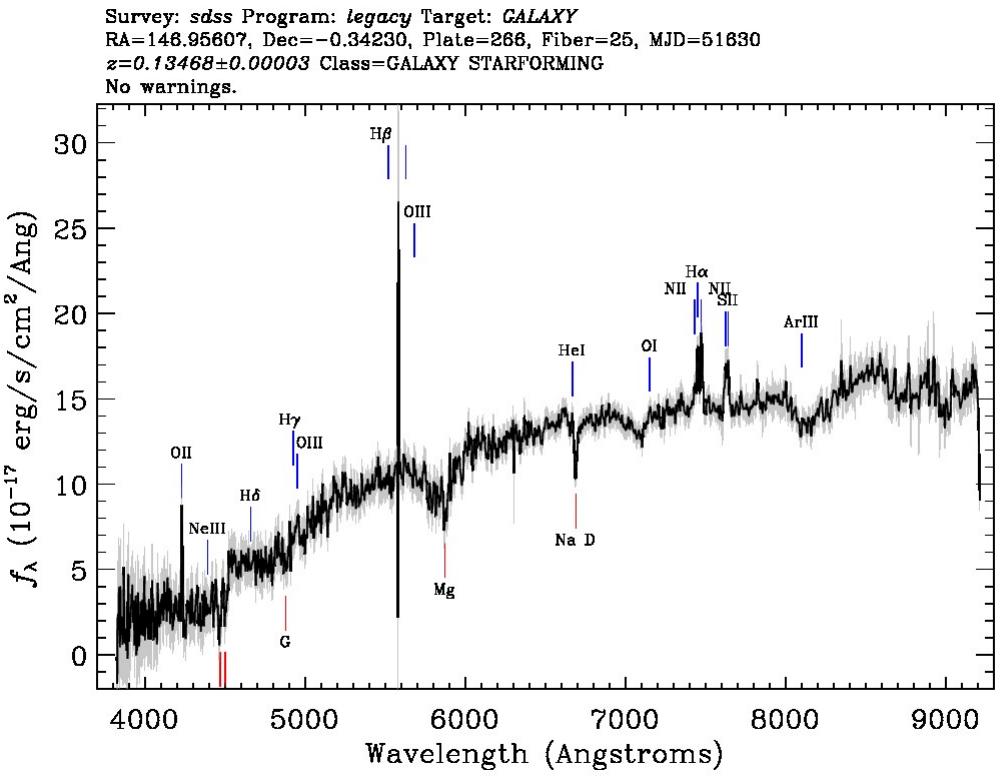

Figure 1.2: Example of an SDSS galaxy spectrum and the identified emission and absorption lines. The redshift of this galaxy is found to be 0.13468. Source: `https://skyserver.sdss.org/dr12/en/tools/explore/Summary.aspx?id=1237650795683512507`

1.2.6 Absolute magnitude

The apparent brightness of an object will change depending on how far away it is. For ease of brightness comparison we define the absolute magnitude as the brightness of an object if it was exactly 10pc away. The absolute magnitude, M, is defined in terms of the apparent magnitude, m, and the luminosity distance of the object, D_L, as

$$M = m - 5\log_{10}\left(\frac{D_L}{10}\right) \tag{1.2.13}$$

However, the section of the galaxy spectrum that overlaps with a given imaging filter will change depending on the redshift of the object. The difference between the measurement had it been made in the rest frame (if it were at z=0) and the measurement in the observer frame (as it is actually measured) is called the k-correction. The absolute magnitude calculation can be re-written to include this k-correction term, k, as

$$M = m - 5\log_{10}\left(\frac{D_L}{10}\right) - k\,, \tag{1.2.14}$$

where the absolute magnitude is now the value that would be found had the galaxy been observed at redshift 0. Depending on the data available the k-correction could be estimated as the same for all objects, inferred from simulations, parameterised in terms of a object colour (proxy for spectral energy distribution (SED) slope), or estimated object by object.

1.3 Statistical probes of observational cosmology and astrophysics

Here we introduce different means used to measure and quantify the galaxy distribution relevant to this thesis. We include the section on the cosmological distance ladder (section 1.3.1) for its historical context. Some notable probes not included in this chapter include lensing, CMB measurements, cluster analysis, gravitational wave detection and statistical descriptions of environment beyond groups such as structure finding.

1.3.1 Cosmological distance ladder

We have seen how different cosmological models result in different distances for the same redshift value. If the measurements of distance by other means can be obtained then we can place constraints on the cosmological model.

One of the most common methods is through the use of a standard candle. A standard candle is an object for which the absolute or intrinsic luminosity is believed to be known, so that the distance to an object can be inferred from the difference in absolute and apparent magnitudes. Often, these methods require calibration using their overlap with other methods which are applied at smaller distances, hence the term "cosmological distance ladder". Riess et al. (1998) used type 1a supernova as standard candles to show that the expansion of the universe was accelerating and that a form of dark energy or cosmological constant was required in any viable cosmological model.

1.3.2 1-point statistics

1 point statistics encompass statistics based on normalised counts of galaxies as a function of one or more properties. I will mention three important examples here. The first example, and the simplest, is number counts. A galaxy imaging survey can count the number of objects detected in a particular band as a function of apparent magnitude. Number counts are filter dependent but do not require galaxy redshift measurements.

The second of these is the luminosity function. The luminosity function gives the number of galaxies per unit volume as a function of absolute magnitude. It is once again a filter dependent measurement, but redshift measurement are now required to infer absolute magnitudes and to assign galaxies to redshift ranges. Figure 1.3 shows an example of the r band luminosity function for the low redshift GAMA survey (Driver et al., 2011) galaxies taken from Loveday et al. (2012). The luminosity function is typically fit by a Schechter function. This function follows a power law distribution for faint galaxies and falls off exponentially for galaxies brighter than the free parameter M^*. The luminosity function is well fit by a Schechter function

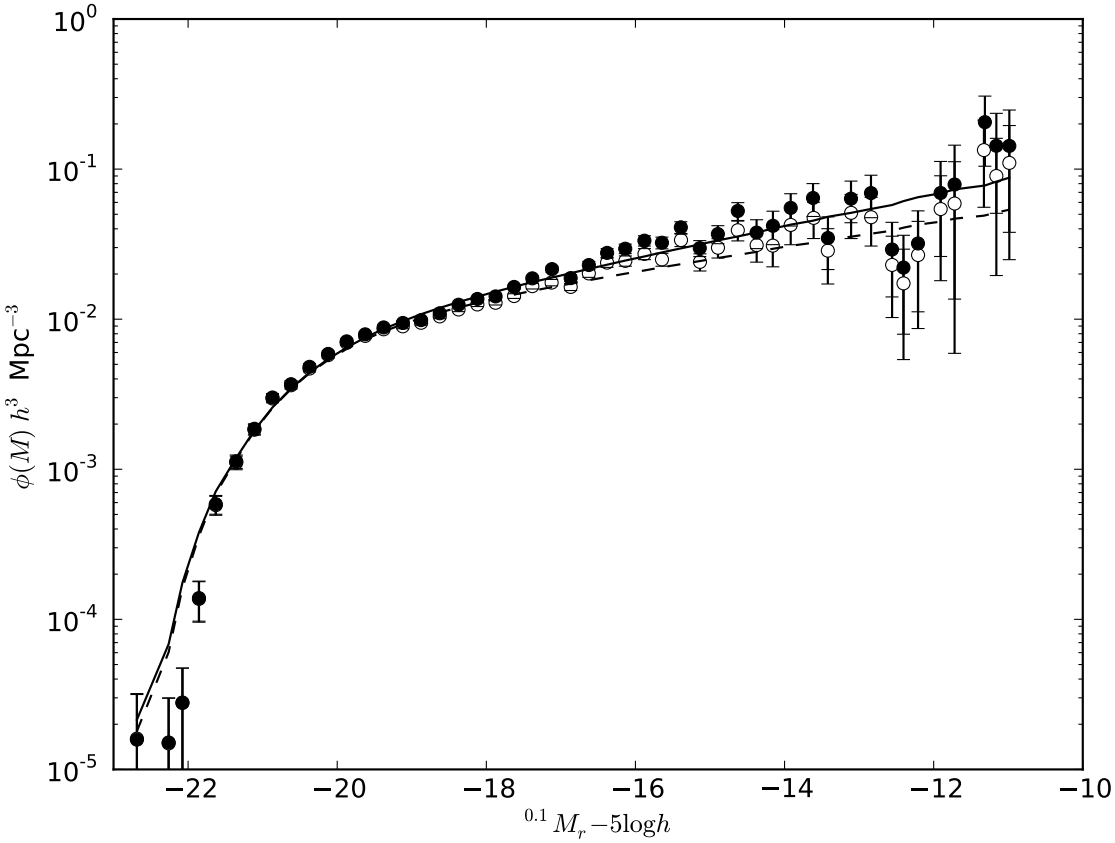

Figure 1.3: Low redshift (z < 0.1) r band luminosity function from the GAMA survey taken from Loveday et al. (2012). Solid symbols and line (Open circles and dashed line) shows the luminosity function with (without) correction for imaging completeness. Lines show the best fit Schechter function.

as the distribution of galaxy luminosities is closely related to the distribution of halo masses, which is itself well described by a Schechter function (Schechter, 1976).

Lastly, we can infer the stellar mass function. This measures the number of galaxies per unit volume as a function of the total stellar mass of galaxies. The stellar mass function should be independent of the filter set used to derive it but in practice this may not be the case. The stellar mass function can also be fit by a Schechter function. It is more difficult to measure observationally than the luminosity function as it requires estimations of the stellar masses of galaxies, which are rather model dependent and can lead to large systematic uncertainties (Mitchell et al., 2013). However, the stellar mass function often requires fewer assumptions to calculate in simulations than the luminosity functions do. This is because the total stellar mass for a galaxy is often known, whereas the luminosity in a given band requires calculation given a particular distribution of stars and gas. For example, the EAGLE simulations (Schaye et al., 2015) are tuned to match the present day stellar mass function.

1.3.3 2-point statistics

The two point correlation function, $\xi(\underline{r})$, is defined as the excess probability of finding a galaxy at a separation $\underline{r}$ from another galaxy. The term "two point" comes from the fact that this is a pairwise statistic rather than counts of single galaxies as in one point statistics. The average probability, $\mathrm{d}P$, of finding a galaxy at a separation $\underline{r}$ from another, can be given in terms of the mean density $\langle\rho\rangle$ and an infinitesimal volume element $\mathrm{d}V$ as (Peebles, 1980)

$$\mathrm{d}P = \langle\rho\rangle\,(1+\xi(\underline{r}))\mathrm{d}V\,. \tag{1.3.15}$$

A zero two point correlation function at a particular scale means that pairs at that scale are randomly distributed. A two point correlation function of greater (less) than zero implies the pairs are overdense (underdense) compared to random. $\xi(\underline{r})$ has a value between -1 and infinity. The two point correlation function is isotropic in real space if the cosmological principle holds, and redshift space measurements provide information about the velocity field.

The two point correlation function provides two of the primary cosmological probes through the Baryon Acoustic Oscillation peak (BAO) and Redshift space distortions (RSD). The BAO peak, first detected in 2dFGRS (Cole et al., 2005) and SDSS (Eisenstein et al., 2005) galaxy redshift surveys, is an overdensity in the distribution of matter in the Universe at a particular scale as a result of sound wave propagation in the early Universe (Eisenstein, 2005). Redshift space distortions measure the impact of large scale infall on the anisotropy of the two point correlation function (Kaiser, 1987).

Further, the two point correlation function provides a significant amount of information on small scales that can be used to infer galaxy formation physics. One popular family of models are Halo Occupation Distribution (HOD) models (e.g. Benson, 2001; Scoccimarro et al., 2001; Berlind & Weinberg, 2002; Cooray & Sheth, 2002). HOD models separate the two point correlation function into a "one halo term", which models the small separations at which most pairs of galaxies lie within the same dark matter halo, and a two halo term which models the large scales where pairs lie between two halos. HOD models provide an estimate of the correlation function starting from the mean number of galaxies in a halo.

1.3.4 Higher order statistics

Further to two-point statistics, work has also been done to analyse the three-point galaxy correlation function (e.g. Gaztañaga et al., 2005; Nichol et al., 2006). The three point correlation function measures the excess probability of finding certain triangle configurations. The probability of finding a particular triangle configuration, $\mathrm{d}P$, is given by

$$\mathrm{d}P = \langle\rho\rangle \left(1 + \xi(\underline{r}_{12}) + \xi(\underline{r}_{13}) + \xi(\underline{r}_{23}) + \zeta(\underline{r}_{12}, \underline{r}_{13}, \underline{r}_{23})\right)\mathrm{d}V\,, \tag{1.3.16}$$

where ξ is the two point galaxy correlation function and ζ is the three point function. The three point function is harder to measure than the two point function in terms of both computational complexity and the level of statistical noise. The three point correlation function is zero in any model that looks at the linear growth of small Gaussian perturbations (Berlind & Weinberg, 2002). It is therefore useful to show

where the linear model breaks down. This non-linearity makes providing analytic predictions for the three point function difficult. The form of higher order correlation functions could prove a useful probe of gravity (Hellwing et al., 2017).

The general N point correlation function would look at different configurations of N points. Little work has been done to consider values of N greater than 3 as a function of the different possible configurations as all the issues that face the three point in terms of difficulty to model and measure only get worse for higher order functions. Therefore, higher order moments are typically probed through the "counts-in-cells" methods (e.g. White, 1979), which are easier to implement (Baugh et al., 1995).

1.3.5 Galaxy groups

A galaxy group is defined as a collection of galaxies that are gravitationally bound within the same dark matter halo. Galaxies within groups can tell us about galaxy interactions and how galaxy properties and small scale clustering depend on local environment (Schneider et al., 2013; Barsanti et al., 2018). One example of the galaxy formation physics that can be inferred from groups is the quenching of the star formation in galaxies as they fall into dark matter halos and become satellite galaxies (Treyer et al., 2018). Galaxy groups, being proxies for dark matter halos, are also important tracers of large scale structure and are often used in galaxy clustering (Wang et al., 2008; Berlind et al., 2006a) or lensing analysis (van Uitert et al., 2017).

1.4 Galaxy surveys

This section gives an overview of past, present and future galaxy surveys, with a focus on the galaxy redshift surveys relevant to this work.

1.4.1 A brief recent history

Galaxy redshift surveys aim to measure galaxy redshifts for a large number of homogeneously selected galaxies. Normally they follow up galaxy imaging surveys and

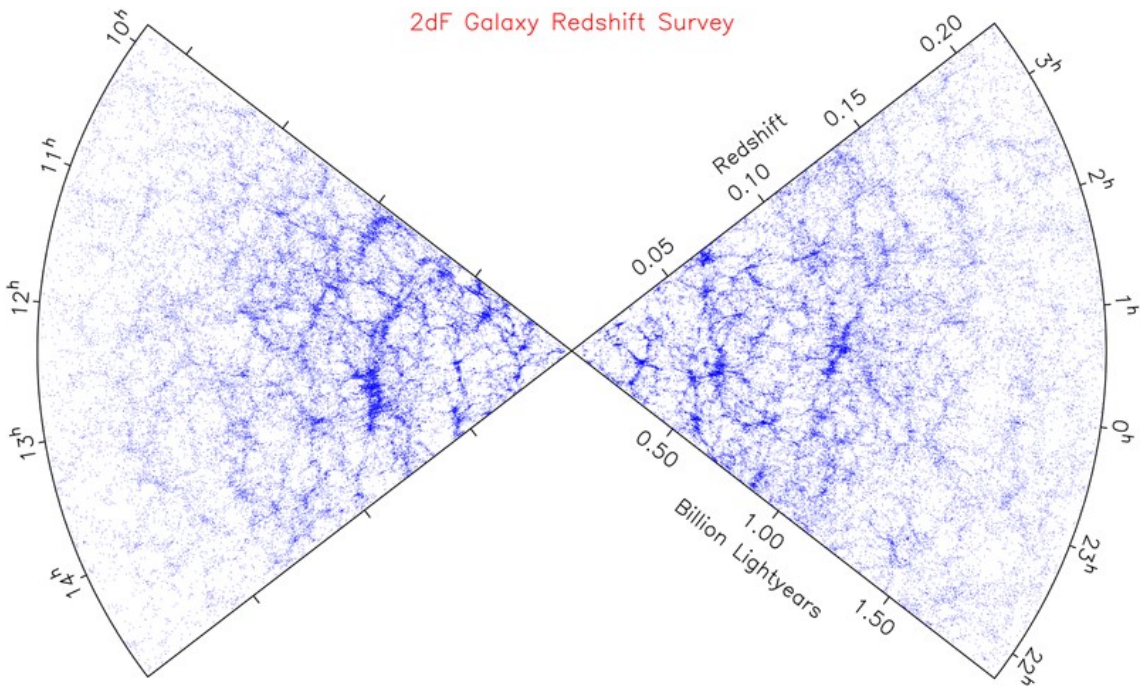

Figure 1.4: Cone plot of the 2dF galaxy redshift survey. The cosmic web and redshift space effects can clearly be seen. Source: `http://www.2dfgrs.net/Public/Pics/2dFzcone.gif`.

select galaxies for which to measure redshifts based on one or more photometric properties. Surveys have a finite amount of telescope time so will combine survey area, survey depth or target completeness to best complete their aims in this finite time. They typically fall into two categories, large wide shallow surveys and small narrow deep surveys. The large solid angle surveys can be used for cosmological purposes by measuring the position of the BAO peak and the shape of the redshift space distortions, while the small solid angle surveys are used to investigate redshift evolution and small scale galaxy interactions and environmental effects.

The Two degree Field Galaxy Redshift Survey (2dFGRS) (Colless et al., 2001) was one of the first galaxy redshift surveys used to measure the BAO peak (Cole et al., 2005). Figure 1.4 shows a slice of the 2dFGRS lightcone. The cosmic web can clearly be seen, as can the smearing of structures due to redshift space distortions. 2dFGRS measures redshifts for ~ 250000 galaxies over ~ 1500 square degrees limited in depth to $b_j < 19.45$.

The Sloan Digital Sky Survey (SDSS) legacy survey (York et al., 2000) is a

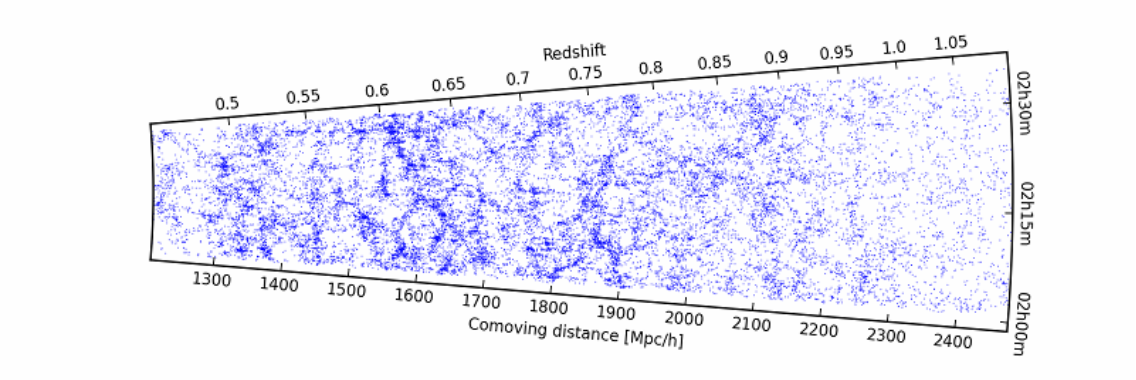

Figure 1.5: Cone plot of the W1 field of the VIPERS survey. The solid angle is lower and the density is higher than seen for 2dFGRS in Figure 1.4. Source: `http://vipers.inaf.it/rel-pdr1.html`.

galaxy redshift survey measuring nearly a million redshifts covering ~ 7500 square degrees to a magnitude limit of $r < 17.77$. The large area also makes SDSS perfect for cosmology measurements, e.g Eisenstein et al. (2005). The Baryon Oscillation Spectroscopic Survey (BOSS) (Dawson et al., 2013) is the successor to SDSS and used a colour cut to select luminous red galaxies at a higher mean redshift than SDSS primarily for cosmology measurements.

An example of a deeper survey with smaller solid angle is the GAMA survey (Driver et al., 2011). GAMA surveyed ~ 250 square degrees of sky to a depth of roughly $r < 19.8$. Unique to GAMA amongst large surveys is the high spectroscopic completeness. Often, only one galaxy of a pair lying very close in angle on the sky can have a fibre placed on them due to the physical restrictions of placing a fibre on each object. GAMA reobserved regions multiple times to reach a high completeness (98%) even in the high density regions. The GAMA survey is good for analysis of the redshift evolution of galaxies that are the earlier analogues of SDSS galaxies due to the greater depth of GAMA. Its high completeness makes it ideal for small scale analysis such as galaxy groups (Robotham et al., 2011).

Figure 1.5 shows one of the two fields of the VIMOS Public Extragalactic Redshift Survey (VIPERS) (Guzzo et al., 2014). VIPERS is a survey covering around 25 square degrees of sky to a depth of $i < 22.5$. A colour cut is used to select galaxies above a redshift of ~ 0.4 and the survey is around 40% complete with random

targeting of the selected image catalogue. The completeness and colour cuts are compromises made in order to be able to cover such an area at that depth. The difference between the VIPERS lightcone in Figure 1.5 and the 2dFGRS lightcone in Figure 1.4 can easily be seen. VIPERS extends far deeper over a far smaller area. The science cases of VIPERS are similar to GAMA but at a higher redshift. The lower completeness makes galaxy environment studies more difficult than for GAMA.

1.4.2 Euclid & DESI

Two future surveys that fall into the regime of cosmological studies are Euclid (Laureijs et al., 2011) and the Dark Energy Spectroscopic Instrument (DESI) survey (DESI Collaboration et al., 2016).

Euclid is a space based mission that will observe up to 15000 $\deg^2$ of sky. It will perform both imaging and spectroscopy. Space based imaging allows very accurate shape measurements of galaxies, free from atmospheric distortion. This imaging will allow very accurate lensing measurements to be made. Euclid will also provide redshift measurements for a subset of these objects using slitless spectroscopy. The spectrograph is limited in wavelength range, which limits the redshift ranges over which different emission lines can be seen. These redshift ranges are at a higher redshifts than previously explored in BOSS so will provide interesting results on the evolution of the BAO feature. The number of objects with a redshift measurement (an estimated 30 million (Pozzetti et al., 2016)) and scale of the volume probed will mean Euclid provides the tightest constraints on the parameters of ΛCDM of any galaxy survey so far.

DESI is a more traditional galaxy redshift survey than Euclid that will follow up ground based imaging with ground based spectroscopy. It is split into dark and bright times (bright time is when the moon is up). The dark time will be used to observe luminous red galaxies (LRGs), emission line galaxies (ELGs) and quasars with the primary goal of providing accurate BAO and RSD measurements over a large redshift range, $0.5 < z < 3.5$. The bright time will perform a bright galaxy survey (BGS) which is a magnitude limited survey of galaxies at a depth

very comparable to GAMA, $r < 20$, and a Milky Way star survey.

1.4.3 The PAU Survey (PAUS)

The PAU Survey (PAUS) is a narrow band imaging survey covering up to 100 square degrees in 40 narrow bands of width 130Å, spaced 100Å apart in the wavelength range 4500-8500Å. The narrow band imaging is done through forced photometry on previously detected objects from CFHTLenS (Heymans et al., 2012) so the requirement in signal to noise ratio is not as high as is needed for object detection. Narrow band imaging will allow more accurate photometric redshift measurements than photometric redshift measurements using traditional broad band surveys, estimated from simulations to be 0.35% for PAUS vs ∼3% for good broad band photometry (Martí et al., 2014a). Current data measurements achieve this accuracy for a significant fraction of objects to $i < 22.5$, and will achieve $\sim 1\%$ accuracy for all objects to that magnitude limit (Eriksen et al. (in prep)). Pipeline revisions currently underway hope to improve this. This accuracy will be sufficient to perform galaxy clustering measurements but it will be more difficult to measure and model redshift space effects. PAUS has similar science goals to VIPERS, being at a similar depth, but measures a redshift for 100% of objects and covers a larger area. The high completeness will allow for more complete small scale environment studies than VIPERS and the larger area and accurate shape measurements from the parent catalogue will allow for a competitive measurement of the intrinsic alignment signal to be made, which is a common systematic uncertainty of lensing measurements. This will be particularly relevant to lensing measurements at the precision that Euclid will provide.

1.5 Cosmological simulations

The work presented in this thesis is mostly using mock galaxy catalogues of the universe. This section will briefly introduce their construction and explain their usefulness.

1.5.1 Dark matter only simulations

In the currently favoured cosmological paradigm, cold dark matter is thought to be the dominant contribution to mass in the universe, so a good approximation to large scale structure in the universe can be achieved by examining the case of a universe made solely of collisionless matter. Simulating such a universe is done through N-body methods. In the N-body approach, the dark matter in a volume is quantised into computational particles, and the time evolution of these quantised elements is followed as they interact gravitationally. The simulations are typically saved at various snapshots of cosmic time. Given a particle distribution, dark matter halos can be identified and merger trees calculated. Merger trees track dark matter halos between snapshots, making it easy to identify which halos merged to form the current halos. This forms a tree because each halo will branch into its direct progenitors, and each of those can branch out in turn. Each leaf halo (a halo with no progenitors) is formed solely from the gravitational collapse.

Springel et al. (2005) ran the Millennium simulation, which simulated the universe using 2160^3 particles in a cubic box with side length $500h^{-1}$Mpc from redshift 127 to the present day. 64 snapshots were saved and used to form dark matter halo merger trees. The work in this thesis uses the MR7 simulation (Guo et al., 2013), this is very similar to the Millennium simulation, but saves 61 snapshots of a universe with WMAP7 cosmology (Hinshaw et al., 2013).

1.5.2 Galaxy simulations

There are two approaches commonly used to build a physical model of galaxies, hydrodynamic and semi-analytic modeling. Hydrodynamic simulations are N-body simulations that include a gas component as well as the dark matter and simulate the creation of galaxies by attempting to include the physics of the gas particles. Semi-analytic simulations use physically motivated empirical schemes to populate a previously calculated catalogue of dark matter halos with galaxies.

Both of these approaches must make approximations to physics that they cannot resolve. In a hydrodynamical simulation this is done through the sub-grid physics

model. The sub-grid model will attempt to simulate the aggregate effects of physical processes that occur below the resolution of the simulation (Crain et al., 2015). One example is star formation; a simulation will not resolve the collapse of gas into stars, but instead set conditions for when star particles (representing a stellar population) are formed from gas particles. In a semi-analytic model, all internal galaxy processes, and the merging of galaxies, are treated in a sub-grid fashion, as each galaxy is treated as a single object. In both cases these physically motivated sub-grid models may have free parameters that can be tuned to try to make the simulation match selected observations.

Extending an N-body simulation to include baryonic particles and hydrodynamics is difficult and time consuming. A state of the art hydrodynamic simulation, EAGLE (Schaye et al., 2015), only simulates a volume of 100 Mpc^{-1}, ~ 320 times smaller than the Millennium simulation, despite being run over a decade later. A semi analytic model, such as the Durham `GALFORM` model (e.g. Lacey et al., 2016; Gonzalez-Perez et al., 2013; Cole et al., 2000), built on top of N-body simulations like the Millennium simulation, will take only a fraction of the time needed to run a hydrodynamical simulation. Semi analytic models are therefore the model of choice when simulations comparable to the size of large galaxy redshift surveys are needed.

1.5.3 Mock catalogues for galaxy surveys

Simulations of the universe are often saved at snapshots in redshift. This is in contrast to the continuous observations of a galaxy survey which could span a significant fraction of cosmic history. We would like to be able to take galaxy catalogue snapshots and mimic a particular galaxy survey. Merson et al. (2013) provide a method of building mock lightcones from `GALFORM` snapshots by interpolating the positions and luminosities of galaxies. Galaxies are interpolated between snapshots to find out when they cross the observer's past lightcone and if they lie within the mock survey sky area. Large survey simulations will often cover volumes much larger than the N-body simulation used to generate the galaxy catalogue, so the simulation volume is replicated as many times as is necessary to build a volume large enough to fill the survey volume. This will mean that in large surveys the same galaxy may be

replicated multiple times at different redshifts and angles on the sky, which will act to artificially reduce the cosmic variance in the mock catalogue.

1.6 Thesis outline

This thesis focuses on modeling and measuring pairwise statistics in large galaxy redshift surveys. Chapter 2 uses point processes to build catalogues with analytically known two and three point correlation functions. Chapter 3 presents and summarises the two point correlation function code `2PCF` and reports the work of Smith et al. (2018) who use it to recover the two point correlation function in a DESI BGS mock galaxy catalogue. Chapter 4 presents a mock galaxy catalogue for the PAU Survey that is built on an N-body simulation using the semi-analytic galaxy formation model `GALFORM`. We use it to quantify the competitiveness of the narrow band imaging for measuring novel spectral features and galaxy clustering. Chapter 5 presents and investigates a novel new approach to galaxy group finding, Markov Clustering. Chapter 6 concludes.

Chapter 2

Point processes and clustering

This chapter explores the use of point processes to generate mock catalogues with known two point correlation functions. This work summarises my contribution to the two point clustering validation team in the OULE3 work package of the European Space Agency's Euclid mission. In particular, this chapter focuses on extending the known literature results of two common Neyman Scott point processes, the segment Cox process and the Thomas process, to produce catalogues with known higher order multipoles of the two point correlation function. These predictions are then tested and successfully validated. The result for the one cluster term of the N point correlation function of a generalised Thomas process is derived. This is used to provide a specific prediction for the three point correlation function of the isotropic 3D Thomas process.

2.1 Introduction

The two point correlation function, ξ, introduced in section 1.3.3, is one of the main statistical measures of the spatial distribution of galaxies. Through the cosmological principle (homogeneity and isotropy of the universe), the two point correlation function is isotropic in real space, i.e it depends only on $|\underline{r}|$. However, we measure galaxy positions in redshift space, and redshift space distortions make the clustering of galaxies along the line of sight different to that perpendicular to the line of sight. As a result, the two point correlation function is often given as a function of the

transverse and radial separations to the line of sight, r_p and π, or as a function of separation s and the cosine of the angle the separation vector makes with the vector pointing to the mean position of the galaxies, μ. These quantities are defined in terms of the two galaxy position vectors, $\underline{x}_1$ and $\underline{x}_2$, as

$$\underline{r} = \underline{x}_1 - \underline{x}_2 \tag{2.1.1}$$

$$s = |\underline{r}| \tag{2.1.2}$$

$$\pi = \underline{r}.\left(\frac{\underline{x}_1 + \underline{x}_2}{2|\underline{x}_1 + \underline{x}_2|}\right) \tag{2.1.3}$$

$$r_p = \sqrt{s^2 - \pi^2} \tag{2.1.4}$$

$$\mu = \pi/s\,. \tag{2.1.5}$$

The multipoles of the two point correlation function are then defined as

$$\xi_n(s) = \frac{2n+1}{2}\int_{-1}^{1} P_n(\mu)\xi(s,\mu)\mathrm{d}\mu\,, \tag{2.1.6}$$

where the function $P_n(\mu)$ is the n^{th} Legendre polynomial. These functions provide an orthogonal basis with which to express the two point correlation function[1]. The functions are orthogonal over the range -1 to 1,

$$\int_{-1}^{1} P_i(\mu)P_j(\mu)\mathrm{d}\mu = \frac{2}{2n+1}\delta_{ij}\,, \tag{2.1.7}$$

where the Kronecker symbol δ_{ij} is defined as

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise}\,. \end{cases} \tag{2.1.8}$$

In the linear regime, coherent infall leaves all multiples above $n = 4$ unchanged, (Hamilton, 1992). On non-linear scales, higher order multipoles are not expected to be zero. Often only the first few multipoles are measured and modeled, e.g. Hawkins et al. (2003). Higher order multipoles are generally too noisy.

Due to the anisotropic nature of galaxy clustering the correlation function is often projected onto the transverse axis by integrating along the line of sight,

$$w_p(r_p) = 2\int_0^{\pi_{\max}} \xi(r_p,\pi)\mathrm{d}\pi\,. \tag{2.1.9}$$

[1]The first three are given by: $P_0(\mu) = 1$, $P_1(\mu) = \mu$, $P_2(\mu) = 0.5(3\mu^2 - 1)$.

where the upper limit of the integral $\pi_{\max}$ should in theory be infinity, but in practice must be set to a large finite value because of the finite dimensions of a galaxy catalogue. Too large a value of $\pi_{\max}$ and the projected clustering measurement would become too noisy. In the plane-parallel approximation, and if $\pi_{\max}$ is large enough, this statistic is the same if measured in real or redshift space, so provides a measurement that is independent of redshift space distortions. In the true plane-parallel approximation, the direction onto which the galaxy separation vector $\underline{r}$ should be projected to find the cartesian decomposition r_p and π would be the same for all pairs in the volume. This approximation works in a simulated volume but clearly fails for galaxy surveys with large solid angles as two pairs of galaxies could be separated by 90 degrees on the sky. We therefore use the local plane-parallel approximation, which now states that the two lines joining the observer to a pair of galaxies are parallel, but the lines to different pairs may not be parallel. In practice, this approximation means that changes in the radial distance to one or both of the pair of galaxies only changes π and leaves r_p unchanged.

Measuring the two point correlation function for a galaxy survey requires calculating the distribution of galaxy pair distances, $DD(\underline{r})$, and comparing them to the distributions of Data-Random pairs, $DR(\underline{r})$, and Random-Random pairs, $RR(\underline{r})$, for a random catalogue with the same density distribution as the data but without spatial correlation. The generation of a random catalogue is necessary to estimate the pair distances of random points in a complicated survey volume. The most commonly used estimator, also the one adopted throughout this thesis, is defined in Landy & Szalay (1993a)

$$\xi(\underline{r}) = \frac{DD(\underline{r}) - 2DR(\underline{r}) + RR(\underline{r})}{RR(\underline{r})}\,, \tag{2.1.10}$$

where the pair count distributions should be appropriately normalised.

It can be seen that a naive approach to calculating the pair counts for N points requires $N(N-1)/2$ pair calculations. This can be said to scale as $\mathcal{O}(N^2)$. For modern galaxy surveys that will potentially measure tens of millions of galaxy redshifts, such as DESI, (DESI Collaboration et al., 2016), and Euclid, (Laureijs et al., 2011), a naive approach becomes computationally unfeasible, so methods must be found to speed up the pair count calculations. At the same time as requiring faster calcula-

tions, the precision required in order that the errors from the codes are sub-dominant to the statistical errors of the measurements in these surveys is significantly increasing. I will present my own code to do this in chapter 3.

Further to two-point statistics, work is also often done to analyse the three-point galaxy correlation function, first introduced in section 1.3.4. This statistic is found with triplet counts rather than pair counts, which means a naive implementation now scales as $\mathcal{O}(N^3)$. Even more so than the two point calculations, this requires improved algorithms to be able to fully explore this statistic, such as the one presented in Slepian & Eisenstein (2015).

The material in this chapter stems from work I did as part of the Euclid two-point galaxy clustering validation team, part of the OU-LE3 validation activity. In order to test the accuracy and precision of the Euclid two-point statistics pipeline code, a catalogue with analytically known two-point multipoles was required. My role within the team was to investigate point processes as a means of generating these catalogues. A point process, or point field, is a series of points that lie in some mathematical space. A point process is often chosen to model a particular dataset whose points exhibit some sort of spatial correlation.

In particular, we consider Neyman-Scott processes (Neyman & Scott, 1958). These have previously been used to model the "one-halo term" in Halo Occupation Distribution (HOD) models (Benson, 2001; Scoccimarro et al., 2001; Berlind & Weinberg, 2002; Cooray & Sheth, 2002). We consider two common point processes in the literature: the segment Cox process (Stoyan et al., 1995), and the Thomas process (Thomas, 1949), which produce known monopoles and zero higher order multipoles. The latter are then extended to produce known non-isotropic correlation functions so that non-null higher order multipole results can be validated and tested. We provide analytic projected correlation function predictions where known. We also provide analytic calculations for the higher order correlation functions of a generalised Thomas process for potential use in validating higher order statistics algorithms.

Section 2.2 provides an overview of isotropic Neyman-Scott processes and provides comprehensive results and validation for the segment Cox process and the

Thomas process. Section 2.3 extends the Cox process and Thomas process models to produce known non-zero higher order multipoles and validates the predictions. Section 2.4 provides specific predictions for the three point correlation function of the isotropic Thomas process. Section 2.5 gives the conclusions.

2.2 Isotropic Neyman-Scott processes

A Neyman-Scott point process is a point process that randomly assigns points to randomly placed clusters which have a known cluster profile. The procedure for this point process is,

- Place N_c cluster centres randomly in a volume V.
- For N_p total points, randomly assign each to a cluster and sample from the cluster pdf to place the points relative to their cluster centres.

Each cluster will not necessarily contain the same number of points, only an average of N_p/N_c points. The random choice of cluster, then of position in a cluster, makes a Neyman-Scott process a "doubly stochastic" point process. A doubly stochastic point process is called a Cox point process (Cox, 1955), so a Neyman-Scott process is a subset of a Cox process. The choice of cluster profile will change the clustering of the points in the catalogue. The clusters in a Neyman-Scott process will sometimes overlap due to the completely random placement of the cluster centres. Forcing them not to overlap would change the clustering result.

Neyman-Scott processes are useful to describe datasets which exhibit some form of local clustering. They were first introduced to model the clustering of galaxies but have applications beyond astrophysics; a particular example is the use of Neyman-Scott point processes to model observations of whales (Hagen & Schweder, 1995).

I will now outline how to calculate the two point correlation function for Neyman-Scott process. This expands on the partial derivation presented in Stoyan et al. (1995). The K-function, $K(r)$, is defined in N dimensions as the average number of points contained within a hypersphere of radius r from any randomly chosen point

in the catalogue. It is calculated from the density field $\rho(\underline{s})$ of a catalogue by

$$K(r) = \Big(\int_V d^N s \int_{|r'|<r} \mathrm{d}^N r' \rho(\underline{s}) \rho(\underline{s} + \underline{r'}) \Big) \Big(\int_V \mathrm{d}^N s \rho(\underline{s}) \Big)^{-1} . \tag{2.2.11}$$

For a finite catalogue the denominator is simply equal to the number of points in the catalogue. The two point correlation function is related to the K-function by

$$1 + \xi(r) = \Big(\frac{\mathrm{d}K(r)_{\mathrm{random}}}{\mathrm{d}r} \Big)^{-1} \frac{\mathrm{d}K(r)}{\mathrm{d}r} , \tag{2.2.12}$$

where $K(r)_{\mathrm{random}}$ is the K-function of a random catalogue lying in the same volume. In 2D this reduces to,

$$1 + \xi(r) = \frac{1}{2\pi r \langle \rho \rangle} \frac{\mathrm{d}K(r)}{\mathrm{d}r} , \tag{2.2.13}$$

and in 3D,

$$1 + \xi(r) = \frac{1}{4\pi r^2 \langle \rho \rangle} \frac{\mathrm{d}K(r)}{\mathrm{d}r} , \tag{2.2.14}$$

where $\langle \rho \rangle$ is the average density of the catalogue, defined by $\langle \rho \rangle = N_p / V$ for V the volume containing the catalogue and N_p is the total number of points.

For a Neyman Scott process with N_c clusters, each with density profile $\rho_c(\underline{s})$, the density in the volume, $\rho(\underline{s})$, is given by

$$\rho(\underline{s}) = \sum_{i=1}^{N_c} \rho_c(\underline{s} - \underline{s}_i) . \tag{2.2.15}$$

Plugging this into equation 2.2.11 for the K-function and choosing N=3 dimensions gives

$$K(r) = \frac{1}{N_p} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' \rho_c(\underline{s} - \underline{s}_i) \rho_c(\underline{s} - \underline{s}_j + \underline{r'}) . \tag{2.2.16}$$

It can be seen that there are two types of contribution to this sum: one where $i = j$, i.e the "one cluster term" which performs a double integral over single cluster profiles and one where $i \neq j$, i.e a "two cluster term" which sums over pairs of points lying in different clusters. There are N_c identical one halo terms which can be centred on zero without a loss in generality, resulting in

$$\begin{aligned} K(r) =& \frac{N_c}{N_p} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' \rho_c(\underline{s}) \rho_c(\underline{s} + \underline{r'}) \\ &+ \frac{1}{N_p} \sum_{i=1}^{N_c} \sum_{j \neq i}^{N_c} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' \rho_c(\underline{s} - \underline{s}_i) \rho_c(\underline{s} - \underline{s}_j + \underline{r'}) . \end{aligned} \tag{2.2.17}$$

Points lying in two different clusters are randomly distributed with respect to each other as the clusters themselves are randomly distributed, so their contribution can be said to be the same as that of a random catalogue. For large N_c such that the number of two halo terms $(N_c - 1)N_c$ can be approximated as N_c^2, equation 2.2.17 becomes

$$\begin{aligned} K(r) =& \frac{N_c}{N_p} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' \rho_c(\underline{s}) \rho_c(\underline{s} + \underline{r'}) \\ &+ \frac{4}{3} \pi r^3 \langle \rho \rangle \ . \end{aligned} \tag{2.2.18}$$

Plugging this relationship into equation 2.2.14 gives

$$\xi(r) = \frac{N_c}{4\pi r^2 N_p \langle \rho \rangle} \frac{\mathrm{d}}{\mathrm{d}r} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' \rho_c(\underline{s}) \rho_c(\underline{s} + \underline{r'}) \, . \tag{2.2.19}$$

This can be written in terms of the probability density function of the cluster, $p_c(\underline{s})$, which is the density of the cluster normalised such that the integral over the whole cluster profile is unity. It is related to the density of a cluster through

$$\rho_c(\underline{s}) = \frac{N_p}{N_c} p_c(\underline{s}) \, , \tag{2.2.20}$$

giving

$$\xi(r) = \frac{N_p}{4\pi r^2 N_c \langle \rho \rangle} \frac{\mathrm{d}}{\mathrm{d}r} \int_V \mathrm{d}^3 s \int_{|r'|<r} \mathrm{d}^3 r' p_c(\underline{s}) p_c(\underline{s} + \underline{r'}) \, . \tag{2.2.21}$$

Equation 2.2.21 provides a method of calculating the analytic correlation function of a Neyman-Scott process given a cluster density probability distribution.

2.2.1 Isotropic segment Cox process

The first Neyman-Scott process to look at is the isotropic segment Cox process (Stoyan et al., 1995). This point process is used for Euclid pipeline validation, and was the first to be extended to provide known non-zero higher multipoles. Stoyan et al. (1995) provides a result for the two point correlation function monopole for this process but no derivation is included, so it is written out here for completeness.

The isotropic segment Cox process sets the cluster profile as lines of fixed length L with random direction. Figure 2.1 visualises this process in the 2D and 3D cases in periodic volumes for 30 lines each of length 200 h^{-1}Mpc. The segment Cox process

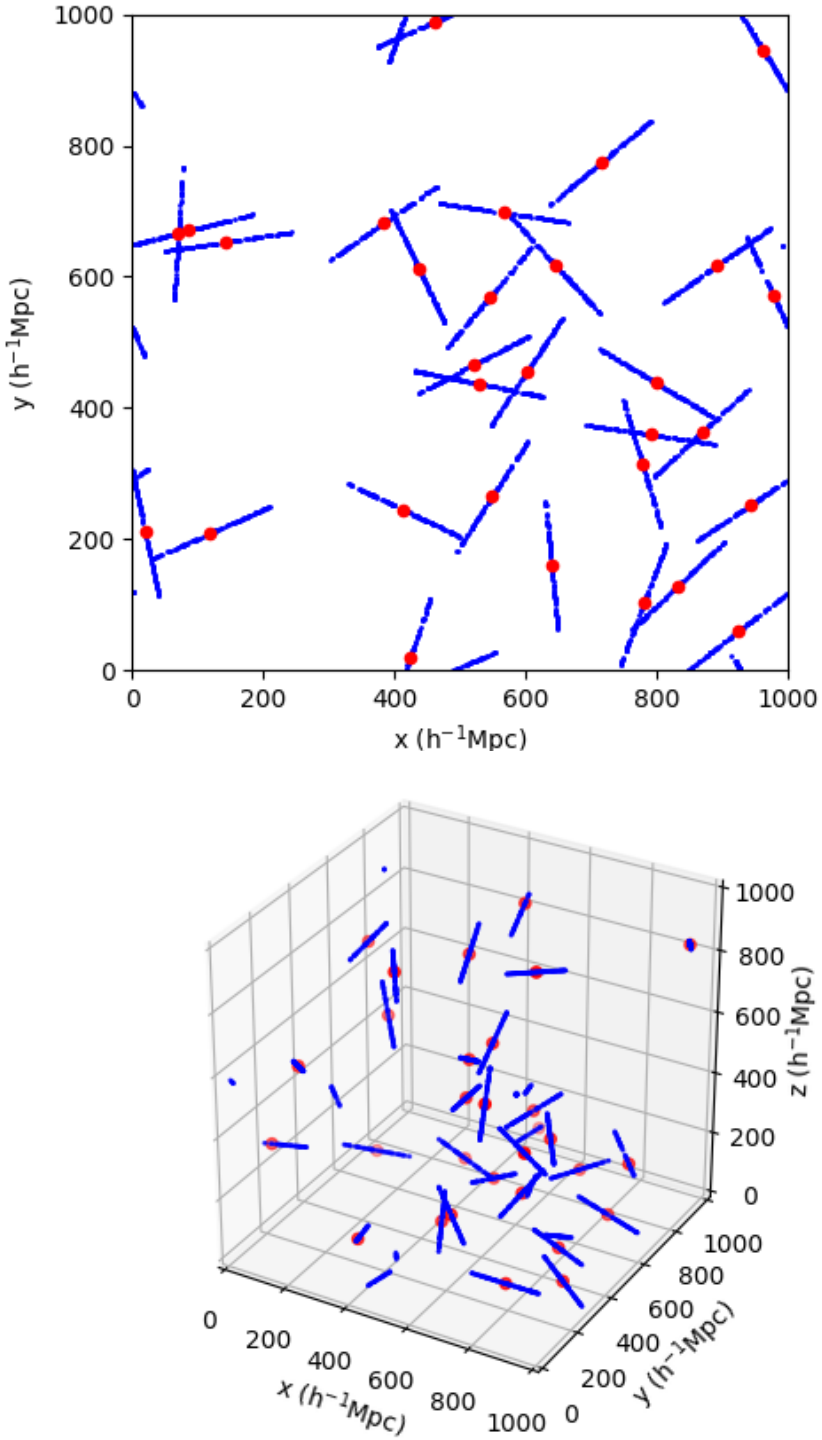

Figure 2.1: Visualisation of the 2D (top panel) and 3D (bottom panel) isotropic segment Cox process for line length of 200 h^{-1}Mpc in periodic volumes of L_{box} = 1000 h^{-1}Mpc. Red points are the 30 cluster centres and blue point samplings from the point process. The mean number of points per cluster is 100.

cluster density probability distribution is given by

$$p(r) = \frac{\theta(r)\theta(L-r)}{L}\,, \tag{2.2.22}$$

with a length of segment L and the heavyside step function $\theta(r)$ defined by

$$\theta(r) = \begin{cases} 1, & \text{if } r \geq 0 \\ 0, & \text{otherwise}\,. \end{cases} \tag{2.2.23}$$

The K-function for the segment Cox process can then be calculated from equation 2.2.21. The integrals are reduced to one dimension due to the one dimensional nature of the cluster density profile and the integral over the catalogue volume is reduced to an integral over the finite size of a cluster. This gives

$$K(r) = \frac{4}{3}\pi r^3 \langle\rho\rangle + \frac{N_p}{L^2 N_c}\int_0^L \mathrm{d}s \int_{-r}^{r} \mathrm{d}r'\theta(s+r')\theta(L-s-r')\,. \tag{2.2.24}$$

The integral can be solved graphically (see Figure 2.2) to give

$$K(r) = \begin{cases} \frac{4}{3}\pi r^3 \langle\rho\rangle + \frac{N_p}{N_c}\left(\frac{2r}{L} - \frac{r^2}{L^2}\right) & \text{if } r \leq L \\ \frac{4}{3}\pi r^3 \langle\rho\rangle & \text{otherwise}\,. \end{cases} \tag{2.2.25}$$

In other than three dimensions the $(4/3)\pi r^3 \langle\rho\rangle$ terms will change to the average number of points in a randomly placed hypersphere of radius r rather than in a three dimensional sphere, other terms are unchanged. This can then be used along with equation 2.2.12 to give the correlation function in N dimensions. In all cases the correlation function is zero on scales larger than the line length and non-zero below it. In two dimensions,

$$\xi(r) = \begin{cases} \frac{1}{\pi\lambda_s}\left(\frac{1}{rL} - \frac{1}{L^2}\right) & \text{if } r \leq L \\ 0 & \text{otherwise}\,, \end{cases} \tag{2.2.26}$$

where λ_s is the density of clusters in the volume, in this case in 2D, given by $\lambda_s = N_c/V$ for V the volume containing the catalogue. In three dimensions,

$$\xi(r) = \begin{cases} \frac{1}{2\pi\lambda_s}\left(\frac{1}{r^2 L} - \frac{1}{rL^2}\right) & \text{if } r \leq L \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2.27}$$

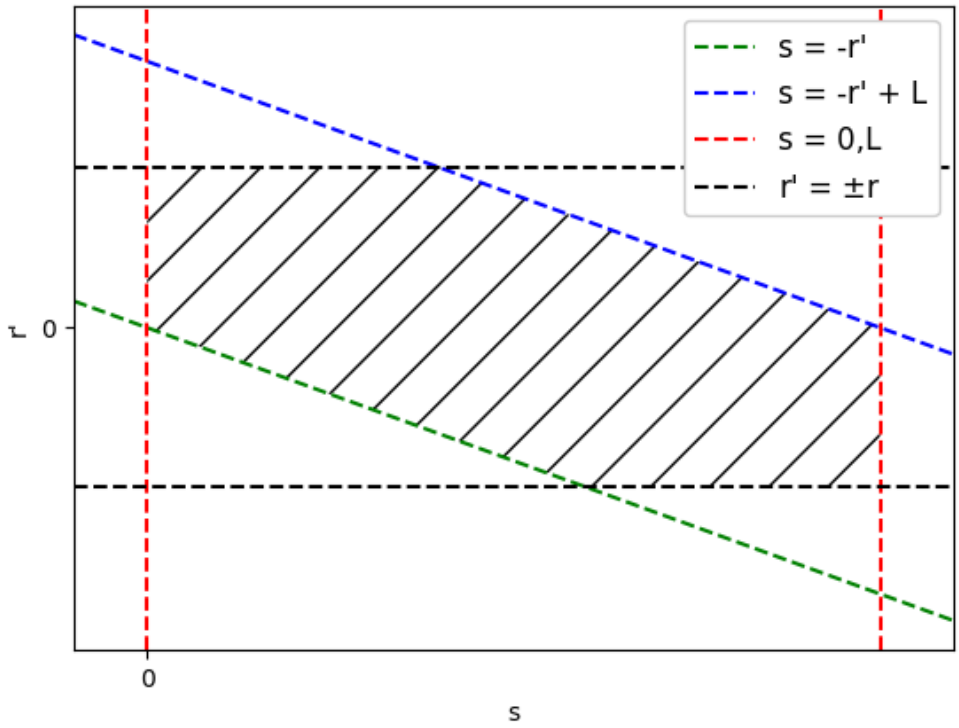

Figure 2.2: Graphical solution to the integral given in equation 2.2.24 for the K function of the segment Cox process. The area to integrate lies within all dashed lines. The black and red dashed lines come from the integration limits and the green and blue lines from the step functions in the integrand.

On small scales this expression acts like a power law that scales as $\sim r^{-\gamma}$ with $\gamma = 2$. Snethlage et al. (2002) showed how this slope can be changed so that $\gamma < 2$ through applying random shifts to the point field.

In order to calculate the projected correlation function the monopole result can be expressed in terms of projected component, r_p, and line of sight component, π, as

$$\xi(r_p, \pi) = \begin{cases} \frac{1}{2\pi\lambda_s}\left(\frac{1}{(r_p^2+\pi^2)L} - \frac{1}{\sqrt{r_p^2+\pi^2}L^2}\right) & \text{if } r_p^2 + \pi^2 \leq L^2 \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2.28}$$

We can define π_0 as the value of π at each r_p for which the correlation function drops to zero. It is given by

$$\pi_0(r_p) = \begin{cases} \sqrt{L^2 - r_p^2} & \text{if } r_p \leq L \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2.29}$$

The projected correlation function using equation 2.1.9 becomes

$$w_p(r_p) = \begin{cases} \frac{2}{\pi\lambda_s}\left(\frac{1}{Lr_p}\arctan(\frac{\pi_{max}}{r_p}) - \frac{1}{L^2}\arcsin(\frac{\pi_{max}}{r_p})\right) & \text{if } r_p \leq L \text{ and } \pi_{max} \leq \pi_0(r_p) \\ \frac{2}{\pi\lambda_s}\left(\frac{1}{Lr_p}\arctan(\frac{\pi_0(r_p)}{r_p}) - \frac{1}{L^2}\arcsin(\frac{\pi_0(r_p)}{r_p})\right) & \text{if } r_p \leq L \text{ and } \pi_{max} > \pi_0(r_p) \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2.30}$$

Verification of the 2D and 3D segment Cox process monopole results are shown in Figure 2.3. In each case 10^4 lines with an average of 100 points per line are used in a periodic square(2D)/box(3D) with side length 1000 h^{-1}Mpc. We generate a uniform random catalogue and use the Landay-Szalay estimator (Landy & Szalay, 1993a) for calculating the correlation function as the code used does not support periodic volumes. The number of randoms is set to ten times the number of data points. The number of randoms needs to be sufficient that there are enough random pairs in the smallest bins that the error on the random pair counts in those bins does not impact the final correlation function measurement. In this case we decide that ten times is sufficient based on the good agreement between the measured and theoretical results. For a measurement where the result isn't predicted beforehand the result can be retested with a new realisation of the same number of randoms

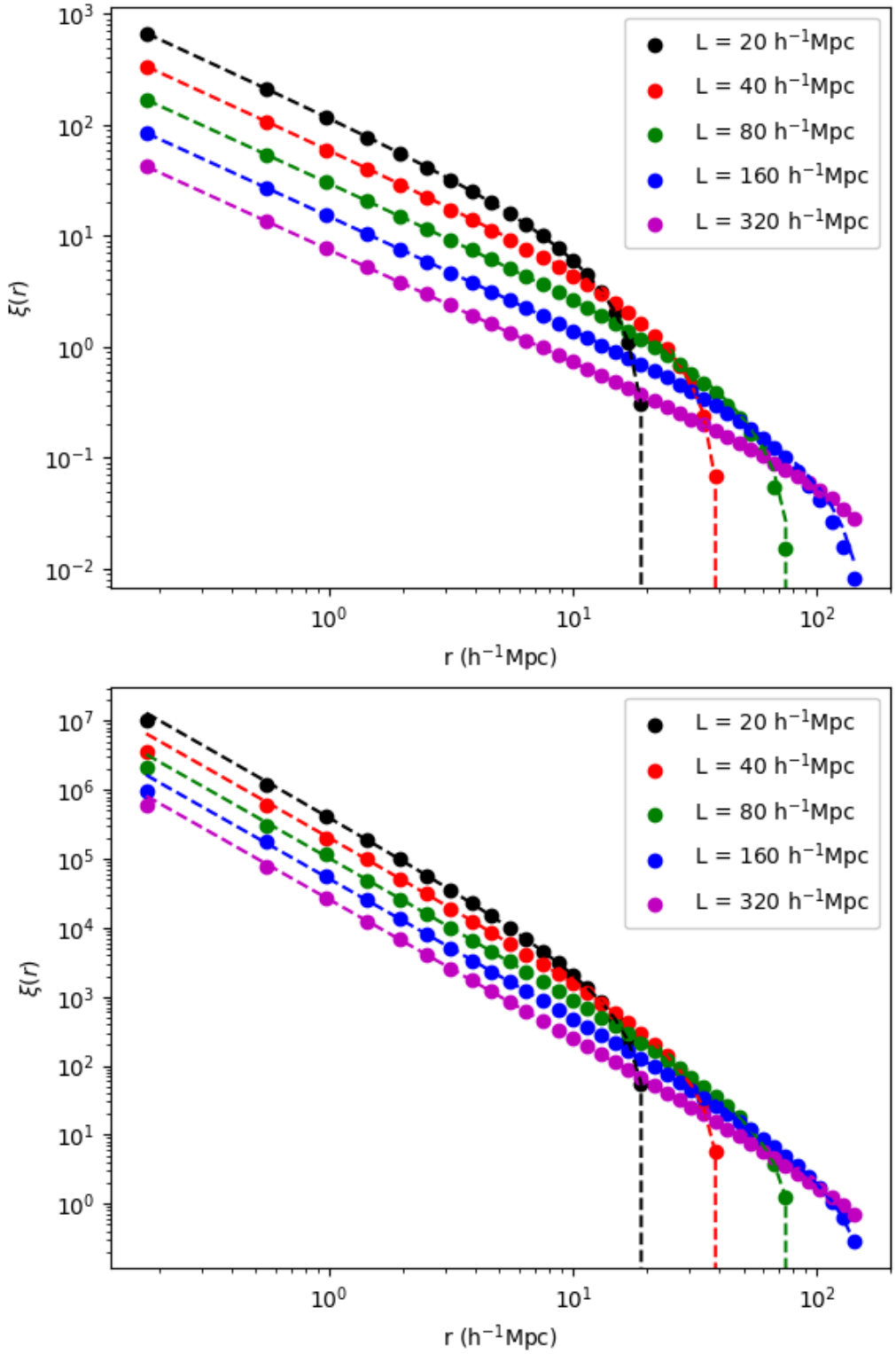

Figure 2.3: Measurement (dots) and expectation (lines) for the correlation function of the 2D (top panel) and 3D (bottom panel) isotropic segment Cox process for different values of the line length. Details of the test are given in the text.

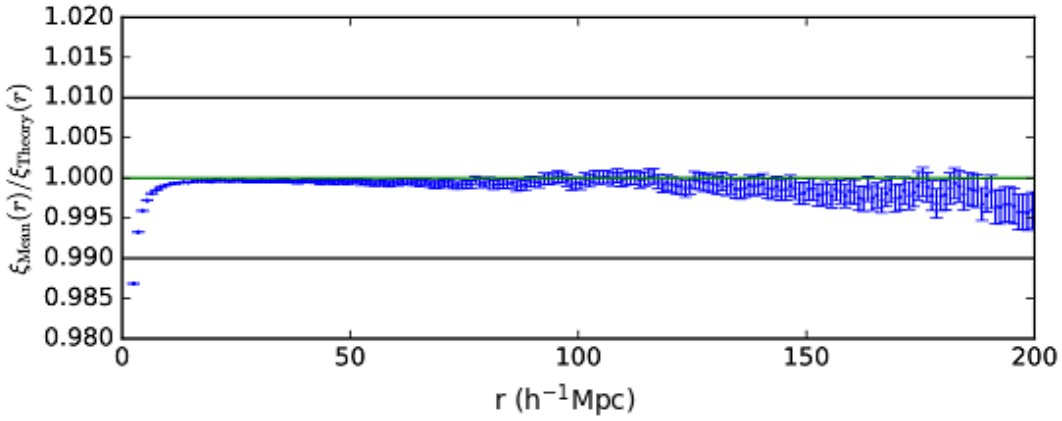

Figure 2.4: Ratio of mean of measured correlation function for 1000 Cox process mocks to Cox process theory, calculated by the Euclid 2 point correlation function pipeline for validation purposes. The errors shown correspond to errors on the mean measurement. Plot provided by Viola Allevato.

to assess its robustness. The theoretical prediction lines up well with the measured correlation function for all tested values of the line length at all scales below the line lengths.

Figure 2.4 shows validation of the Euclid 2 point correlation function pipeline using the segment Cox process. The brief mandated that the code was validated to the sub percent level in 200 linear bins between 0 and 200 h^{-1}Mpc. This plot shows this is achieved for all bins except those on the smallest scales. This is because the theoretical function varies quickly over the width of the bin so the approximation that the average value over the bin will equal the theoretical value at the centre of the bin breaks down, and because the theoretical expression is divergent as the separation approaches zero. Integrating the theoretical expression over each bin fixes the small scale discrepancy in all but the first bin, where the pole at zero makes this integration impossible. There is a small but significant disagreement on large scales, where the theoretical result is larger than the estimated one. This bias gets worse as the separation approaches the line length and the correlation function approaches zero. The line length should be set to be significantly larger than the largest relevant separation in order for the estimate to agree well with the theoretical prediction. Here the line length being 2.5 times larger than the largest separation keeps this bias

below the 1% required accuracy. In order to reach the accuracy shown the mean of 1000 Cox process mocks was calculated. Each of the mocks used 10^6 lines of 500 h^{-1}Mpc with an average of 100 points per line in a 5000 h^{-1}Mpc per side cubic box. The large number of points and realisations could be significantly reduced if different scales could be validated with different mocks.

It is worth mentioning the typical amplitude of these Cox process correlation functions. The lower the density of the clusters the higher the amplitude of the correlation function and the higher the signal to noise will be in measuring the correlation function on one cluster/halo scales. The results in figures 2.3 and 2.4 show results for correlation functions with amplitudes far above what is found in the real universe in order that the results have little scatter. The Cox process can produce correlation functions of more realistic amplitudes using higher densities of clusters but the scatter on the final result will be larger.

2.2.2 Thomas process

The Thomas process sets the cluster profile of the Neyman-Scott process to a Gaussian (Thomas, 1949). This process is investigated because the two point correlation function in 2D and 3D for an isotropic Gaussian is known from the literature (Stoyan et al., 1995; Moller & Waagepetersen, 2004). This point process is visualised in Figure 2.5 in the 2D case in a periodic box with 30 clusters, each with Gaussian standard deviation of 30 h^{-1}Mpc. The cluster centres are the same as the 2D case in figure 2.1 but the clusters are now theoretically of infinite extent.

In 2 dimensions, for a Gaussian cluster profile with standard deviation σ, the cluster density is

$$\rho_c(r) = \frac{N_p}{N_c}\sqrt{\frac{1}{(2\pi\sigma^2)^2}}\exp\left(\frac{-r^2}{2\sigma^2}\right), \tag{2.2.31}$$

and the corresponding correlation function is

$$\xi(r) = \frac{1}{\lambda_s}\sqrt{\frac{1}{(4\pi\sigma^2)^2}}\exp\left(\frac{-r^2}{4\sigma^2}\right), \tag{2.2.32}$$

where λ_s is the cluster density in either 2D or 3D. In 3D the cluster profile is given

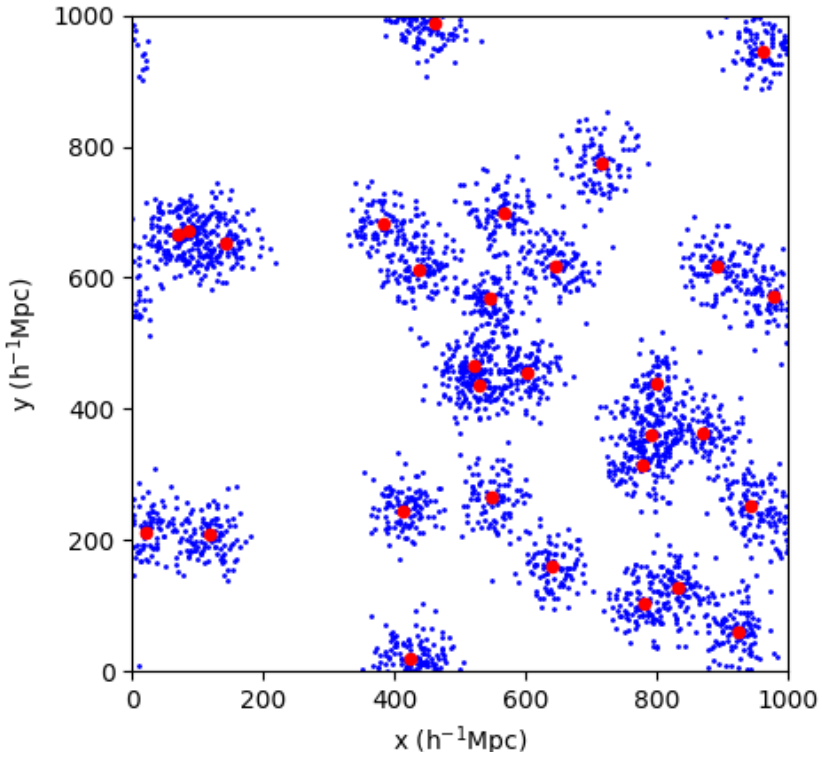

Figure 2.5: Visualisation of the 2D isotropic Thomas process for 30 Gaussian clusters with $\sigma = 30\ h^{-1}$Mpc and an average of 30 points per cluster in a periodic box of side length 1000 h^{-1}Mpc. Red points show cluster centres and blue points show samplings from the point process.

by

$$\rho_c(r) = \frac{N_p}{N_c}\sqrt{\frac{1}{(2\pi\sigma^2)^3}}\exp\left(\frac{-r^2}{2\sigma^2}\right), \tag{2.2.33}$$

which gives a correlation function of

$$\xi(r) = \frac{1}{\lambda_s}\sqrt{\frac{1}{(4\pi\sigma^2)^3}}\exp\left(\frac{-r^2}{4\sigma^2}\right). \tag{2.2.34}$$

The derivations of these results are special cases of the generalised Thomas process derivation that will be given in section 2.3.2. Verification of the 2D and 3D Thomas process results are shown in Figure 2.6. 10^4 clusters with an average of 100 points per cluster are used in a periodic square(2D)/box(3D) with side length 1000 h^{-1}Mpc. The number of randoms is set to ten times the number of data points. The theoretical prediction lines up well with the measured correlation function below $\sim 5\sigma$ for all cluster sizes tested. The open circles in the plot represent where the value of $\xi(r)$ falls below 10^{-2} and does not match well with the expectation. Beyond $\sim 5\sigma$, the correlation function moves rapidly toward zero and there are too few two-halo pairs in this low cluster density catalogue to accurately approximate such a low amplitude correlation function. The density of clusters could be increased to combat this, but the scatter in one cluster regime would increase.

2.2.3 Other examples from the literature

The Matern process is included here for completeness as it is a common result from the literature (Stoyan et al., 1995). Unlike the segment Cox process and the Thomas process this point process will not be extended to produce a known anisotropy. The Matern process is a point process in which the cluster profile is given by a sphere with uniform density. The cluster density probability distribution for a cluster of radius R with centre $\underline{s}_c$ is given by

$$p(\underline{s}) = \frac{3\theta(R - |\underline{s} - \underline{s}_c|)}{4\pi R^3}. \tag{2.2.35}$$

This point process is visualised for 30 clusters of radius 30 h^{-1}Mpc in a periodic volume in Figure 2.7. The cluster centres are shared with the 2D example in Figure 2.1 and with Figure 2.5. Compared to the Thomas process clusters shown in Figure 2.5 the clusters now have a finite extent.

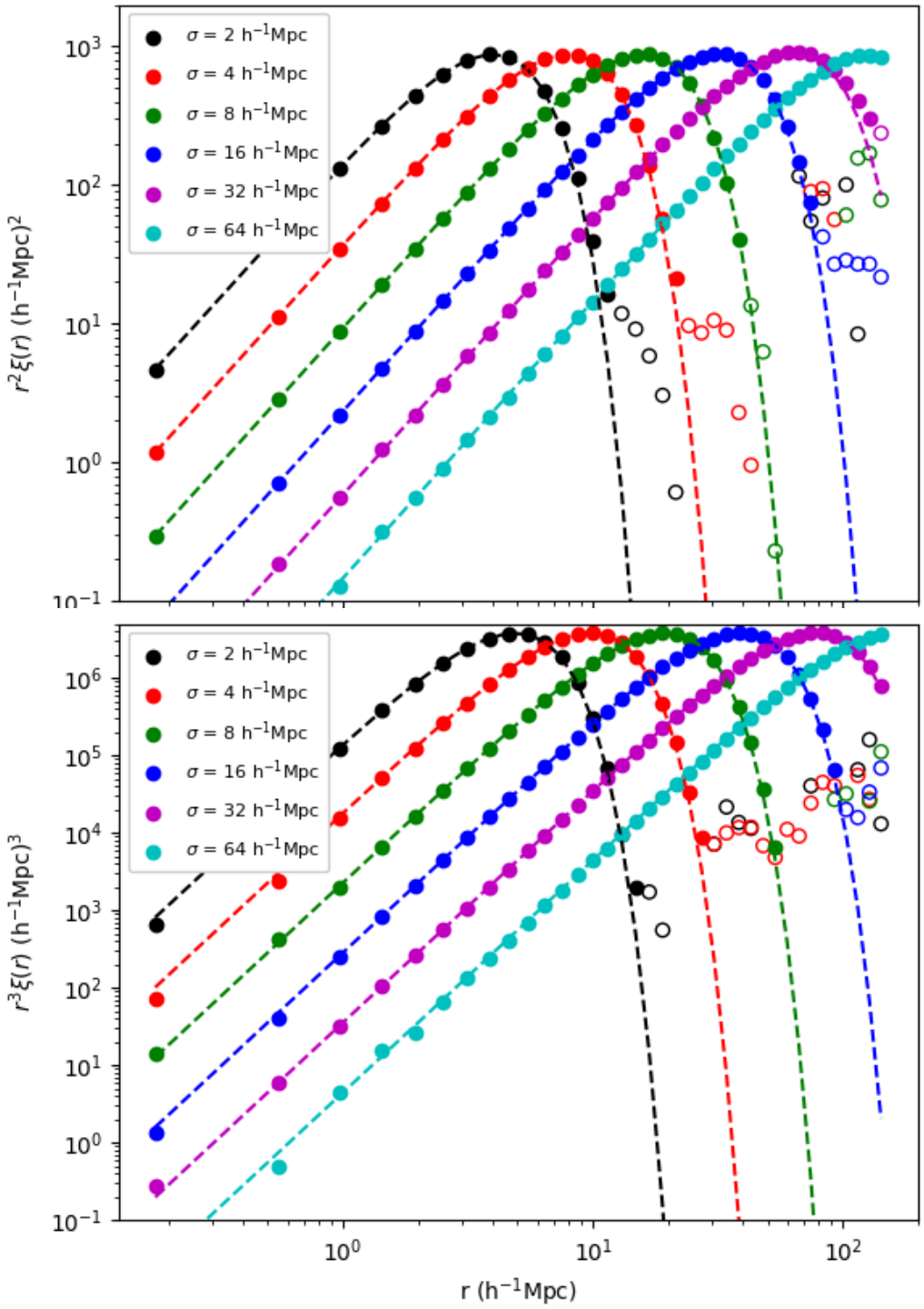

Figure 2.6: Measurement (dots) and expectation (dashed lines) for the correlation function of the 2D (top panel) and 3D (bottom panel) isotropic Thomas process for different values of the Gaussian cluster size. Open circles represent measurements in a regime where the measurement does not match the expectation well. Note the different y-axis scales.

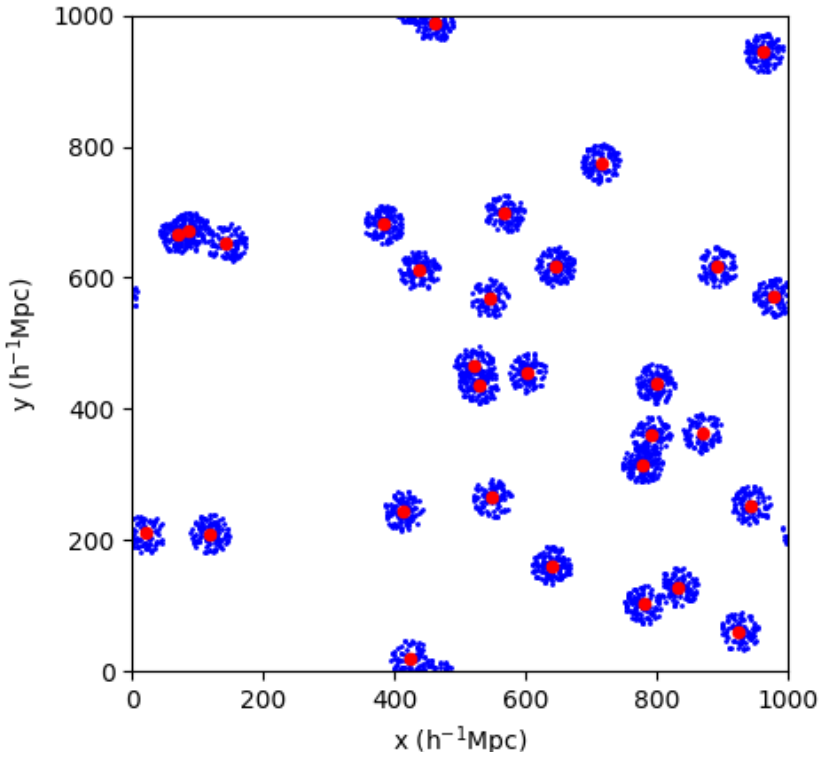

Figure 2.7: Visualisation of the 2D Matern process for 30 circles of radius 30 h^{-1}Mpc with an average of 30 points each in a periodic box of side length 1000 h^{-1}Mpc. Red points are cluster centres and the blue points are the samples drawn from the point process.

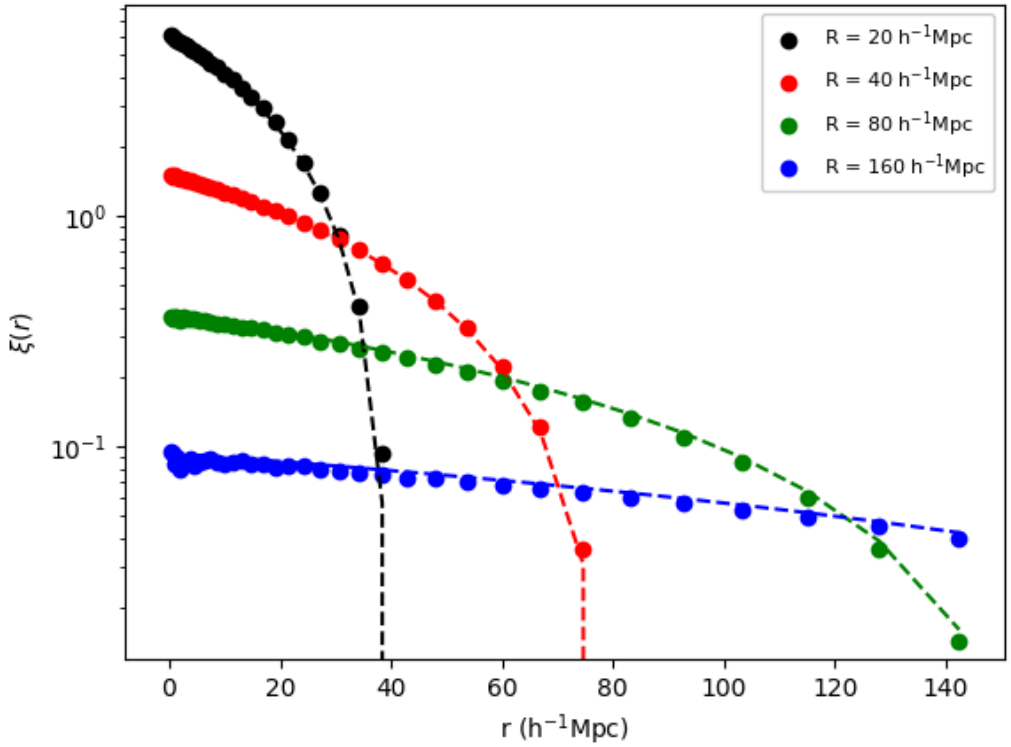

Figure 2.8: Measurement (dots) and expectation (dashed lines) for the correlation function of the 2D Matern process for different values of the sphere radius.

The correlation function for the 2D Matern cluster process is given by the expression (Stoyan et al., 1995),

$$\xi(r) = \begin{cases} \frac{2}{\pi^2 R^2 \lambda_s} \left[\arccos\left(\frac{r}{2R}\right) - \frac{r}{2R}\sqrt{1 - \frac{r^2}{4R^2}} \right] & \text{if } r \leq 2R \\ 0 & \text{elsewhere}\,, \end{cases} \tag{2.2.36}$$

with λ_s the density of clusters in the catalogue.

Verification of this result is shown in Figure 2.8 for the same volume as used in the isotropic Cox process and Thomas process. 10^4 clusters with an average of 100 points per cluster were used. The number of randoms was set to be ten times the number of data points. The theoretical predictions line up well with the measured correlation function for all tested values of the cluster radius at all scales below twice the cluster radius, or where a non-zero correlation function is predicted.

See the appendix of Sheth et al. (2001) for analytic solutions to the one halo correlation function in the case of the cluster profile of a truncated isothermal sphere,

a Hernquist profile and a truncated NFW profile.

2.3 Extending the models to non-isotropic cases

The models considered so far all produce known monopole correlation functions but have zero higher order multipoles on all scales. This section will extend the segment Cox process and the Thomas process so that they can produce known non-zero anisotropies and provide analytic predictions for the higher order multipoles of the two point correlation function. Only even multipoles will be non-zero, odd multipoles are zero through symmetry.

2.3.1 Anisotropic segment cox process

In order to extend the segment Cox process such that it has non-zero higher order multipoles, the orientation of the lines can be drawn from the desired distribution of line pair angles rather than completely at random[2]. Changing the orientation of the lines has no impact on the radial shape of the correlation function as this is determined uniquely by pairs lying on single lines, the "one cluster" term. In fact changing the angle of the lines leaves the monopole correlation function completely unchanged. This decoupling between the angular and radial components means that if line angles are drawn from a probability distribution $f(\mu)$, which must be greater than or equal to zero for $-1 < \mu < 1$, then the correlation function $\xi(r, \mu)$ is given simply by

$$\xi(r, \mu) = \xi_{cox}(r) f(\mu)\,, \tag{2.3.37}$$

where $\xi_{cox}(r)$ is the monopole correlation of the isotropic segment Cox process (equations 2.2.26 and 2.2.27 give the 2D and 3D results respectively). The multipoles are then found by

$$\xi_n(r) = \frac{2n+1}{2} \int_{-1}^{1} P_n(\mu) \xi_{Cox}(r) f(\mu) \mathrm{d}\mu\,. \tag{2.3.38}$$

[2]Inspiration for drawing the angles of the lines from a known distribution was taken from a suggestion in Stoyan et al. (1995) that "the line orientations do not need to be random".

We validate this prediction by expressing $f(\mu)$ as a linear combination of the Legendre polynomials,

$$f(\mu) = \sum_{i=0}^{\infty} a_i P_i(\mu)\,. \tag{2.3.39}$$

The reason for this choice is the simple expression that arises for the higher order multipoles due to the definition of $\xi_n(r)$ and the orthogonality of the Legendre polynomials,

$$\xi_n(r) = \frac{2n+1}{2}\int_{-1}^{1} P_n(\mu)\xi_{Cox}(r)\sum_{i=0}^{\infty} a_i P_i(\mu)\mathrm{d}\mu = \frac{a_n}{a_0}\xi_{cox}(r)\,. \tag{2.3.40}$$

The higher order multipoles are simply a scalar multiple of the monopole with the scalar being equal to the coefficient of the corresponding Legendre polynomial used in the choice of angular distribution of the lines divided by the zeroth coefficient.

In this choice of $f(\mu)$ there is a range of values for the coefficients a_i such that $f(\mu)$ is greater than zero for all possible values of μ. For the case of only non-zero monopole and quadrupole, the following condition,

$$a_0 + \frac{a_2}{2}(3\mu^2 - 1) \geq 0\,, \tag{2.3.41}$$

must be true for all values of μ in the range $-1 < \mu < 1$. This leads to constraints on the coefficients of

$$\begin{aligned} a_0 &\geq \frac{a_2}{2} \\ a_0 &\geq -a_2\,. \end{aligned} \tag{2.3.42}$$

The necessary condition becomes more complicated on including more non-zero multipoles. Graphing $f(\mu)$ to check for zero crossings is the easiest solution to finding acceptable parameters in this situation.

Figure 2.9 visualises this point process in 2D in a concentric circular volume with a random line orientation (top panel) and then an anisotropic line orientation with $a_2/a_0 = 2$ (bottom panel). The position of the cluster centres is the same in both panels to aid visualisation of the anisotropy. In the anisotropic case the lines are more radially distributed than in the random case, as $P_2(\mu)$ has a max at $\mu = 1$ and a minimum at $\mu = 0$.

This process is tested in a spherical volume. The angle μ for each line is defined relative to the origin. The local plane parallel approximation is required for the

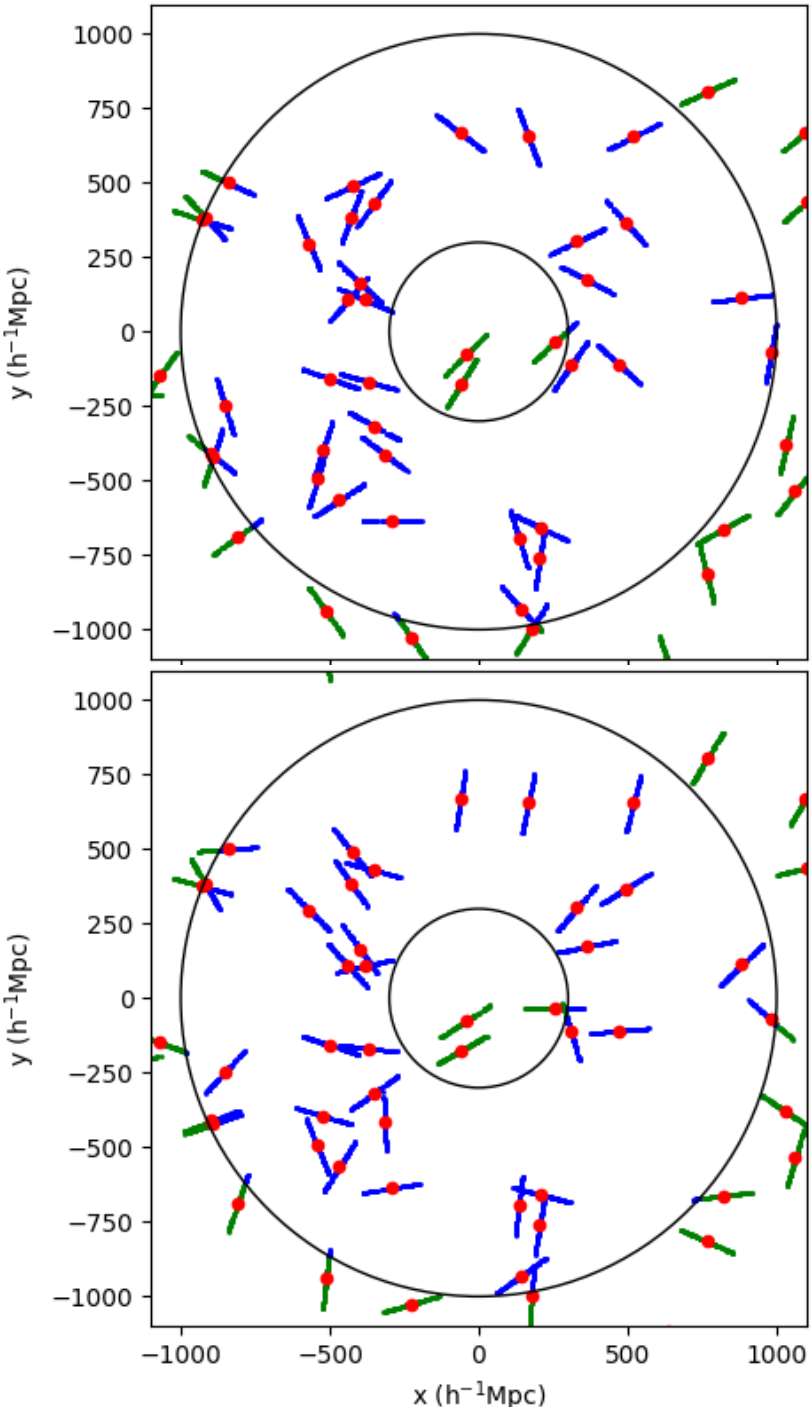

Figure 2.9: Visualisation of the 2D isotropic (top panel) and anisotropic (bottom panel) segment Cox process with L = 200 h^{-1}Mpc. Line angles in the top panel are random, while in the bottom panel they are drawn from $1 + 2P_l(\mu)$ (more radially aligned than random). Red points are cluster centres, equally positioned in both panels, blue points are accepted points in the catalogue and green points are rejected. The black circles show the volume of the final catalogue.

direction from the origin to both ends of a Cox process line to be approximately equal. In order for this approximation to hold, the distance to the line from the origin must be significantly longer than the length of the line, so cluster centres are not generated within a fixed distance from the origin. This leaves a concentric spherical volume. The volume shown in figure 2.9 is only illustrative, the inner radius is too small for the local plane-parallel approximation to hold. Care must also be taken to account for points assigned to clusters that lie near the volume boundaries that are placed outside the catalogue volume. This is accounted for by extending the inner and outer radius such that points may scatter into the volume as frequently as they scatter out of it. The number of cluster centres and points generated must also be increased so that the average density in the extended volume will match what was intended originally. The extra volume should be large enough so that no point attached to a cluster inside the original volume may scatter outside the extended volume. Points lying outside the original volume are then masked. The average number of points in a realisation will be equal to what was intended but each random realisation may contain different numbers of cluster centres and points. If this extension process is not done, a bias will be introduced into the results because of the one way scatter of the points.

Figure 2.10 shows the validation of this prediction in 3D in the case of $a_2/a_0 = 2$ for 10^4 lines each with an average of 100 points per line in a spherical volume with inner radius 1000 h^{-1}Mpc and outer radius 5000 h^{-1}Mpc. Good agreement with the prediction is seen in the monopole and quadrupole for all scales below the line length with the exception of the first data point. The first data point contains two errors, one of the breaking of the approximation that the value of the analytic function at the centre of the bin is equal to the average value over the bin, and the other that the binning in μ as well as r requires significantly greater numbers of data and random points to resolve properly. The test could be rerun with significantly larger numbers of data and random points but this will take significantly more CPU time and will still fail to resolve the first issue. It is safer to assume that the first bin may not agree well with the analytic prediction. The monopole is unchanged from the isotropic case and the quadrupole values are all simply twice the corresponding

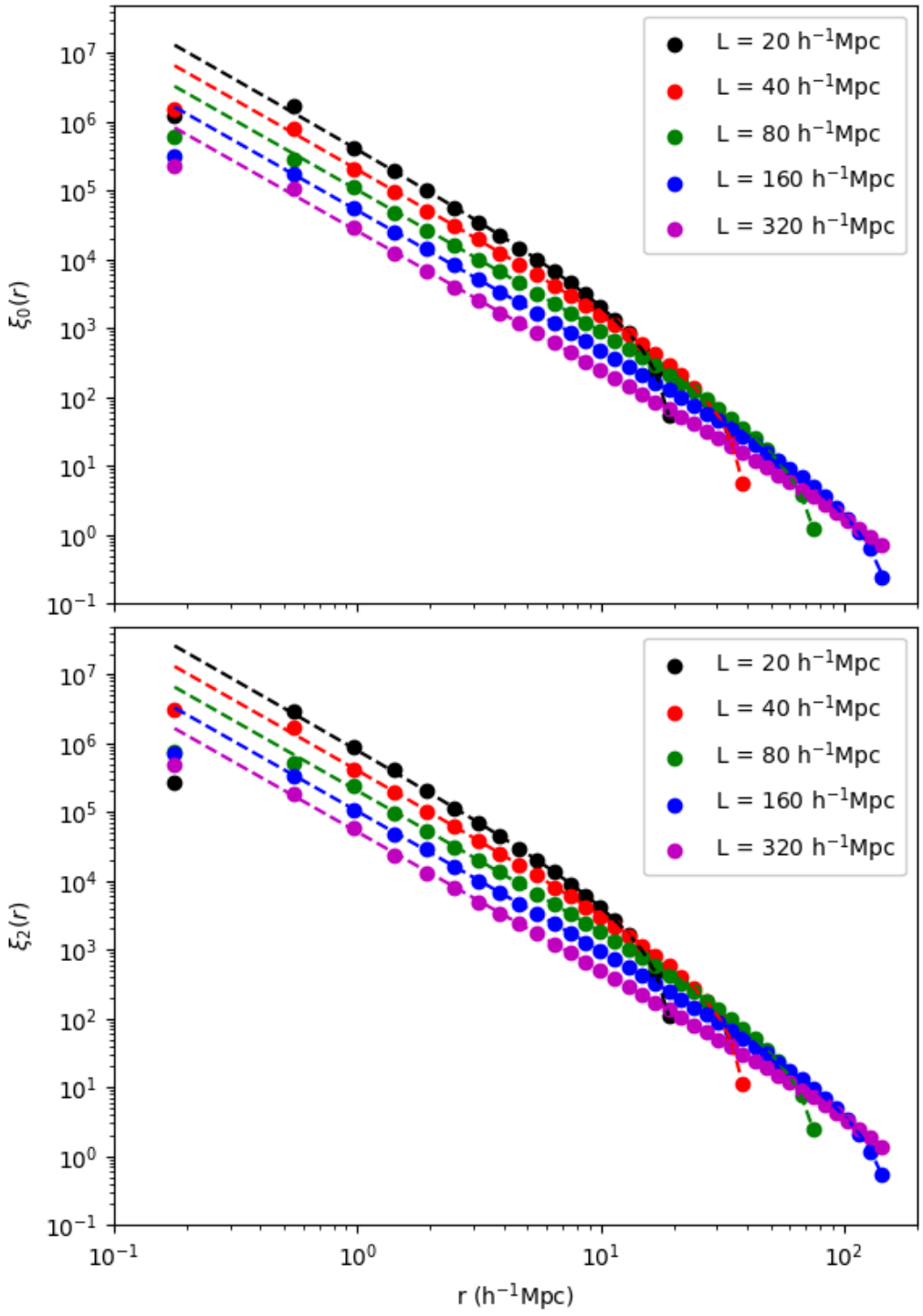

Figure 2.10: Measurements (dots) and expectations (lines) for the monopole (top panel) and quadrupole (bottom panel) of the anisotropic segment Cox process for different values of the line length. Line angles are sampled from $1 + 2P_l(\mu)$, which results in $\xi_n(r) = 0$ for $n \geq 3$.

monopole.

2.3.2 Generalised Thomas process

In order to extend the results of the Thomas process to include non-zero higher order multipoles of the two point correlation function we first show that the two point correlation function of a Neyman Scott process depends only on the one cluster term. The density, $\rho(\underline{x})$, of a Neyman-Scott point process can be written as a sum of N_c randomly placed clusters,

$$\rho(\underline{x}) = \sum_{i=0}^{N_c} \rho_i(\underline{x}) \,, \tag{2.3.43}$$

where $\rho_i(\underline{x})$ is defined as the density profile of the i^{th} cluster given by

$$\rho_i(\underline{x}) = \rho_c(\underline{x} - \underline{x}_i) \,, \tag{2.3.44}$$

for ρ_c the cluster density profile and $\underline{x}_i$ the centre of the i^{th} cluster in M dimensions.

The two point correlation function can be written,

$$1 + \xi(\underline{x}_{12}) = \frac{1}{\langle\rho\rangle^2 V} \int_V \mathrm{d}^M x \rho(\underline{x} - \underline{x}_1)\rho(\underline{x} - \underline{x}_2) \,, \tag{2.3.45}$$

where V is the volume and $\langle\rho\rangle$ is the mean density of the catalogue respectively. Due to homogeneity, ξ is only a function of $\underline{x}_{12}$, where $\underline{x}_{ij}$ is the vector separation between points i and j, given by $\underline{x}_j - \underline{x}_i$ (note $\underline{x}_{ii} = 0$). Without a loss of generality the integral can be centred on $\underline{x}_1$ to give

$$1 + \xi(\underline{x}_{12}) = \frac{1}{\langle\rho\rangle^2 V} \int_V \mathrm{d}^M x \rho(\underline{x})\rho(\underline{x} - \underline{x}_{12}) \,. \tag{2.3.46}$$

Substituting in equation 2.3.43 for a Neyman-Scott process gives

$$1 + \xi(\underline{x}_{12}) = \frac{1}{\langle\rho\rangle^2 V} \sum_{i=0}^{N_c} \sum_{j=0}^{N_c} \int_V \mathrm{d}^M x \rho_i(\underline{x})\rho_j(\underline{x} - \underline{x}_{12}) \,. \tag{2.3.47}$$

The sum of the two halo terms ($i \neq j$) tends to 1 in the limit of large N_c due to the random placing of each cluster. This cancels with the 1 one the left hand side and leaves only the contribution from the one halo term,

$$\xi(\underline{x}_{12}) = \frac{1}{\langle\rho\rangle^2 V} \sum_{i=0}^{N_c} \int_V \mathrm{d}^M x \rho_i(\underline{x})\rho_i(\underline{x} - \underline{x}_{12}) \,. \tag{2.3.48}$$

All the N_c one halo terms are identical, so this expression reduces to

$$\xi(\underline{x}_{12}) = \frac{N_c}{\langle\rho\rangle^2 V}\int_V \mathrm{d}^M x \rho_i(\underline{x})\rho_i(\underline{x}-\underline{x}_{12})\,. \tag{2.3.49}$$

We will calculate the expression for the one cluster term for the N point correlation function in M dimensions, $\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N})$, as this will also prove useful for the three point correlation function of the Thomas process explored in section 2.4. There are always N_c identical one cluster terms so $\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N})$ is given by a generalisation of equation 2.3.49,

$$\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N}) = \frac{N_c}{\langle\rho\rangle^N V}\int_V \mathrm{d}^M x \prod_{i=1}^{N}\rho_c(\underline{x}-\underline{x}_{1i})\,. \tag{2.3.50}$$

To apply this to the Thomas process we look at the case where the cluster profile is given by a generalised Gaussian profile with symmetric covariance matrix A,

$$\rho_c(\underline{x}) = \frac{N_p}{N_c}\sqrt{\frac{\det(\mathrm{A})}{\pi^M}}\exp(-\underline{x}^T A\underline{x})\,. \tag{2.3.51}$$

For example, setting

$$A = \begin{bmatrix}(2\sigma^2)^{-1} & 0\\ 0 & (2\sigma^2)^{-1}\end{bmatrix} \tag{2.3.52}$$

recovers the expression for a 2D isotropic Gaussian cluster given in equation 2.2.31. We define λ_s as the mean density of clusters, given by N_c/V. For this generalised Gaussian cluster profile equation 2.3.50 becomes

$$\begin{aligned}
&\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N})\\
&= \frac{N_c}{\langle\rho\rangle^N V}\frac{N_p^N}{N_c^N}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{N/2}\int_V \mathrm{d}^M x\exp(-\sum_{i=1}^{N}(\underline{x}-\underline{x}_{1i})^T A(\underline{x}-\underline{x}_{1i}))\\
&= \frac{1}{\lambda_s^{N-1}}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{N/2}\int_V \mathrm{d}^M x\exp(-\sum_{i=1}^{N}(\underline{x}-\underline{x}_{1i})^T A(\underline{x}-\underline{x}_{1i}))\\
&= \frac{1}{\lambda_s^{N-1}}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{N/2}\int_V \mathrm{d}^M x\exp(-N\underline{x}^T A\underline{x}+2\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x}-\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x}_{1i})\\
&= \frac{1}{\lambda_s^{N-1}}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{N/2}\exp(-\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x}_{1i})\int_V \mathrm{d}^M x\exp(-N\underline{x}^T A\underline{x}+2\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x})\,.
\end{aligned} \tag{2.3.53}$$

Making a substitution

$$\underline{x} = \frac{\underline{y}}{\sqrt{N}}\,, \tag{2.3.54}$$

leads to

$$\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N}) = \frac{1}{\lambda_s^{N-1}N^{M/2}}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{N/2}\exp(-\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x}_{1i})\int_V \mathrm{d}^M y\exp(-\underline{y}^T A\underline{y}+\frac{2}{\sqrt{N}}\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{y})\,. \tag{2.3.55}$$

The solution to this integral, a Gaussian integral with linear term, is known. This gives the final result for the one halo term of the N point correlation function in M dimensions as

$$\eta_{1H}(\underline{x}_{12},...,\underline{x}_{1,N}) = \frac{1}{\lambda_s^{N-1}N^{M/2}}\left(\frac{\det(\mathrm{A})}{\pi^M}\right)^{(N-1)/2}\exp(-\sum_{i=1}^{N}\underline{x}_{1i}^T A\underline{x}_{1i})\exp(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\underline{x}_{1i}^T A\underline{x}_{1j})\,. \tag{2.3.56}$$

For the two point correlation function, equation 2.3.56 reduces to (using $\underline{x}_{ii}=0$),

$$\xi(\underline{x}_{12}) = \frac{1}{\lambda_s}\sqrt{\frac{\det(\mathrm{A})}{(2\pi)^M}}\exp(-\frac{\underline{x}_{12}^T A\underline{x}_{12}}{2})\,. \tag{2.3.57}$$

To test this prediction we look at the isotropic case to see if we recover the literature result. Using the matrix given in equation 2.3.52 for the 2D isotropic Thomas process cluster profile, equation 2.3.57 reduces to

$$\xi(r) = \frac{1}{\lambda_s}\sqrt{\frac{1}{(4\pi\sigma^2)^2}}\exp\left(\frac{-r^2}{4\sigma^2}\right). \tag{2.3.58}$$

This is the same result as quoted from the literature in equation 2.2.32. The isotropic 3D Thomas process result is also recovered using the covariance matrix

$$A = \begin{bmatrix}(2\sigma^2)^{-1} & 0 & 0\\ 0 & (2\sigma^2)^{-1} & 0\\ 0 & 0 & (2\sigma^2)^{-1}\end{bmatrix}. \tag{2.3.59}$$

To build an anisotropic model the scale length of the Gaussian can be made larger in one dimension than the other two. This can be taken to be similar to the astrophysical case of smearing along the radial direction. In 2 dimensions this is done by setting the Gaussian covariance matrix as

$$A = \begin{bmatrix}(2\sigma_T^2)^{-1} & 0\\ 0 & (2\sigma_r^2)^{-1}\end{bmatrix}, \tag{2.3.60}$$

where σ_T is the transverse Gaussian scale and σ_r the radial scale. We once again assume that the local plane-parallel approximation holds in that the direction from the origin to all parts of a cluster is the same. A visualisation of this point process in the isotropic and the anisotropic 2D case is shown in figure 2.11. Thirty clusters with a transverse scale of 30 h^{-1}Mpc are shown. In the anisotropic case the cluster length scale was set to three times longer in the radial direction than in the transverse direction, so that the smearing of the clusters can be clearly seen. The same volume and volume extension process is used as with the visualisation of the anisotropic Cox process. Here the extension is large enough that no part of the original volume is closer than 5 σ to the edge of the extended volume. This means the probability of a point attached to a cluster centre inside the original volume has a very low probability of falling outside the extended volume.

We provide an analytic expression for the 3D case, which is equivalent to setting the Gaussian covariance matrix to

$$A = \begin{bmatrix} (2\sigma_T^2)^{-1} & 0 & 0 \\ 0 & (2\sigma_T^2)^{-1} & 0 \\ 0 & 0 & (2\sigma_r^2)^{-1} \end{bmatrix}, \tag{2.3.61}$$

which leads to a simple expression for the correlation function as a function of r_p and π in the local plane-parallel approximation of

$$\xi(r_p, \pi) = \frac{1}{8\pi^{3/2}\sigma_T^2\sigma_r\lambda_s} \exp\left(-\frac{r_p^2}{4\sigma_T^2} - \frac{\pi^2}{4\sigma_r^2} \right). \tag{2.3.62}$$

This can also be expressed in terms of r and μ (using $r_p^2 = r^2 - \pi^2$ and $\pi^2 = r^2\mu^2$) by

$$\begin{aligned} \xi(r, \mu) &= \frac{1}{8\pi^{3/2}\sigma_T^2\sigma_r\lambda_s} \exp\left(-\frac{r^2}{4\sigma_T^2} - r^2\mu^2\Big(\frac{1}{4\sigma_r^2} - \frac{1}{4\sigma_T^2}\Big) \right) \\ &= \frac{1}{8\pi^{3/2}\sigma_T^2\sigma_r\lambda_s} \exp\left(-\frac{r^2}{4\sigma_T^2} \right) \exp\left(-r^2\mu^2\Big(\frac{1}{4\sigma_r^2} - \frac{1}{4\sigma_T^2}\Big) \right) \\ &\equiv B(r) \exp(\alpha(r)^2\mu^2), \end{aligned} \tag{2.3.63}$$

for $B(r)$ and $\alpha(r)^2$ defined as

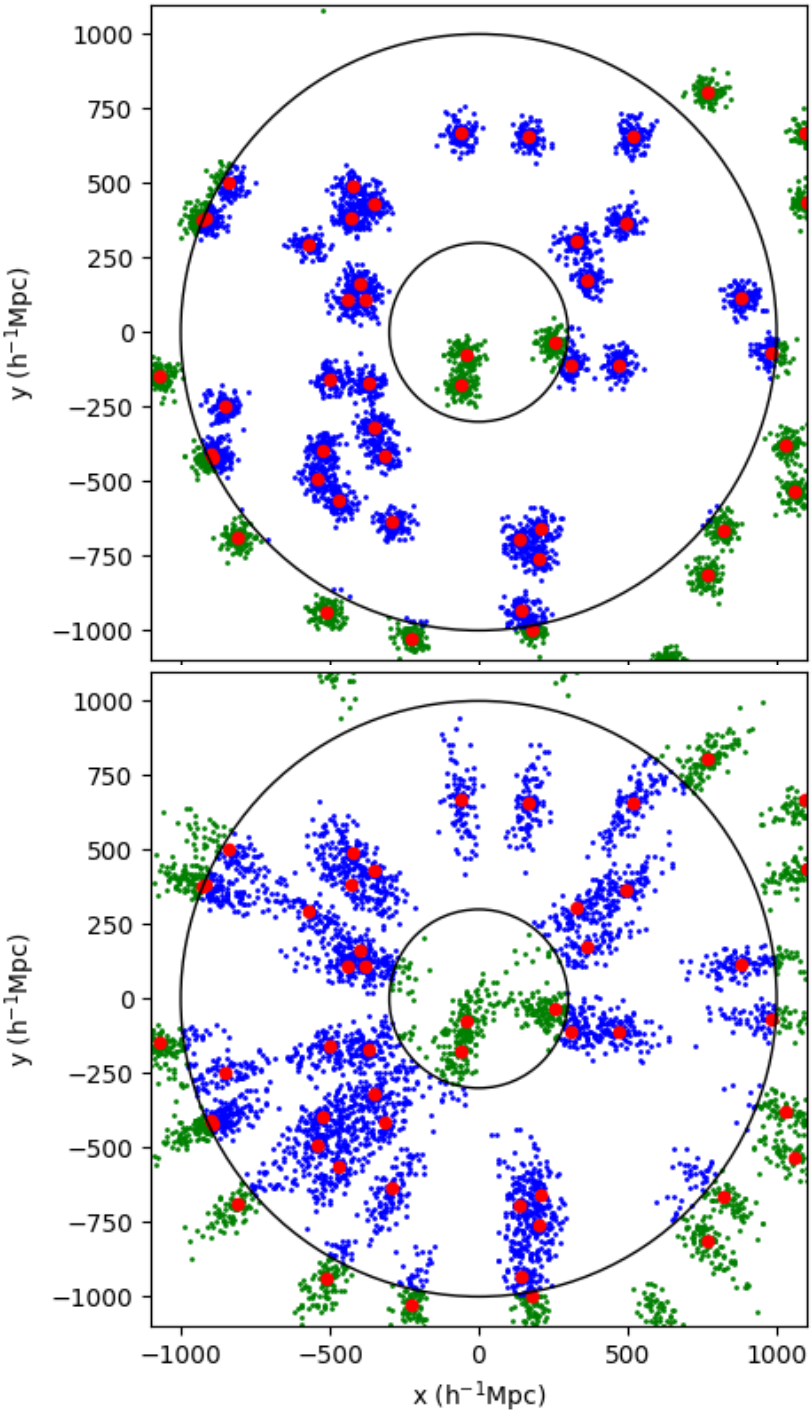

Figure 2.11: Visualisation of the 2D Thomas process for the isotropic case (top panel) and the anisotropic case (bottom panel). In the anisotropic case the radial scale is set to be three times larger than the transverse scale. Colour coding and circles have the same meaning as in figure 2.9. Cluster centre positions are shared between panels and are the same as in figure 2.9. See text for discussion.

$$B(r) \equiv \frac{1}{8\pi^{3/2}\sigma_T^2\sigma_r\lambda_s}\exp\left(-\frac{r^2}{4\sigma_T^2}\right), \tag{2.3.64}$$

$$\alpha(r)^2 \equiv -r^2\left(\frac{1}{4\sigma_r^2}-\frac{1}{4\sigma_T^2}\right). \tag{2.3.65}$$

$\alpha(r)^2$ is positive for the case we are interested in ($\sigma_r > \sigma_T$). Analytic expressions for the multipoles of this point process can then be calculated using

$$\xi_n(r) = \frac{2n+1}{2}B(r)\int_{-1}^{1}\mathrm{d}\mu\exp(\alpha(r)^2\mu^2)P(\mu)\,. \tag{2.3.66}$$

The monopole and quadrupole results are

$$\xi_0(r) = \frac{\sqrt{\pi}}{2\alpha(r)}B(r)\mathrm{erfi}(\alpha(\mathrm{r})) \tag{2.3.67}$$

$$\xi_2(r) = \frac{5}{8\alpha(r)^3}B(r)\Big(6\alpha(r)\exp(\alpha(r)^2)-\sqrt{\pi}(3+2\alpha(r)^2)\mathrm{erfi}(\alpha(\mathrm{r}))\Big) \tag{2.3.68}$$

where erfi is the imaginary error function, defined as

$$\mathrm{erfi(x)} = \frac{2}{\sqrt{\pi}}\int_0^{\mathrm{x}}\exp(\mathrm{t}^2)\mathrm{dt}\,. \tag{2.3.69}$$

The higher order multipoles are also non-zero but their results are not presented here.

The projected correlation function for this point process is the same as in the isotropic case as the transverse projection of the cluster profile is unchanged. It is given by

$$w_p(r_p) = \frac{1}{4\pi\sigma_T^2\lambda_s}\exp\left(-\frac{r_p^2}{4\sigma_T^2}\right). \tag{2.3.70}$$

These analytic predictions for the monopole and quadrupole are tested for multiple values of the transverse length scale in figure 2.12 for 10^4 clusters with an average of 100 points per cluster in the same concentric spherical volume as the anisotropic Cox process validation. The radial scale of the cluster is twice the transverse scale. Good agreement is seen in the case of the monopole below 10 times the transverse Gaussian scale, except for the first point, where similar issues arise as in the isotropic case. Good agreement is also seen in the quadrupole below scales of 10 times the transverse length scale, but there is more noise on smaller scales

than the monopole. This is somewhat to be expected as the quadrupole is typically more difficult to measure than the monopole. The open circles indicate points where either the monopole or quadrupole falls below 0.03, which is chosen to mask noisy points on large scales. This cut fails on small scales in the quadrupole. A further cut on results that are below half the transverse Gaussian scale is suggested for the quadrupole for this test. This value could be extended lower if a significantly larger volume test was used.

2.3.3 Point pair generation

This section summarises another point process that was investigated and validated, which we name "point pair generation". This is similar to the scheme first laid out in Stoyan (1994). This process can produce any correlation function with any chosen anisotropy by placing individual pairs of points draw from a chosen distribution. Placing only pairs means all higher order correlation functions of this process are zero. This process was investigated for potential use in the Euclid validation work package but was ultimately not used as it was difficult to reach the desired accuracy, and the Cox process had already proven sufficient. Results are included as they are relevant to this work. The scheme goes as follows:

- Randomly choose N_c pair centres in the volume.
- For each pair centre place a pair of points a distance r apart oriented such that the dot product of their separation vector and the line from the origin to the pair centre is μ. The distance r and value of μ are samples from an input probability distribution $f(r, \mu)$.

In the local plane parallel approximation this scheme produces a correlation function of

$$\xi(r, \mu) = \frac{f(r, \mu)}{4\pi r^2 \lambda_s}, \tag{2.3.71}$$

where λ_s is the density of the pairs in the volume equal to N_c/V for N_c the number of clusters and V the volume of the catalogue. As the function $f(r, \mu)$ is a probability distribution it must be equal to or greater than zero for all values of r and μ. The

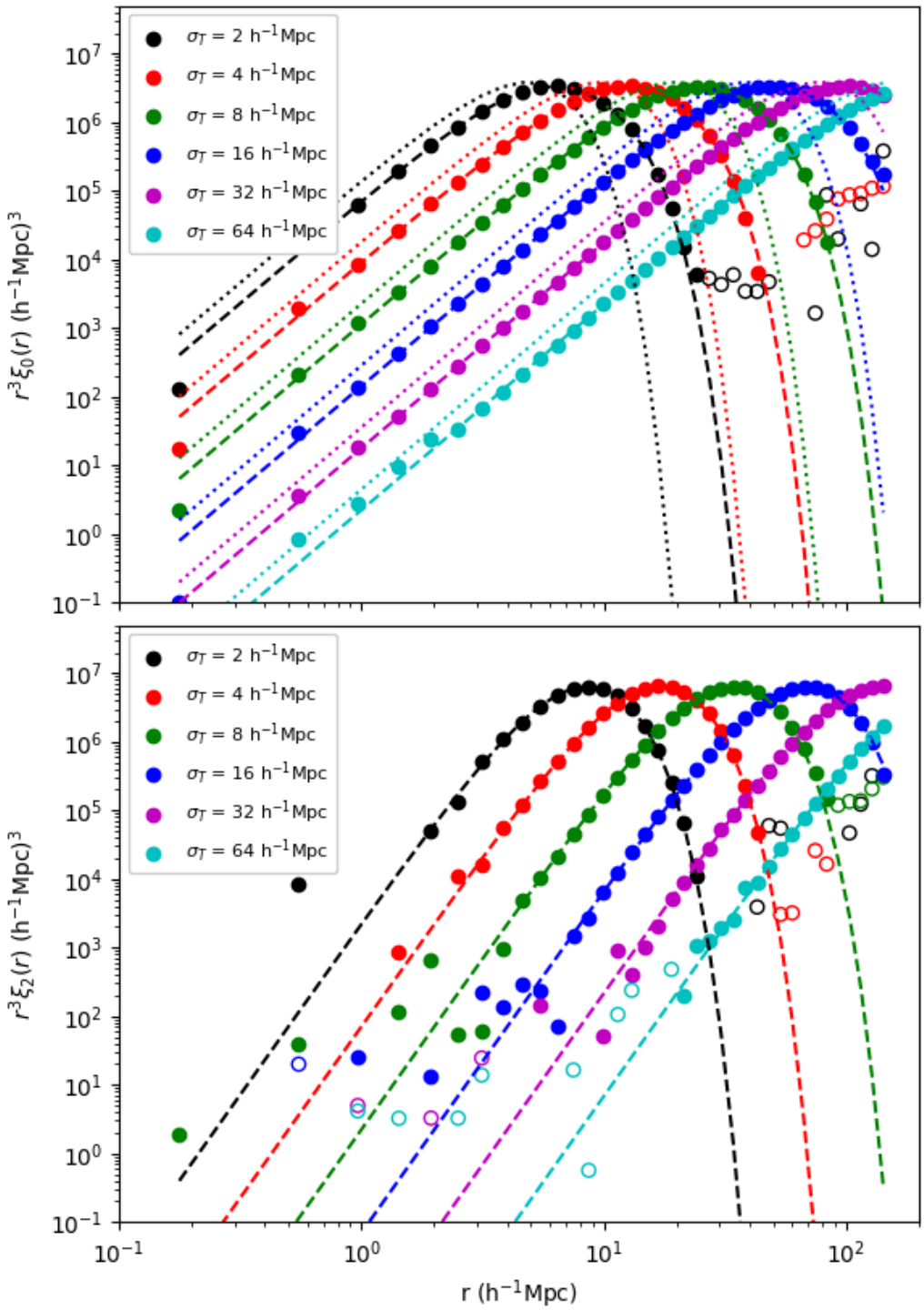

Figure 2.12: Monopole (top panel) and quadrupole (bottom panel) measurements (dots) and expectations (dashed lines) for the anisotropic Thomas process ($\sigma_r = 2\sigma_T$) for different values of the transverse Gaussian scale. The dotted line shows the corresponding monopole result for the isotropic case ($\sigma_r = \sigma_T$). Open circles show points where the value of either multipole falls below 0.03.

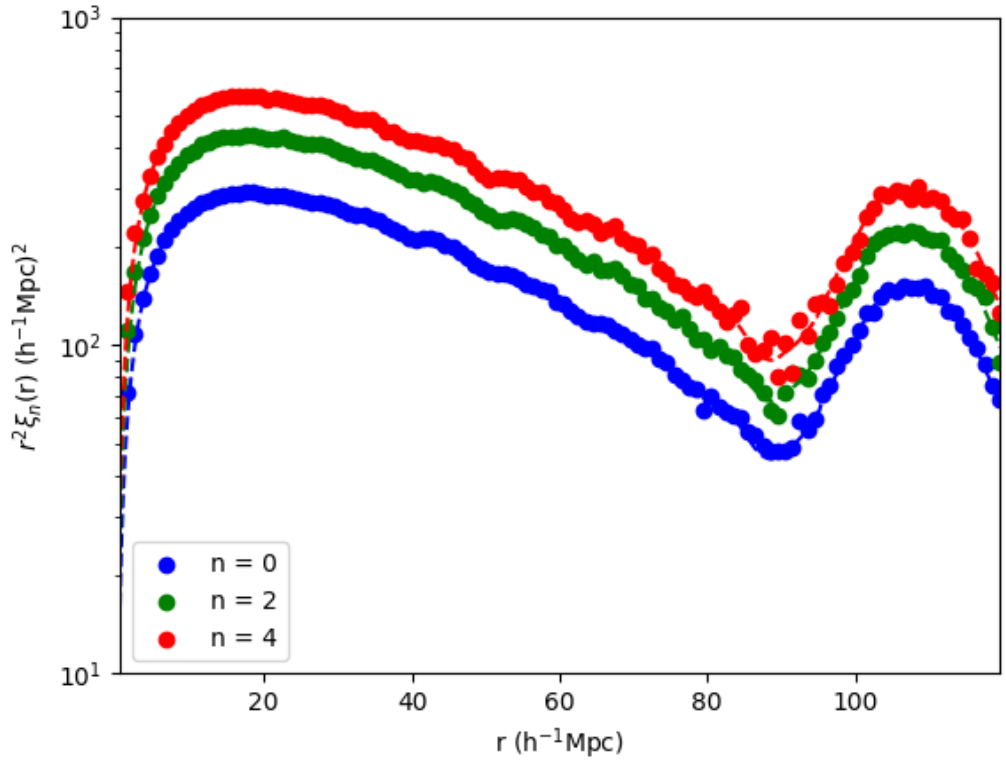

Figure 2.13: Monopole, quadrupole and hexadecapole results for the mean of 500 realisations of the pair generation process tuned to have the same shape correlation functions as the MR7 simulation initial conditions but with larger amplitude. Dashed lines show the input correlation function and dots the measured results.

function $f(r, \mu)$ sets the shape of the correlation function, and the density of pairs λ_s sets the normalisation.

We demonstrate this process by choosing $f(r, \mu)$ and λ_s such that the monopole is similar to the correlation function of SDSS LRGs found in Eisenstein et al. (2005), and the quadrupole and hexadecapole are simply multiples of this correlation function. The shape is replicated roughly by using the shape of the correlation function from the initial conditions of the Millenium MR7 simulation (Guo et al., 2013). Any values less than zero were set to zero so we could sample from the resulting $f(r, \mu)$. If we call the input monopole correlation function $\xi(r)_{MR7}$, the choice of $f(r, \mu)$ used was

$$f(r, \mu) = r^2 \xi(r)_{MR7} g(\mu)\,, \tag{2.3.72}$$

where $g(\mu)$ is the angular distribution of the lines, and is given by

$$g(\mu) = \frac{1}{2}(P_0(\mu) + 2P_2(\mu) + 3P_4(\mu)) \,, \tag{2.3.73}$$

where $P_l(\mu)$ is the l^{th} Legendre polynomial. This choice of $f(r,\mu)$ means $\xi_0(r)$ is proportional to $\xi(r)_{MR7}$. The multipole ratios are

$$\begin{aligned} \xi_2(r) &= 2\xi_0(r) \\ \xi_4(r) &= 3\xi_0(r) \,. \end{aligned} \tag{2.3.74}$$

All higher order multipoles are zero by construction. The density of pairs was then tuned such that the monopole roughly matched that of Eisenstein et al. (2005). Figure 2.13 shows the first three multipole results of this process for the mean of 500 runs of 2x10^6 points in the same concentric spherical volume as the other tests. 500 runs are needed to reach a reasonable signal to noise ratio at a more realistic amplitude than the other tests. Good agreement is seen between the theory and the measurements, and even a more complicated feature such as the BAO peak can be recovered. This process provides more flexibility than the previous models considered, but it is not particularly realistic due to only a single pair lying in each cluster.

2.4 Higher order correlation functions

This section will provide an analytic expression for the three point function of the isotropic 3D Thomas process. Extending analytic predictions to higher order correlation functions is more difficult than for the two point function. Soneira & Peebles (1978) designed a fractal process that produced an analytic prediction for higher order correlation functions but so far no analytic prediction exists for any Neyman Scott process.

We first show that the three point function of a Neyman Scott process, as with the two point function, only depends on the one cluster term. The three point correlation function is given by (Bernardeau et al., 2002),

$$\begin{aligned} &1 + \xi(\underline{x}_{12}) + \xi(\underline{x}_{13}) + \xi(\underline{x}_{23}) + \zeta(\underline{x}_{12}, \underline{x}_{13}, \underline{x}_{23}) = \\ &\frac{1}{\langle\rho\rangle^3 V} \int_V \mathrm{d}^M x \rho(\underline{x}) \rho(\underline{x} - \underline{x}_{12}) \rho(\underline{x} - \underline{x}_{13}) \,, \end{aligned} \tag{2.4.75}$$

with V the volume and $\langle\rho\rangle$ the mean density of the catalogue. $\rho(\underline{x})$ is the density of the catalogue at position $\underline{x}$. We have centred the integral on $\underline{x}_1$. We can substitute the expression for the density of a Neyman Scott process with N_c clusters (equation 2.3.43), to give

$$\begin{aligned} &1+\xi(\underline{x}_{12})+\xi(\underline{x}_{13})+\xi(\underline{x}_{23})+\zeta(\underline{x}_{12},\underline{x}_{13},\underline{x}_{23}) = \\ &\frac{1}{\langle\rho\rangle^3 V}\sum_{i=0}^{N_c}\sum_{j=0}^{N_c}\sum_{k=0}^{N_c}\int_V \mathrm{d}^M x\rho_i(\underline{x})\rho_j(\underline{x}-\underline{x}_{12})\rho_k(\underline{x}-\underline{x}_{13})\,. \end{aligned} \tag{2.4.76}$$

The expression on the right hand side now splits into one, two and three cluster terms, defined as where the three points in the triangle configuration lie between one, two and three clusters respectively. Similar to the two halo term for the two point function, the three halo term ($i \neq j \neq k \neq i$) tends to unity in the limit of large N_c as the clusters are randomly distributed. Looking at one of the cases of the two halo term in the limit of large N_c where $i = j \neq k$,

$$\frac{1}{\langle\rho\rangle^3 V}\sum_{i=0}^{N_c}\int_V \mathrm{d}^M x\rho_i(\underline{x})\rho_i(\underline{x}-\underline{x}_{12})\rho(\underline{x}-\underline{x}_{13})\,, \tag{2.4.77}$$

where the sum over k clusters has been collapsed into the $\rho(\underline{x}-\underline{x}_{13})$ term. This term can be seen to be equal to multiplying by $\langle\rho\rangle$ as it is randomly distributed relative to the other clusters. This reduces this two halo term to

$$\frac{1}{\langle\rho\rangle^2 V}\sum_{i=0}^{N_c}\int_V \mathrm{d}^M x\rho_i(\underline{x})\rho_i(\underline{x}-\underline{x}_{12})\,. \tag{2.4.78}$$

This term is equal to our definition for $\xi(\underline{x}_{12})$ of a Neyman Scott process from equation 2.3.57. For the two cluster terms where $i \neq j = k$,

$$\frac{1}{\langle\rho\rangle^3 V}\sum_{j=0}^{N_c}\int_V \mathrm{d}^M x\rho(\underline{x})\rho_j(\underline{x}-\underline{x}_{12})\rho_j(\underline{x}-\underline{x}_{13})\,. \tag{2.4.79}$$

We can recentre this integral so it reads

$$\frac{1}{\langle\rho\rangle^3 V}\sum_{j=0}^{N_c}\int_V \mathrm{d}^M x\rho(\underline{x}+\underline{x}_{12})\rho_j(\underline{x})\rho_j(\underline{x}-\underline{x}_{23})\,. \tag{2.4.80}$$

The $\rho(\underline{x}+\underline{x}_{12})$ is equivalent to multiplying by $\langle\rho\rangle$ as it is randomly distributed relative to the other clusters, so we are left with

$$\frac{1}{\langle\rho\rangle^2 V}\sum_{j=0}^{N_c}\int_V \mathrm{d}^M x\rho_j(\underline{x})\rho_j(\underline{x}-\underline{x}_{23})\,, \tag{2.4.81}$$

which is equal to the definition for the two point correlation term $\xi(\underline{x}_{23})$. In a similar way the final two cluster terms ($i = k \neq j$) can be shown to be equal to $\xi(\underline{x}_{13})$ to leave only the one halo term to contribute to the value of the three point correlation function,

$$\zeta(\underline{x}_{12}, \underline{x}_{13}, \underline{x}_{23}) = \frac{N_c}{\langle\rho\rangle^3 V} \int_V \mathrm{d}^M x \rho_c(\underline{x}) \rho_c(\underline{x} + \underline{x}_{12}) \rho_c(\underline{x} + \underline{x}_{13}) \,, \tag{2.4.82}$$

where $\rho_c(\underline{x})$ is the cluster density profile. It is possible that this generalises such that the N-point function of a Neyman Scott process only depends on the one cluster term but this is not investigated.

We can now use the result for the one halo term of the N point correlation function in M dimensions of a generalised Thomas process (2.3.56), requoted here,

$$\begin{aligned} &\eta(\underline{x}_{12}, ...\underline{x}_{N-1,N}) \\ &= \frac{1}{\lambda_s^{N-1} N^{M/2}} \left(\frac{\det(\mathrm{A})}{\pi^M} \right)^{(N-1)/2} \exp(-\sum_{i=1}^{N} \underline{x}_{1i}^T A \underline{x}_{1i}) \exp(\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \underline{x}_{1i}^T A \underline{x}_{1j}) \,, \end{aligned} \tag{2.4.83}$$

where λ_s is the density of clusters defined as N_c/V and A is the symmetric covariance matrix of the generalised Gaussian cluster profile. Setting N=3 and using that $x_{ii} = 0$ reduces this to

$$\begin{aligned} &\zeta(\underline{x}_{12}, \underline{x}_{13}) = \\ &= \frac{\det(\mathrm{A})}{\lambda_s^2 3^{M/2} \pi^M} \exp(-\underline{x}_{12}^T A \underline{x}_{12}) \exp(-\underline{x}_{13}^T A \underline{x}_{13}) \exp(\frac{1}{3} (\underline{x}_{12} + \underline{x}_{13})^T A (\underline{x}_{12} + \underline{x}_{13})) \,. \end{aligned} \tag{2.4.84}$$

We look at the simple case of an isotropic Gaussian cluster in three dimensions. This choice leads to a three point correlation function of

$$\zeta(|\underline{x}_{12}|, |\underline{x}_{13}|, \theta) = \frac{1}{\lambda_s^2 3^{3/2} \pi^3 8 \sigma^6} \exp\Big(-\frac{1}{3\sigma^2} (|\underline{x}_{12}|^2 + |\underline{x}_{13}|^2 - |\underline{x}_{12}||\underline{x}_{13}| \cos\theta) \Big) \,, \tag{2.4.85}$$

where θ is the opening angle of the triangle i.e. the angle between $\underline{x}_{12}$ and $\underline{x}_{13}$. The reduced three point function is defined as (Bernardeau et al., 2002)

$$q(|\underline{x}_{12}|, |\underline{x}_{13}|, \theta) = \frac{\zeta(|\underline{x}_{12}|, |\underline{x}_{13}|, \theta)}{\xi(|\underline{x}_{12}|)\xi(|\underline{x}_{13}|) + \xi(|\underline{x}_{12}|)\xi(|\underline{x}_{23}|) + \xi(|\underline{x}_{13}|)\xi(|\underline{x}_{23}|)} \,, \tag{2.4.86}$$

where $\xi(|\underline{x}_{12}|)$ is the monopole of the corresponding two point correlation function. For the case here the reduced correlation function is given by

$$q(|\underline{x}_{12}|, |\underline{x}_{13}|, \theta) = \frac{8}{3^{3/2}} \frac{\exp\Big(-\frac{1}{3\sigma^2}(|\underline{x}_{12}|^2 + |\underline{x}_{13}|^2 - |\underline{x}_{12}||\underline{x}_{13}| \cos\theta)\Big)}{\exp\Big(-\frac{1}{4\sigma^2}(|\underline{x}_{12}|^2 + |\underline{x}_{13}|^2)\Big) + ...}, \qquad (2.4.87)$$

where the two terms not written in the denominator look the same as the one that is given but with the other two cyclic permutations of $|\underline{x}_{12}|$, $|\underline{x}_{13}|$, $|\underline{x}_{23}|$. The benefit of the reduced function that can be seen here is that most of the scale terms at the front of the expression have cancelled. The reduced three point correlation function (also known as the third hierarchical amplitude) is also a useful statistic in the real universe as it is scale independent in the weakly linear regime (Bernardeau et al., 2002).

We will look at the results for $\sigma = 1\ h^{-1}$Mpc for the two cases where $|\underline{x}_{12}| = |\underline{x}_{13}|$ and $|\underline{x}_{12}| = 2|\underline{x}_{13}|$. Figure 2.14 plots the reduced three point function in both cases for multiple values of $|\underline{x}_{13}|$ as a function of opening angle θ. Results for equilateral triangles are scale independent, but the reduced three point function falls rapidly with triangle scale for other configurations. For equilateral triangles, setting $|\underline{x}_{12}| = |\underline{x}_{13}| = |\underline{x}_{23}|$ and $\theta = \pi/3$ in equation 2.4.87 results in $q(|\underline{x}_{12}|, |\underline{x}_{13}|, \theta) = 8/3^{(5/2)}$, which explains the scale independence of this configuration. For two fixed side lengths, q has a maximum at a value of θ between 0 and π. This is in contrast to results for the reduced three point function in the real universe that show that the value of the three point function for fixed $|\underline{x}_{12}|$ and $|\underline{x}_{13}|$ often has a minimum in that range (McBride et al., 2011). This minimum in the real universe is postulated to be a result of filamentary structure that boosts the likelihood of finding triplets of galaxies aligned rather than at other angles to each other. The spherical structure for which we provide an example above does not boost these aligned triangles in the same way. Analytic expressions for cylindrically symmetric cluster profiles rather than spherical could provide an avenue to model the three point function in the future.

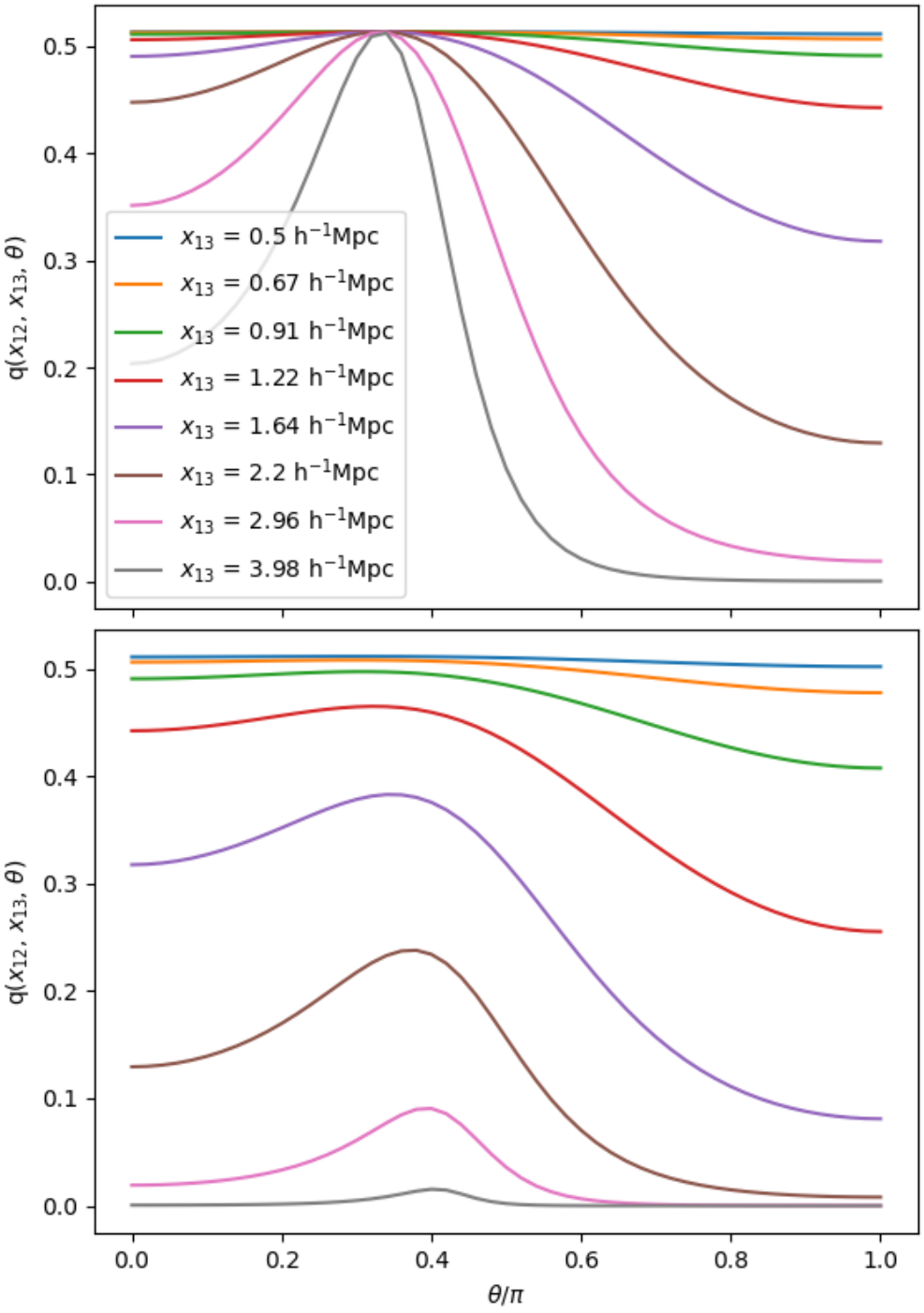

Figure 2.14: Analytic results for the reduced three point correlation function for the 3D isotropic Thomas process with $\sigma = 1$ h^{-1}Mpc. The cases of $|\underline{x}_{12}| = |\underline{x}_{13}|$ (top panel) and $|\underline{x}_{12}| = 2|\underline{x}_{13}|$ (bottom panel) are shown for multiple values of $|\underline{x}_{13}|$ as a function of the opening angle of the triangle θ. The key applies to both panels.

2.5 Conclusion

This work extends two common Neyman Scott point processes, the segment Cox process and the Thomas process, such that their higher order multipoles can be non-zero and predicted analytically.

This work first summarises and validates the literature results for common point processes as the current literature in cosmology is sparse. The segment Cox process (Stoyan et al., 1995), a point process that places points on randomly oriented lines of fixed length, is derived and successfully validated. The Thomas process (Thomas, 1949), a point process that places points in gaussian clusters of fixed size is also successfully validated. Results and validation are also presented for the 2D Matern process (Stoyan et al., 1995).

We present the successful validation of the Euclid two point statistics monopole pipeline using 1000 segment Cox process mocks. The sub-percent accuracy required is reached for scales below 200 h^{-1}Mpc.

The segment Cox process is extended to produce a known anisotropy by sampling the angles of the lines relative to the origin from a non-uniform distribution. This is validated successfully for the case that the angles of the lines are drawn from linear combinations of Legendre polynomials, which produces higher order multipoles proportional to the literature monopole result.

The Thomas process result is extended to the case of a general Gaussian cluster profile rather than an isotropic one. The result for the one cluster term of the N point correlation function of this generalised process is derived in M dimensions. We show that this term is the only contribution to the two point function of a Neyman Scott process. This result reduces to the literature results for the two point correlation function by setting N = 2 and M = 2 or 3. Results for the monopole and quadrupole are presented for the three dimensional case of a Gaussian cluster profile where the scale is extended in one dimension similar to the smearing of structure from the random motions of galaxies. The results are validated and scales where the predictions are reliable for the test done in this work are provided.

We show that, like the two point function, the only contribution to the three point correlation function of a Neyman Scott process is the one halo term. The result for

the one cluster term of the N point correlation function of the Thomas process is then used to give an analytic prediction for the 3D three point and reduced three point correlation function of an isotropic Gaussian cluster. Unlike results in the real universe, the reduced three point function of this process is only scale independent for equilateral triangles and has a maximum for a value of θ between 0 and π for all configurations. This is the first analytic prediction for a higher order correlation function of a Neyman Scott process.

The higher order correlation function predictions for the Thomas process are yet to be validated and could be explored in future work. Future work could also extend the work done here to look at the results for higher order correlation functions of a Neyman Scott process with a more realistic cluster profile for use in one halo term results of HOD modeling. The finite extent of some clusters such as the truncated NFW profile (Navarro et al., 1996), makes analytic predictions for higher order correlation functions significantly more difficult to derive than for the Thomas process. Further, spherically symmetric cluster profiles fail to boost the likelihood of aligned triangles like is seen in the three point function in the real universe, which is most likely as a result of filamentary structure (McBride et al., 2011). Using cylindrically symmetric cluster profiles rather than spherically symmetric ones could provide an interesting avenue for exploration into analytically modeling the three point function.

Chapter 3

Two point correlation function code 2PCF

This chapter presents the publicly available C++ two point clustering statistics code 2PCF. The code is similar in scope to CUTE (Alonso, 2012), but includes more flexible IO, on the fly jackknife calculations, a flexible binning scheme and implements the pair upweighting scheme of Bianchi & Percival (2017). An extension to this scheme used to account for realisations of different survey footprint positions is presented. The scaling performance of the code is presented and the code is shown to scale nearly ideally on a multi-threaded CPU. The results of Smith et al. (2018) are presented, who successfully use this implementation of the pair upweighting scheme to correct for the targeting incompleteness of a mock DESI BGS galaxy catalogue.

3.1 Introduction

The two point correlation function is an important tool for studying the spatial distribution of objects in the universe. It can help place constraints on cosmological models through BAO or RSD measurements and can help infer galaxy physics and dynamics on small scales. It is important that measurements of this quantity are accurate, precise and free of systematic biases. Upcoming surveys such as DESI (DESI Collaboration et al., 2016) and Euclid (Laureijs et al., 2011), that plan to

measure redshifts for tens of millions of objects, will require very high precision calculations for the errors in the calculation to be significantly smaller than the statistical errors from the data.

The most common method of estimating the two point correlation function is defined in Landy & Szalay (1993a),

$$\xi(\underline{r}) = \frac{DD(\underline{r}) - 2DR(\underline{r}) + RR(\underline{r})}{RR(\underline{r})}\,, \tag{3.1.1}$$

where DD, DR and RR are normalised data-data, data-random and random-random pair counts, for a random catalogue with the same density distribution as the data but Poisson distributed. The random catalogue is used to include the effects of complicated survey geometries. Calculating pair counts is a problem that scales as $\mathcal{O}(N^2)$ with N the number of points in the catalogue. With galaxy surveys becoming larger, the computational requirements are significantly increasing.

Two point correlation function codes must also be easily adaptable to include new measurement methods. One example of this is the pair upweighting scheme presented in Bianchi & Percival (2017) used to correct for biases due to missing observations. Any two point correlation function code implementation must be precise, fast and flexible.

Publicly available codes exist to estimate the two point correlation function. CUTE (Alonso, 2012) is a C based code that supports multiple types of output of the two point correlation function. The code is fast, runs in parallel and is reasonably intuitive to use. It does however lack some features that could be useful. It is restricted in IO to solely ASCII files, it only supports linear binning in the r_p, π decomposition, can not automatically calculate resampling errors and has no support for any specific missing observation correction methods. Attempts at adding in new features have proven difficult, one reason being the more time consuming nature of developing C code vs a more modern C++ implementation.

Two other publicly available codes are TreeCorr, first presented in Jarvis et al. (2004) and the two point correlation function component of CosmoBolognaLib (CBL) (Marulli et al., 2016). Both of these codes are broader in scope than CUTE. TreeCorr has the ability to perform two and three point correlation function measurements

and lensing measurements and CBL aims to provide a common framework for measuring and modeling many common cosmological statistics. Both of these provide more modern C++ approaches with options to use python wrappers for simpler interfacing, but TreeCorr still lacks some of the specific two point functionality that CUTE does. CBL provides much more functionality but still lacks support for the missing observation correction of Bianchi & Percival (2017). Rather than try to implement such a scheme on such a large and broad project it was decided it was easier to write a code of smaller scope from scratch that fulfills the requirements of being a fast, precise, feature rich and flexible code.

This chapter presents a two point correlation function code, `2PCF`, that is publicly available at `https://github.com/lstothert/two_pcf`. The approach taken is very similar to that of CUTE (Alonso, 2012) in that it focuses solely on two point correlation function statistics using nearest neighbour cell searching to speed up performance, but is written in C++ and provides more user flexibility and features.

This code is used throughout this thesis for two point clustering measurements and is part of the basis for building a galaxy group detection algorithm in chapter 5. The code has been used to investigate constraints on f(R) modified gravity models using marked correlation functions (Hernández-Aguayo et al., 2018). The code implements the pair weighting scheme from Bianchi & Percival (2017), a feature that is used in Smith et al. (2018) to correct clustering measurements from a simulation of the DESI Bright Galaxy Survey (BGS) (DESI Collaboration et al., 2016).

Section 3.2 provides a summary of the features of the code. Section 3.3 then goes into more detail about the implementation of the main features of the code and section 3.4 summarises the scaling performance of the code. Section 3.5 of this chapter will then focus on the application of this code in Smith et al. (2018) in correcting for DESI BGS fibre collision effects. Section 3.6 presents the conclusions. This chapter will not provide details of how to use the code, this is covered in the README file included with the code.

3.2 Feature summary

3.2.1 Output

The code provides estimates of many two point correlation statistics:

- The spherically averaged 2pt correlation function $\xi_0(r)$.
- The 2D cartesian decomposition of the 2pt function $\xi(r_p, \pi)$.
- The "spherical" decomposition of the 2pt function $\xi(s, \mu)$.
- The angular correlation function, w(θ).

All calculations are exact, i.e no approximations are used. Only the angular correlation function must be run separately the others can be toggled on or off and output together in a single run. A separate python script is included to calculate the projected correlation function. See section 3.3.2 for more on the 2D decomposition calculations. Linear and logarithmic bins are supported in all cases (section 3.3.3). Shared memory parallelisation is supported (section 3.3.5). The code will automatically calculate jackknife region statistics with little extra computational expense (section 3.3.4). Individual galaxy weights are supported. The pair upweighting scheme presented in Bianchi & Percival (2017) is supported (section 3.3.6).

3.2.2 Input

The code calculates two point statistics for any catalogue of galaxies and a corresponding catalogue of randoms. Periodic boundary conditions are not currently supported.

An ASCII parameter file must be provided to set the chosen code options. An example is provided with the code.
The following data file formats are supported:

- ASCII
- hdf5 [1]

[1]The pair upweighting scheme currently only supports hdf5 files.

The following coordinate systems are supported as input:

- Equatorial (ra, dec, redshift)
- Cartesian (x, y, z)

Currently only a flat ΛCDM cosmology is supported when internally converting from equatorial to cartesian coordinates; future versions could extend this.

3.3 Implementation

3.3.1 Local cell search

As discussed above, a simple estimate of every pair distance for a catalogue of N points requires $N(N-1)/2$ distance calculations, which for large catalogues soon outstrips computational feasibility. However, for most use cases the scale of the survey far exceeds the maximum scale for which the user wishes to calculate the correlation function, so a large majority of pair calculations are unnecessary. To reduce their number, the code splits the galaxy survey into cells so that for each galaxy, only galaxies lying in the same or neighboring cells need to be considered, similar to the scheme used in Alonso (2012). The scaling of the calculation time with an increase in survey volume at fixed density should now scale as $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$. The scaling with galaxy density at fixed volume is still $\mathcal{O}(N^2)$. The measurements and practical limitations of these scalings are given in section 3.4. The speed of the code could be increased by using a tree method that can place entire tree branches or leaves into single bins but the development time needed for the implementation of such a structure was not justified.

The splitting of the catalogue into cells is performed differently for 3D calculations than for angular pair counts.

3D cells

Galaxies in a 3D catalogue are partitioned into cubic cells each with side length equal to the maximum required distance. Hence, only galaxies lying in the same or

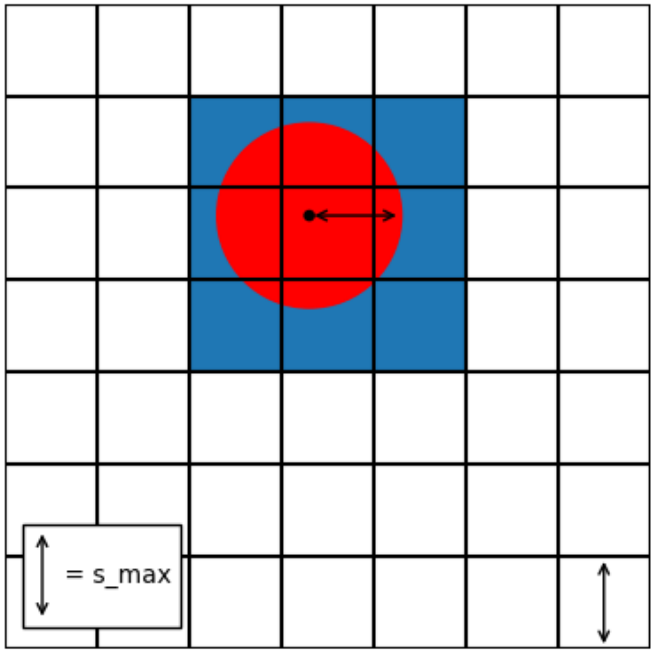

Figure 3.1: A partitioning of a 2D density field into cells. For the black dot representing a galaxy, all pair distances less than s_{max} are found by considering just the boxes shaded in blue and the cell containing the black dot, rather than an exhaustive search of the entire grid.

neighbouring cells need to be considered. Figure 3.1 shows this setup in 2D for a maximum distance s_{max}. In the case of the monopole or the (s, μ) decomposition the largest distance is simply taken to be the largest radial separation requested, i.e s_{max}. In the case of the $r_p - \pi$ decomposition the largest radial distance is given by $\sqrt{r_{p,\max}^2 + \pi_{\max}^2}$.

2D Galaxy pixels

To efficiently calculate angular pair counts, galaxies are assigned to pixels on the sky defined with the HEALPix C++ package[2]. HEALPix partitions the entire sky

[2] `https://healpix.jpl.nasa.gov/` and `http://healpix.sourceforge.net/html/Healpix_cxx/index.html`

into equal area pixels. The number of pixels is $3(2^{2n})$ with n being the user set integer parameter called the HEALPix order. These pixels are used in a slightly different way to the 3D method, as any HEALPix order can be chosen, and then only pixel pairs which may contain galaxy pairs within the maximum desired angular separation are considered. The `query_disc` function from the HEALPix package returns all pixels whose centres lie within a given angular distance from a given pointing on the sky. From the centre of a pixel, all pixels whose centres lie within a disc of angular radius θ_{query}, where

$$\theta_{query} = \theta_{max} + 2\theta_{pixel}\,, \tag{3.3.2}$$

should be considered to guarantee all pairs are found. θ_{max} is maximum angular scale on which $w(\theta)$ is to be estimated and θ_{pixel} is the maximum healpix pixel radius. θ_{pixel} is an output of the function `max_pixrad` from the HEALPix package and is unique to each choice of HEALPix order. The number of unnecessary galaxy pair distances calculated is lower for a higher choice of HEALPix order due to the smaller pixel size. However, a larger choice of HEALPix order will come with a larger over-head in finding pixel pairs. The optimum value for the HEALPix order will vary from catalogue to catalogue, with higher density catalogues preferring larger values of the HEALPix order. The default value in the code is 5, giving a pixel area of $\sim$ 13.5 square degrees. For a survey such as the GAMA survey (Driver et al., 2011) this corresponds to $\sim$ 13500 galaxies per pixel.

3.3.2 2D decomposition

The separation between pairs of galaxies is often projected into components perpendicular and parallel to the line of sight, labeled r_p and π respectively (Figure 3.2). In the distant observer approximation, only the π component is affected by redshift space distortions and redshift uncertainties. This allows for an integration over this component to produce a projected correlation function which is independent of redshift space effects,

$$w_p(r_p) = 2\int_0^{\pi_{\max}} \xi(r_p, \pi)\mathrm{d}\pi \tag{3.3.3}$$

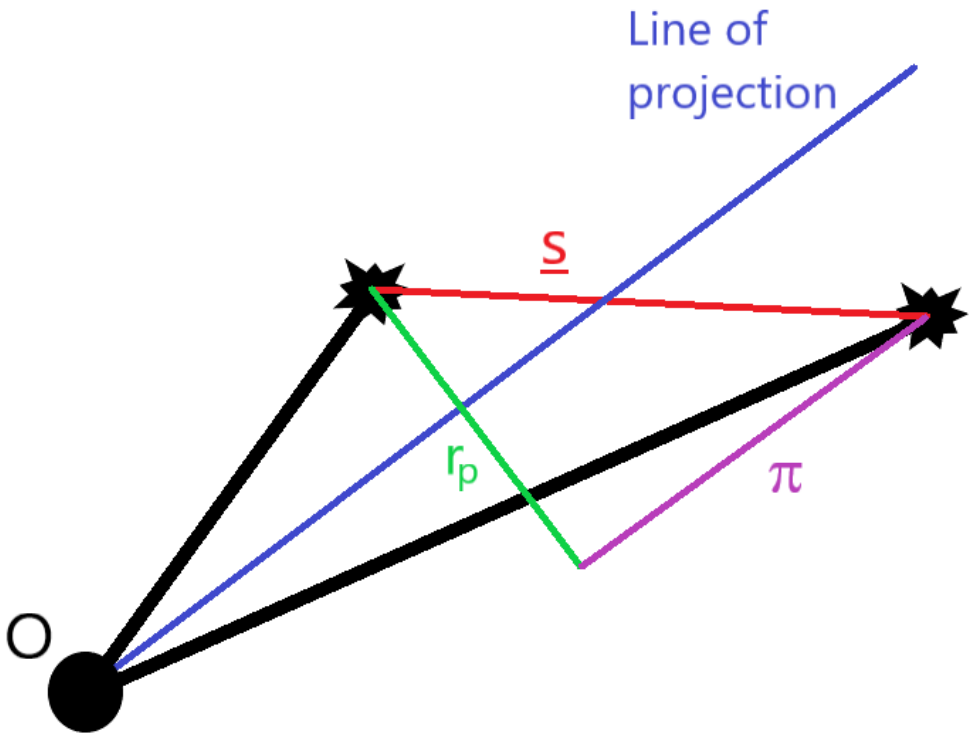

Figure 3.2: Diagram showing the definitions of r_p and π. π is the component of s parallel to the line of projection, and r_p the component perpendicular to the line of projection. The two choices of line of sight projection vector are given in equations 3.3.6 and 3.3.7.

The upper limit of the integral $\pi_{\max}$ should in theory be infinite, but in practice this cannot be the case due to the finite size of a catalogue volume. Also, choosing a value of $\pi_{\max}$ too large will increase the noise of the measurement so that value should be chosen to be as large as is needed for the result to converge and not necessarily larger. The value of π is defined using the line of projection $\underline{p}$ and the two galaxy vectors $\underline{x}_1$ and $\underline{x}_2$ as

$$\pi = |\underline{p}.(\underline{x}_1 - \underline{x}_2)|\,, \tag{3.3.4}$$

leaving r_p to be defined as

$$r_p = \sqrt{(\underline{x}_1 - \underline{x}_2)^2 - \pi^2}\,. \tag{3.3.5}$$

A common choice in the literature is to project the pair separation onto the

direction of the average position of the two galaxies,

$$\underline{p} = \widehat{\underline{x}_1 + \underline{x}_2} = \frac{\underline{x}_1 + \underline{x}_2}{|\underline{x}_1 + \underline{x}_2|} \,. \tag{3.3.6}$$

This definition is not the default choice used in this code. Instead the default choice is to project onto the direction bisecting the two galaxies on the sky. That is to project onto the average of the two galaxy unit directions,

$$\underline{p} = \frac{\hat{\underline{x}}_1 + \hat{\underline{x}}_2}{2} \,. \tag{3.3.7}$$

There are two benefits to this definition. The first one is of speed. In the first definition the value of $|\underline{x}_1 + \underline{x}_2|$ must be recalculated for every pair, a calculation that includes a square root, a function that requires many CPU instruction cycles. With the second definition, each galaxy unit direction can be precomputed, reducing the scaling of the number of norm vector calculations from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. This does come with an increase in memory use, which must be taken into account if memory usage is an issue for extremely large catalogues. This increase in memory usage also means fewer CPU cache hits, but this effect is outweighed by the speed up due to the fewer vector norm calculations.

The second benefit is that this definition of the direction of projection is not impacted by redshift space effects as it only depends on the angles on the sky. This may simplify modeling that does not make the plane-parallel approximation. For comoving distance errors on galaxies 1 and 2 of ϵ_1 and ϵ_2 the plane-parallel approximation means the error in the radial direction $\Delta\pi$ is given by

$$\Delta\pi = \epsilon_1 - \epsilon_2 \,, \tag{3.3.8}$$

whereas the true change in π when using the definition given in equation 3.3.7 is

$$\Delta\pi = \frac{1}{2}(1 + \hat{\underline{x}}_1.\hat{\underline{x}}_2)(\epsilon_1 - \epsilon_2) \,. \tag{3.3.9}$$

If using the standard definition given by equation 3.3.6 the analytic expression for this quantity is significantly more complicated.

In the code the default definition is given by equation 3.3.7, but the definition given by 3.3.6 can be used by commenting out the compile time option `_ANGULAR_PI_DEF` in the makefile and recompiling.

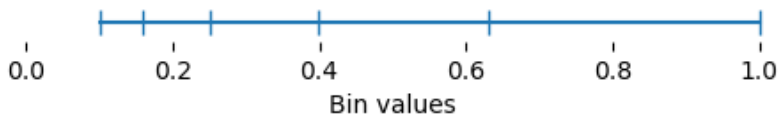

Figure 3.3: Visualisation of the naive log binning scheme for 5 bins between 0.1 and 1. Note that the bins do not extend down to zero.

3.3.3 Flexible binning scheme

It is often desirable when calculating correlation functions to use a binning scheme other than linear. The contrast between the data and random catalogues at larger scales is smaller than that at small scales, and the variation as a function of scale is larger at small scales than large scales, so larger bins at larger scales are often used.

Achieving larger bins at larger scales is typically achieved through using a logarithmic binning scheme. A standard implementation of a logarithmic binning scheme must have a histogram minimum value greater than zero. Equation 3.3.10 shows an example of a simple logarithmic binning scheme which returns the integer bin value B given a value x, a minimum x_{min}, a maximum x_{max}, and a number of bins N,

$$B = \text{int}\left(N\frac{\ln(x) - \ln(x_{min})}{\ln(x_{max}) - \ln(x_{min})}\right). \tag{3.3.10}$$

This is suitable for many cases but not all. For example, in the calculation of the projected correlation function, the integral of equation 3.3.3 starts at a π value of zero, so a non-zero choice for the smallest value of π could lead to a biased result. Figure 3.3 visualises this standard scheme for 5 bins between 0.1 and 1.

Instead this code uses an improved logarithmic binning scheme that allows any arbitrary limits and allows the user to choose how aggressively the bin sizes will scale, i.e. how much bigger each subsequent bin will be compared to the previous one. A scaling factor, the log base b, is introduced, which represents the scaling factor between the size of one bin and the next. For this scheme, the integer bin value B is now calculated using

$$B = \text{int}\left(\log_b\left(1 + \frac{x - x_{min}}{x_{max} - x_{min}}(b^N - 1)\right)\right). \tag{3.3.11}$$

In order to understand this binning scheme it is useful to examine the term

$$\frac{x - x_{min}}{x_{max} - x_{min}}(b^N - 1)\,. \tag{3.3.12}$$

This term takes on values from 0 to b^N-1, 0 where $x = x_{min}$, and b^N-1 where $x = x_{max}$. This binning scheme therefore first creates a linear binning scheme with b^N - 1 bins, and takes the log base b of one plus this linear bin value to map onto a value of 0 to N as required. An instructive example of this is if two bins (N=2) and a value of b of 2 is chosen. The second bin is twice the size of the first, so three intermediate linear bins are created (b^N - 1 = 3). A pair lying in the first intermediate linear bin will map onto the first bin and a pair lying in either of the next two intermediate linear bins will map onto the second bin. This scheme also works for non integer values of b but this is harder to visualise.

Figure 3.4 shows the bins in this scheme for multiple choices of the log base with 5 bins between 0 and 1. In contrast to the standard binning shown in Figure 3.3 the binning extends down to a value of zero. This plot also shows the flexibility that this scheme provides, the bins with larger values of the log base scale far more aggressively in size. It can be seen that linear binning is a subset of this binning scheme, corresponding to a log base equal to 1, with each bin being equal in size to the last. However, equation 3.3.11 returns zero for all x for a value of $b = 1$, and only actually reduces to linear bins in the limit as b tends to 1. To avoid this, whenever $b < 1.05$, the code adopts linear binning instead. In this scheme, bins are uniquely defined by a min, a max a number of bins and a log base.

In the code, this flexible scheme is available for use with any of the output options. In the case of 2D decompositions, separate binning schemes can be used for each dimension. One might choose to use bins closer to linear along the line of sight than perpendicular to it, as more of the signal is spread out along the line of sight due to redshift space effects. In the case of binning in s-μ the μ direction is fixed to linear bins. To help the user with this non-trivial binning scheme the code output includes each bin centre and each bin width.

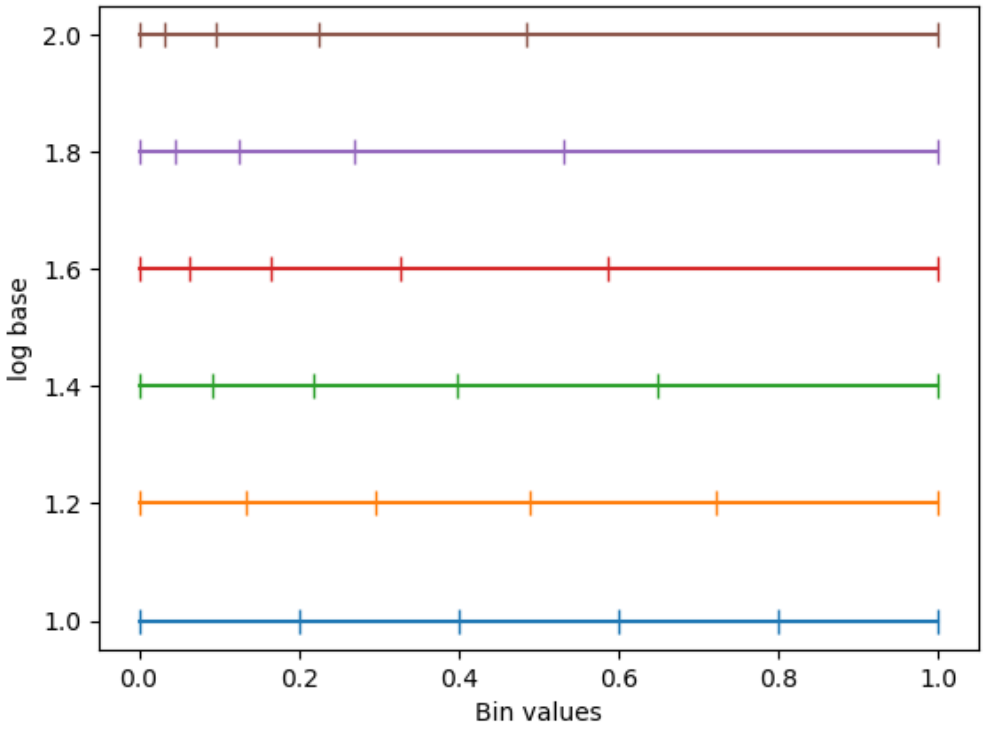

Figure 3.4: Visualisation of the flexible binning scheme for different values of the log base in the case of 5 bins between 0 and 1. Note in all cases the bins can extend down to zero and how aggressively the bins scale can be changed by changing the log base parameter.

3.3.4 On the fly jackknife calculations

One of the methods of calculating uncertainties on galaxy correlation functions is the process of jackknifing (Zehavi et al., 2002; Norberg et al., 2009a). The typical procedure used to do this is outlined below,

- Calculate the galaxy correlation function, ξ, for the entire sample.
- Separate the galaxy survey on the sky into N regions of similar sky area.
- Mask region i and calculate the galaxy correlation function ξ_i.
- Repeat the process N times each time masking a different region in turn.
- The variance of the correlation function is then calculated using

$$Var(\xi) = \frac{N-1}{N}\sum_{i=1}^{N}(\xi_i - \xi)^2\,, \tag{3.3.13}$$

One method of creating regions of similar sky area is to split the survey with straight line cuts in RA and Dec such that each region contains the same number of points in the random catalogue. First split the survey by RA into strips with equal numbers of randoms, then split each of these strips with cuts in Dec to achieve the equal area regions. A python script in included with the code to assign regions using this scheme. This script also contains the option to rotate the survey about the survey centre before the scheme is applied so that different partionings of equal area can be tested for robustness. Figure 3.5 shows an example partitioning with this scheme for a region similar to a GAMA survey equatorial patch for the cases of no rotation and a rotation by 45 degrees.

The issue with the standard jackknife procedure is that it requires recalculation of the correlation function when masking each jackknife region, so increases computation time by a factor of roughly N. For large area surveys many jackknife regions may be used so this significantly impacts the computational feasibility of the calculation.

It can be seen with the standard approach that many galaxy pair distances are unnecessarily recalculated each time the correlation function is recalculated. Pair

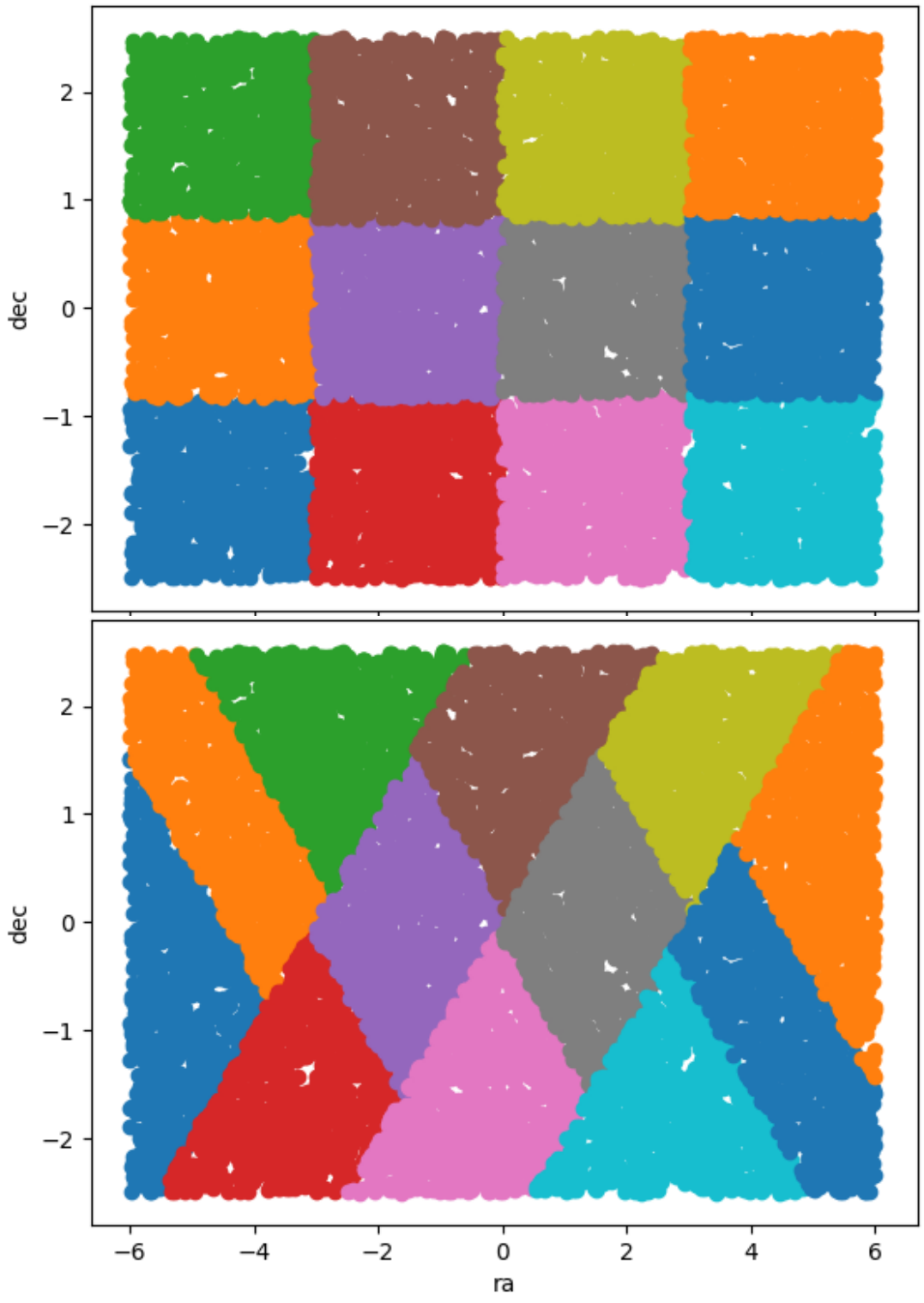

Figure 3.5: An example of an equal-area jack-knife partioning of a region similar to a GAMA equatorial region using the included python script. The different colours show the different regions. The bottom panel shows the results when the survey is rotated by 45 degrees before running and the top when no rotation is performed.

distances for galaxies that lie in a single jackknife region will be recalculated N - 1 times (Only not in the case where that region is masked), despite their separation not changing between each run. The approach taken in this code removes the need for this unnecessary recalculation of galaxy pairs. It also allows the user to calculate all jackknife region calculations by running the code once rather than multiple times, redefining the catalogue each time. The improved procedure used in this code for an example of a single pair count i.e DD of a survey is,

- User specifies the number of jackknife regions N in the parameter file, jackknife results are not output if $N <= 1$.
- User specifies a jackknife region for each galaxy with integers running from 0 to $N - 1$.
- Hold one histogram for the pair count result with the whole catalogue, DD , and one sub-histogram for each jackknife region, labeled $\mathrm{DD}_{sub,i}$, for i ranging from 0 to $N - 1$, with each containing the number of pairs where at least one of the pair of galaxies overlaps with the associated jackknife region.
- An individual jackknife pair count DD_i found from masking one region is the calculated from the total pair count and the sub-histograms using

$$DD_i = DD - DD_{sub,i}\,. \tag{3.3.14}$$

- The value of the correlation function on masking each jackknife region can now be calculated from knowing the pair counts found when masking each region, and equation 3.3.13 can be used as normal to calculate measurement errors.

The performance overhead for this procedure compared to a single run is modest. The more jackknife regions used, the larger the memory usage will be, but the increase will typically be small compared to the memory needed to hold the catalogue. When no jackknife errors are needed, the code will still internally follow the above procedure, i.e no special case is coded for when no resampling errors are needed. The computational overhead of this implementation does not scale with the number of jackknife regions as in the standard implementation, so the performance gain in the case of multiple jackknife regions is significant.

3.3.5 Parallelisation

A calculation that can be performed in parallel with none of the individual processes depending on one another is called "embarrassingly parallel". In the case of counting galaxy pairs, each pair calculation is independent of one another, the only thing shared being the resultant histogram, so this process is said to be "nearly embarrassingly parallel". What this means is that the parallelisation, that is, using multiple CPU threads to accelerate calculation of galaxy pair counts, should be a relatively simple process. Indeed the splitting of the galaxy catalogue into cells (section 3.3.1), lends itself to a very simple prescription for splitting the job across multiple threads. The galaxy cells can be split between the threads such that each thread calculates the pairs in which the first galaxy in the pair lies within its assigned cells.

This is implemented in the code using the OpenMP C++ API for shared memory parallel programming[3]. Shared memory means that some of the job memory can be accessed by all threads without the need for replication. In this case the galaxy catalogue and final histogram are shared between threads. OpenMP allows for scaling across multiple cores on a single CPU but does not support scaling across multiple CPUs. This limitation could be overcome by simultaneously using the message passing interface MPI[4] but computation times for the catalogues tested in this work were fast enough solely using OpenMP that extra development time could not be justified to extend the parallelisation across multiple CPUs.

As the result histogram is shared between the threads, care must be taken so that two or more threads do not try to simultaneously update the same part of the histogram, as this can corrupt the result. Two options present themselves for solving this problem. The first, is to place the histogram incrementation behind an "atomic" barrier, which is a region of code through which the threads must pass one at a time. This option was found to significantly decrease multi-core scaling performance, as each thread must queue and pass the barrier one at a time, a problem which becomes worse the more threads are used. The second is to create

[3] `http://www.openmp.org/`

[4] `https://en.wikipedia.org/wiki/Message_Passing_Interface`

a private histogram for each thread, then combine them after all pair calculations have finished. This option was found to be significantly faster, and the memory overhead from holding one histogram for each thread is typically small.

3.3.6 Pair upweighting scheme

In order to correct for missing correlated data the pair upweighting scheme presented in Bianchi & Percival (2017) is implemented. This scheme gives a weight to pairs of galaxies equal to the inverse probability that both galaxies were selected during target selection, p_{ij}[5]. This increased weight accounts for similar pairs of galaxies that were not actually targeted. The weight of a pair of galaxies, w_{ij}, is given by

$$w_{ij} = \frac{1}{p_{ij}} \,. \tag{3.3.15}$$

It is important to note that p_{ij} is only equal to the product of the individual galaxy selection probabilities, $p_i p_j$, in the case where the selection of the two galaxies is uncorrelated. Often, due to limitations in positioning of spectroscopic fibres, only one of two galaxies in a close pair can be observed, so their targeting probabilities are highly correlated. In these cases using a weight of

$$w_{ij} = \frac{1}{p_i p_j} \,, \tag{3.3.16}$$

will lead to a biased result.

It is impractical to estimate and to save a weight for each pair of galaxies, so binary masks of random realisations of the targeting algorithm are saved instead. If $\underline{b}_i$ is the vector of binary values of length N (the number of realisations of the targeting algorithm) with 1 representing when galaxy i was observed in a realisation and 0 when it wasn't, equation (3.3.15) can be written

$$w_{ij} = \frac{N}{\underline{b}_i . \underline{b}_j} \,, \tag{3.3.17}$$

where the denominator is the dot product of the two binary mask vectors.

[5]By construction this scheme only works if there is a non-zero probability of targeting every pair of galaxies in the parent sample.

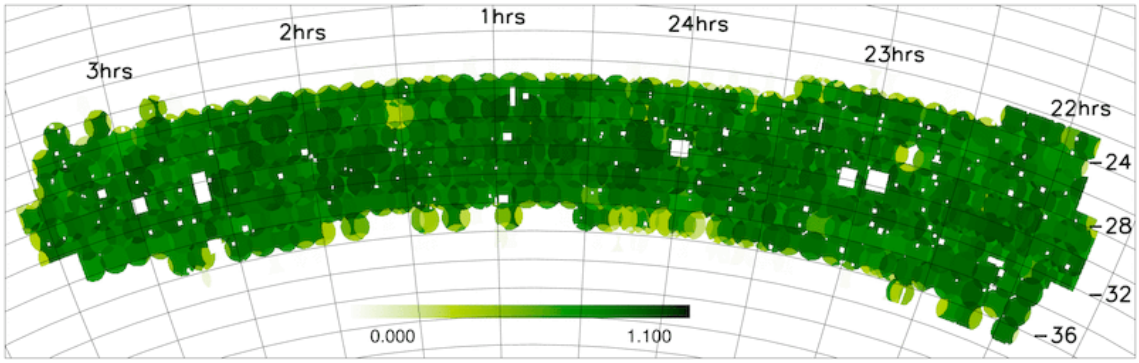

Figure 3.6: Redshift completeness mask of the 2dF Galaxy Redshift Survey North Galactic Plane strip. Each 2dF pointing is visible as a circle. Note the variation in completeness from region to region and the higher average completion in regions of overlapping pointings. Source: `http://magnum.anu.edu.au/~TDFgg/Public/Release/Masks/`

A complication that often must be accounted for in real surveys is the fact that different regions of the survey have different target selection statistics. Figure 3.6 shows the redshift completeness mask for the North strip of the 2dF Galaxy Redshift Survey (Colless et al., 2001). Significant variation in completeness can be seen over the survey. One variation in particular is that regions lying in the overlap of multiple 2dF pointings have, on average, higher completeness. For 2dFGRS, adaptive tiling was used, i.e. more pointings were done in regions of higher density. This means that the survey geometry is correlated with the galaxy density field in such a way that is very difficult to define algorithmically, so the pair upweighting scheme would most likely fail to recover the correct clustering. DESI BGS tiling on the other hand, is random with respect to the background density, so this scheme can be used. This randomness of tiling needs to be taking into account in the scheme. If DESI BGS starts in a slightly different sky position, galaxies that would have been in a region of poor completeness could be in regions of high completeness and vice-versa. Smith et al. (2018) accounted for this in DESI BGS targeting by randomly shifting the survey before each rerun of the targeting algorithm such that each galaxy had an equal chance of falling into the various types of overlap region generated by the structure of the survey.

Applying this solution to the exact scheme defined in Bianchi & Percival (2017) will produce a bias if the shifts of the survey are not random each time on the whole sky. This is because pairs near the edge of the survey will often be missed due to one or both galaxies not lying in the shifted footprint, i.e. they were never candidates for observation during some realisations. This would artificially upweight pairs near the edge of the survey. If the shifting covers the whole sky this excess upweighting is the same for all pairs in the survey, so the bias disappears. For a large survey such as DESI BGS, which covers a third of the sky, this may be possible, but for smaller surveys the vast majority of the targeting realisations would not contain any galaxies at all. In the latter case case, the number of realisations of the targeting algorithm will have to be significantly increased to ensure all pair probabilities are well defined. The solution proposed and implemented in this code is to save a second vector of binary weights, $\underline{c}_i$, of length N, in which 1 represents when galaxy i lies in the shifted footprint and 0 when it doesn't. This can then be used to account for when pairs were not candidates for selection by modifying equation (3.3.17) to become

$$w_{ij} = \frac{\underline{c}_i.\underline{c}_j}{\underline{b}_i.\underline{b}_j} \tag{3.3.18}$$

Previously the maximum value of $\underline{b}_i.\underline{b}_j$ was taken to be N the number of realisations. If a pair of galaxies was seen in every realisation, then no upweighting was needed. Now the new maximum value is replaced with the more accurate $\underline{c}_i.\underline{c}_j$, so the maximum number of times that a pair could be observed is now the number of realisations in which that pair was a candidate for targeting. This modification allows for smaller random shifts between each targeting realisation, which could reduce the numbers of realisations needed to accurately apply this scheme to some surveys by orders of magnitudes.

The angular upweighting suggested by Bianchi & Percival (2017) to lower the variance of the result is also implemented. This makes the final upweighting scheme for the data-data pair count

$$DD(\underline{r}) = \sum w_{ij} \frac{DD^P(\theta)}{DD(\theta)}\,, \tag{3.3.19}$$

where the w_{ij} weights are defined in equation (3.3.18). $DD^P(\theta)$ is the angular pair

count of the parent sample that could have been targeted, and $DD(\theta)$ is the angular pair count of the targeted sample calculated using the pair weights w_{ij} defined in equation (3.3.18). The way to think of this factor is that if the pair upweighting scheme over (under) estimates the angular correlation function, the weights w_{ij} are on average too large (small) by a factor of $DD(\theta)/DD^P(\theta)$ at a particular angular separation. Multiplying the weights in the 3D case by the term $DD^P(\theta)/DD(\theta)$ will slightly downweight (upweight) pairs to try to correct for this.

One caveat to this scheme is that for redshift selected samples of the real data it is difficult to know the parent sample for calculating $DD^P(\theta)$. We note that in the application of Smith et al. (2018) presented in Section 3.5 the correction is exact as $DD^P(\theta)$ is perfectly known for the mock sample considered.

3.4 2PCF Performance

In this section we will explore how the performance of the 2PCF code varies with the volume or density of the catalogue, the maximum scale required and number of cores used. A common catalogue is used across all sub-sections, 14000 square degrees of the HOD mock presented in Smith et al. (2017)[6]. The catalogue is volume limited $M_r - 5\log_{10} h < -20.5$ between $0.15 < z < 0.3$ and covers two patches to roughly mimic the DESI BGS survey area. The catalogue contains $\sim$ 840 000 galaxies and in every test ten times the number of random points than data points are used. When the full catalogue is used around 8.5 million random points are considered. The random catalogue is generated by randomly sampling from the RA, Dec and redshift values of the data in each patch, which works as each patch is bounded by constant values of RA and Dec. One random catalogue is generated that is cut in the same way as the data catalogue is in each test. The results presented here apply to the monopole correlation function, and where the results differ for the angular correlation function it is explicitly stated. When the number of cores is not directly varied, the code is run in parallel on a 12 core CPU[7]. All scales given are in h^{-1}Mpc.

[6]Publically available from the Virgo database at `http://virgo.dur.ac.uk/data.php`

[7]Intel X5650 with 60GByte shared memory.

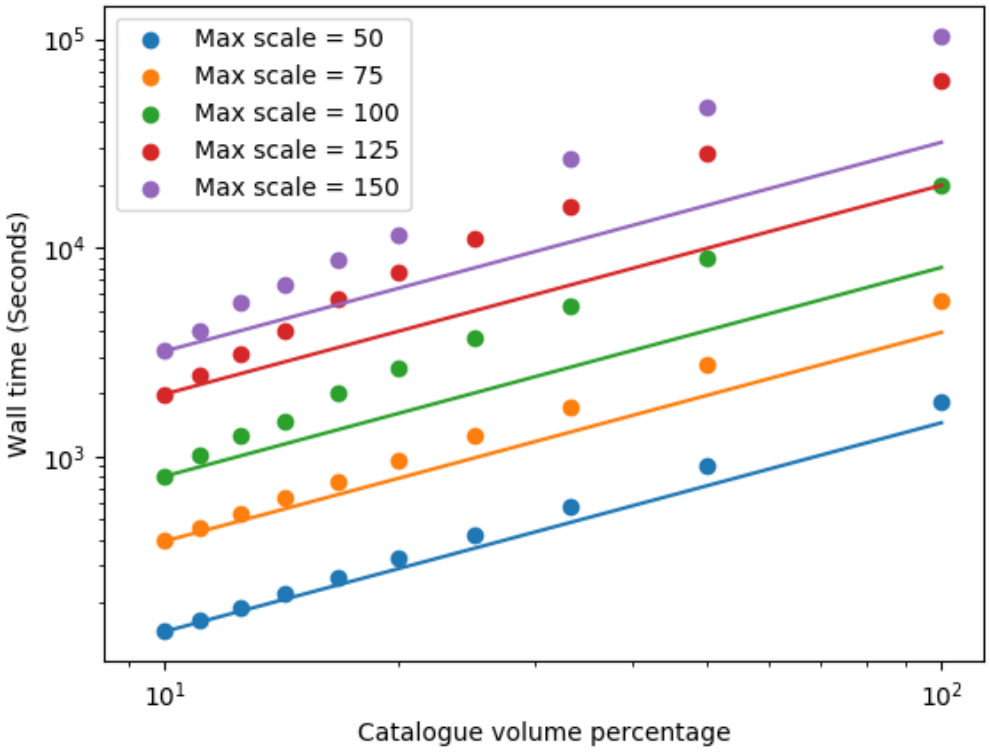

Figure 3.7: Wall time of the code in seconds for different fractions of the full catalogue volume and different maximum scales (in units of h^{-1}Mpc). A 10% catalogue volume percentage means that only 10% of the solid angle of the full catalogue has been used. The dots represent the measured values and the lines show the theoretical $\mathcal{O}(N)$ scaling, extrapolated from the smallest volume for each maximum scale.

The run for the entire catalogue when the max scale is set to 150 h^{-1}Mpc takes ~ 30 hours. This code is therefore capable of dealing with the currently most challenging catalogues in a reasonable time. If many runs of that size are needed, for example to estimate covariance matrices, spreading runs over multiple nodes would be necessary to keep the computation time reasonable.

3.4.1 Volume scaling

The first performance scaling test is to see how the code scales as the volume of the catalogue changes. Smaller volume catalogues are generated by cutting the two patches in ra such that only a particular percentage of the full catalogue remains,

the same cuts are made for the random catalogue. Because of the local cell search for relevant pairs, presented in section 3.3.1, it is expected that the runtime for the code will scale linearly with increasing catalogue size, i.e. $\mathcal{O}(N)$.

Figure 3.7 shows the runtime of the code in seconds for different volume cuts and maximum scales, as well as lines of $\mathcal{O}(N)$ scaling from the smallest volume of each maximum scale test. The runtime of the code significantly increases with an increase in maximum scale. As the maximum scale (cell size) increases, the average number of points inside each cell increases, which significantly increases computation time. In fact, the code scales as $\mathcal{O}(R_{\rm max}^6)$ with $R_{\rm max}$ the maximum scale required.

It can also be seen that the smaller the maximum scale, i.e the smaller the cell size, the closer the volume scaling is to the theoretical linear case. For larger cell sizes, the scaling is significantly worse than expected for smaller catalogues, but tends towards the theoretical $\mathcal{O}(N)$ scaling for larger volume catalogues. It is easy to see why this scaling occurs by considering two extreme cases of the ratio of cell size to survey size, one a pencil beam survey and large cells, and the other a large survey with small cells. When the volume of the pencil beam survey is increased, the survey may at many points still lie within the width of one cell, so rather than increasing the number of cells, the number of points per cell has mostly increased. Increasing the number of points per cell is expected to scale as $\mathcal{O}(N^2)$ rather than $\mathcal{O}(N)$. However in the case of the large survey and small cells, the average number of points per cell is mostly unchanged, and only the number of cells needed has changed, so we now recover the $\mathcal{O}(N)$ scaling we wanted. The conclusion of this is that the code will scale as $\mathcal{O}(N)$ with catalogue volume if the size of the cells is significantly smaller than the smallest scale of any cartesian dimension of the catalogue volume, and will scale as the density scales if the cells are larger than the survey volume.

3.4.2 Density scaling

The second aspect of performance scaling to investigate is how the code scales with density. In this case the full catalogue and the randoms are randomly subsampled to generate the smaller catalogues. It is expected that the code will scale as $\mathcal{O}(N^2)$

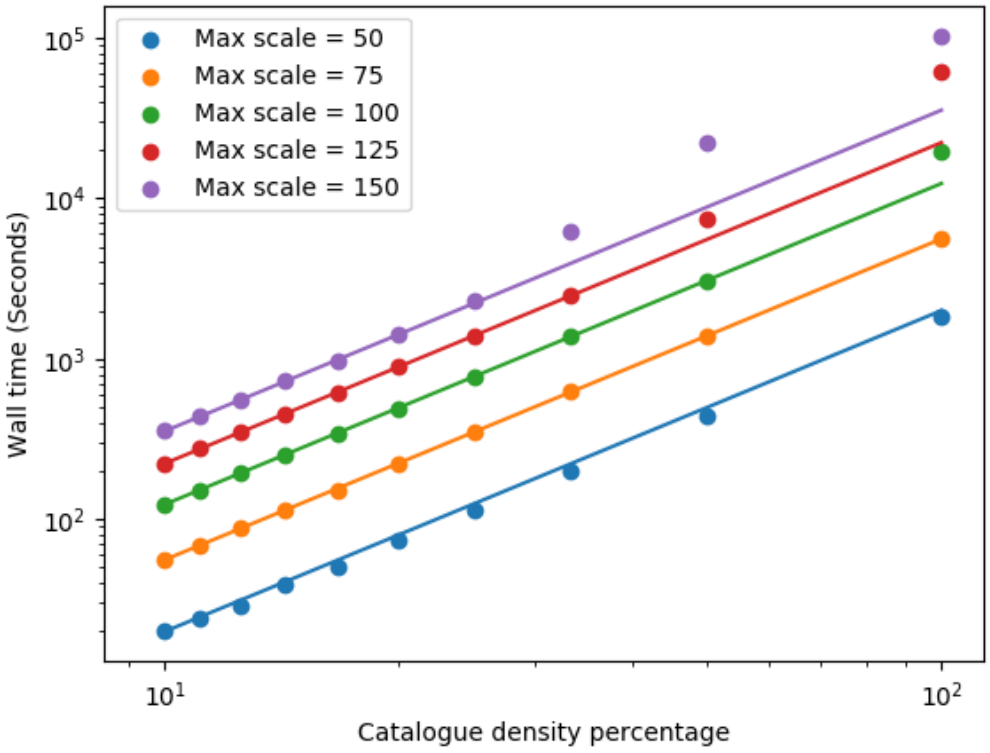

Figure 3.8: Wall time of the code in seconds for different catalogue densities and different maximum scales in units of h^{-1}Mpc. The dots show the measured values and the lines show the theoretical $\mathcal{O}(N^2)$ scaling, extrapolated from the lowest density for each maximum scale.

with changing density as only the number of galaxies per cell is being changed.

Figure 3.8 shows the runtime of the code in seconds for different density subsamples and maximum scales, as well as lines of $\mathcal{O}(N^2)$ scaling from the lowest density of each maximum scale test. For the runs with a smaller maximum scale, the results follow the theoretical scaling prediction well. Runs with high densities and large maximum scales deviate above the expected scaling. All the catalogues that are above the theoretical expectation are the runs that had the largest average number of points per cell. A possible explanation for this deviation is that the cells are now too large to fit in the CPU cache in one go, so many more memory reads are needed, slowing down the code. The point at which this may happen will be architecture dependent. It is postulated that above this value the code will eventually return to

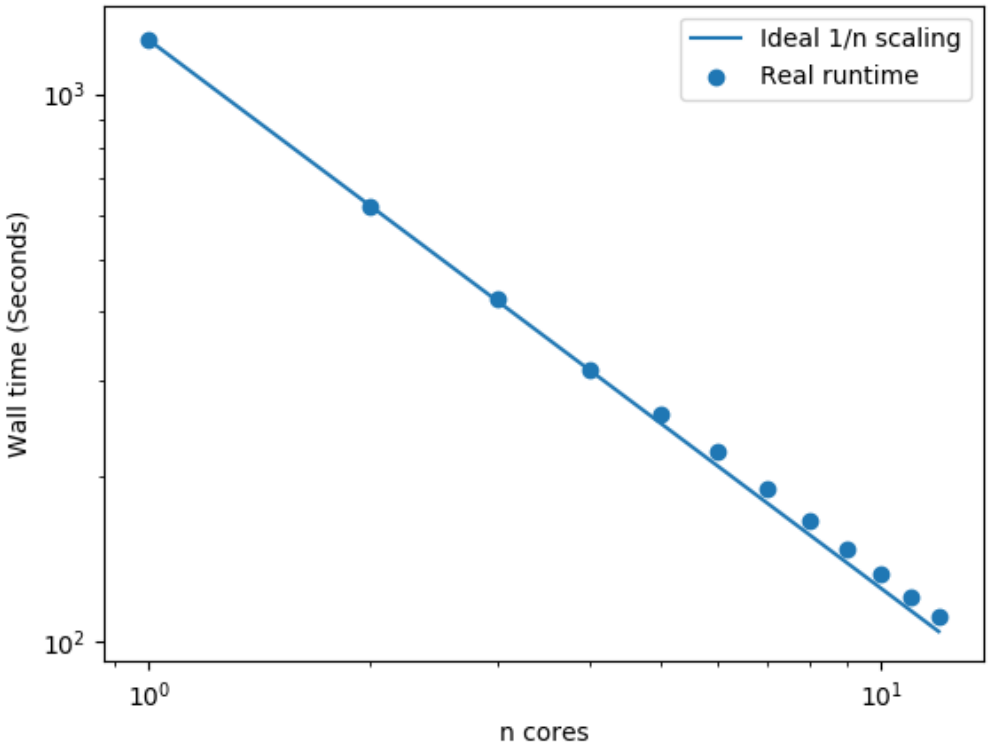

Figure 3.9: Wall time of the code as a function of the number of cores used. The dots show the measured results and the line shows the ideal 1/n scaling extrapolated from the run with a single core. It can be seen the code scales closely to the ideal case.

$\mathcal{O}(N^2)$ scaling but this is not explicitly tested.

3.4.3 Multicore scaling

The final aspect of performance scaling to investigate is how the code scales with increasing the number of cores used. For this the whole catalogue with 1/4 density and a max scale of 50 h^{-1}Mpc is used. Figure 3.9 shows the experimental results along with ideal 1/n scaling and the code is seen to perform very close to the ideal case. This verifies that the parallelisation implementation laid out in section 3.3.5 is efficient and has little overhead compared to the ideal case.

3.5 Application to mock DESI BGS fibre collision correction

This section will summarise the application of the code in Smith et al. (2018) to a mock catalogue for the DESI Bright Galaxy Survey (BGS). This work shows that the missing observation correction presented in section 3.3.6 successfully recovers the true galaxy clustering measurement for a BGS mock catalogue that includes the complicated DESI target selection effects.

3.5.1 DESI BGS

The Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016) will be used to conduct a large spectroscopic survey with the primary science aims of making precision measurements of the baryon acoustic oscillation (BAO) scale and the large scale redshift space distortion (RSD) of galaxy clustering.

The instrument, which is currently being built and assembled, will be installed on the 4-m Mayall Telescope at Kitt Peak, Arizona. DESI will consist of dark-time and bright-time programs. The bright galaxy survey (BGS), part of the bright-time program, is a low redshift, flux limited survey of ~ 10 million galaxies with a median redshift $z_{\rm med} \sim 0.2$. BGS will have two priorities of galaxies, priority 1 ($r < 19.5$), and priority 2 ($19.5 < r < 20.0$). Smith et al. (2018) investigates both samples but here we look only at a subset of the results from the priority 1 galaxies.

3.5.2 Mock catalogue

The BGS mock catalogue used here is from the Millennium-XXL (MXXL) simulation (Smith et al., 2017). This is a halo occupation distribution (HOD) mock, which contains galaxies to $r = 20$ over the same redshift range as the BGS, and is constructed to reproduce the luminosity function and clustering measurements from SDSS (Blanton et al., 2003; Zehavi et al., 2011) and GAMA (Loveday et al., 2012;

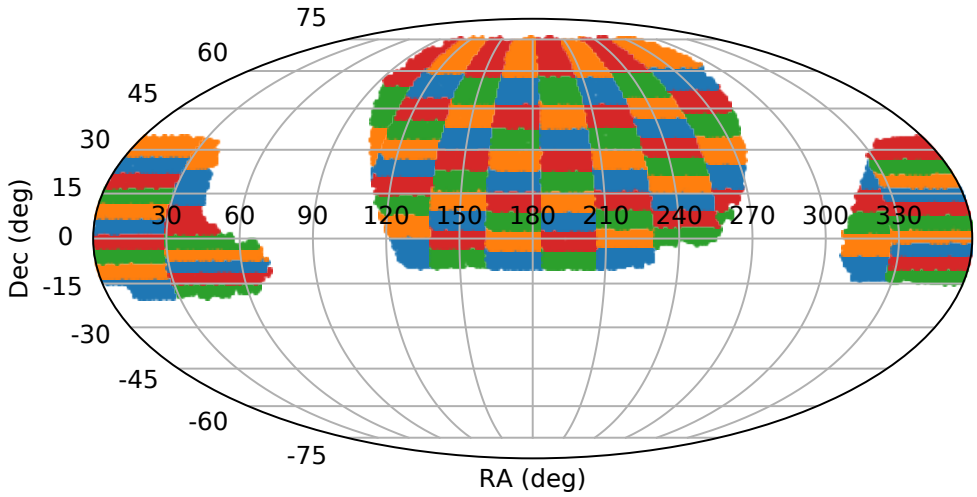

Figure 3.10: Footprint of the DESI BGS, which covers 14,800 square degrees. Colours indicate the 100 jackknife regions. Taken from Smith et al. (2018).

Farrow et al., 2015a).[8]

Figure 3.10 shows the sky footprint of the mock with different colours highlighting the 100 jackknife regions used to generate the errors. The sample used for testing the clustering measurements is a volume limited sample of priority 1 galaxies, $0.09 < z < 0.3$, $-22 < M_r - 5\log h < -21$ that contains ~ 1.5 million galaxies. The randoms are generated by randomly sampling from the values of ra, dec and z of the data and are 8 times more numerous than the data points.

The full DESI BGS targeting algorithm is run over the mock catalogue to generate a realistic selection. The algorithm is then run a further 2048 times to estimate the binary mask vectors needed to estimate the inverse pair probabilities.

3.5.3 Clustering correction

This section looks at the correction of the monopole correlation function using different methods of correction. The tested methods of recovery are as follows:

[8]The MXXL mock is available at `http://icc.dur.ac.uk/data/` and `https://tao.asvo.org.au/tao/`

- **Pair inverse probability (PIP)** - this is the scheme first discussed in Bianchi & Percival (2017) and implemented in section 3.3.6

- **Angular upweighting** - galaxy pairs are upweighted by the factor

$$W(\theta) = \frac{1 + w^{(p)}(\theta)}{1 + w(\theta)}, \tag{3.5.20}$$

  where $w^{(p)}(\theta)$ is the angular correlation function of the complete, parent sample of galaxies, and $w(\theta)$ is the incomplete, targeted sample. This is the method used in the 2dFGRS analysis of Hawkins et al. (2003).

- **Nearest object** - missing galaxies are assigned the redshift of the nearest targeted object on the sky. This approach is taken in most SDSS survey analyses, e.g. Zehavi et al. (2005), Berlind et al. (2006b), Zehavi et al. (2011).

- **Nearest weight** - each galaxy is first given a weight of 1, and the weight of a missing galaxy is added to the nearest targeted object on the sky. This method is used in BAO analysis in the BOSS survey (Anderson et al., 2012, 2014a,b). It was implemented in 2dFGRS but the extra weight was spread over more neighbours (Norberg et al., 2002).

Figure 3.11 shows the monopole result when applying these different methods of correction, along with the parent clustering and the result using no correction on the targeted galaxies.

Not performing any corrections produces a correlation function that is too low on small scales, as pairs of galaxies have been missed due to fibre collisions. The result is also too large on large scales, but this is not statistically significant. The nearest redshift method works well on large scales but significantly overestimates the correlation function on small scales, by up to an order of magnitude. This is because a large number of the galaxies placed at very small separations to their nearest angular neighbour will in reality not be close pairs. Nearest weight also performs well on large scales, but undercorrects the clustering on scales less than ~ 1 h^{-1}Mpc. The result on those scales is however closer than the nearest redshift result. The angular weighting works well on intermediate scales, but overpredicts both on

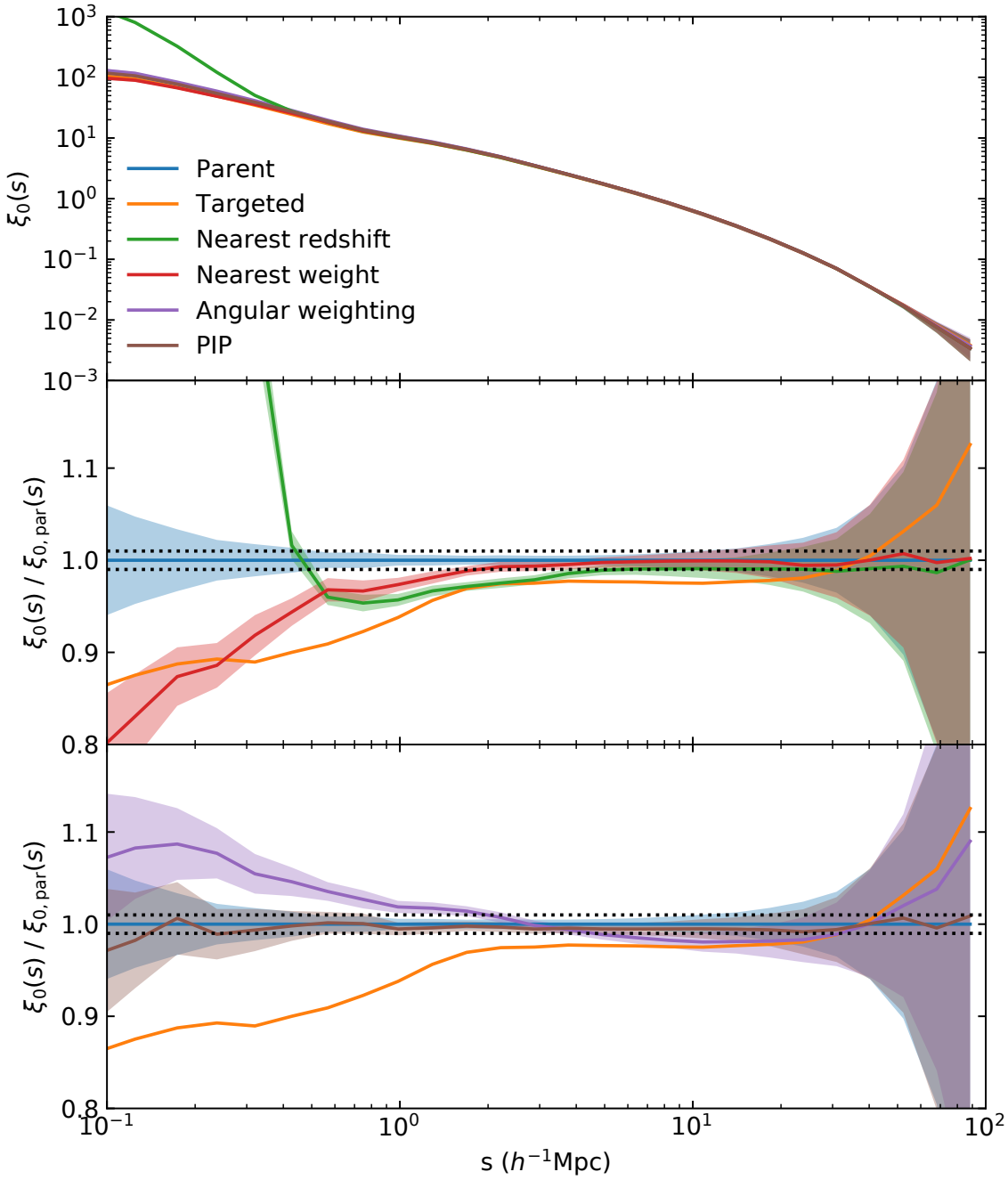

Figure 3.11: Monopole of the redshift space galaxy correlation function of the main volume limited sample, with different corrections defined in the text applied. The two lower panels show the ratio to the complete parent sample. Shaded regions are errors estimated from 100 jackknife samples. Taken from Smith et al. (2018).

small and large scales, with only the small scale overprediction being statistically significant. The PIP scheme corrects the monopole to within the scatter found by the jackknife resampling on all scales. This validates the scheme presented in section 3.3.6.

Smith et al. (2018) also go on to show that this PIP correction, as implemented here, also successfully recovers the higher order multipoles of the correlation function and the projected correlation function. These results are omitted here for brevity.

3.6 Conclusion

This chapter has introduced a publicly available two point correlation function code, `2PCF`, written in C++, that is fast, flexible and contains the features needed for modern galaxy redshift survey clustering statistics. The code is similar in scope and approach to CUTE (Alonso, 2012), but adds flexible binning, on the fly jackknife resampling calculations, more flexible IO and the pairwise upweighting scheme of Bianchi & Percival (2017). An extension of this pair upweighting scheme to account for shifting the survey in each run is also explained and implemented.

This code is shown to scale as $\mathcal{O}(N)$ with N the number of objects in the catalogue if the increase is in volume provided the cell size is small compared to the smallest dimension of the survey. The code is shown to scale as $\mathcal{O}(N^2)$ with increasing density, with an extra penalty added if the number of points per cell exceeds architecture specific cache sizes. The shared memory parallelisation is shown to have little overhead as the code scales close to ideally when the number of cores is increased.

The code has been used throughout this thesis for two point statistics and has formed the base of a galaxy group detection algorithm. The code has been used to investigate constraints on f(R) modified gravity models using marked correlation functions in Hernández-Aguayo et al. (2018). The use of the code in Smith et al. (2018) is summarised, which shows that the implementation of the pair upweighting scheme here is sufficient to correct the clustering statistics of the DESI BGS for the complicated target selection effects.

A list of proposed future developments to the code:

- Periodic box support
- Cross correlation support
- Non-flat cosmology support
- FITS file support
- Automated random catalogue generation

Chapter 4

A mock catalogue for the PAU Survey

This chapter presents the results of Stothert et al. (2018) verbatim.

We present a mock catalogue for the Physics of the Accelerating Universe Survey (PAUS) and use it to quantify the competitiveness of the narrow band imaging for measuring spectral features and galaxy clustering. The mock agrees with observed number count and redshift distribution data. We demonstrate the importance of including emission lines in the narrow band fluxes. We show that PAUCam has sufficient resolution to measure the strength of the 4000Å break to the nominal PAUS depth. We predict the evolution of a narrow band luminosity function and show how this can be affected by the OII emission line. We introduce new rest frame broad bands (UV and blue) that can be derived directly from the narrow band fluxes. We use these bands along with D4000 and redshift to define galaxy samples and provide predictions for galaxy clustering measurements. We show that systematic errors in the recovery of the projected clustering due to photometric redshift errors in PAUS are significantly smaller than the expected statistical errors. The galaxy clustering on two halo scales can be recovered quantitatively without correction, and all qualitative trends seen in the one halo term are recovered. In this analysis mixing between samples reduces the expected contrast between the one halo clustering of red and blue galaxies and demonstrates the importance of a mock

catalogue for interpreting galaxy clustering results. The mock catalogue is available on request at `https://cosmohub.pic.es/home`.

4.1 Introduction

Clustering measurements at low redshifts have been shown to display a dependence on galaxy properties such as stellar mass, luminosity, and colour, which suggests that these properties depend on the mass of the host dark matter halo (e.g. Norberg et al. 2002; Zehavi et al. 2011). Galaxy clustering measurements are therefore not only useful for constraining the cosmological model but also for developing our understanding of galaxy formation physics.

The processes that shape how the efficiency of galaxy formation depend on halo mass may change with redshift, so it is important to extend measurements of galaxy clustering as a function of intrinsic galaxy properties to higher redshift. One clear piece of evidence hinting at evolution in the galaxy formation process is the dramatic change in the amount of star formation activity since $z \sim 1-2$, with roughly ten times less star formation globally by the present day (Lilly et al., 1996; Madau et al., 1996).

The measurement of clustering as a function of galaxy properties poses different challenges to those faced when using large-scale structure to constrain cosmological parameters. In the cosmological case, the aim is to maximize the volume probed whilst maintaining an appropriate number density of galaxies to achieve a moderate signal-to-noise ratio in the power spectrum measurement (e.g. Feldman et al. 1994). The signal-to-noise ratio can be boosted by targeting galaxies with stronger clustering or a larger bias than the average population; beyond this, the selection of the galaxies is not that important in the cosmological case. On the other hand, when using clustering to probe galaxy formation, the desire is for a high number density of galaxies with a uniform selection covering a wide baseline in the intrinsic galaxy property of interest.

Progress towards compiling large-scale structure samples for galaxy formation studies at intermediate redshifts has been made through the Galaxy And Mass

Assembly Survey (GAMA; Driver et al. 2011), which targets galaxies in the r-band brighter than $r = 19.8$, with a median redshift of $z \sim 0.2$ over 286 sq deg with high completeness, and the VIMOS Public Extragalactic Redshift Survey (VIPERS), which obtained redshifts for 86 7765 galaxies with $i_{\rm AB} < 22.5$ over 24 deg^2 at ∼47% completeness (Scodeggio et al., 2018). The PRIsm MUlti-object Survey (PRIMUS; Coil et al. 2011) used slit masks to measure ~ 2500 redshifts in a single telescope pointing, recording 130 000 redshifts over 9.1 deg^2 to $i_{\rm AB} = 23.5$, with a redshift distribution peaking at $z \sim 0.6$. These surveys have been used to carry out a large number of analyses to quantify the galaxy populations and to constrain the cosmological model. Below we highlight some results from these surveys which explicitly focus on using galaxy clustering measurements to probe the physics of galaxy formation. Farrow et al. (2015b) measured galaxy clustering as a function of luminosity and colour using GAMA. Loveday et al. (2018) inferred the pairwise velocity distribution using the small scale galaxy clustering measured from GAMA. In both cases, these observational results were compared to theoretical models of the sort we will use here. Marulli et al. (2013) used VIPERS to measure the dependence of galaxy clustering on stellar mass and luminosity for $0.5 < z < 1.1$. Coupon & Arnouts (2015) combined clustering measurements with a gravitational lensing analysis to constrain the galaxy halo connection. Skibba et al. (2014) measured the clustering of galaxies in PRIMUS as a function of colour and luminosity, Skibba et al. (2015) studied the variation of the clustering amplitude with stellar mass and Bray et al. (2015) examined how the luminosity dependence of clustering depends on pair separation.

A limitation of spectroscopic surveys is the number of redshifts that can be measured in a single telescope pointing. This is set by the number of fibres or slits available to deploy to measure galaxy redshifts in the field of view. The use of some form of aperture to capture the light from a single galaxy also introduces a systematic effect on the clustering measured on small scales. The physical size of the slit or fibre means that in some cases only one member of a pair of galaxies within a particular angular separation can be targeted for a redshift measurement. This "fibre collision" effect can be mitigated by repeat observations of the same field or

by applying a correction to the measured pair counts.

An alternative to using spectroscopy to measure the radial distance to a galaxy is to use photometry taken in a number of bands. A photometric redshift can be assigned to a galaxy by, for example, comparing the observed flux in different bands to that derived from a template spectrum that is shifted in redshift (Benítez, 2000; Bolzonella et al., 2000). The photometric redshift approach has three advantages over spectroscopy: 1) the galaxy selection is homogeneous down to the flux limit, without any bias towards a higher success rate of redshift measurement for galaxies with emission lines (although the precision and catastrophic error rate of photometric redshifts will vary for different populations of galaxies; see e.g. Martí et al. 2014a; Sánchez et al. 2014), 2) there are no 'fibre collisions' that can impact galaxy clustering measurements and 3) there is no requirement to match the surface density galaxies to the number of slits or fibres within the field of view.

Broad band photometry, in which the typical filter width is ~ 1000 Å, is limited to a redshift precision $\Delta z/(1+z)$ (hereafter σ_z) of ∼3-5%. CFHTLS wide, a broad band survey observing in u, g, r, i and z which is 80% complete to $i < 24.8$ reaches $\sigma_z \sim 3\%$ for $i < 24$ with $\sim 4\%$ catastrophic errors (defined as $\sigma_z > 15\%$) (Ilbert, 2012). This level of precision is sufficient to divide galaxies into redshift shells in which the projected clustering can be measured. The error in the radial distance estimate in this case is $\sim 100h^{-1}$ Mpc at z = 0.7.

The accuracy of photometric redshifts can be improved by using narrower filters (Wolf et al., 2004). The Advanced Large, Homogeneous Area Medium Band Redshift Astronomical Survey (ALHAMBRA) Moles et al. (2008), offers a recent example of this by using 20 medium band filters, each ∼300Å in width, to reach an accuracy of $\sigma_z = 1.4\%$ for galaxies with $i < 24.5$ (Molino et al., 2014). Ilbert et al. (2009) reached $\sigma_z = 1.2\%$ for objects with $i < 24$ over the 2 deg^2 COSMOS field using a combination of broad, medium and narrow bands spanning the ultra-violet to the mid-infrared.

The Physics of the Accelerating Universe Survey (PAUS) is a narrow band imaging survey using PAUCam, Padilla et al. (2016), which was commissioned in June 2015, on the 4.2 m William Herschel Telescope, and Padilla et al. (In prep). PAUS

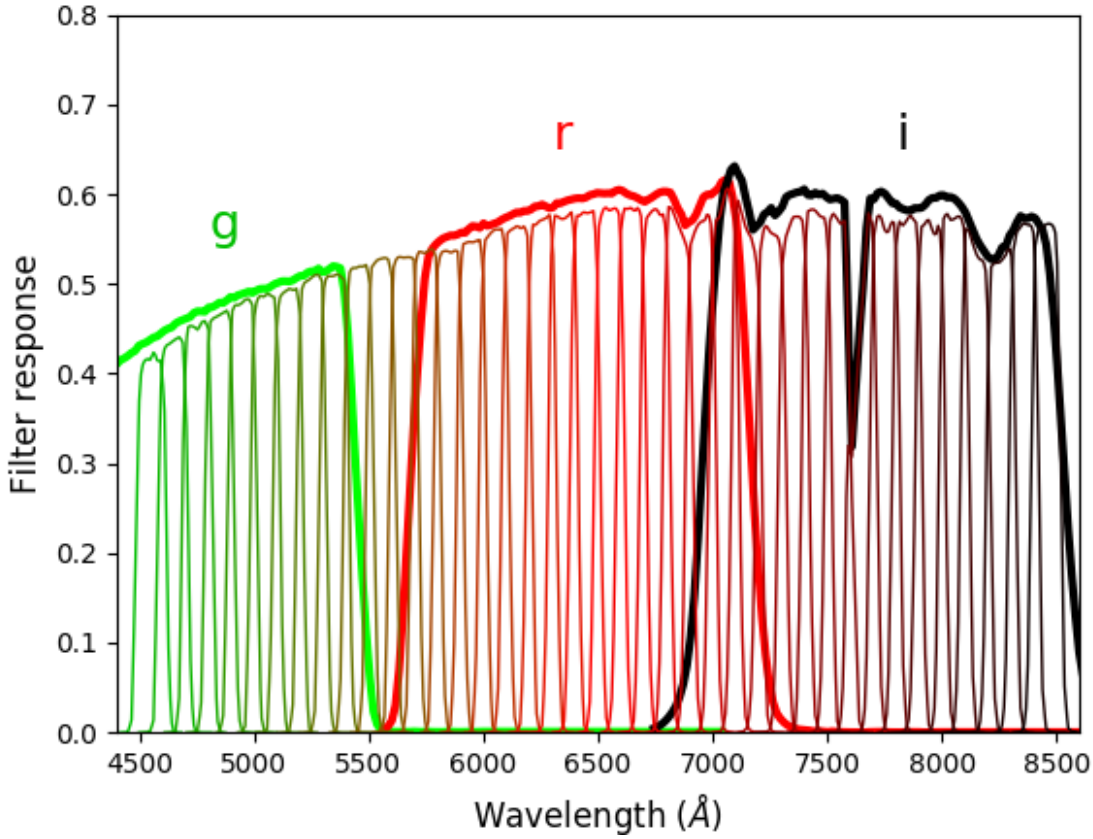

Figure 4.1: Filter response as a function of wavelength for the 40 PAUcam filters (thin lines) compared to CFHT MegaCam broad band filters g, r, i (thick lines). Filter response curves include atmospheric transmission, telescope optics and CCD quantum efficiency.

will measure narrow band fluxes by using forced photometry on objects previously detected in overlapping broad band photometric surveys CFHTLenS (Heymans et al., 2012) and KiDS (Kuijken et al., 2015). PAUS aims to perform forced photometry measurements in 40 narrow bands over 100 deg^2 for objects $i < 23$, and reach signal-to-noise of 3 at narrow band magnitude 23. Each of the 40 narrow band filters have FWHM 130Å and are spaced by 100Å, over the wavelength range of 4500Å to 8500Å (Martí et al., 2014a). Fig. 4.1 shows the PAUCam narrow band filters compared to the g, r and i bands from CFHTLS. 40 narrow bands span the region covered by these three broad band filters. The increased spectral resolution of PAUS imaging will allow for photometric redshift measurements of $\sigma_z = 0.35\%$ for objects $i < 23$ (Martí et al., 2014b). This represents an improvement of nearly an order of magnitude compared with typical broad band redshift measurement uncertainties, and in principle allows the radial distance information to be used in clustering estimates and to infer membership of galaxy groups.

The spectral features of a galaxy encode information about intrinsic properties such as its stellar mass, age and metallicity. Using these properties to define samples for clustering studies can then help us to understand the connection between galaxy properties and the mass of the host dark matter halo. These features include emission lines, absorption features, the 4000Å break and the shape of the continuum. Measuring the spectral features of individual galaxies has largely been in the domain of spectroscopic surveys. Kauffmann et al. (2003) used a combination of the strength of the 4000Å break and the Hδ absorption feature to constrain the stellar age, and contribution to stellar mass from recent star formation events, for a large sample of galaxies drawn from the spectroscopic Sloan Digital Sky Survey. Kriek et al. (2011) used stacking to measure the average values of spectral features using the medium band photometry of 3500 galaxies from the NEWFIRM survey to constrain star formation histories $0.5 < z < 2.0$. One of our goals here is to determine how competitively PAUS can be used to determine spectral features of galaxies, compared to the use of higher resolution spectra e.g. from zCOSMOS (Lilly et al., 2007), allowing for any modifications to the definitions of the spectral features driven by the narrow band photometry and taking into account errors in

the photometry and in the photometric redshift estimation.

Here we use the galaxy formation model GALFORM combined with a large-volume, high-resolution N-body simulation to build a mock catalogue for PAUS. Contreras et al. (2013) demonstrated that semi-analytical models of galaxy formation give robust predictions for galaxy clustering and, where differences exist between the models, they can be traced back to choices made in the treatment of galaxy mergers and the spatial distribution of satellite galaxies (see also Pujol et al. 2017). Farrow et al. (2015a) used the Gonzalez-Perez et al. (2014a) model to interpret GAMA clustering measurements as a function of luminosity, stellar mass and redshift.

The layout of this paper is as follows. Section 4.2 introduces the galaxy formation model and the PAUS mock catalogue, Section 4.3 investigates the use of the PAUS narrow band filters to measure galaxy spectral features, and Section 4.4 gives predictions for the narrow band luminosity functions, other characterisations of the galaxy population in PAUS and galaxy clustering. We conclude with Section 4.5.

4.2 PAUS mock lightcone

Here we describe the N-body simulation and galaxy formation model used (§ 4.2.1), introduce some basic properties of the mock catalogue constructed (§ 4.2.2), discuss the modelling of emission lines and their impact on narrow band fluxes (§ 4.2.3) and set out the treatment of errors in photometry and in photometric redshift errors.

4.2.1 N-body simulation & galaxy formation model

To model the galaxy population observed with PAUS we use the GALFORM semi-analytic galaxy formation model presented in Gonzalez-Perez et al. (2014a) (hereafter GP14). The GALFORM model (Cole et al., 2000) aims to follow the formation and evolution of galaxies in dark matter halos by solving a set of differential equations that describe the transfer of mass and metals between reservoirs of hot gas, cold gas and stars (see the recent extensive description of the model by Lacey et al. 2016 and the reviews by Baugh 2006 and Benson 2010). Due to the complexity and uncertainty of galaxy formation physics, many processes are modelled using equa-

tions which require parameter values to be specified. These are set by requiring the model to reproduce a selection of observations of the galaxy population, mostly at low redshift. The model calculates the star formation and merger history for each galaxy, including all of the resolved progenitors. With an assumption about the stellar initial mass function (IMF) and a choice of stellar populations synthesis (SPS) model, `GALFORM` outputs the flux for each galaxy in the PAUS bands using the composite stellar population obtained from the star formation history (Gonzalez-Perez et al., 2013). This includes a calculation of the attenuation in each band, based on the optical depth calculated from the metallicity of the gas and the size of the disk and bulge components of the galaxy (Gonzalez-Perez et al., 2013)

To build a mock catalogue on an observer's past lightcone with spatial information about the model galaxies, it is necessary to implement the galaxy formation model in an N-body simulation. The dark matter halo merger trees used in the galaxy formation model are also extracted from the N-body simulation (Jiang et al., 2014). The GP14 model is implemented in the Millennium WMAP7 N-body simulation (hereafter MR7, Guo et al. 2013). The MR7 run has a halo mass resolution of $1.86 \times 10^{10}\, h^{-1}\, \mathrm{M}_{\odot}$ (defined by the condition that a halo must consist of at least 20 particles) in a cube of side $500h^{-1}$Mpc. The use of the MR7 run means that the GP14 model is complete to $i < 23$ for $z > 0.2$. This is sufficient for our analysis. GP14 is an update of the model presented in Lagos et al. (2011a) to make it compatible with the WMAP7 cosmology and includes the improved star formation treatment implemented by Lagos et al. (2011b).

4.2.2 Mock catalogue on the observer's past lightcone

The depth of PAUS means that the properties and clustering of galaxies will evolve appreciably over the redshift range covered. Hence it is necessary to take this into account when constructing a mock catalogue for PAUS. The starting point is the galaxy population calculated using `GALFORM` at each of the N-body simulation outputs. Following the lightcone interpolation described in Merson et al. (2013), we construct a mock catalogue of one contiguous 60 sq deg patch. PAUS will target multiple fields but this will make little difference to one point statistics and small

scale clustering results presented here.

It is important to demonstrate that the mock catalogue is in broad agreement with the currently available observational data. The number counts of the PAUS mock compare well with large area photometric surveys as shown by Fig. 4.2, which shows the agreement between the model and the observations from Pan-STARRS (N. Metcalfe, priv. comm) and the Sloan Digital Sky Survey (York et al., 2000). The systematic differences between the data points are partly due to the slightly different i band filters used in each survey. The offset between the mock catalogue and the data is reasonable when considering the systematic differences between the data. The low redshift incompleteness due to finite halo mass resolution of the WM7 simulation does not impact this comparison as the total number of faint objects is dominated by galaxies with $z > 0.2$, which are well resolved in the model.

Fig. 4.3 shows the redshift distributions for the mock lightcones associated with five different galaxy surveys, along with data from VIPERS (de la Torre et al., 2013), and COSMOS photo-z (Ilbert et al., 2009). The choice of the two comparison datasets was made to test the mocks against surveys with flux limits on either side of the nominal PAUS i-band magnitude limit, VIPERS $i < 22.5$ and COSMOS photo-z with $21.5 < i < 24.5$. The model predictions agree reasonably well with the observations. The disagreement with the lowest redshift COSMOS data point is due to incompleteness in the model; this will be less important for the PAUS mock which is shallower than the COSMOS one. There is some disagreement with the high redshift tail of the VIPERS n(z). This suggests that the model under predicts the bright end of the i-band luminosity function at higher redshifts. However, as our analysis is limited to $z < 0.9$, an investigation into the cause and significance of this discrepancy is left to a later date. For $z < 0.9$, the VIPERS mock catalogue agrees well with the observations.

One current limitation of the mock catalogue is that it cannot be used for validation of photometric redshift codes. Tests run using the photo-z code embedded in the PAUS pipeline reveal discreteness in the returned redshifts which are aligned with MR7 snapshots. This issue arises due to the narrow width of the PAUS filters and the associated shift in redshift being smaller than the spacing of the N-body outputs

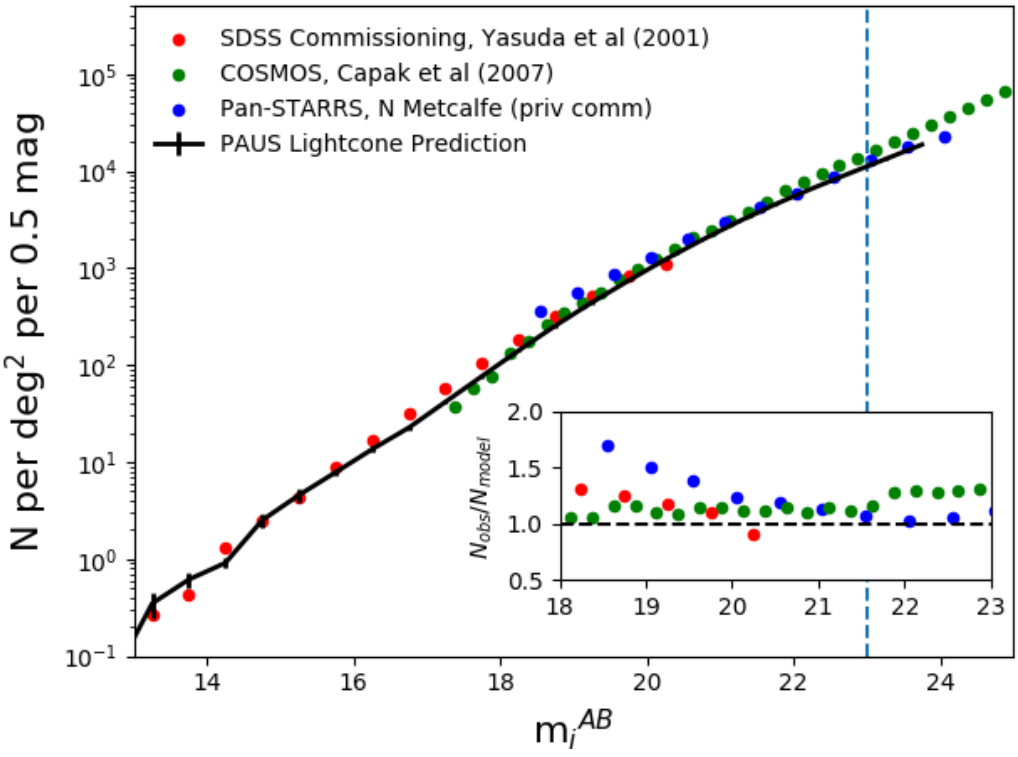

Figure 4.2: The predicted i-band galaxy number counts in the PAUS mock catalogue (solid line) compared with various observations (coloured symbols; see legend). The vertical bars on the solid line show a jackknife estimate of the sample variance on the number counts. We have omitted the errors on the observational estimates of the counts as they come from very different solid angle surveys. The vertical blue dashed line indicates the PAUS magnitude limit $i = 23$. The inset shows, on a linear scale, the result of dividing the observed number counts by the lightcone predictions.

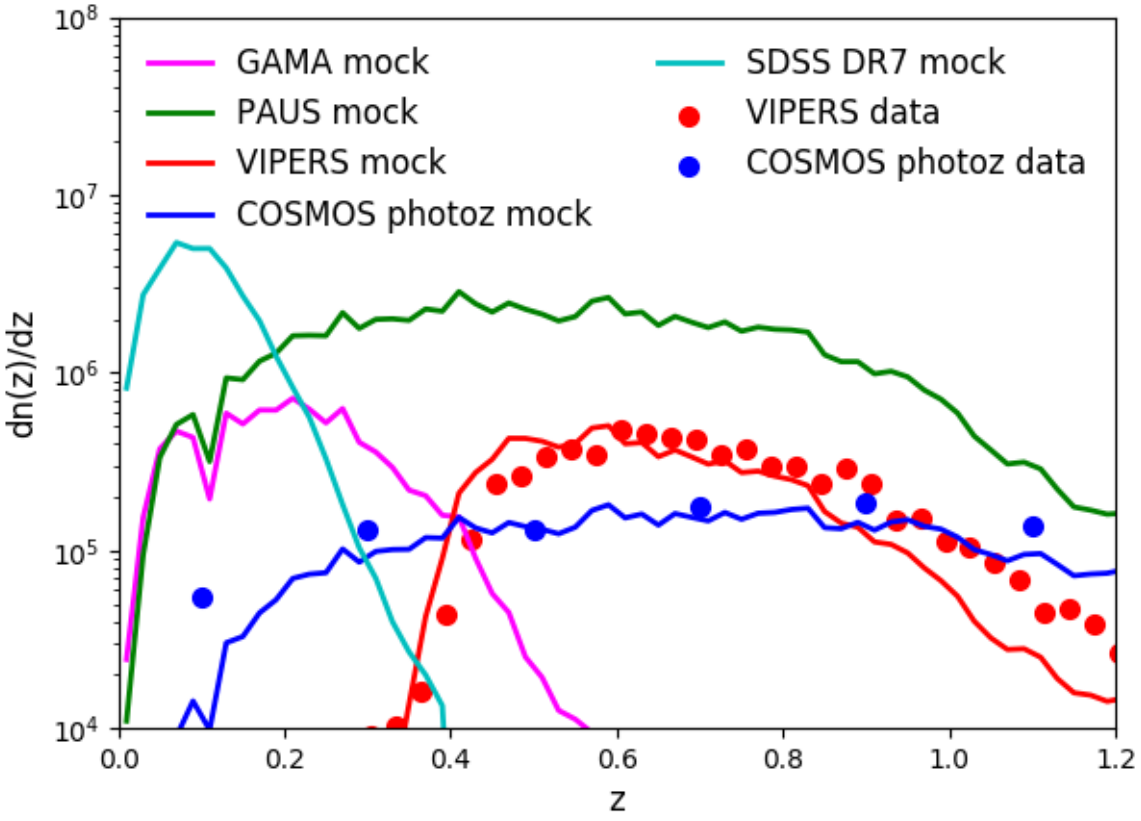

Figure 4.3: The redshift distributions in various mock catalogues (lines) compared to survey data (circles; see legend). The VIPERS data is taken from de la Torre et al. (2013), and the VIPERS mock catalogue is a 24 deg^2 lightcone to $i < 22.5$ with a 65% sampling rate. The mock VIPERS n(z) is then statistically corrected for the colour cut using the empirical relation found in de la Torre et al. The COSMOS photo-z data is taken from Ilbert et al. (2009), and the COSMOS photo-z mock is a 2 deg^2 lightcone retaining galaxies with $21.5 < i < 24.5$. The SDSS mock is a 10000 deg^2 lightcone with $r < 17.77$ and the GAMA lightcone covers 180 deg^2 to $r < 19.8$. These are plotted without an observational comparison to show the relative survey sizes and depths.

in redshift. This is not an issue for broad band photometry or when using multiple adjacent filters for measurements as in this analysis. A catalogue constructed using the P-Millennium simulation (Baugh et al, in prep), will improve both the mass and time resolution of our lightcone mock catalogue.

4.2.3 Impact of emission lines on narrow band fluxes

Emission lines are generally thought to make a negligible contribution to the flux measured in broad band filters, even for high redshift galaxies (Cowley et al., 2017). However, the narrow width of the PAUCam filters means that it is necessary to revisit the contribution of emission lines for PAUS.

GALFORM makes a calculation of the emission line luminosity of each galaxy using the number of Lyman continuum photons, the metallicity of the star-forming gas and a model for HII regions from Stasińska (1990). Gonzalez-Perez et al. (2017) give a recent illustration of this functionality presenting predictions for the abundance and clustering of OII emitters.

Fig. 4.4 shows the contribution emission lines can make to the PAUS narrow band fluxes for a single model galaxy. This illustrates that emission lines can be beneficial not only for the estimation of photometric redshifts, but suggests that PAUS could be used to identify and characterise populations of emission line galaxies. This is particularly relevant for the preparations for upcoming large spectroscopic surveys such as DESI (DESI Collaboration et al., 2016) and Euclid (Laureijs et al., 2011) which will build redshift catalogues from emission line galaxies.

Fig. 4.5 shows the fraction of galaxies whose relevant PAUS filter flux changes by a given percentage due to the contribution of one of the H_α, OII or OIII emission lines. For this calculation we restrict ourselves to a redshift range over which all lines are visible in the PAUCam filter wavelength range (see Table 4.1). The curves show the change in the flux of the filter with peak transmission closest to the observed emission line. Note that as PAUS filters have a FWHM 130Å, a full width of $\sim$ 135Å, and are spaced by 100Å, in a good fraction of cases a line will also contribute significantly to a second narrow band flux measurement.

It can be seen from Fig. 4.5 that for 50% of galaxies in this sample that at least

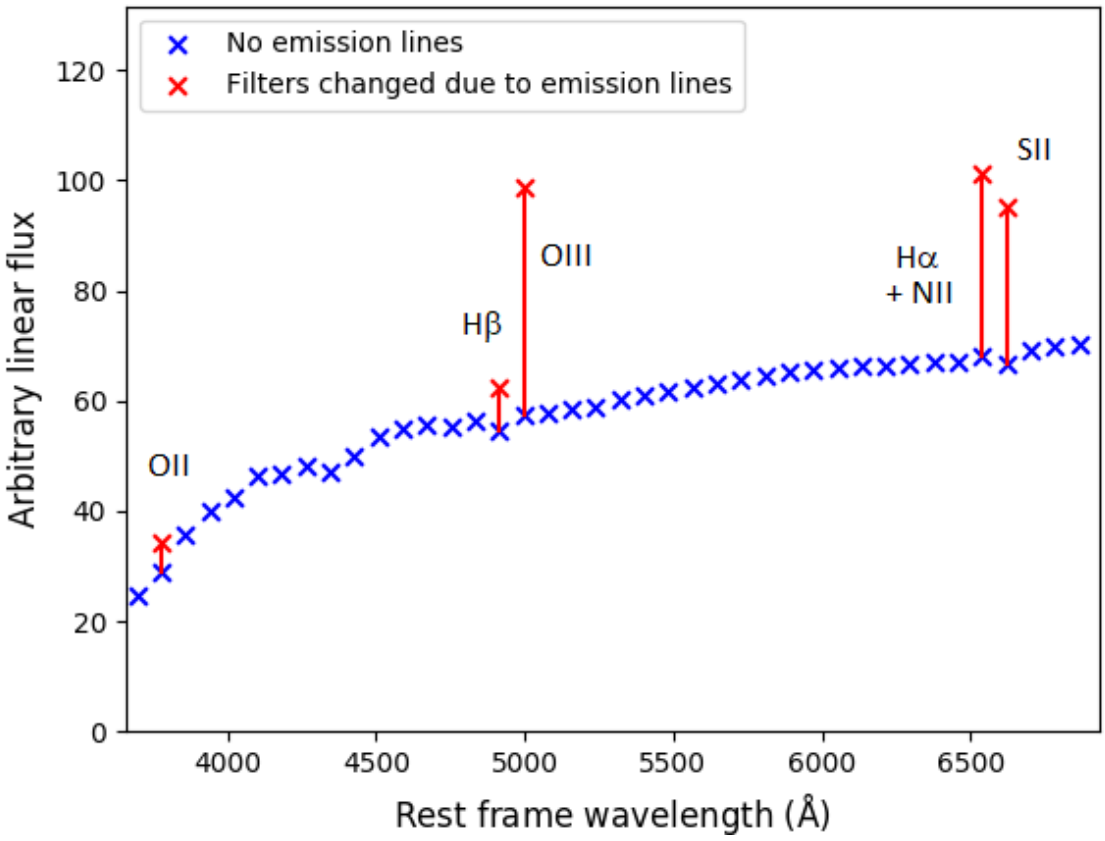

Figure 4.4: PAUCam filter fluxes for an illustrative star-forming galaxy taken from the PAUS mock. All 40 PAUCam filters are plotted. Blue (red) crosses show filter fluxes without (after including) emission lines.

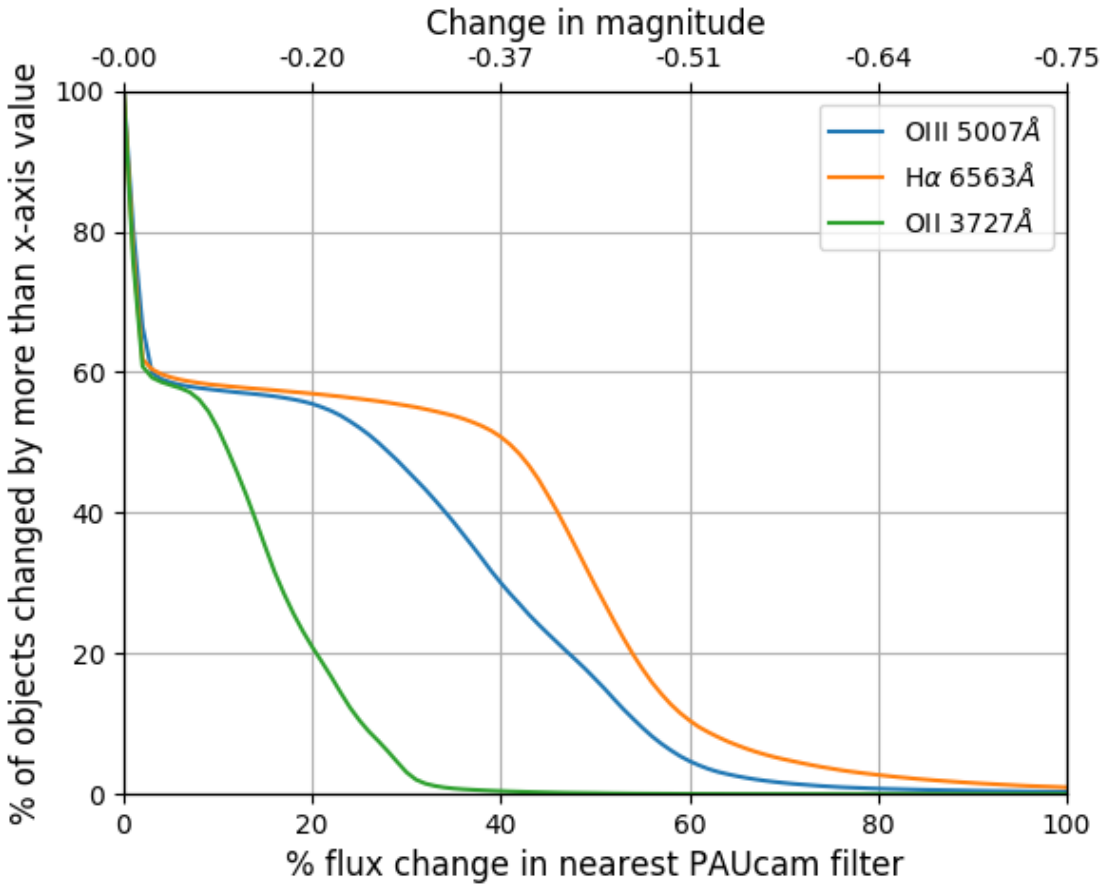

Figure 4.5: Fraction of model galaxies whose flux in nearest PAUCam filter is affected by the inclusion of a specific emission line (as indicated by the key). Only galaxies with redshift $0.21 < z < 0.3$ and magnitude $i < 23$ are shown to preserve a common sample where all lines can be sampled by a PAUCam filter. See section 4.2.3 for a discussion.

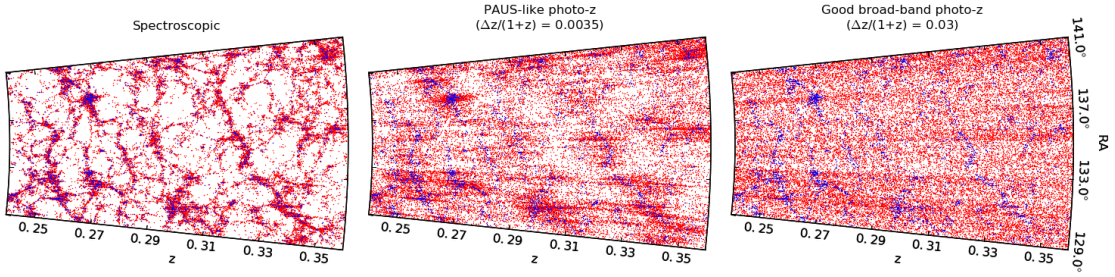

Figure 4.6: The spatial distribution of galaxies in a 1 degree thick slice from the PAUS mock catalogue. The three panels show the spatial distribution with spectroscopic redshift resolution (left), with PAUS-like redshift resolution (centre) and for typical broad band redshift resolution (right). Red points are galaxies brighter than the PAUS magnitude limit $i = 23$, while blue points correspond to GAMA galaxies ($r < 19.8$) with the spectroscopic redshift.

one narrow band flux measurement changes by 40% or more due to the inclusion of emission lines. That fraction falls to 38% for OIII and to 5% for OII, due to the average lower luminosity in these lines compared to that in the Hα line.

4.2.4 Photometry and redshift errors

Photometric redshift errors and photometry errors are added to the mock catalogue in post-processing. Two lightcones are produced, one with perfect photometry and correct redshifts and the other with PAUS-like errors applied. These errors are defined as Gaussian redshift errors of $\sigma_z = 0.35\%$ and Gaussian flux errors equivalent to a signal-to-noise ratio of 3 at magnitude 23 in the narrow band filters. These redshift errors are a simple approximation to PAUS photo-z measurements which will be fully explored in Eriksen et al (in prep). No photometry errors are included in the broad band magnitudes as the sources of the broad band photometry will be at least one to two magnitudes deeper than the nominal depth of PAUS of $i = 23$.

Fig. 4.6 shows the spatial distribution of galaxies in the PAUS mock catalogue and illustrates the impact of different redshift errors on the appearance of the large-scale structure of the universe traced by galaxies. Also shown in Fig. 4.6 are the model galaxies that satisfy the selection criterion for the GAMA survey, $r < 19.8$

(Driver et al. 2011; plotted at their spectroscopic redshift using blue points). The left panel of Fig. 4.6 highlights how much richer structures will be in a spectroscopic PAUS compared with GAMA, due to the deeper flux limit. The middle panel of Fig. 4.6 shows that a significant amount of radial information is retained once the redshifts of the mock galaxies are perturbed by the photometric redshift errors expected for PAUS. At $z \sim 0.3$, the expected photometric redshift errors for PAUS, σ_z of 0.35%, correspond to a comoving distance error of $\sim 13h^{-1}\,\mathrm{Mpc}$. Hence, it will be feasible to extract information about group and cluster membership from PAUS (for an example of group finding in a catalogue with less accurate photometric redshifts than those expected in PAUS, see Jian et al. 2014a). The right panel shows how little radial position information is retained when applying the photometric errors expected for broad band photometry.

4.3 PAUS Galaxy properties

The PAUS narrow band filters cover the wavelength range from 4500-8500Å in which certain spectral features can be observed. Over the range in which PAUS will make the greatest contribution to clustering measurements, $0.2 < z < 0.9$, the rest frame wavelengths from 3700Å to 4470Å are always accessible with PAUS photometry. Table 4.1 lists the spectral features in the PAUS wavelength range that are investigated here. We assess the direct observation of these features given a galaxy with PAUS-like uncertainties in photometry and redshift. By direct observation, we mean that we calculate the value of a feature by integrating over the appropriate PAUS filter fluxes, assuming that a redshift (of appropriate accuracy) has been measured by the photometric redshift code. An alternative approach would be to extract the spectral information by integrating over the appropriate range of the best fitting template spectral energy distribution obtained as part of the photometric redshift estimation. Using the templates in this way could reduce the statistical error, as this approach uses information from all of the filters that are available for a given galaxy. However, this would introduce a systematic error through restricting the results to be derived from combinations of a limited number of templates. It will in

fact be best to switch to using templates for measurements whose statistical errors exceed a certain threshold. The exact threshold is unknown as it depends on the unquantifiable systematic of template incompleteness, but this analysis can be used to define the point at which direct measurements become unfit for purpose, i.e when must we switch to using templates.

We restrict ourselves to features that are measures of the SED. Galaxy properties such as stellar mass and star formation rate require further modelling. Preliminary results attempting to recover stellar mass and star formation rate values of the simulated galaxies showed that without a k band measurement it is difficult to distinguish which galaxies are intrinsically red and which are red from extinction, causing a degeneracy in the inferred quantities.

4.3.1 Rest-frame defined broad bands

We define rest frame broad bands to best utilise the narrow band information from PAUS. These quantities are calculated by integrating the interpolated low resolution spectrum provided by the narrow bands. This type of direct rest frame measurement is possible because each of the PAUCam filters is flux calibrated, something which is often not the case with higher resolution spectra.

As can be seen from Fig. 4.7 and Table 4.1, the PAUS UV band has been chosen to be blue-wards of the 4000Å break, and hence is sensitive to very young stars in the composite stellar population of a galaxy. Conversely, the PAUS Blue band is chosen to be red-wards of the break, and thereby probes somewhat older stellar content. PAUS UV is chosen to be wider than PAUS Blue to increase its signal-to-noise ratio. This is important particularly for the UV band due to the typical shape of an i-band selected galaxy SED meaning that, on average, the UV is fainter than the Blue. PAUS Blue can only be directly measured up to z = 0.9.

There are several benefits to using these new rest frame broad bands over and above single narrow bands or traditional broad bands:

- These bands cover multiple narrow band filters, increasing the signal-to-noise ratio of an individual measurement compared with using a single narrow band.

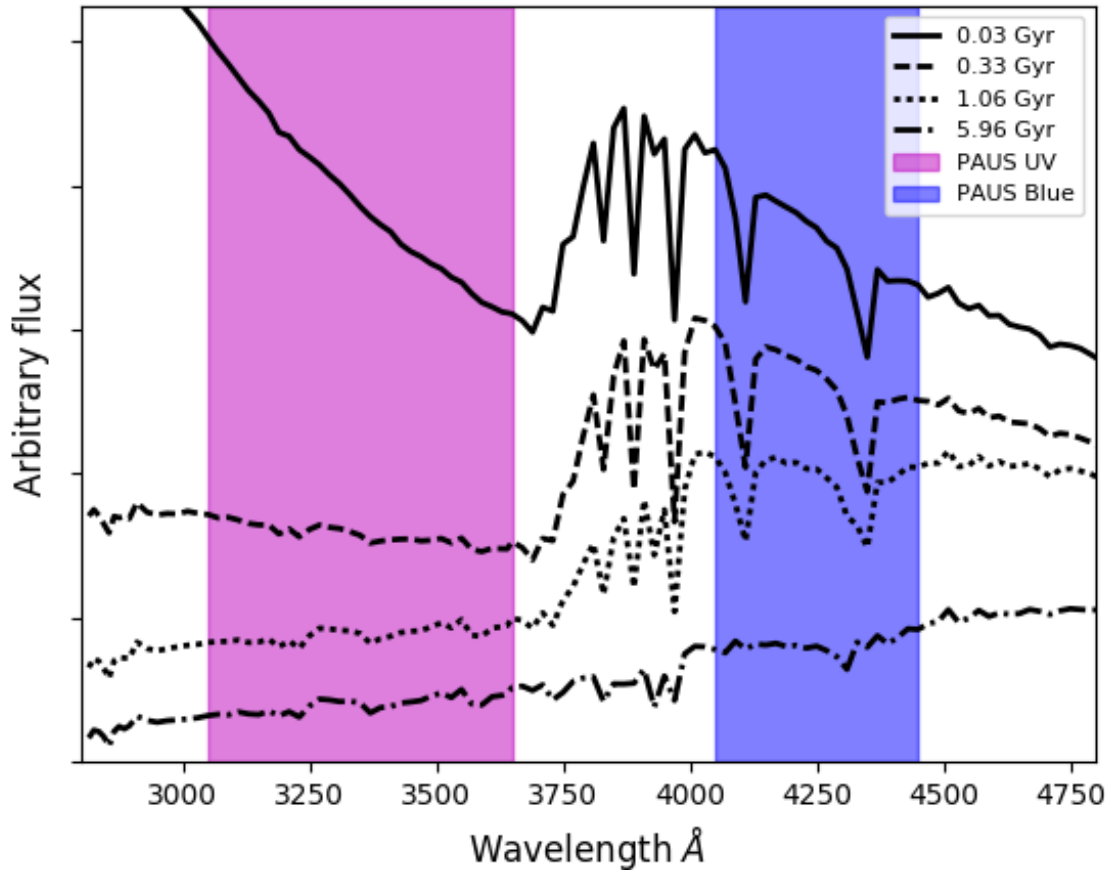

Figure 4.7: The definition of new rest frame broad bands, PAUS UV (magenta) and PAUS Blue (blue). At $z = 0.6$, PAUS UV overlaps with 9.6 PAUCam filters and PAUS Blue overlaps with 6.4 PAUCam filters. The curves shown are some of the SEDs for single age stellar populations that are used in the construction of the mock catalogue. In all cases these are for one quarter solar metallicity, with ages given in the key.

Feature	Wavelength Range Å	Redshift Range
OII	3727	0.21 - 1.28
OIII	4959/5007	0.0 - 0.70
Hα	6563	0.0 - 0.29
$D4000_N$	3850-3950 , 4000-4100	0.17 - 1.07
$D4000_W$	3750-3950 , 4050-4250	0.20 - 1.00
PAUS UV (M^h_{UV})	3050-3650	0.48 - 1.39
PAUS Blue (M^h_B)	4050-4450	0.11 - 0.90

Table 4.1: Wavelength and redshift ranges over which PAUCam filters (4500-8500Å) are sensitive to some common spectral features. The table is limited to the main features observable over the redshift range $0.2 < z < 0.9$. See Fig. 4.7 for the definitions of the PAUS UV and PAUS Blue bands and see Fig. 4.9 for the definitions of D4000. Note that $M^h \equiv M - 5\log_{10}h$.

- They are near direct measurements of galaxy rest frame SEDs and so do not require average k-corrections that broad band colour selections often require.
- They can be chosen to sample desirable sections of a galaxy SED precisely.
- Similar analyses can be performed on other photometrically calibrated spectra.
- The filter wavelengths are fixed in the observer frame but sample a wavelength range in the rest frame that shrinks as $1/(1+z)$ with increasing redshift. This means that the rest frame magnitudes we have defined are measured using filters that become more closely spaced as the redshift of the source increases. Hence the rest frame magnitudes are better sampled with increasing redshift, which partly offsets the typical decrease in the signal-to-noise as sources get fainter.

Fig. 4.8 shows how PAUS redshift and photometry errors propagate into errors in the PAU UV and PAU Blue magnitudes for a sample of mock galaxies with redshifts in the range $0.5 < z < 0.63$ and $i < 23$. For 80 % of model galaxies at $i = 23$ PAUS Blue can be measured to within ± 0.2 mags and PAUS UV to within

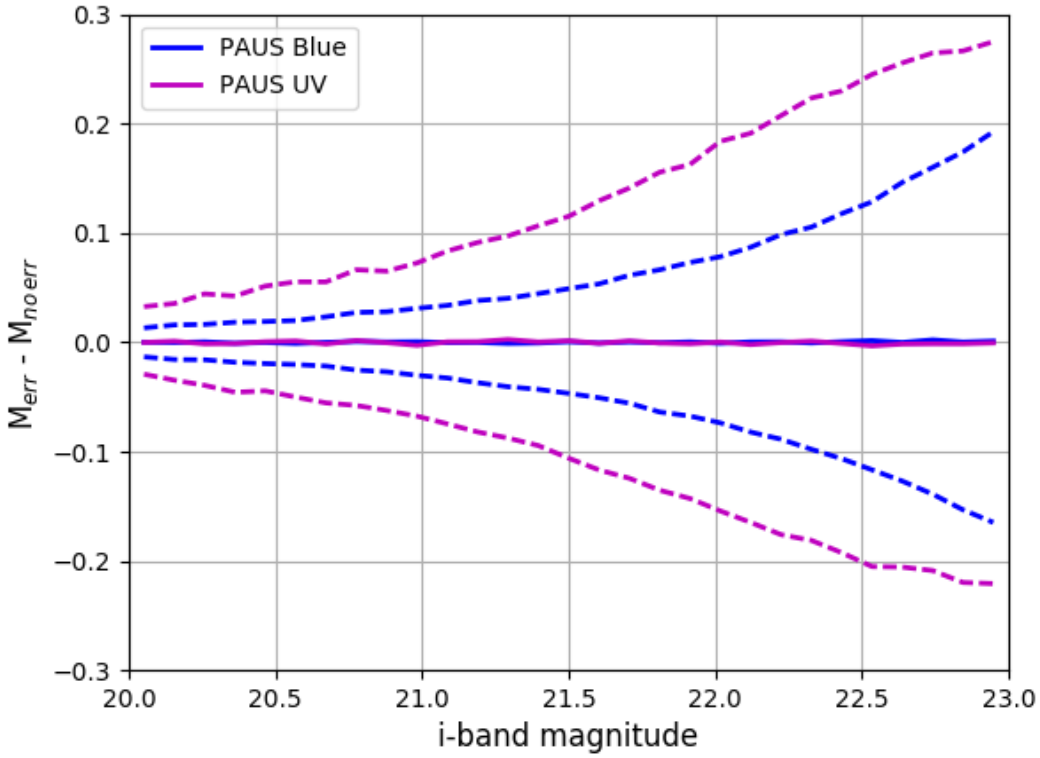

Figure 4.8: Statistical uncertainty in the PAUS UV and PAUS Blue magnitudes as a function of i-band magnitude for mock galaxies with $0.5 < z < 0.63$. The uncertainty includes redshift and photometry errors as described in Section 4.2.4. Solid lines show the median error and the dashed lines show the 10−90 percentile range.

± 0.25 magnitudes. There is also no bias in the measurement at all values of i-band magnitude. Other redshift selections give similar errors and also show no bias.

4.3.2 The 4000Å break

The 4000Å break is driven by a combination of CaII absorption lines and CN bands in the spectra of old stars. The quantity D4000 is the ratio of average flux in one spectral region at wavelengths just above 4000Å and that in a region just below in wavelength. The literature defines this quantity in two ways, D4000 narrow defined in Balogh et al. (1999) and D4000 wide defined in Bruzual (1983). The two flux bands used are different in each case and are visualised in Fig. 4.9. We first investigate if PAUCam has high enough resolution in a high signal to noise scenario to measure D4000 wide and narrow and then separately investigate D4000 measurements of PAUS mock galaxies.

Measuring the 4000Å break strength with PAUCam spectral resolution

In order to test the measurement of the D4000 feature we look at a sample of 4500 SDSS DR12 galaxy spectra, selected around $z = 0.1$ (Alam et al., 2015; Smee et al., 2013). We consider SDSS spectra for this test as the SPS used in `GALFORM` are limited to 20Å resolution. The SDSS galaxies were each randomly uniformally placed at a redshift in the range $0.2 < z < 0.9$ so that the different ways in which the PAUS filter can trace the feature are taken into account. The fluxes in the 40 PAUCam narrow bands were calculated for each galaxy. D4000 was then calculated using both the full resolution SDSS spectra, and then again by integrating a linear interpolation of the PAUS filter measurements. Both definitions of D4000 from the literature were calculated and results are presented with and without PAUS-like redshift errors, as defined in Section 4.2.4. We do not include photometry errors, as first we want to check if PAUCam has sufficient resolution to measure D4000 in a high signal to noise scenario.

Fig. 4.10 shows how well interpolating between the PAUCam filters recovers the spectroscopic result for both the wide and narrow D4000 definitions from the literature. Both definitions of D4000 are biased due to the effective smoothing

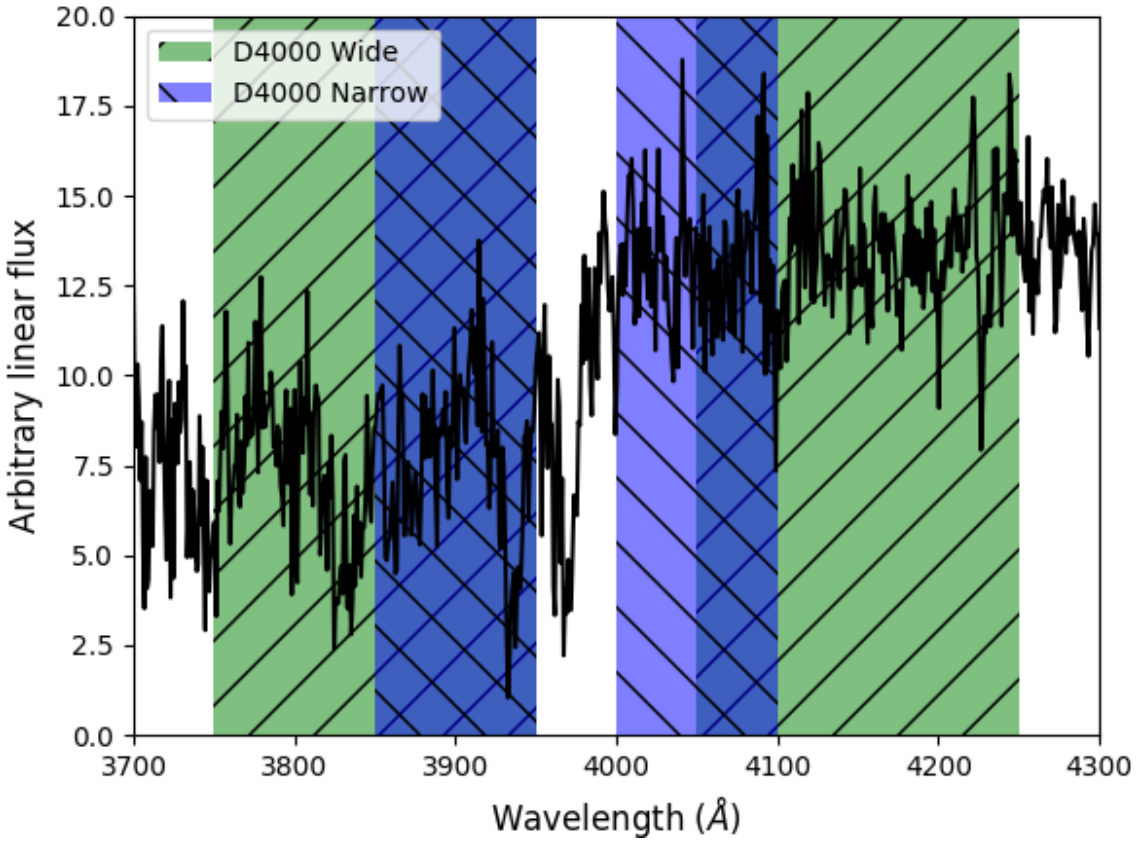

Figure 4.9: Definitions of D4000 wide and D4000 narrow overlaid on a randomly selected, de-redshifted, SDSS DR10 galaxy. The green shaded region represents the wide definition (3750-3950 and 4050-4250Å) from Bruzual (1983), and the blue the narrow (3850-3950 and 4000-4100Å) from Balogh et al. (1999).

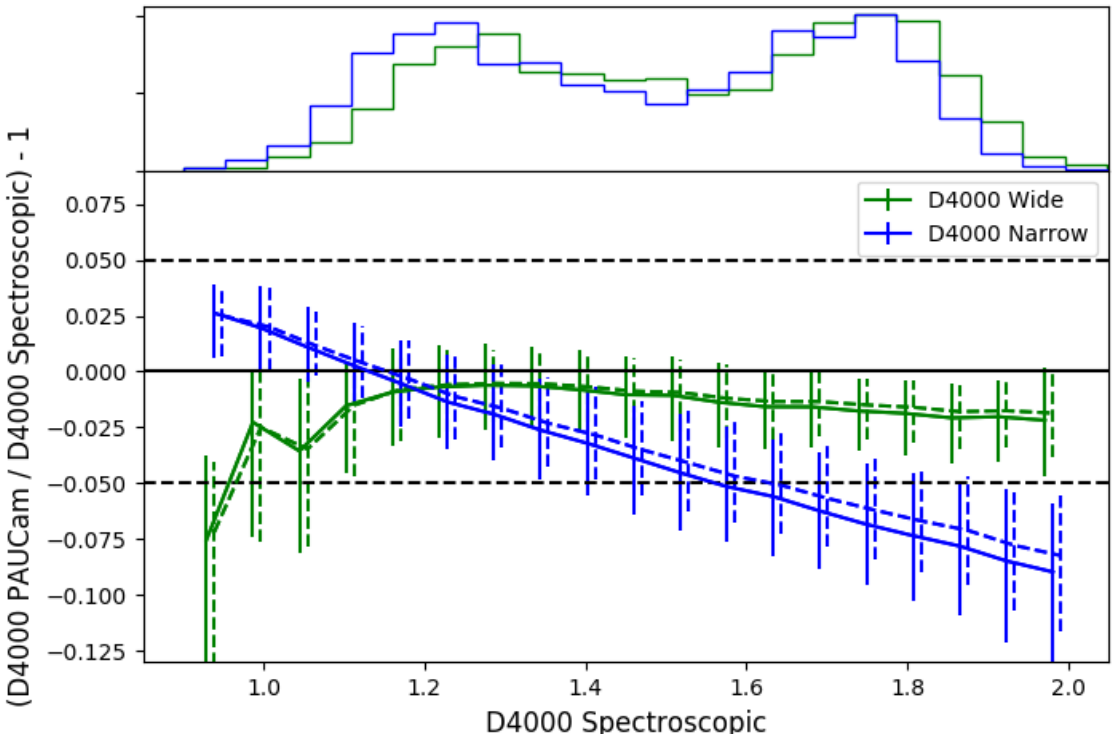

Figure 4.10: Relative accuracy with which D4000 can be recovered using PAUCam, as a function of the strength of D4000, measured using 4500 SDSS spectra observed at z $\sim$ 0.1 and redshifted over the interval $0.2 < z < 0.9$. D4000 spectroscopic is measured using the full spectra information while D4000 PAUCam uses the PAUS filters. The green line shows the result for D4000 wide and the blue for D4000 narrow. Solid lines and error bars (which indicate the 10-90 percentile range) include a PAUS-like photo-z error while the dotted lines and error bars do not. Dashed lines are displaced in the x direction by 0.01 to make the error bars visible. The top panel shows the distributions of D4000 values for the sample.

of a sharp spectral feature due to the finite width wavelength intervals used to calculate D4000. $D4000_n$ is affected by this bias more than $D4000_w$. The $D4000_n$ bias also scales as a function of the spectroscopic value for D4000 whereas the bias of $D4000_w$ is nearly constant with respect to this ideal. The $D4000_w$ measurement is biased by ∼2%. This bias is not corrected for in later analysis as we will see in section 4.3.2 that it is small compared to the random errors on PAUS mock galaxies. Once photometric redshift errors are included the error bars on both measurements increase only slightly. The error bars on $D4000_w$ are also smaller than those of $D4000_n$, $\sim \pm 2\%$ and $\sim \pm 4\%$ respectively. The superior recoverability of $D4000_w$ suggests this 4000Å break definition should be used for PAUS measurements. The superior bias and noise performance of $D4000_w$ is to be expected as it overlaps with more PAUCam filters than $D4000_n$ does at a given redshift.

The redshift dependence of the $D4000_w$ measurement bias was investigated, as at each redshift the filters will trace the break in a different manner. The extreme scenarios are that the D4000 break lies mid-way across a filter or exactly in between two filters. It was found that the bias of $D4000_w$ varies by less than 1% as a function of redshift. It is therefore not necessary to model this redshift dependence.

4000Å break strength in PAUS

To investigate the ability of using the PAUS photometry to measure $D4000_w$, this quantity is measured in both the mock catalogue with no errors and in the one with redshift and photometric errors introduced in section 4.2.4. Fig. 4.11 shows the relative error in $D4000_w$ for redshift slices as a function of i-band magnitude. 80 percent of galaxies at $i = 23$ lie within 50% of the true value of $D4000_w$. Photometric uncertainty is therefore the dominant source of error for PAUS galaxies. Looking at the population histogram in Fig. 4.10 it can be seen that the majority of galaxies have values of $D4000_w$ between 1.0 and 2.0, with a bimodal distribution peaking at 1.2 and 1.75. An error of 50% is therefore very large compared to the range of $D4000_w$. Galaxies with $i = 21.5$ and $z = 0.55$, however, are expected to have just a 15% error in $D4000_w$, showing that direct $D4000_w$ measurements for a bright subset of PAUS objects are feasible. $D4000_w$ errors are smaller for higher redshift

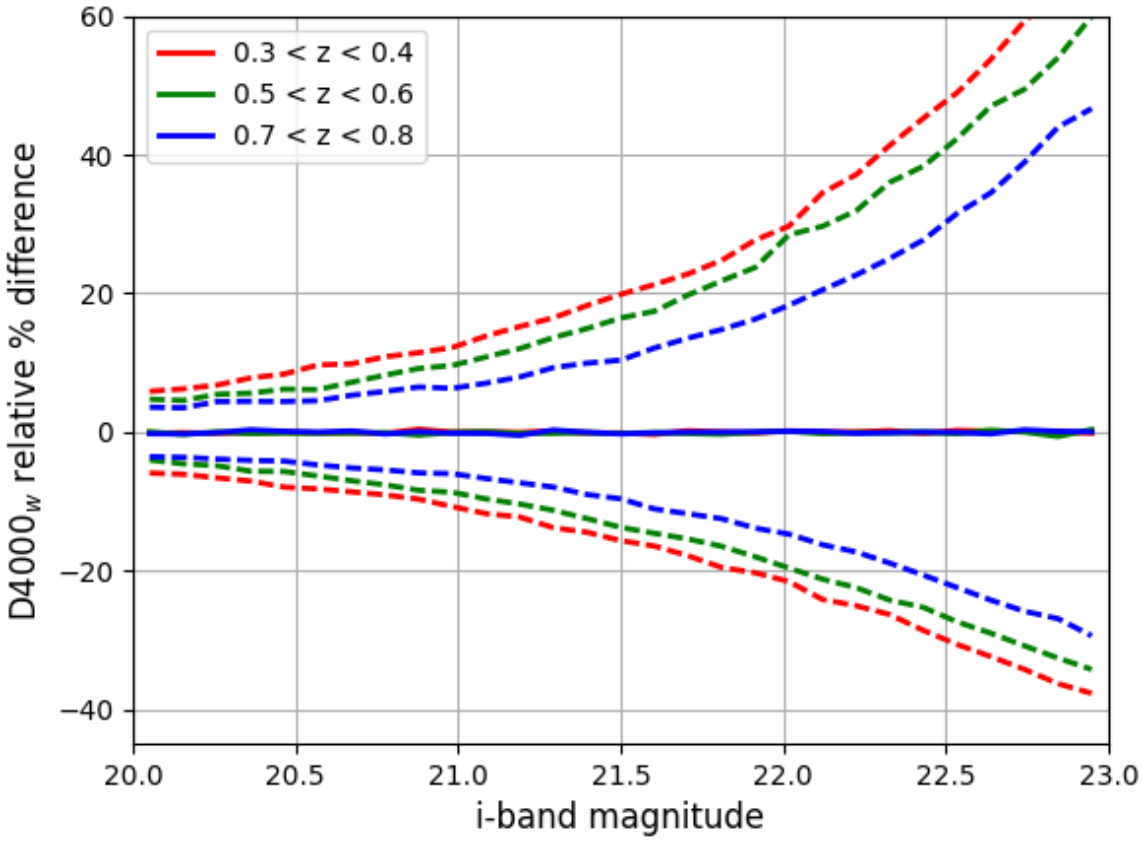

Figure 4.11: Relative percentage difference in D4000$_w$ as a function of i-band magnitude for different redshift slices. The relative percentage difference is defined as $100 \times (D4000_{\rm err} - D4000_{\rm true})/D4000_{\rm true}$, where the subscript err(true) refers to measurements made in the catalogue with(without) PAUS simulated redshift and photometric errors.

galaxies at a fixed i-band magnitude as the rest frame defined $D4000_w$ bands overlap with more PAUCam filters in this case than at lower redshifts. Individual studies will need to define a tolerable error for this quantity. Bimodal population cuts for example will be able to use a large subset of data and retain completeness and purity, whereas studies on the ages of individual galaxies may need to use a significantly restricted subset of the catalogue. One could also stack populations of galaxies and make a measurement on a mean spectra to reduce statistical error.

4.4 Results

In this section we review various properties of the galaxy population that we expect PAUS will be able to measure based on the predictions made using our mock catalogues.

4.4.1 Narrow band luminosity functions

The parameters in the `GALFORM` model are calibrated to match low redshift observations, which are mainly one-point statistics such as the luminosity function. One of the applications of PAUS is to provide improved constraints on the model parameters by providing measurements of the narrow band luminosity function over a significant baseline in redshift.

We have seen that individual PAUCam narrow band magnitudes can be significantly affected by the emission line flux from a galaxy, so here we investigate the sensitivity of the narrow band luminosity functions to the inclusion of emission lines in the GP14 model (see Gonzalez-Perez et al. 2017 for a further discussion of model predictions for OII emitters). Fig. 4.12 shows how a narrow band luminosity function of PAUCam like filter chosen to overlap in the rest frame with the OII emission line changes when the flux from the line is included. Measurements are made in the simulation snapshots. Inference of this quantity from observer frame measurements would require accurate k-corrections, which will be an output of the photo-z code. It can be seen that neither redshift evolution nor inclusion of emission line flux change the faint end slope of the luminosity function in the GP14 model. The value of M^*

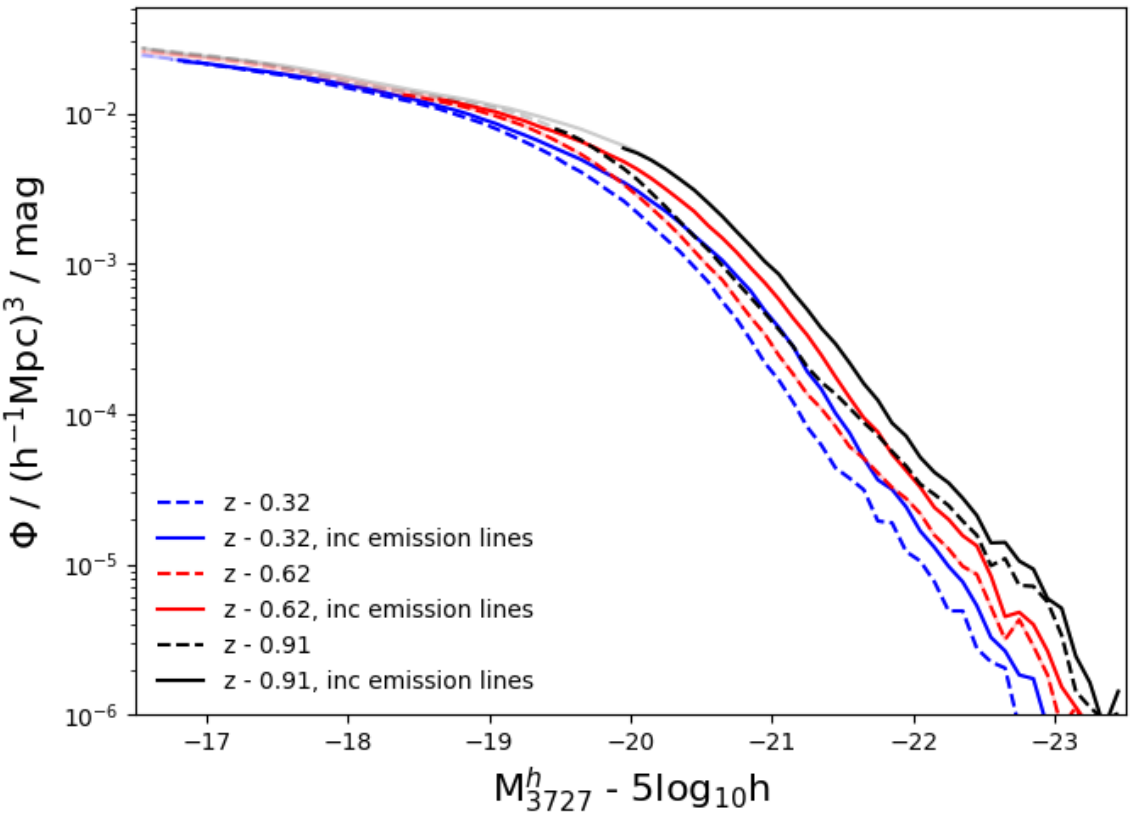

Figure 4.12: Luminosity functions at several snapshot redshifts (as labelled) of a PAUS filter at rest frame wavelength of 3727Å ± 62.5Å. A different PAUS filter is used at each redshift, chosen to overlap with the OII emission line. Solid lines show the prediction including the emission line flux and dashed lines do not. The plotted curves become fainter when they fall below 95% completeness at $i < 23$.

(the knee of the luminosity function), however, increases with both redshift, as a result of the increasing star formation, and also with the inclusion of OII line flux. The contribution of the stellar continuum to the flux in this band can be estimated by averaging the flux in bands placed at either side of the band that contains the OII emission, providing a constraint on the amount of emission line flux and its evolution with redshift.

4.4.2 Characterisation of the galaxy population

One desirable objective for studying the evolution of the galaxy population is the ability to separate galaxies by colour in a consistent way across the redshift range

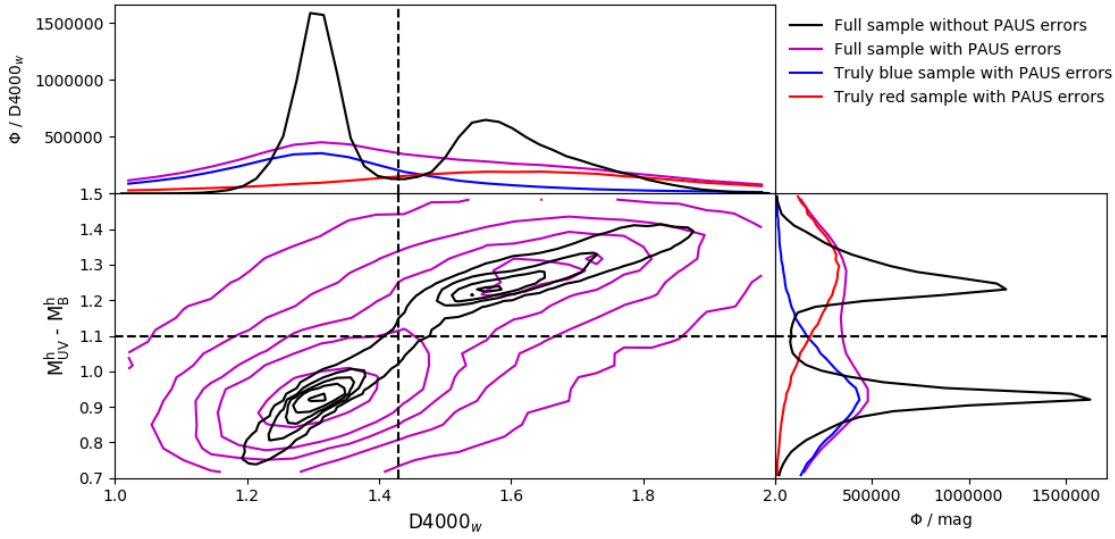

Figure 4.13: Distribution of galaxies with $i < 23$ and $0.5 < z < 0.63$ in the D4000$_w$ and M^h_{UV} - M^h_B colour plane, with and without simulated PAUS errors. The contours contain 10, 30, 50, 70 and 90% of the sample. The two histograms, labelled Φ/D4000$_w$ and Φ/mag, are the counts per unit D4000 and per unit M^h_{UV} - M^h_B respectively. The solid black lines show the distributions for the full sample without errors and the magenta ones show the full sample with errors. The red (blue) curves show the distribution of galaxies that are intrinsically red (blue) in each measure when errors are included.

sampled by PAUS. This objective can be achieved by using a cut in D4000 at $z < 0.5$ and a cut in M^h_{UV} - M^h_B above redshift 0.5. We could define a band further into the red to make a colour cut at lower redshifts, as M^h_{UV} cannot be defined below z $\sim$ 0.5, see Table 4.1, but a cut in a different section of a galaxy SED may non-trivially select galaxies differently than the M^h_{UV} - M^h_B cut. In particular, the use of a redder colour selection might mix galaxies with different recent star formation histories, making clustering comparisons across redshift ranges less informative. The use of $D4000_w$ means that we are making a colour cut centred on the same portion of the SED as a cut in the colour M^h_{UV} - M^h_B.

Fig. 4.13 shows the distributions of $D4000_w$ and M^h_{UV} - M^h_B for a redshift range in which both can be measured. Both quantities show a bimodal distribution, which we can loosely refer to as 'red' and 'blue' populations. A cut is made at $D4000_w = 1.42$ and M^h_{UV} - $M^h_B = 1.1$. Before photometric errors are added, disagreements in red-blue classification when using the two measures are at the sub-percent level. That is, very few galaxies are different classifications according to the two measures, which can be seen by how little of the black contours lie in the top-left and bottom-right sections of the plot compared to the other sections. The cut in M^h_{UV} - M^h_B is appropriate to split the bimodal population at higher redshifts, as is the cut in $D4000_w$ for lower redshifts. Comparisons carried out using the model rest frame bands show that these colour cuts are similar to a traditional broad band rest frame cut in $u - g$. When including photometric errors, mixing between the red and blue populations is more severe when using D4000 than with the rest frame magnitudes due to the larger fractional error in $D4000_w$ at a fixed i band magnitude (see Sections 4.3.1 and 4.3.2). Errors on the M^h_{UV} - M^h_B colour are driven largely by errors in the UV magnitude.

4.4.3 Galaxy clustering

We select volume limited galaxy samples for clustering measurements based on redshift, PAUS blue luminosity and rest frame colour (as defined in Section 4.4.2). We choose not to split samples based on inferred quantities such as star formation rate or stellar mass as the inference of these properties from narrow band photometry

is left to future work. Inferring these properties has also been shown to introduce biases based on the assumptions made in these inferences (Mitchell et al., 2013). In the mock including simulated PAUS errors the cuts are made after all sources of error are included. See Appendix A.1 for clustering definitions, details of the calculations and open source code links, and Appendix A.2 for more information on sample selection. All errors in this section are calculated by using a jackknife over 12 regions in the simulated survey, see e.g Norberg et al. (2009b).

We estimate the galaxy bias from the ratio of the projected galaxy clustering to the projected clustering of the MR7 dark matter at the median redshift of the sample in question. The values of the correlation function for the MR7 snapshots were taken from McCullagh et al. (2016). This quantity allows us to separate the evolution of the dark matter over time from the evolution of the galaxy population. On large scales this quantity is equal to the linear bias. More specifically we define projected galaxy bias as

$$b(r_p, z) = \sqrt{\frac{w_p(r_p, z)}{w_p(r_p, z)_{DM}}}, \tag{4.4.1}$$

where $w_p(r_p, z)$ is the projected correlation function defined in Eqn. 3.3.3.

Impact of photometric uncertainty

Fig. 4.14 shows the bias measured for one mock PAUS sample (-19.5 $< \mathrm{M_B^h} <$ −19.0) in the redshift range $0.5 < z < 0.63$, both with and without PAUS magnitude and photometric redshift errors. The value of $\pi_{\max}$ used was $100h^{-1}$Mpc. Fig. A.1 in the Appendix shows the recovery of the projected correlation as a function of different photometric redshift errors. A value of $\pi_{\max}$ of $50h^{-1}$Mpc would have been sufficient for the photometric redshift errors assumed in this work, and would have slightly reduced the statistical noise, but the real survey will have a distribution of photometric redshift errors so the conservative value of $100h^{-1}$Mpc was chosen.

For the sample selected only on redshift and $\mathrm{M_B^h}$, the black lines in Fig. 4.14, the projected clustering signal is recovered without systematic error when including PAUS-like errors. The jackknife statistical errors only slightly increase when compared with the ideal case. This demonstrates that the PAUS photo-z measurements

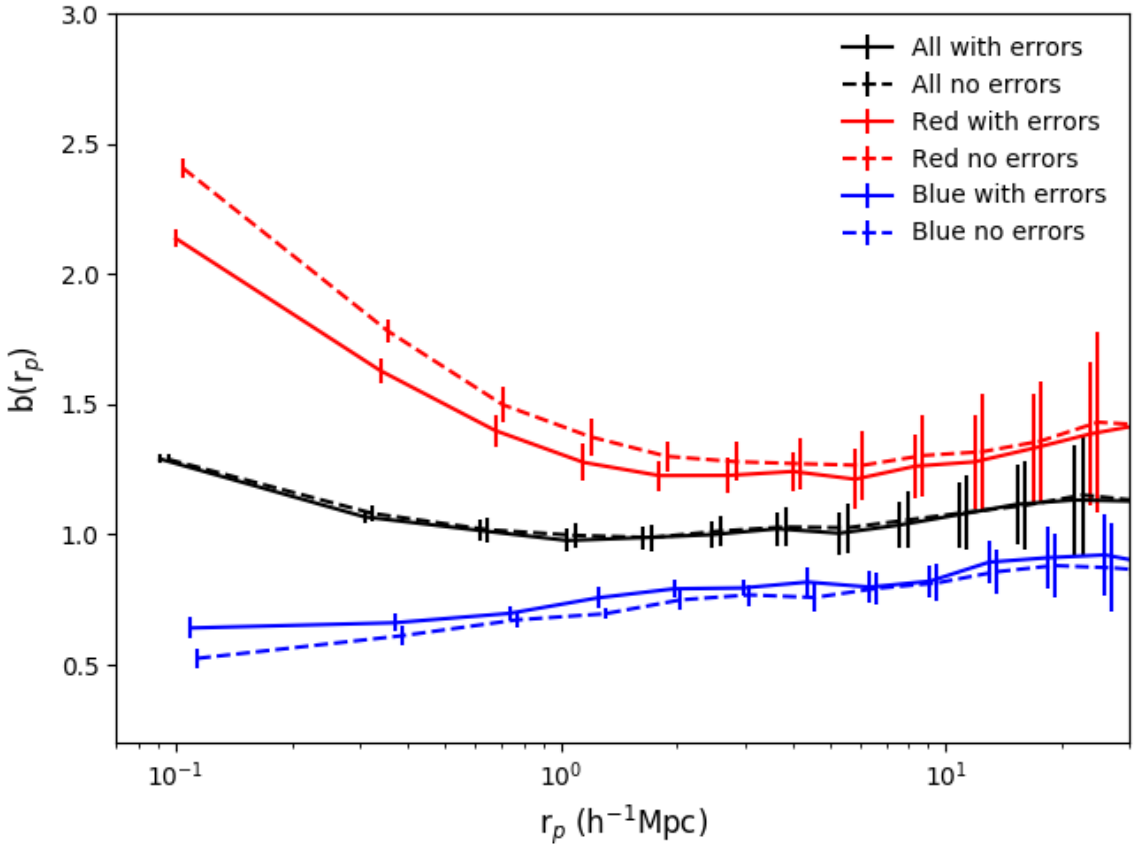

Figure 4.14: Projected galaxy bias (Eqn. 4.4.1) for a typical PAUS sample ($0.5 < z < 0.63$, $-19.5 < \mathrm{M_B^h} < -19.0$). The full galaxy sample is shown in black and the results on splitting the sample into red and blue populations are shown in these colours. Solid lines show the results using the lightcone with redshift and photometry errors taken into account and the dashed lines the results without including these uncertainties. Errors were calculated using jackknife resampling.

are sufficient to calculate the projected galaxy clustering without systematic error. Table A.1 in the appendix shows that the sample with PAUS-like errors is over 90% pure and complete. For the same sample with photo-z errors only and no photometry error these numbers both rise above 96%, showing that mixing between samples due to photometric redshift errors is minimal.

Once a colour cut is applied to the full magnitude limited galaxy sample, a significant difference can be seen in the projected bias measurements for the red and blue populations. Errors in the photometry introduce mixing between the red and blue populations which leads to a small reduction in the difference between the one-halo ($\sim< 1h^{-1}$Mpc) scale projected bias of red and blue galaxies. Nevertheless the difference between the clustering measurements for these populations remains significant. Systematics on two-halo scales ($\sim> 1h^{-1}$Mpc) are within the statistical uncertainties. This confirms that the most significant source of systematic error in this analysis will be on one-halo scales and come from the misclassification of galaxies into red or blue sub-samples using these direct rest frame measurements. This systematic error shows up here as there is a large contrast between the one-halo clustering of red and blue samples, and PAUS will have small statistical errors on those scales. Again, statistical colour errors could be reduced by using the best fit photo-z SED inferred colours for fainter samples, but this is not tested here. This highlights the importance of understanding sample selection and the role of mock catalogues in interpreting clustering results.

The redshift evolution of clustering

Fig. 4.15 shows the predicted redshift evolution of projected galaxy bias measured for samples of red and blue galaxies with $-19.5 < \mathrm{M_B^h} < -19.0$. Our estimate of the bias naturally takes into account the evolution of the clustering of the dark matter over this redshift interval. For all redshift bins red galaxies show stronger clustering than blue galaxies. This difference becomes larger for pair separations below $\sim 1h^{-1}$Mpc corresponding to pairs within common dark matter halos. The bias also increases with redshift for both red and blue samples. This trend is also seen in all the other luminosity bins we have explored. This result, the decline in the bias

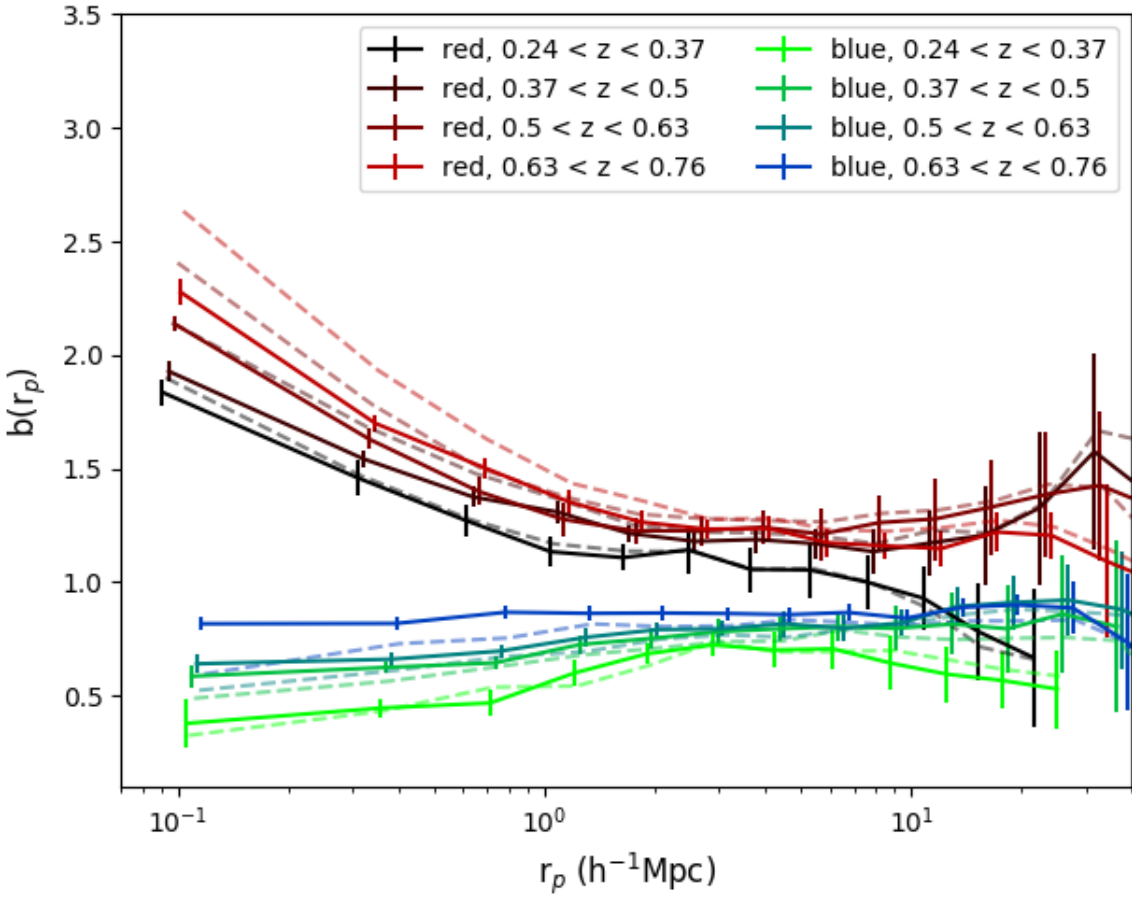

Figure 4.15: Projected galaxy bias (Eqn. 4.4.1) inferred from the projected correlation function measured for samples with $-19.5 < \mathrm{M}_{\mathrm{B}}^{\mathrm{h}} < -19.0$, split by colour and redshift. Solid lines show the results using the lightcone including redshift and photometry errors and the dashed lines show the results without these uncertainties. Errors, from jackknife resampling, are only shown for PAUS-like sample.

as the universe ages, is due to faster growth of the dark matter correlation function compared with that of the galaxy correlation function over the same period, see e.g. Baugh et al. (1999). Again the systematic errors on two-halo scales are within statistical uncertainties. Qualitative trends seen on one-halo scales are preserved once errors are included, but the contrast between red and blue one-halo clustering is reduced due to colour mixing.

The luminosity dependence of galaxy clustering

Fig. 4.16 shows the model prediction for the luminosity dependence of galaxy clustering. The split between the red and blue galaxies is once again very evident. As commented above, the red samples have stronger clustering than their blue counterparts. There is little luminosity dependence of the clustering measure for the blue samples (see also Kim et al. 2009 for a discussion of the luminosity dependence of clustering in an earlier version of the `GALFORM` model used here). On the other hand, the clustering of the red samples shows a moderate dependence on luminosity which weakens on large scales and does not preserve the same ordering with luminosity that is displayed on small scales. Once again two-halo scale results are recovered within statistical errors.

One reason for the inverted trend of clustering decreasing with luminosity seen on small scales is due to the dominance of satellite galaxies in the lower luminosity red samples. This can be seen in Fig. 4.17, which shows the satellite fractions of the clustering samples (Number of galaxies with satellite label in a sample divided by the total number of galaxies in the sample). Note that measuring this with the data would require significant modeling work. This figure also illustrates the impact of colour mixing on the satellite fraction of the samples. The lower luminosity bins at the lowest redshifts are significantly affected by mixing between central and satellites. These lower luminosity and redshift samples have the largest difference in satellite fraction between the red and blue populations and are the most likely to be misclassified in colour. This mixing error will either need to be modeled using mocks or we will have to rely instead on inferred colours extracted from an SED template, allowing for template incompleteness as a systematic error.

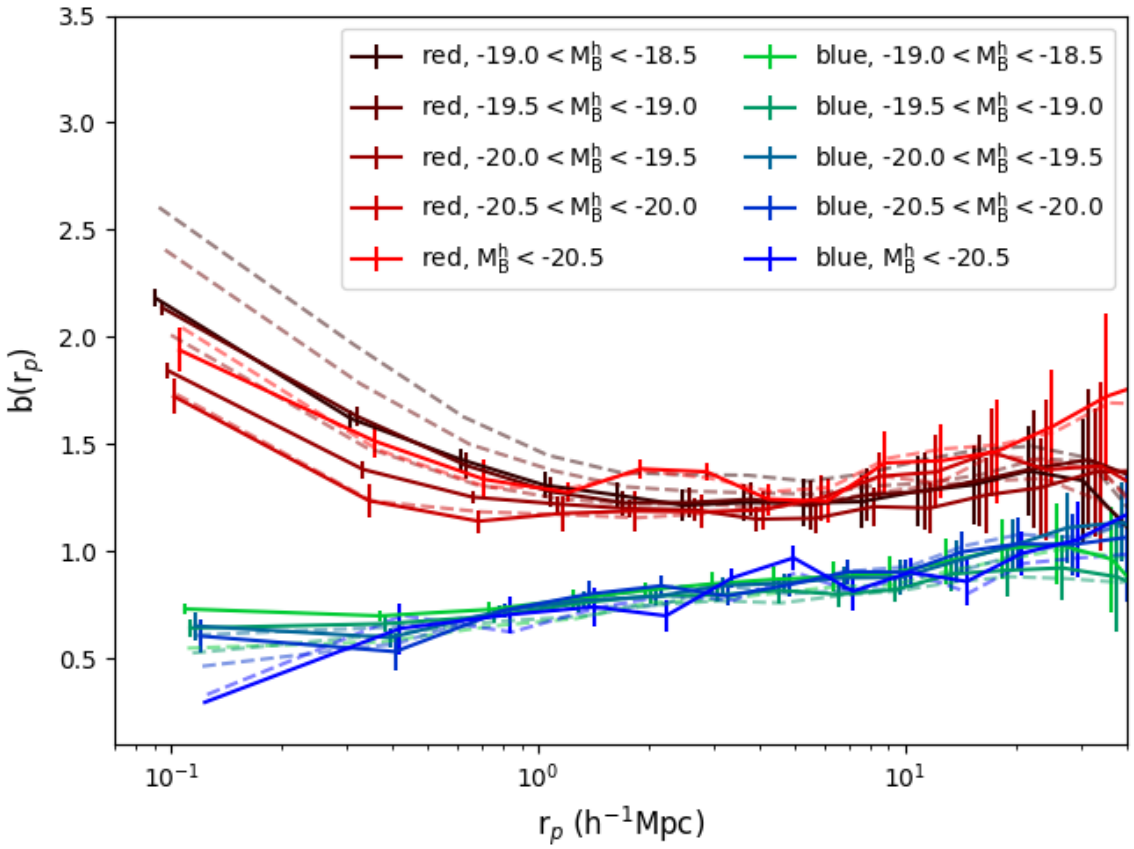

Figure 4.16: Projected galaxy bias (Eqn. 4.4.1) inferred from the projected clustering measured for samples $0.5 < z < 0.63$, split by colour and M_B^h. Line types as in Figure 4.15.

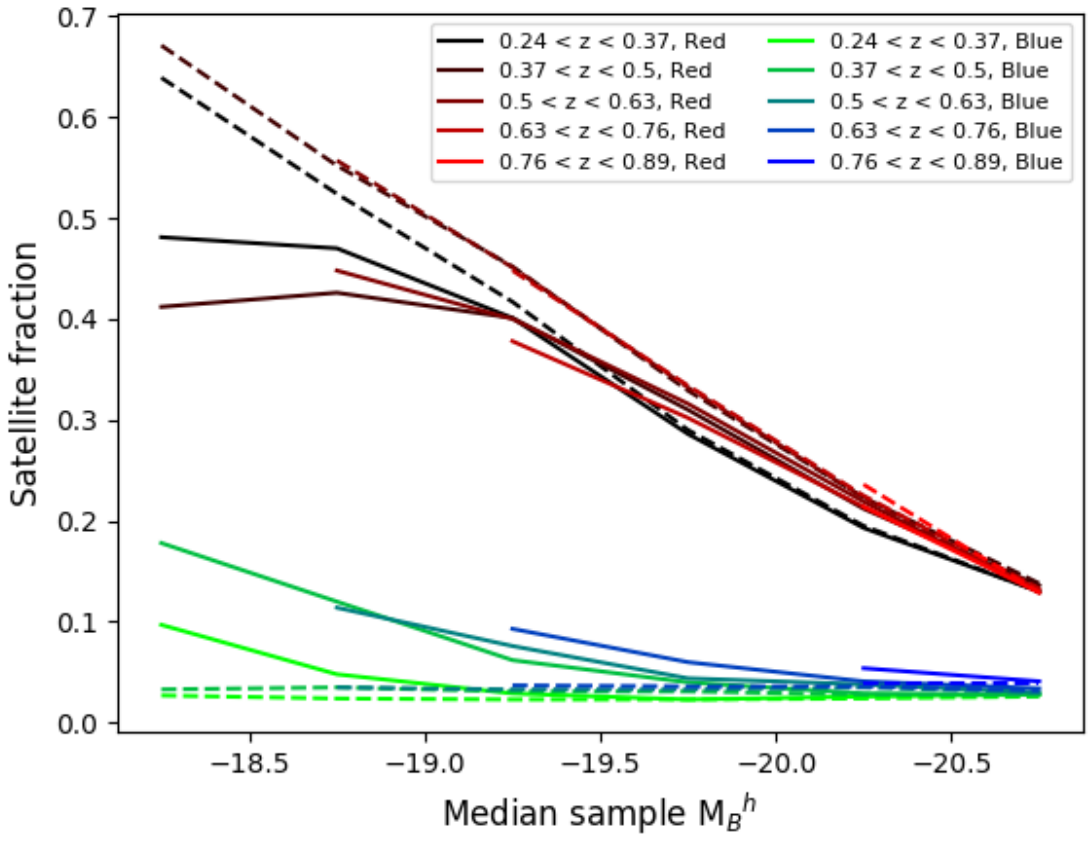

Figure 4.17: Satellite fraction as function of M_B^h for galaxy samples split by colour and redshift. Line types as in Figure 4.15.

4.5 Conclusions

We have introduced a mock catalogue built from a semi-analytical model of galaxy formation implemented in an N-body simulation for use in conjunction with the Physics of the Accelerating Universe Survey (PAUS). PAUS is a novel narrow band imaging survey which is underway on the William Herschel Telescope. The width of the PAUS filters means that photometric redshifts of unprecedented accuracy will become available for a homogeneously selected sample of galaxies down to $i = 23$. The PAUS mock is built using the GP14 `GALFORM` model (Gonzalez-Perez et al., 2014a), which is run on the MR7 N-body simulation (Guo et al., 2013). The galaxy snapshots produced at the output times of the MR7 run are then used to construct a mock catalogue on an observer's past lightcone, which predicts the evolution of the clustering of galaxies and their properties (Merson et al., 2013). The mock catalogue is available on request at CosmoHub[1] (Carretero et al., 2017).

The resulting mock catalogue agrees with observed galaxy number counts to within the scatter between different surveys. Over the redshift range in which PAUS is expected to make the largest impact, $0.2 < z < 0.9$, the mock is in good agreement with the redshift distributions from COSMOS photo-z and VIPERS. There is some tension at $z > 1$ where the mock under predicts the VIPERS n(z), but this redshift range is less relevant for PAUS, and the observational errors are large at these redshifts (de la Torre et al., 2013).

We include galaxy emission lines in the predicted PAUS measurements and show that this has a significant effect on PAUS narrow band fluxes. We show how the rest-frame narrow band luminosity function changes when emission lines are included by choosing a rest frame narrow band that overlaps with the OII emission line. The GP14 `GALFORM` model predicts no change in the faint end slope of the narrow band luminosity function with or without emission line flux included and as a function of redshift. It does, however, predict an increase in M* with both redshift and on the inclusion of emission lines.

We define rest frame broad bands calculated directly from narrow band fluxes

[1] `https://cosmohub.pic.es/home`

and predict that a PAUS Blue (PAUS UV) flux can be directly measured with an error of $\pm$ 0.15 ($\pm$0.25 mags) down to $i = 22.5$. These provide rest-frame measurements without needing to make any of the assumptions that come with average k-corrections used with broad band measurements. These rest-frame measurements are only possible because the PAUS narrow band measurements are flux calibrated. We show that the PAUCam filter set has sufficient resolution to measure the strength of the 4000Å break, D4000. We predict that $D4000_w$ can be directly measured in PAUS to better than $\pm \sim 10\%$ precision for galaxies with $i < 21.5$. Providing errors on these quantities as a function of i-band magnitude will allow the PAUS data analysis pipeline to decide when to switch from directly measuring a quantity using the observed PAUCam filters to integrating over the best fitting SED assigned by a photometric redshift code. The latter incorporates statistical information from all filters but restricts results to a linear combination of SED templates, and is not explored here.

We explore galaxy clustering measurements over a redshift range of 0.2 to 0.9 for multiple luminosities and colours using the rest frame colours, $D4000_w$ and redshift. PAUS will provide a unique sample spanning this redshift range over a larger area than previously possible, with nearly 100% completeness. No close galaxy pairs are missed as is often the case in spectroscopic surveys.

We show that systematic errors in projected clustering recovery due to PAUS photometric redshift errors are significantly smaller than statistical errors. All two-halo scale projected clustering results are recovered within statistical errors once PAUS redshift and photometry errors are included. One-halo scale clustering shows the same qualitative trends as measurements made in the ideal case but there is a loss of contrast between the one-halo scale clustering of red and blue galaxies caused by colour misclassification. This demonstrates the importance of a mock catalogue to interpret galaxy clustering results, particularly in the case of PAUS results on small scales, where statistical errors are small and any systematics are likely to be the dominant source of error.

We provide testable predictions for the mock catalogue that the measured galaxy clustering will evolve more slowly with redshift than the redshift evolution in the

dark matter, especially for the one-halo term. The mock also predicts that red galaxies will cluster more strongly than blue galaxies. We also predict that fainter galaxies will cluster more strongly than brighter galaxies on small scales due to their larger satellite fraction, and that this trend will be particularly strong for red galaxies.

This work provides a tantalising illustration of the science that will be possible with PAUS, particularly with a view to constraining the galaxy - dark matter halo connection.

Chapter 5

Galaxy group identification with Markov Clustering (MCL)

Galaxy groups are the observable counterparts to dark matter halos, so detecting galaxy groups can help us infer more about the galaxy-halo connection. We introduce a new framework for finding galaxy groups, Markov Clustering (MCL) (Van Dongen, 2000). We explain that the widely used friends-of-friends (FOF) algorithm is a subset of MCL. We test the MCL algorithm in real space on a mock galaxy catalogue constructed from an N-body simulation using the `GALFORM` semi-analytic model. With a fixed linking length the FOF algorithm produces the best group catalogues as measured by the variation of information statistic (Meilă, 2003). We use the local galaxy density to modify the linking length which improves both the FOF and the MCL group catalogues, with the latter being superior to FOF. The MCL group catalogue recovers accurately the group multiplicity function (to within 7%) across all multiplicities. It has better and more consistent purity and completeness values as a function of multiplicity than the comparable FOF catalogue. MCL allows probabilistic pairwise connection amplitudes which could prove very useful for galaxy catalogues with mixed redshift precision such as PAUS. We propose a model to extend this work to redshift space and photometric redshift space and demonstrate how this method connects galaxy pairs of different separations and position uncertainties.

5.1 Introduction

A galaxy group is defined as a collection of galaxies that are gravitationally bound within the same dark matter halo. Galaxies within groups can tell us about galaxy interactions and how galaxy properties and small scale clustering depend on local environment (Schneider et al., 2013; Treyer et al., 2018; Barsanti et al., 2018). Galaxy groups, being proxies for dark matter halos, are also important tracers of large scale structure and are often used in galaxy clustering measurements (Wang et al., 2008; Berlind et al., 2006a) or lensing analysis (van Uitert et al., 2017). If we can estimate the masses of groups, we can investigate the mass to light ratio in different structures, which can help us to better understand the galaxy-halo connection. We can also test the limits of our assumptions that the clustering of galaxy halos only depends on halo mass by searching for a signal of 'assembly bias' (e.g. Gao et al., 2005). Wang et al. (2013) claimed to detect this assembly bias for SDSS galaxy groups.

Identifying galaxy groups requires estimation of which galaxies lie in the same dark matter halo. There are multiple ways to do this. Two common methods are the friends-of-friends (FOF) based approach and the halo based approach. Eke et al. (2004) constructed a FOF group catalogue (2PIGG) containing 190 000 galaxies from the 2dF Galaxy Redshift Survey (Colless et al., 2001). Robotham et al. (2011) constructed a FOF group catalogue (G^3Cv3) of $\sim$45000 galaxies from the GAMA survey (Driver et al., 2011). Liu et al. (2008) extended the friends of friends method to galaxies with photometric redshift measurements (pFOF), which was later tested and applied to the Pan-STARRS1 medium deep surveys (Jian et al., 2014b). Yang et al. (2005) developed and tested a halo based group finder that was later used to construct a galaxy group catalogue of SDSS galaxies (Yang et al., 2007).

This chapter frames the FOF approach to galaxy group finding as a particular solution to the graph clustering problem (Schaeffer, 2007). Graph clustering aims to find clusters of points given the pairwise connection amplitudes between them. It is a problem that occurs in many situations, such as detecting communities in social networks (Liu et al., 2014). We explain that the FOF algorithm is a subset of the Markov graph clustering algorithm MCL (Van Dongen, 2000) and investigate the

application of the MCL algorithm to the problem of galaxy group detection. MCL has been widely used in the field of bioinformatics in detecting groups of proteins based on their pairwise interactions (Vlasblom & Wodak, 2009).

This work is carried out with a view to constructing a group catalogue using the PAU Survey data (Castander et al., 2012). A PAUS group catalogue would be significantly deeper than an SDSS (York et al., 2000) or GAMA (Driver et al., 2011) group catalogue, and would have a larger area and better completeness in both sampling and redshift than a group catalogue constructed using similar depth surveys such as zCOSMOS (Lilly et al., 2007) or VIPERS (Guzzo et al., 2014). A PAUS group catalogue would therefore better probe the redshift evolution of galaxy groups and better probe further down the galaxy luminosity function and the halo mass function at low redshift than has previously been possible. The challenge of constructing a PAUS group catalogue will be identifying how to deal with the varying redshift precision of the narrow band photometric redshift measurements. MCL is a promising approach as it allows probabilistic pairwise connections, something that could be useful in a PAUS catalogue where it is more natural to frame pairwise connections as probabilities than as binary links.

While working on Markov Clustering, Tempel et al. (2018) proposed a new Bayesian group finder based on marked point processes. Both approaches are vastly different in nature, but might have similar positive properties in the sense that they can deal with probabilistic spatial information.

Section 5.2 presents the MCL algorithm and explains its relation to the FOF algorithm. Section 5.3 presents the mock catalogue we use to test the algorithm. Section 5.4 summarises the metrics we use to assess group finding performance. Section 5.5 presents the results in real space. Section 5.6 proposes a scheme to extend this work to redshift space and demonstrates how it works with a toy model. Section 5.7 concludes. Note that in this work we refer to a 'clustering' of galaxies interchangeably with a 'grouping' of galaxies. We will refer to the two point correlation function if we discuss galaxy clustering in the more classical/typical context.

5.2 Markov clustering algorithm

We briefly introduce the concept of graph clustering (Schaeffer, 2007). A graph is a structure that gives pairwise connection amplitudes between points. The most obvious and instructive example of a graph would be people connected on a social network (Liu et al., 2014). Here users are 'friends' with other users. The entire friendship network can be represented by a symmetric binary matrix, which we will call the pairwise connection matrix w_{ij}, which contains a 1 if two users are friends and a 0 if they are not. A graph clustering algorithm aims to detect communities within this structure. We can understand the common friends-of-friends algorithm (FOF) as an approach to solving this problem that defines communities as containing users that can be connected in any way at all on the graph. We can also see how this algorithm is flawed in this context, as it is most likely that the majority of the network would be connected in a large single group, despite most people in the group having little connection to most of the other users in the group. Other approaches have been developed to tackle this problem. One of them is the widely used 'highly connected substructure' (HCS) algorithm (Hartuv & Shamir, 2000), which, as the name suggests, detects substructure that is highly rather than loosely connected.

In the astrophysical case, we first have a connection criterion that sets the values of w_{ij}. This is normally based on the distance between galaxies, setting w_{ij} to 1 if the galaxies are close to each other and 0 if they are not. The w_{ij} matrix is then typically used along with an FOF algorithm to detect groups, but there is nothing to stop us using a different graph clustering algorithm once we have decided our connection criterion.

The Markov clustering algorithm (MCL) was developed by S.Van Dongen as a fast and scalable approach to graph clustering (Van Dongen, 2000). The code is publicly available at `http://micans.org/mcl/`. MCL is an algorithm that takes w_{ij}, as an input, and assigns points to clusters. It does this by simulating a random walk on the graph using w_{ij} as transition probabilities to determine which points are most bound. A random walk will get temporarily stuck, more so in a structure that is tightly bound, only rarely jumping between structures. MCL has one free parameter, inflation (Γ), which is used to trim connections that are used for these

rare inter-cluster jumps. MCL is an iterative process that repeats matrix operations on w_{ij} until the matrix converges and the clusters can be read off. It is implemented as follows:

1. Normalise w_{ij} column-wise such that the sum of each column is 1.
2. Square the matrix w_{ij}[1].
3. Raise every element of w_{ij} to the power of Γ and renormalise.
4. Go to step 2 again if w_{ij} has yet to converge. (Convergence defined when all elements of w_{ij} have changed by less than a constant factor)
5. Read the groups from the converged w_{ij}.

Raising the elements of w_{ij} to the power of Γ before renormalising is designed to boost the more traveled connections and reduce the value of the less traveled inter-cluster connections. The larger the value of Γ the quicker the fall in the the lower probability connections and the more that MCL will split structure into smaller parts. A Γ value of 1 will simulate an infinite random walk, and will join any structure that has any path connecting it, just as in an FOF algorithm. The FOF algorithm is therefore a subset of the MCL algorithm corresponding to a Γ of 1[2]. Γ has no maximum value, but there will be a value of Γ above which the catalogue stops splitting, as all clusters are fully connected subgraphs, i.e. all points in clusters are connected to all other points in the same cluster. MCL was chosen for investigation here because it has a limit of the common FOF algorithm, and because it supports pairwise connection matrices that contain probabilities rather than just binary links.

Figure 5.1 shows how the algorithm works on an example graph that contains two clusters with three fully connected points each, with one link between the clusters[3]. A Γ value of less than 1.47 fails to correctly identify the two clusters and instead

[1]This is not strictly a random walk as the initial matrix is not saved. See Van Dongen (2000) for a discussion on why this produces a similar result and why this exact procedure was chosen.

[2]We tested this using MCL and an FOF algorithm and it was found to be correct.

[3]For the astrophysically minded, one can consider all links to be of equal distance.

$$w_{ij} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

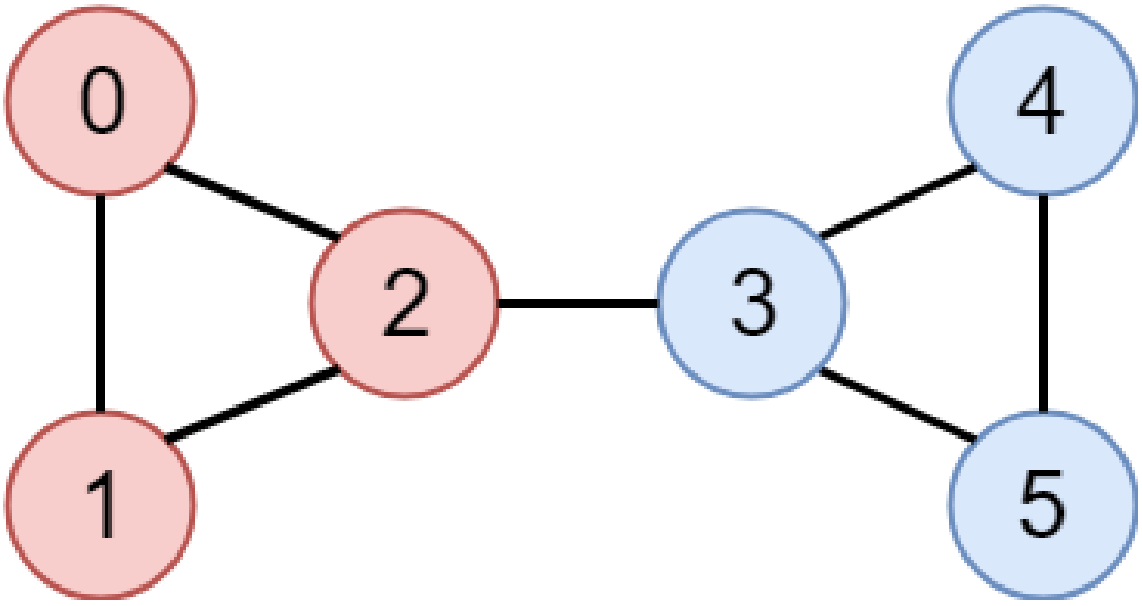

Figure 5.1: An example graph of two clusters, coloured red and blue, each containing three fully connected points, with one link connecting the two structures. The pairwise connection matrix w_{ij} represents the binary links between points, colour coded such that red-red links are red, blue-blue links are blue, and red-blue links are black (links also shown by connections on the graph). MCL recovers the correct clustering for a value of $\Gamma > 1.47$. For $\Gamma < 1.47$, MCL incorrectly connects all points in one large group, as does an FOF algorithm.

joins all points in one large cluster, as does an FOF approach. A value of Γ above 1.47 correctly separates the graph into two clusters as the larger value of inflation has cut the link between points 2 and 3. Larger values of inflation do not further split the clusters as they are fully connected subgraphs.

A further example is shown in Figure 5.2 that has two fully connected subgraphs each with five points rather than the three in figure 5.1. The minimum value of inflation necessary to correctly identify the two clusters has now fallen to 1.28. A lower value of inflation is needed to split larger clusters because jumps between clusters on a random walk are less frequent for large clusters than for smaller ones. A FOF approach still joins all points into a single cluster, despite the clusters being immediately obvious by eye.

FOF requires that the pairwise joining matrix w_{ij} contain binary links, i.e 0 or 1 for the connection amplitude, whereas MCL allows any value of connection amplitudes. Figure 5.3 shows an example graph that demonstrates the benefit of having more freedom in choosing the connection amplitudes. Two matrices are shown; the top one is a binary connection matrix that contains only 1s and 0s, while the bottom matrix is a possible probabilistic connection matrix, where larger values represent stronger connections.

Using the binary connection matrix, FOF once again fails to identify the clusters correctly. A value of inflation below 1.28 finds one cluster, above 1.28 but below 2.7 correctly identifies the two clusters, but the points incorrectly get placed into 4 clusters if $\Gamma > 2.7$. For the larger values of Γ, the 4 clusters found are points 0 and 1, points 4 and 5, and 2 and 3 are clusters by themselves. This shows the hierarchical nature of MCL when Γ is increased. To find more than two clusters, one might think that finding three clusters where 2 and 3 are connected is a sensible answer, but points 2 and 3 have already been disconnected for values of Γ large enough to find two clusters, and they cannot be reconnected if it is increased further. We can explain why the cut is made at the link between galaxies 2 and 3 first by considering a random walk on the graph that starts with equal weight at all points. Weight at the centre will spread out in both directions but weight at the ends will spread towards the centre in one direction only. So links near the ends of the chain are used

$$w_{ij} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

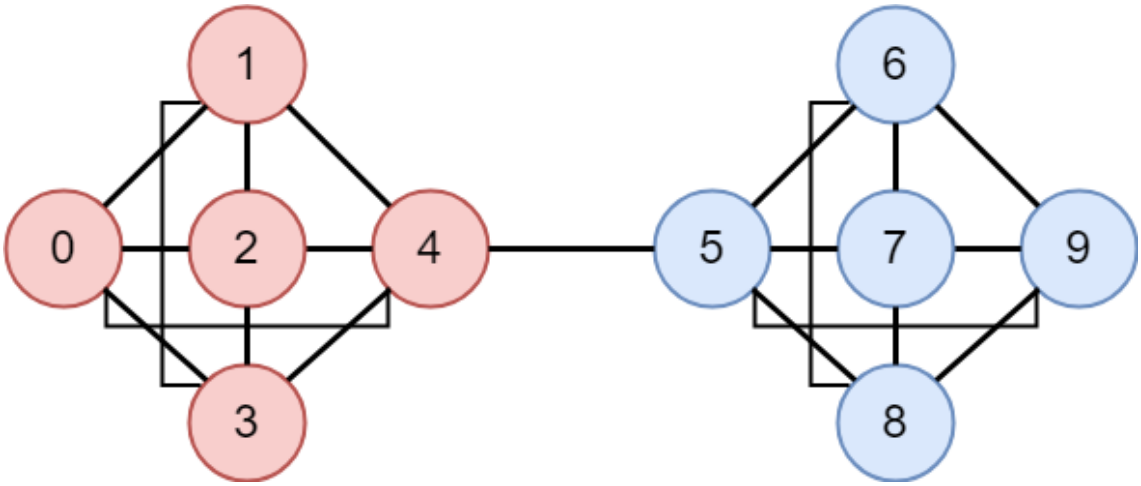

Figure 5.2: An example graph of two clusters, coloured red and blue, each containing five fully connected points, with one link connecting the two structures. The pairwise connection matrix w_{ij} represents the binary links between points colour coded as in Figure 5.1 (also shown by the lines on the graph). MCL recovers the correct clustering for a value of $\Gamma > 1.28$. For $\Gamma < 1.28$, MCL incorrectly connects all points in one large group, as does an FOF algorithm.

$$w_{ij} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$w_{ij} = \begin{bmatrix} 1 & 1 & 0.7 & 0.1 & 0 & 0 \\ 1 & 1 & 1 & 0.2 & 0 & 0 \\ 0.7 & 1 & 1 & 1 & 0.2 & 0.1 \\ 0.1 & 0.2 & 1 & 1 & 1 & 0.7 \\ 0 & 0 & 0.2 & 1 & 1 & 1 \\ 0 & 0 & 0.1 & 0.7 & 1 & 1 \end{bmatrix}$$

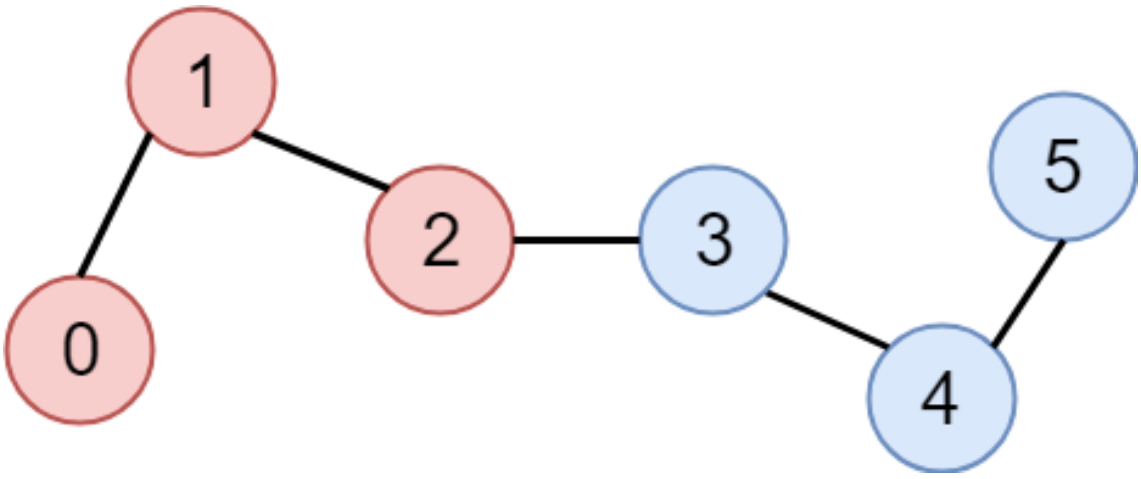

Figure 5.3: An example graph of two clusters, coloured red and blue, each containing three points. The top matrix shows a possible binary connection matrix for the graph, also shown with the links on the graph. The bottom matrix is a possible probabilistic connection matrix, with larger values representing stronger connections. The matrix colour coding is the same as in Figure 5.1 Using the binary matrix, MCL recovers the correct clustering for $1.28 < \Gamma < 2.7$, but using the probabilistic matrix increases the range of values for which the correct clustering is found to $1.69 < \Gamma < 8$. FOF again connects all the points in a single cluster.

more so than the ones near the centre after a few iterations. This means MCL will always try to split chains down the middle first. In this case this splits the clusters correctly, but if there were three clusters in a chain end on end the middle cluster would be split incorrectly.

Using the probabilistic connection matrix increases the range of inflation values for which the correct clustering is found to $1.69 < \Gamma < 8$. The minimum value of Γ has increased slightly compared to the binary case. This is because compared to the binary connection matrix, the probabilistic matrix has changed some zero values to values between 0 and 1, increasing the overall connectedness of the points, so a slightly larger inflation value is needed to split the points. Forcing the matrix to be binary results in a loss of information that the MCL algorithm in this case could use to determine the correct clustering.

5.3 Mock catalogue

To test this novel approach to galaxy group finding we apply it to a realistic mock galaxy catalogue. We use the $z = 0$ snapshot of a `GALFORM` mock catalogue, specifically the model presented in Gonzalez-Perez et al. (2017), built on top of the 125 h^{-1}Mpc per side MilliGas simulation cube. We note that this simulation has the same cosmology and number of snapshots as the $500h^{-1}$Mpc MR7 simulation (Guo et al., 2013). The MilliGas simulation also has runs with many more snapshots, so would be ideal to use for lightcone generation. Even if the smaller simulation is unlikely to contain any structures as large as the largest found in the MR7 simulation, we use a smaller N-body simulation to speed up the calculations, as deciding between methods of linking galaxies and optimisation of free parameters will require running the group finder many times. The scaling with volume would be linear, so running MCL on the larger simulation would take roughly 64 times longer per run than of the smaller one. Each run only takes a few minutes for the MilliGas mock catalogue, but performing grid searches requires hundreds of runs. The scaling with density is quadratic, as it scales with the number of pairwise connections on small scales. We could have chosen to do the tests on a lightcone mock catalogue, such

as the one presented in Chapter 4 for the PAU Survey, but this work can also be applied more generally than for one specific survey, with a lightcone catalogue introducing unnecessary complications such as changing number density[4]. We use the halos identified using the particles in the simulation as the 'truth' to which we will compare our galaxy group finders. There are multiple ways to identify halos from simulation particles but the disagreements between these methods will be negligible when compared to the errors we introduce when using galaxies as tracers, as the simulation particles are far more numerous than our galaxy catalogue.

Figure 5.4 shows a 25 h^{-1}Mpc thick slice of the mock catalogue at $z = 0$. The catalogue is limited in rest frame SDSS r band magnitude to $M_r - 5\log h < -20.0$ and contains ~ 20000 objects, corresponding to a galaxy density of 10^{-2} $(h^{-1}\text{Mpc})^{-3}$. The catalogue has comparable density to galaxies in the GAMA survey at $z \sim 0.15$. The catalogue is continued periodically on each side such that we do not have to deal with any edge effects, as the code used to produce the pairwise connection matrix does not support periodic boundaries as it is built from the `2PCF` code presented in Chapter 3. MCL takes a sparse matrix as input. A sparse matrix format is one which lists the value and location of the non-zero elements of the matrix. This saves a lot of memory in this example, as most of the pairwise connections in a galaxy catalogue will be zero. Extending the simulation periodically will mean all relevant pairs are included in this sparse matrix at least once, but will also produce some repeated pairwise connnections. The MCL algorithm provides a preprocessing step to deal with repeated entries in multiple ways. In our case taking the maximum value of repeated entries is the correct approach.

5.4 "Goodness of clustering" measures

We require a method of quantifying the quality of a galaxy grouping. To do this we need to quantify how similar a galaxy grouping, G, is to the underlying halos, H,

[4]One must also consider the limitations in the construction of the lightcone. The interpolation scheme is unlikely to perfectly place galaxies where they would have been found if a snapshot of the simulation had been taken at that point.

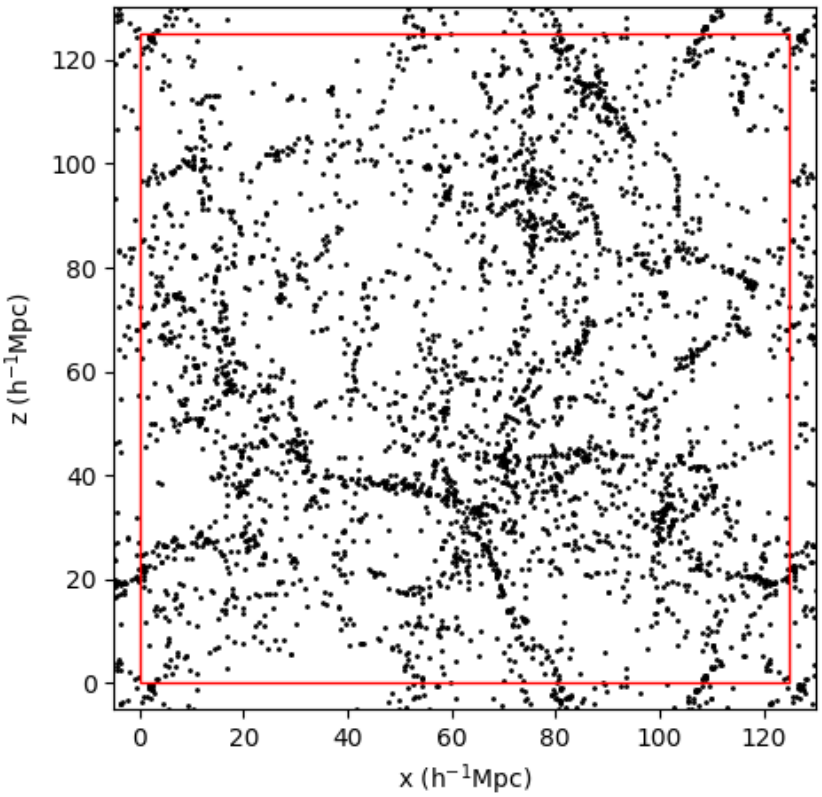

Figure 5.4: A 25 h^{-1}Mpc thick slice of the mock catalogue in real space. The catalogue is periodically extended by 5 h^{-1}Mpc on each side (much larger than any link between galaxies in real space). The red box encloses the original catalogue before periodic extension.

to which the galaxies belong. In the astrophysics literature, this is typically done through the measures of completeness and purity, although their definition varies drastically between works. Many metrics exist to decide the goodness of a clustering. See Wagner & Wagner (2007) for an overview of the most common metrics used in the mathematics and computer science communities.

5.4.1 Completeness and purity

We provide general expressions for the 'one way' matching completeness and purity. They are one way because by these definitions pure groups do not need to correspond to complete halos and vice versa. One could place constraints that a pure group can only be counted as pure if its corresponding group is complete, which would be a 'two way' or 'bijective' match. We use purity and completeness as tools to understand a particular grouping, and not as optimisation criteria, and therefore consider the more intuitive one way matching. See Gerke et al. (2005) and Knobel et al. (2009, 2012) for detailed discussion on one and two way matching.

The completeness quantifies the extent to which galaxies in the same halos are placed in the same galaxy groups. We define it using weight functions f and g as

$$C^*(f,g) = \frac{1}{\sum_{j=1}^{N_H} f(n_{\Sigma j}) g(n_{\Sigma j})} \sum_{j=1}^{N_H} f(n_{\Sigma j}) g(\max_i n_{ij}) \,. \tag{5.4.1}$$

The purity quantifies the extent to which galaxies in the same groups are actually in the same halo, defined as

$$P^*(f,g) = \frac{1}{\sum_{i=1}^{N_G} f(n_{i\Sigma}) g(n_{i\Sigma})} \sum_{i=1}^{N_G} f(n_{i\Sigma}) g(\max_j n_{ij}) \,. \tag{5.4.2}$$

$n_{\Sigma j}$ is the number of galaxies in halo H_j, $n_{i\Sigma}$ is the number of galaxies in group G_i and $n_{\Sigma\Sigma}$ is the total number of galaxies. These are calculated from the number of

	H_1	H_2	...	H_j	...	H_{N_H}	Σ
G_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1N_H}	$n_{1\Sigma}$
G_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2N_H}	$n_{2\Sigma}$
.	.	.	...	...	...	.	.
G_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iN_H}	$n_{i\Sigma}$
.	.	.	...	...	...	.	.
G_{N_G}	n_{N_G1}	n_{N_G2}	...	n_{N_Gj}	...	$n_{N_GN_H}$	$n_{N_G\Sigma}$
Σ	$n_{\Sigma1}$	$n_{\Sigma2}$	...	$n_{\Sigma j}$	...	$n_{\Sigma N_h}$	$n_{\Sigma\Sigma}$

Table 5.1: Contingency matrix for a clustering of $n_{\Sigma\Sigma}$ points into N_G groups, G_1 to G_{N_G}, that attempt to identify the N_H halos, H_1 to H_{N_H}. A perfect group catalogue would result in a diagonal matrix, representing perfect agreement between the groups and the halos. We can write galaxy group statistics as a function of all or parts of this contingency matrix.

galaxies in group G_i and halo H_j, n_{ij}, using

$$n_{\Sigma j} = \sum_{i=1}^{N_G} n_{ij} \tag{5.4.3}$$

$$n_{i\Sigma} = \sum_{j=1}^{N_H} n_{ij} \tag{5.4.4}$$

$$n_{\Sigma\Sigma} = \sum_{i=1}^{N_G}\sum_{j=1}^{N_H} n_{ij}\,, \tag{5.4.5}$$

where N_G and N_H are the number of groups and halos respectively. This information can be summarised neatly in the contingency matrix shown in table 5.1.This matrix can be written generally for any clustering problem and quantifies the overlap between N_H true halos and N_G identified galaxy groups. A perfect clustering should appear as a diagonal matrix.

The weight functions f and g can be used to change the weightings of groups or halos of different multiplicities, i.e different numbers of members. The function f can be changed so the statistic only looks at the quality of halos or groups of a particular multiplicity. The function g is a weighting function that can be used to change the penalty for incorrect groupings. For example, say we have chosen f so

as to only look at the completeness of halos of a specific multiplicity, the function g will decide how much the completeness will fall on missing a certain number of galaxies. Here we will always set $g(n) = n$, so the purity and completeness will fall linearly with the number of incorrectly placed galaxies. Another sensible choice might be a step function such that a halo is considered as being complete if more than half of its members are placed into the same group. This is the choice used in Eke et al. (2004).

We will look at cases where $f(n)$ is a step function that sets the minimum group or halo multiplicity considered:

$$f(n) = \begin{cases} 1, & \text{if } n \geq N \\ 0, & \text{otherwise}\,. \end{cases} \tag{5.4.6}$$

We define P(≥N) and C(≥N) as the purity $P^*(f, g)$ and completeness $C^*(f, g)$ in the case $g(n) = n$ and $f(n)$ is given by equation (5.4.6). Hence P(≥N) is the purity of groups whose multiplicity is at least N, and C(≥N) is the completeness of halos whose multiplicity is at least N. It is important to note with these measures that we talk about the purity of groups and the completeness of halos. We will only consider values of P(≥N) and C(≥N) for $N \geq 2$, as single groups are always pure and single halos are always complete, resulting in the statistics P(≥1) and C(≥1) being dominated by P(=1) = 1 and C(=1) = 1.

5.4.2 Optimisation metric

In order to optimise the parameters of a given method a single statistic is required to decide which clustering is most adequate. Unfortunately, there is no definite answer as to how to decide if one clustering is better than another. If solving a particular problem, one might want to design a bespoke statistic to find an optimal group finding model. For example, one of the science cases for the PAU Survey group catalogue is to identify Milky Way analogues for spectroscopic follow up. In this case, we may wish to maximise the probability that spectroscopically observed groups are actually Milky Way analogues. The issue with this is that it would require a new group catalogue to be built for each science case, which might not be

practical.

Here we would like a problem agnostic measure to build a generally 'optimal' group catalogue. Most astrophysical applications use combinations of bijective measures of completeness and purity to define such a quantity (Gerke et al., 2005; Robotham et al., 2011; Knobel et al., 2012; Jian et al., 2014b). We follow to the work of Wu et al. (2009) who tested multiple goodness of fit metrics and choose to use the variation of information (Meilă, 2003).

The variation of information, also called the shared information distance, quantifies the distance between two clusterings by looking at the amount of information in each clustering that cannot be inferred by the other clustering. Figure 5.5 visualises this for halos with entropy E(H), and a detected group catalgoue with entropy E(G)[5]. These are calculated as

$$E(H) = -\sum_{j=1}^{N_H} p_{\Sigma j} \ln(p_{\Sigma j}) \tag{5.4.7}$$

$$E(G) = -\sum_{i=1}^{N_G} p_{i\Sigma} \ln(p_{i\Sigma}) \,, \tag{5.4.8}$$

where $p_{xy} = n_{xy}/n_{\Sigma\Sigma}$ for any x or y, with the values of the n_{xy} taken from the contingency matrix shown in Table 5.1. We can see the similarity with the standard definition of entropy from statistical physics which is proportional to $\sum_i p_i \ln p_i$. The overlap of the two clusterings, the 'mutual information', is given by I(H, G), which is calculated as

$$I(H,G) = \sum_{i=1}^{N_G} \sum_{j=1}^{N_H} p_{ij} \ln \left(\frac{p_{ij}}{p_{i\Sigma} p_{\Sigma j}} \right) . \tag{5.4.9}$$

The variation of information is given by the information contained in the parts of the Venn diagram that do not overlap. This can be calculated as

$$\begin{aligned} VI(H,G) &= \big(E(H) - I(H,G)\big) + \big(E(G) - I(H,G)\big) \\ &= E(H) + E(G) - 2I(H,G) \,. \end{aligned} \tag{5.4.10}$$

[5]We note the entropy is typically given by H in the graph clustering literature, but we use E so as to avoid confusion with our halo catalogue.

For perfect overlap, the mutual information and the entropy of both clusterings will be equal, and the variation of information will fall to zero. The VI is therefore given by

$$\begin{aligned} VI(H,G) = &- \sum_{j=1}^{N_H} p_{\Sigma j} \ln(p_{\Sigma j}) - \sum_{i=1}^{N_G} p_{i\Sigma} \ln(p_{i\Sigma}) \\ &- 2 \sum_{i=1}^{N_G} \sum_{j=1}^{N_H} p_{ij} \ln \left(\frac{p_{ij}}{p_{i\Sigma} p_{\Sigma j}} \right). \end{aligned} \tag{5.4.11}$$

Figure 5.6 shows the variation of information and three values of P(≥N) and C(≥N) as a function of the linking length for a FOF group algorithm run on the mock catalogue. This is a simple FOF approach that uses the same linking length for all pairs of galaxies. We can see that the minimum value of VI produces a catalogue that is well balanced between completeness and purity. The minimum value of VI also agrees with the value of the linking length relative to the mean galaxy separation found in Eke et al. (2004). This shows that the mock catalogue and choice of minimisation statistic are sensible, and produce results comparable to those found in previous work.

5.5 Testing the Markov Clustering method

This section will test the Markov clustering algorithm on the mock catalogue, using multiple methods of assigning pairwise connection amplitudes, to see if it provides improvements over a FOF approach. In choosing the best model, we must also consider the insensitivity of the model to changes in the free parameters. If we add complexity to the model and find a new minimum in VI, but in doing so significantly narrow the parameter space for which we find reasonable results, we should not necessarily consider the model as an improvement. The mock catalogue on which the model is tuned is unlikely to be perfectly representative of the real Universe, so finding a robust model is as important as finding the minimum value of VI. All measurements made in this section use the catalogue described in section 5.3.

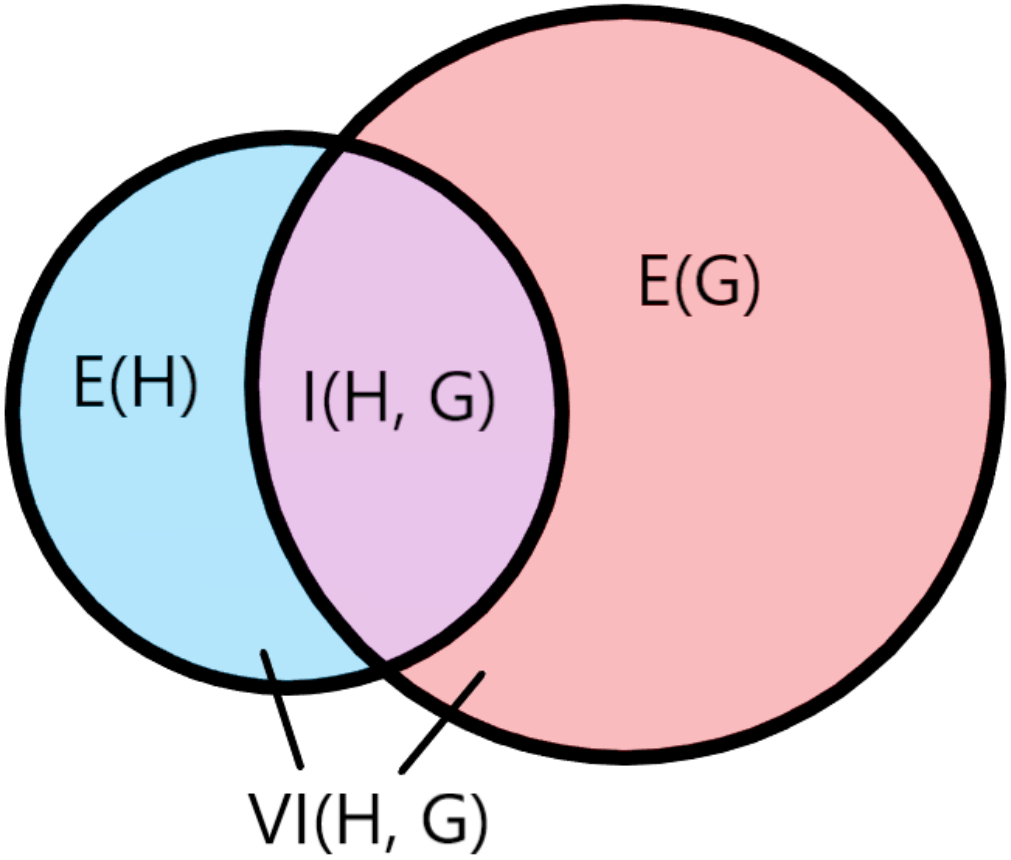

Figure 5.5: Schematic explaining the meaning of the variation of information measure. The entropy of the halos E(H) and of the groups E(G) represent the total information in each full circle. The VI is given by the total area, excluding the overlap region of mutual information I(H, G). The more the circles overlap, the larger the mutual information will be, and the smaller the value of VI will be. This plot was adapted from a similar plot in Meilă (2003).

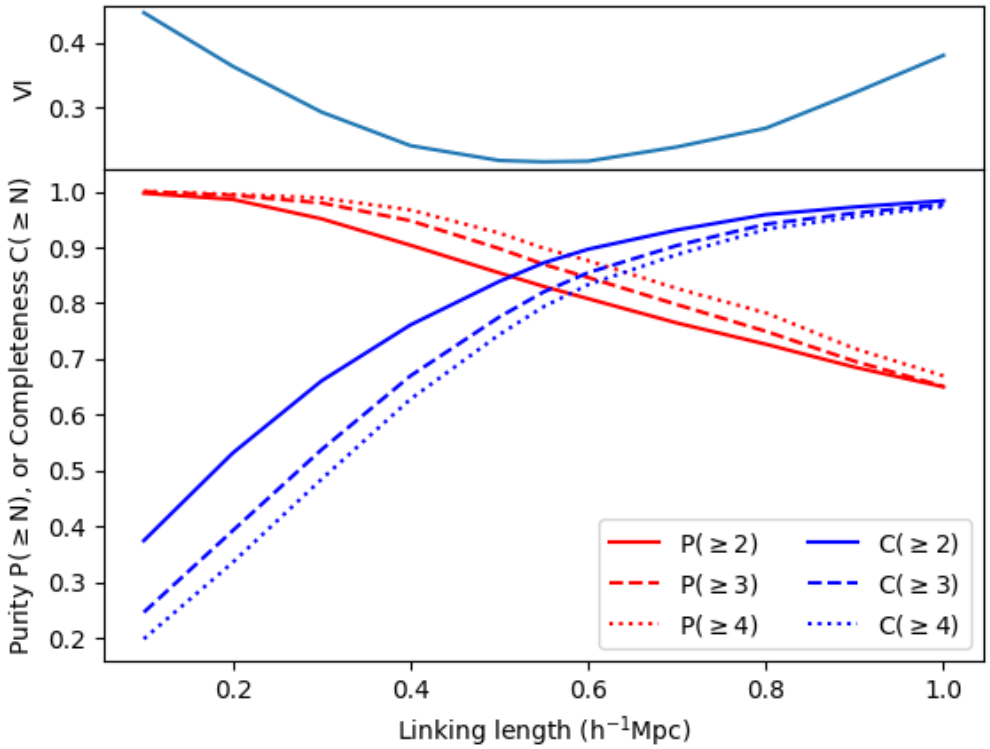

Figure 5.6: Variation of information (top panel), purity and completeness (bottom panel) as a function of linking length in a pure FOF approach to galaxy group finding for three values of minimum group or halo multiplicity, N. The best fitting value of linking length relative to the mean galaxy separation found in Eke et al. (2004) ($b \sim 0.13$) corresponds to a linking length of 0.6 h^{-1}Mpc in this catalogue, which roughly agrees with our minimum variation of information value.

5.5.1 Constant linking length

The simplest method of creating pairwise connection amplitudes is to create binary connections between galaxies that are 1 if the separation is less than a constant linking length L and 0 if not. That is

$$w_{ij} = \begin{cases} 1, & \text{if } r_{ij} \leq L \\ 0, & \text{otherwise}\,, \end{cases} \tag{5.5.12}$$

for the pairwise separation, r_{ij}, between galaxies i and j.

The linking length L and the value of inflation Γ are varied. When the inflation value is set to 1, MCL acts like an FOF algorithm, but also takes many more iterations of the random walk to converge, so the lowest value of inflation used here is 1.01 to avoid such CPU intensive runs. A value of 1.01 produces clusterings very similar to, and in most cases identical to, those found in a standard FOF run.

Figure 5.7 shows the values of VI as a function of linking length and inflation in the case of w_{ij} as specified by Equation (5.5.12). It is clear that the best value of VI is found as Γ tends towards 1, the value where the MCL algorithm mimics an FOF approach. Figure 5.8 visualises why this is the case. It shows the cumulative multiplicity function T($\geq$N) for the true halos, and for three group catalogues using the same linking length but different values of inflation. T($\geq$N) gives the number of halos of multiplicity greater than or equal to N. In the FOF approach with this simple joining scheme there are too many low multiplicity groups, many of which will be spurious, and too few large groups. The low multiplicity groups are not very pure, and the large groups are not very complete, which was seen in Figure 5.6. Increasing the value of the inflation parameter only makes this problem worse, as it is less likely to split small tightly bound structures and more likely to further split loosely connected large groups. So when inflation is increased there are even more low multiplicity groups and even fewer high multiplicity groups compared to the FOF case.

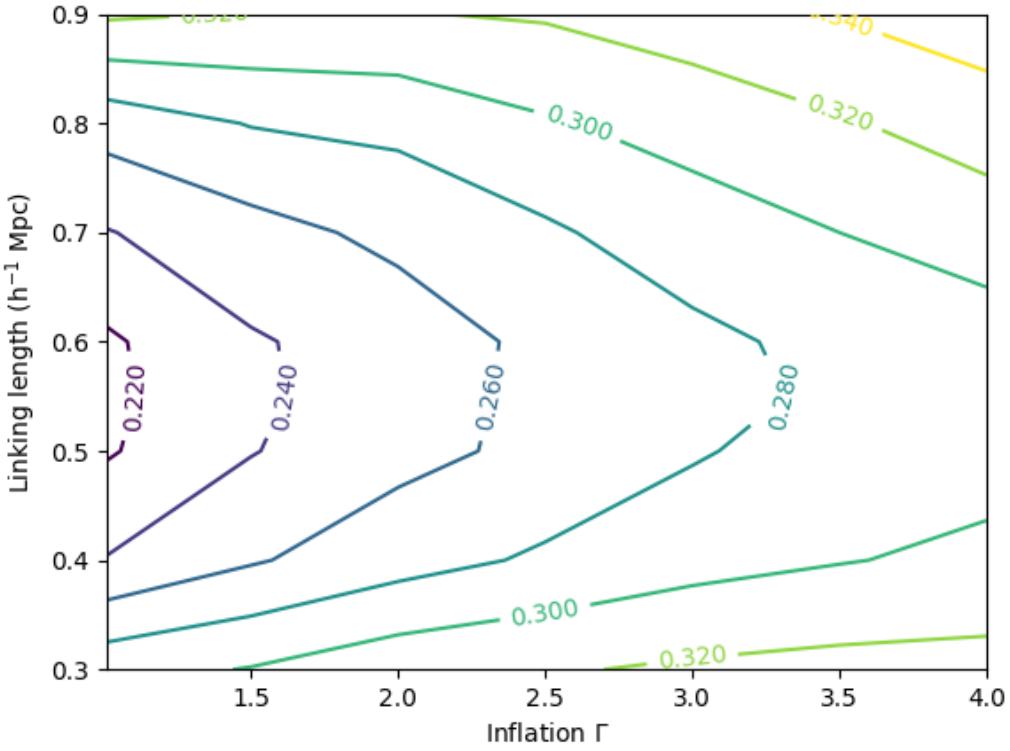

Figure 5.7: VI contours as a function of the linking length L and the value of inflation Γ for w_{ij} as given by Equation (5.5.12). The minimum value of VI is found as the inflation value approaches 1, which is the value where MCL acts like a FOF algorithm.

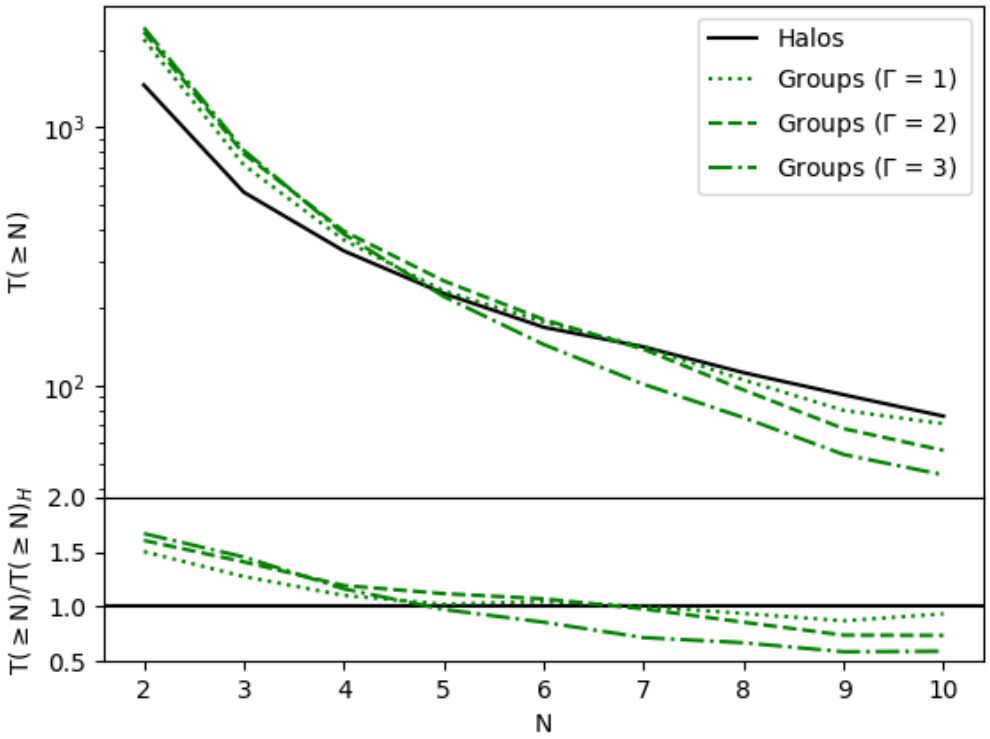

Figure 5.8: The cumulative multiplicity function, T(≥N), for halos, and for three groups catalogues with different values of inflation as labeled. The top panel shows the multiplicity function and the bottom panel normalises the results by the multiplicity function of the underlying halos $T(\geq N)_H$. All group catalogues use the optimal linking length for a pure FOF approach. The FOF approach produces too many small halos and too few large halos. Increasing the inflation parameter makes this problem worse.

5.5.2 Local density enhancement

In order to reduce the variation in purity and completeness with group multiplicity we propose to vary the linking length based on the local density. As larger groups have poor completeness when using a constant linking length we would like the linking length to be larger in overdense regions and smaller in underdense regions. There is a precedent for this in the literature, with both Eke et al. (2004) and Robotham et al. (2011) modifying the linking length based on the local density and finding that it improves the group identification.

We calculate the local galaxy density in real space, ρ_i, at the position of galaxy i using a 3D Gaussian kernel with $\sigma = 1h^{-1}$Mpc truncated at 4σ. Other reasonable values of the smoothing scale were tested and no significant improvement over this value was found. The connection scheme becomes,

$$w_{ij} = \begin{cases} 1, & \text{if } r_{ij} \leq L_{ij} \\ 0, & \text{otherwise}\,, \end{cases} \tag{5.5.13}$$

with the linking length L_0 modified by the geometric mean of the local densities of the two galaxies, giving

$$L_{ij} = L_0 \left(\frac{\sqrt{\rho_i \rho_j}}{\langle \rho \rangle \, (r_{ij})} \right)^{\beta}. \tag{5.5.14}$$

L_0 and β are free parameters and $\langle \rho \rangle \, (r)$ is the mean value of the geometric mean of the pairwise local densities at a separation r, calculated as

$$\langle \rho \rangle \, (r) \equiv \frac{\sum_i \sum_j \Delta(|r_{ij}|) \sqrt{\rho_i \rho_j}}{\sum_i \sum_j \Delta(|r_{ij}|)}\,, \tag{5.5.15}$$

where the sums are over all galaxies and Δ is a function that is unity if the value of r_{ij} lies in the same bin as the value of r. In doing this, we are extending the linking length if the pair lies in an overdense region relative to other pairs of similar separation. This is important because for a pair of galaxies with small separation the product of the galaxy local densities will on average be larger than for galaxies of larger separations. This is because the local density of each galaxy will include some contribution from the other galaxy in the pair, and because of the clustered nature of galaxies. The simple scheme used in section 5.5.1 is recovered when $\beta = 0$. Eke et al.

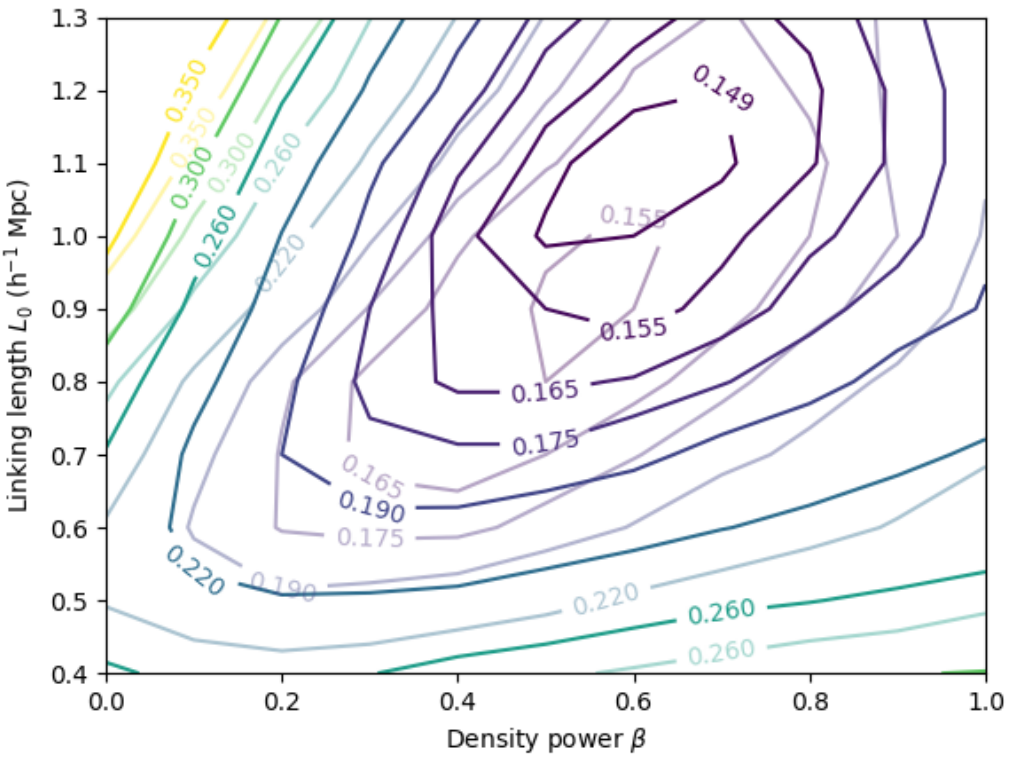

Figure 5.9: VI contours as a function of linking length L_0 and density power β for the case of $\Gamma = 1.6$ (Solid lines) and $\Gamma = 1.01$ (Translucent lines). Both cases show that density enhancement can be used to produce a catalogue with a lower value of VI than one built using a simple connection scheme, which was ~ 0.21. The minimum value of VI of ~ 0.144 lies at $L_0 = 1.1\ h^{-1}$Mpc, $\beta = 0.6$ and $\Gamma = 1.6$.

(2004) and Robotham et al. (2011) normalise the density using a free parameter, but the present method provides a means to measure the local density normalisation from the data. It is important to try to reduce the number of free parameters in case the mock does not perfectly represent the real Universe. If extending this to real observations, the density at each galaxy would have to first be normalised by the density of a random catalogue at the same point measured using the same kernel to account for the selection function. We also tested measuring the local density at the centre of each pair of galaxies but it provided no measurable improvement in results and was computationally far more challenging than this scheme.

Figure 5.9 shows VI contours as a function of linking length L_0 and density power

β for the case of an inflation value Γ of close to unity, the FOF case (translucent lines), and for an inflation value of 1.6 (solid lines). For the FOF with density enhancement a new minimum of VI is found for a value of β of ~ 0.6. This tells us that even in the FOF case, the density enhancement has helped to improve the galaxy group catalogue. The best fit value of the linking length with this value of β has increased from $\sim 0.55\ h^{-1}$Mpc to $\sim 0.9\ h^{-1}$Mpc. When inflation is allowed to vary, the minimum value is found when $\Gamma \sim 1.6$. Therefore, when the linking length is allowed to vary with the local density, the FOF algorithm is no longer the optimal choice. The best fitting value of L_0 in this case is larger than in the FOF case, which shows that the Markov algorithm is working as expected, in allowing more galaxy pairs to be connected before poorly connected structure is split apart. The optimal linking length further increases for values of Γ above 1.6 but the minimum value of VI also increases. The larger the value of Γ, the more galaxy pairs must be connected to compensate for inflation splitting apart more structure.

Figure 5.10 shows why the density enhancement helps improve group finding. It shows the fraction of pairs at a particular separation and pairwise density that lie within the same halo in the catalogue. We can recast the linking criterion so we can include it on this plot. Pairs of galaxies are connected if their separation r_{ij} meets the condition

$$\log_{10}(r_{ij}) < \log_{10}(L_0) + \beta \log_{10}(\sqrt{\rho_i \rho_j}/\left\langle \rho \right\rangle (r_{ij}))\,. \tag{5.5.16}$$

The cuts that the best fit simple FOF, FOF with density enhancement and MCL with density enhancement make are shown by the black lines. Pairs to the left of each of the lines are connected by those schemes. It is clear by eye that both algorithms that vary the linking length with density better split the pairs that do or do not lie in the same halo compared with the simple FOF approach. The roughly straight line split on this diagram in fact helped motivate the power law form of the density enhancement. This plot also shows why L_0 is larger with density enhancement, because it allows the scheme to avoid incorrectly connecting underdense pairs, which can be seen to happen with a fixed linking length. The MCL best fit catalogue connects more galaxy pairs than its FOF counterpart because it then splits poorly connected structure using the inflation parameter.

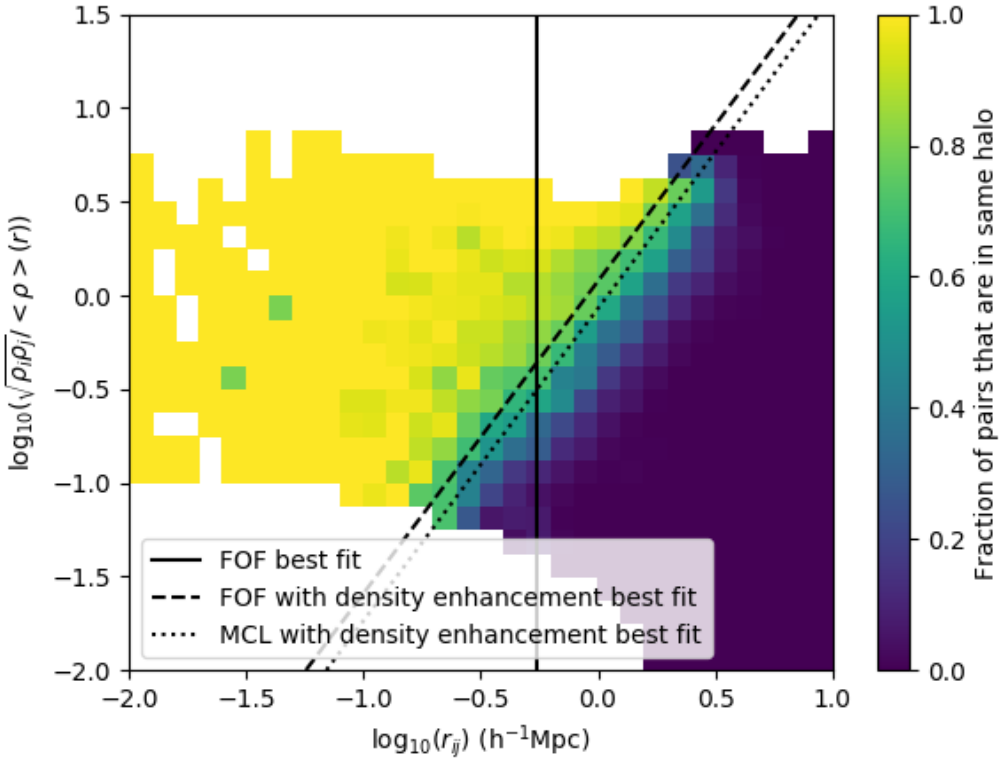

Figure 5.10: Fraction of pairs that lie in the same halo as a function of pair separation and normalised pairwise density. The lines show the cuts made by the best fit schemes of simple FOF (solid), FOF with density enhancement (dashed) and MCL with density enhancement (dotted). The region to the left of each line would be connected in each scheme. The simple FOF scheme can only yield a vertical cut, which is likely to poorly separate pairs that do and do not lie in the same halo. The cuts with density enhanced linking lengths make more sensible cuts.

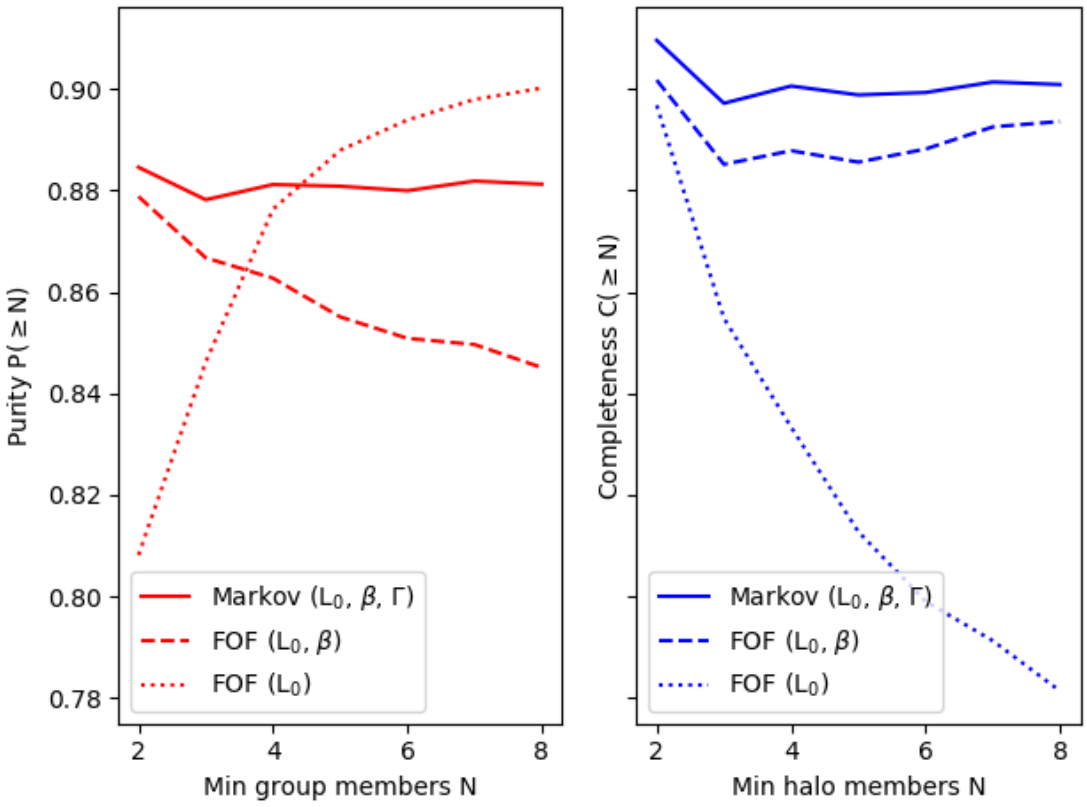

Figure 5.11: Purity, P(≥N) (left panel), and completeness, C(≥N) (right panel), as a function of minimum multiplicity, N, for the catalogue defined by the best fit simple FOF (dotted), FOF with density enhancement (dashed) and MCL with density enhancement group catalogues (solid). The purity and completeness of the MCL group catalogue is the most consistent as a function of multiplicity, and clearly has the best completeness.

Figure 5.11 shows the values of P(≥N) and C(≥N) as a function of N for the catalogues defined by the best fit simple FOF, FOF with density enhancement and MCL with density enhancement. As mentioned before, the simple FOF case has low purity for small groups (P(≥2) ∼ 0.8) and poor completeness for large groups (C(≥2) ∼ 0.78). The FOF with density enhancement is significantly better, but still overjoins some of the larger groups, as the purity falls with multiplicity. The MCL algorithm improves on both aspects, and produces a purity and completeness that is largely independent of the multiplicity. A catalogue with purity and completeness that vary little with multiplicity is preferred over one in which they vary a lot. It also produces a catalogue that has better purity and completeness for all multiplicities tested here than the FOF catalogue with density enhancement. The purity of high multiplicity groups is larger for the simple FOF case, but only because the completeness is so poor.

Figure 5.12 shows cumulative multiplicity function, T(≥N), for the halos and three group catalogues. The three group catalogues are the best fit simple FOF, FOF with density enhancement and MCL with density enhancement. It can be seen that using density enhancement to improve the FOF algorithm significantly improves the estimation of the number of small groups, but it still underestimates the number of large groups. The MCL algorithm with density enhancement impressively recovers the correct numbers of groups at all multiplicities tested here to better than 7%, and most to better than 3%, compared to the best FOF algorithm which underestimates the number of halos by as much as 25% at $N >= 5$ and ∼15% for most multiplicities. It is worth once again noting that these results were not used to tune the catalogues, which were solely tuned using VI.

These results show that the MCL algorithm can address the problem of bridges connecting large structure in an FOF algorithm. An FOF approach must be more cautious about the connection criterion as there is a large penalty if even one link is found between two large structures, whereas the MCL algorithm reduces this penalty by using inflation to break these bridges. These links cause the underestimation in the number of high multiplicity groups found by the FOF algorithm that can be seen in Figure 5.12, and the corresponding poor purity values for these groups seen

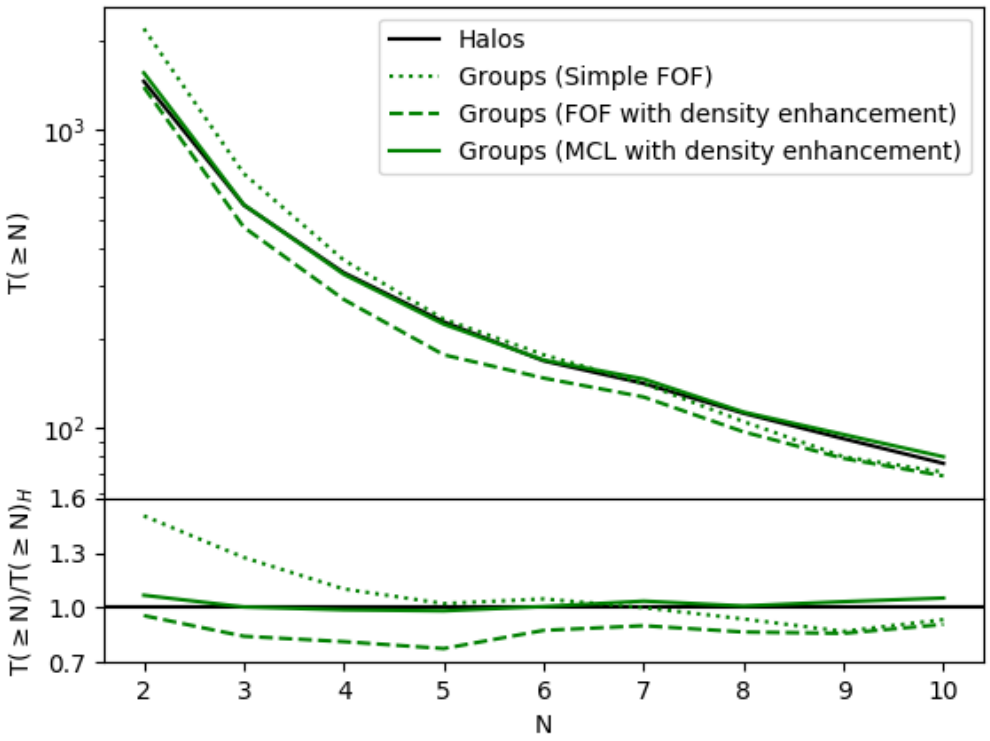

Figure 5.12: The cumulative multiplicity function, T(≥N), as a function of the multiplicity, N, for halos and three groups catalogues as labeled. The top panel shows the multiplicity function and the bottom panel shows the ratio to the true halo measurement $T(\geq N)_H$. The three group catalogues (green lines) are the best fitting simple FOF (dotted), FOF with density enhancement (dashed) and MCL with density enhancement (solid).

in Figure 5.11. Both of these are improved when using the MCL algorithm.

5.5.3 Fractional connection amplitudes

This section briefly mentions tests done with a scheme that produces fractional pairwise connection amplitudes. This is done to test if providing MCL with more information about the actual pairwise separations can provide any improvement. This connection scheme sets values of the connection matrix as,

$$w_{ij} = \left(1 + \left(\frac{r_{ij}}{L_{ij}}\right)^{\alpha}\right)^{-1}, \tag{5.5.17}$$

for a free parameter α and a linking length modified by the local density as in Equation (5.5.14). This form of w_{ij} tends to 1 for close pairs and to 0 for distant pairs. The larger the value of α, the sharper this transition becomes around the linking length L_{ij}. This tends towards a binary connection scheme in the limit of α tending towards infinity. Tests found no better values of VI, and the best fitting values were for values of α such that this scheme was indistinguishable from a binary connection scheme. So for the real space scenarios considered so far, the MCL algorithm works best if the matrix w_{ij} contains binary links.

5.6 Extension to redshift space and photometric redshifts

This section proposes a method of extending this work to catalogues with mixed redshift precision.

5.6.1 Model

The pFOF scheme laid out in Liu et al. (2008) uses the full redshift probability distribution of the galaxies to extend the FOF model to catalogues including photometric redshifts. We will use this scheme as a starting point. The pFOF scheme connects two galaxies if the probability that they lie within the linking length of each other is greater than a free parameter, the threshold probability p_{thresh}. Galaxies i

and j with redshift probability distributions $p_i(z)$ and $p_j(z)$ are connected if

$$\int \mathrm{d}z_1 \int \mathrm{d}z_2 \; p_i(z_1) p_j(z_2) w_{ij}(\underline{r}_1, \underline{r}_2) > p_{\mathrm{thresh}} \,, \tag{5.6.18}$$

where $w_{ij}(\underline{r}_1, \underline{r}_2)$ (hereafter w_{ij}), is the pairwise connection condition, which is a function of the relative positions of the two galaxies, and will change as the two integration variables change. This could be a binary condition like the one given by Equation (5.5.13) or a smooth function as given in Equation (5.5.17). We have seen examples in real space where w_{ij} varied with the total separation of the two galaxies, but in redshift space the projected separation is usually treated differently to the line of sight separation. The linking length along the line of sight is made to be longer than the projected linking length such as to recover structure that has been spread out by redshift space distortions. This scheme has the attractive property that in the limit where all redshift measurements are exact, i.e. all redshift probability distributions are given by delta functions, this scheme mimics a typical FOF approach to group finding.

There is nothing to stop the binary links defined in this scheme being used with the MCL algorithm. As we have shown in real space, even with binary connection matrices the MCL algorithm can provide a significant improvement over a FOF approach. However, as MCL allows probabilistic connections we can define the connection matrix as simply

$$\langle w_{ij} \rangle = \int \mathrm{d}z_1 \int \mathrm{d}z_2 p_i(z_1) p_j(z_2) w_{ij} \,. \tag{5.6.19}$$

This has removed the free parameter p_{thresh} from the pFOF scheme, but MCL adds a free parameter in inflation (Γ) as part of the MCL pipeline.

In this scheme, we still require the linking length along the line of sight to be longer than in projection. We extend the linking length to deal with smearing of structure, and only need to do so because of the clustered nature of galaxies. We propose here an extension to the model that uses the clustering signal of galaxies as a means to recover the lost signal due to redshift space distortions and redshift space uncertainties. The two point correlation function is defined as the excess probability of finding two galaxies at a particular separation from each other, so we define the

average connection matrix between a pair of galaxies i and j as

$$\langle w_{ij} \rangle = \frac{\int \mathrm{d}z_1 \int \mathrm{d}z_2 p_i(z_1) p_j(z_2)(1 + \xi_{ij}(r_{12}))w_{ij}}{\int \mathrm{d}z_1 \int \mathrm{d}z_2 p_i(z_1) p_j(z_2)(1 + \xi_{ij}(r_{12}))} , \tag{5.6.20}$$

where ξ_{ij} is the auto/cross correlation function between galaxies i and j[6]. In practice this could be the real or redshift space correlation function, that choice is discussed later in section 5.6.3. If redshift measurements are exact and in real space, this scheme recovers the deterministic value of w_{ij}, i.e. $\langle w_{ij} \rangle = w_{ij}$.

We explore and justify this scheme with a thought experiment. Suppose we know the position of one galaxy exactly in real space. Then, suppose that another galaxy is close by in projection, but its redshift probability distribution is binary, such that it is 50% likely to lie next to the first galaxy and 50% likely to be very far away. For any reasonable line of sight linking length and a binary form of w_{ij}, the pFOF scheme of Equation (5.6.18) will always give a probability of the two galaxies being connected as 0.5, because according to that scheme there is a 50% chance that the galaxies are close enough to be connected. However, intuition tells us that this number should be higher. It is more likely that a galaxy in a group has had 50% of its redshift probability placed at a random point in the field, rather than a randomly placed field galaxy having had 50% of its redshift probability measured to be directly next to another galaxy. The clustering weighting scheme of equation (5.6.20) will find a probability of connection of greater than 0.5 if the two galaxies come from a clustered galaxy sample as the nearby separation will have its weight increased by a factor of $(1 + \xi_{ij})$. If value of ξ_{ij} was ~ 100 when at the nearby separation and near 0 at the distant separation the probability of connection would in fact be ~ 0.99.

We can extend this thought experiment further by considering two scenarios, one where both galaxies are red, and one where both galaxies are blue. Using the cross correlation function between different samples allows the two scenarios to have two different connection probabilities. The connection probability in the red case will be larger. The amplitude of the correlation function on one halo scales for

[6]Typically the sample is split into one or several samples, and ξ_{ij} is the auto/cross correlation function of the samples containing galaxies i and j respectively.

red galaxies is higher than for blue galaxies, as red galaxies are more likely to be satellite galaxies than blue galaxies. This scheme therefore treats the recovery of the connection amplitudes differently for galaxies with different properties in a manner that is data driven. One way to think of this is that we are replacing lost line of sight information with the average information about that type of pair of galaxies.

5.6.2 Testing with a toy model

We will look at what probabilities equation (5.6.20) assigns to galaxy pairs of different line of sight and projected separations for different values of positional uncertainty and correlation function amplitude. We will look at the situation where both galaxies have the same Gaussian position uncertainties, σ, with different peak positions, and both belong to a sample with a power law two-point correlation function of the form

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma}, \tag{5.6.21}$$

for separation r and free parameters γ and r_0. This setup is visualised in Figure 5.13. The larger the uncertainty σ, or the smaller the peak separation π, the greater the overlap between the two distributions along the line of sight. Setting $r_0 = 0$ corresponds to the case of zero correlation function.

Figure 5.14 shows the probability of connection as a function of the line of sight separation of the peaks of the two Gaussian distributions, π, and their width, σ, for the case of a zero correlation function. The value of the projected separation is fixed at 0.4 h^{-1}Mpc. At low values of σ, the scheme approaches the deterministic cut shown by the black dashed line. For larger values of σ, the transition becomes broader and the maximum possible probability (at small values of π) falls. The top right hand part of the contours extend beyond the black dashed line, which shows that this scheme can give a chance of connection even when the median separation of the galaxies along the line of sight is greater than $\sqrt{L^2 - r_p^2}$, although in this scheme those probabilities are small. If we look across at a value of $\sigma = 10h^{-1}$Mpc (roughly the error on two well measured PAUS galaxies at low redshift) we can see the problem if no correlation function is included, as the galaxies would have a very

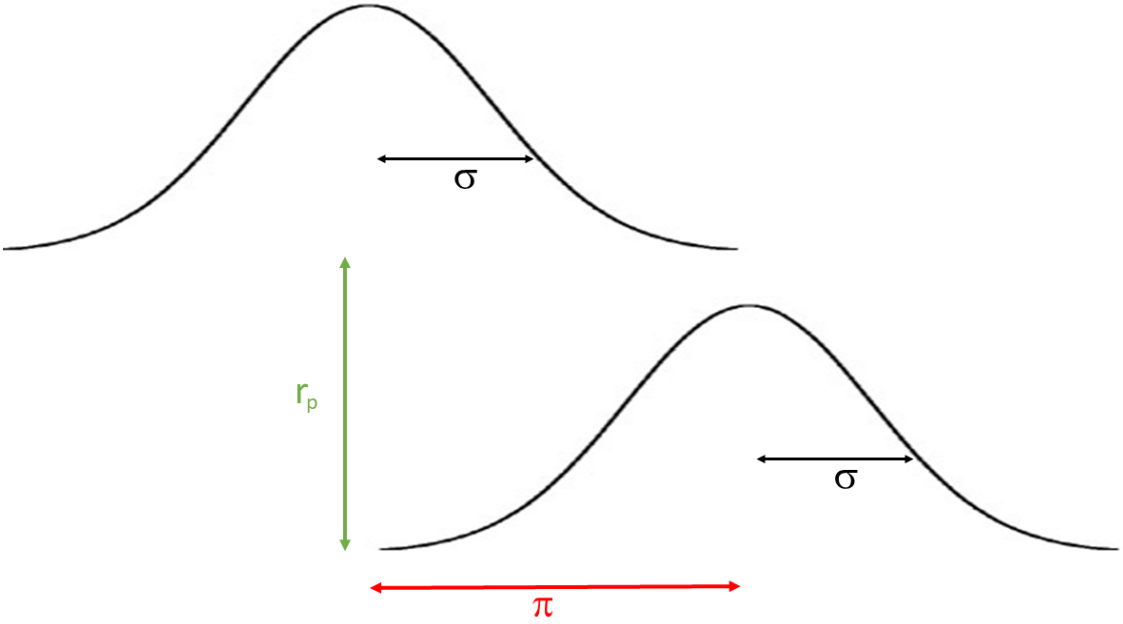

Figure 5.13: Schematic drawing showing the toy model of a pair of galaxies, both with line of sight position pdfs given by Gaussian distributions of width σ. The galaxies are separated by r_p in projection and the peak of their line of sight distributions are separated by π. In the distant observer approximation r_p is unchanged by changes in redshift of the two galaxies.

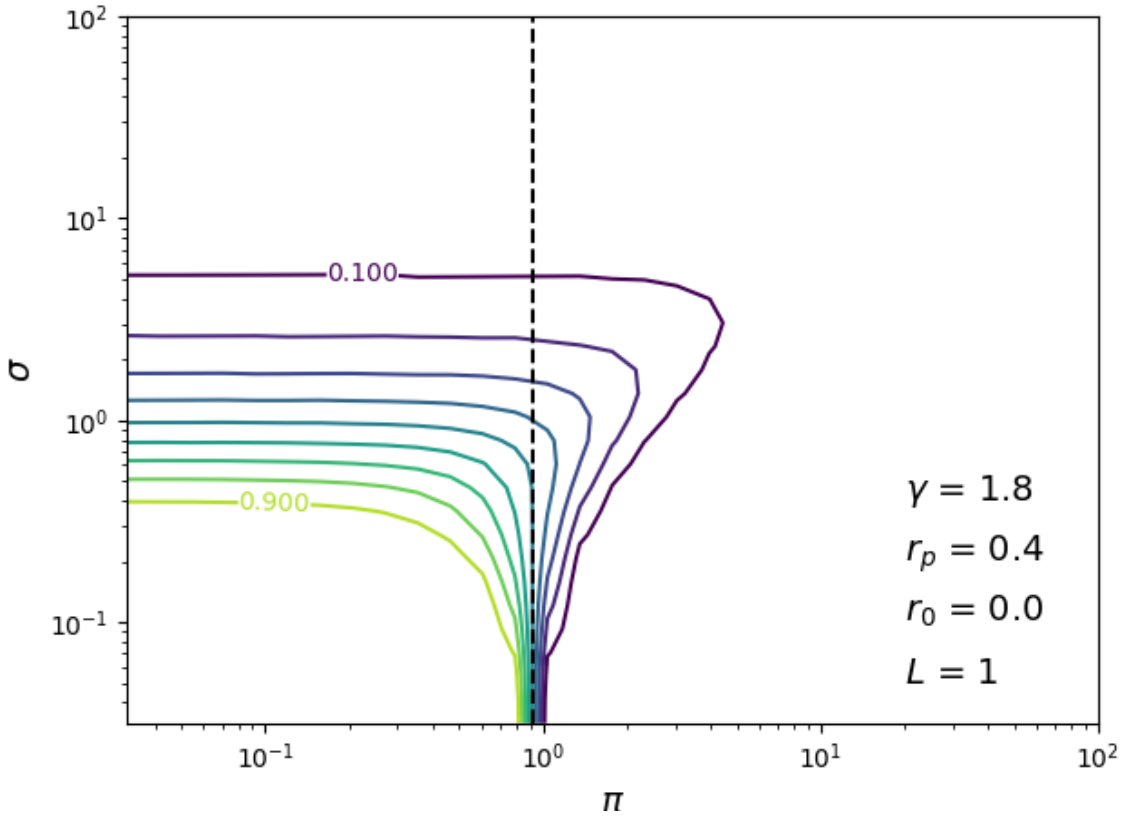

Figure 5.14: Contours of average connection probability given by Equation (5.6.20) for a pair of galaxies in the setup as shown in Figure 5.13 as a function of Gaussian width σ and line of sight peak separation π for the case of no correlation function. The contours have values every 0.1, with the innermost and outermost contours are labeled. The black dashed line shows the maximum line of sight separation that would result in a connection if $\sigma = 0$, given by $\sqrt{L^2 - r_p^2}$. All length scales are in units of h^{-1}Mpc.

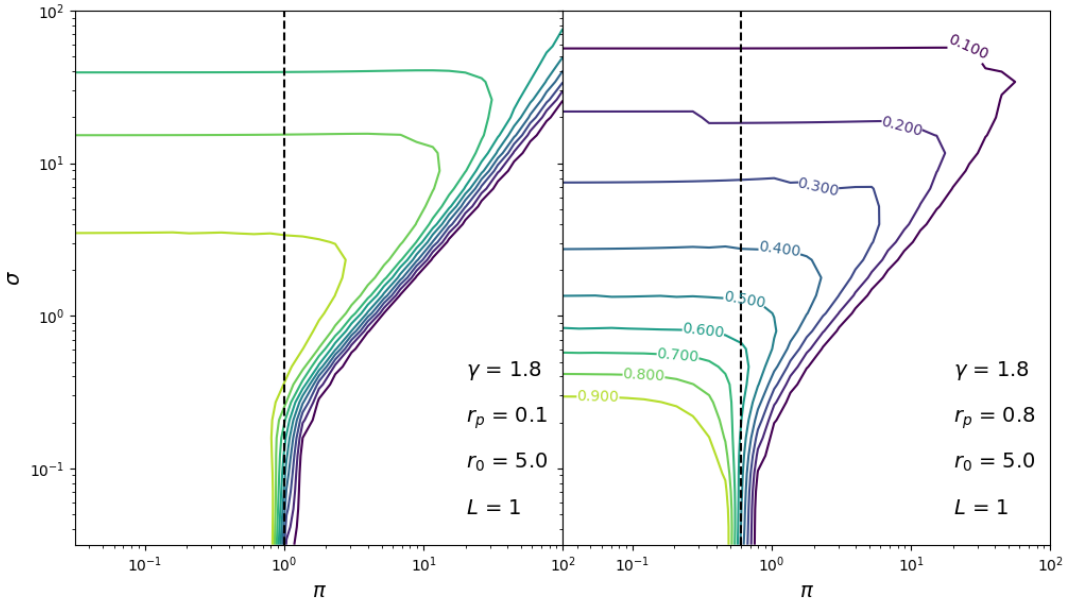

Figure 5.15: Same plot as in Figure 5.14 but for two values of r_p (left and right panels) and a power law correlation function. The contours are labeled in the right panel and have the same values in both panels. All length scales are in units of h^{-1}Mpc.

low probability of connection, even in the case where their median positions are exactly the same along the line of sight.

Figure 5.15 shows the impact of including the correlation function weighting, using a power law correlation function with $r_0 = 5h^{-1}$Mpc and $\gamma = 1.8$. We show this for two values of r_p, 0.1 h^{-1}Mpc (left panel) and 0.8 h^{-1}Mpc (right panel). When the value of σ tends towards zero, we once again tend towards a deterministic connection scheme. Now, compared with not including correlation function weighting as in Figure 5.14, the fall in connection probability at small values of π as σ increases is significantly slower. This fall happens faster in the case of larger projected separation compared with smaller projected separations, as the galaxies need to be closer to each other along the line of sight to be connected when their projected separations are larger. We can now see that pairs with large uncertainties and large line of sight separations that would be disconnected without correlation function weighting can now have a significant connection amplitude, which is larger for smaller projected separations. Now PAUS like galaxies may have significant chance of connection.

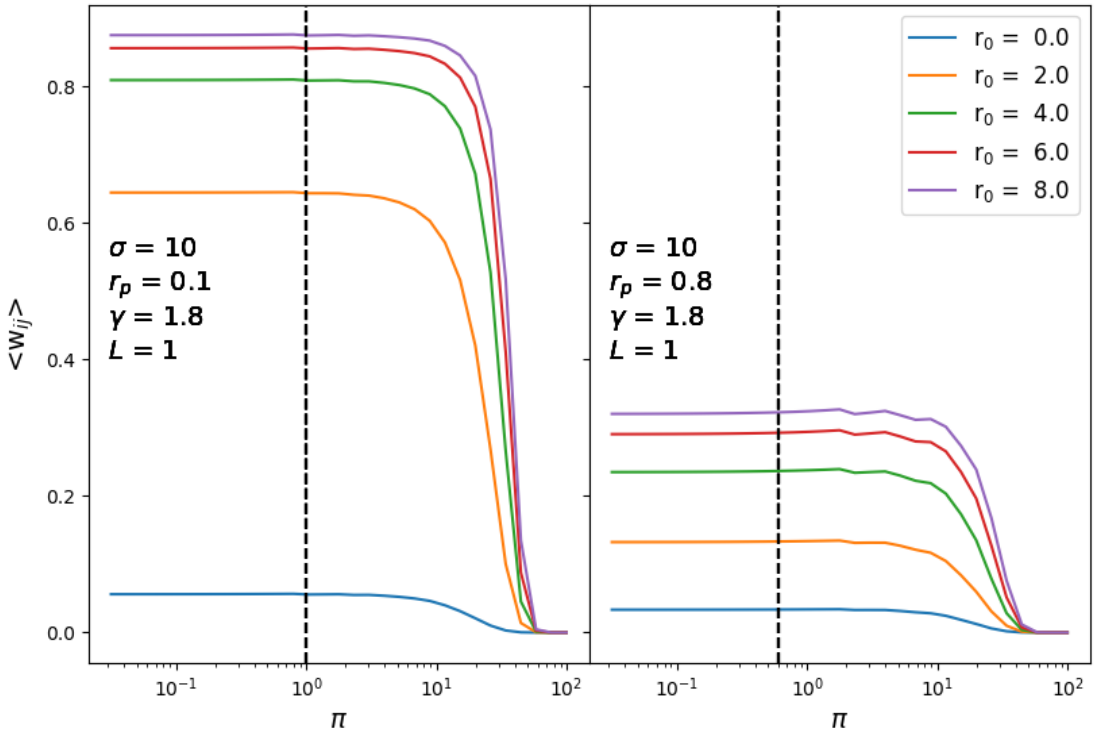

Figure 5.16: Average connection probability given by Equation (5.6.20) for a pair of galaxies in the setup shown in Figure 5.13 with $\sigma = 10$ as a function of the line of sight peak separation π for different values of correlation function amplitude r_0 and two values of r_p (left and right panels). The black dashed lines show the maximum line of sight separations that would result in a connection if $\sigma = 0$, given by $\sqrt{L^2 - r_p^2}$. All length scales are in units of h^{-1}Mpc.

Figure 5.16 shows the values of w_{ij} as a function of line of sight peak separation π for two galaxies with $\sigma = 10\ h^{-1}$Mpc for different value of the correlation function amplitude r_0, and for two values of projected separation r_p. γ is fixed at 1.8. The uncertainty of $\sigma = 10\ h^{-1}$Mpc is chosen to roughly mimic two well measured PAUS galaxies at low redshift. We can see that the connection probabilities are nearly step functions as a function of π. The transition of this step function is sharper with larger correlation function amplitudes. Larger correlation function amplitudes lead to larger peak heights and larger scales at which the probability falls toward zero. Smaller projected separations lead to smaller peak connection probabilities and larger scales at which the probabilities fall to zero. The dotted black lines show the maximum line of sight separation for which the two galaxies would be connected if $\sigma = 0$. It is clear from these plots how this scheme has provided an effective extension to the connection criterion along the line of sight in a data driven manner. This effective extension is larger in the case where the two galaxies are from highly clustered samples.

5.6.3 Discussion

So far we have not addressed the issue of redshift space, as we have only shown demonstrative tests in real space to avoid this complication. There are two ways in which we suggest dealing with redshift space effects, an anisotropic connection criterion, or treating redshift space effects as further position uncertainties.

First the case of an anisotropic connection criterion. This is similar to previous literature approaches and would extend the linking length along the line of sight to deal with redshift space effects. The scheme laid out in Equation (5.6.20) would in this case only deal with uncertainties in the redshift space measurement, and would use the redshift space correlation function in its probability calculations. In a survey with larger redshift uncertainties, such as PAUS, the redshift space correlation function may prove difficult to estimate accurately, so we would need to use measurements from a spectroscopic survey such as VIPERS. Requiring a spectroscopic survey as deep as the photometric survey seems to defeat the point but a measurement of the correlation function can be made with sparse sampling,

as in VIPERS, whereas groups are far more sensitive to survey completeness. This measurement could also be made in a smaller solid angle survey such as zCOSMOS. The weighting is sensitive to the clustering but the dominant effect comes from moving from no clustering at all to some realistic clustering values. This means the clustering result used may not need to be exactly the same as for the galaxies in PAUS to still give a reasonable result.

The second method would be to treat redshift space effects as further line of sight position uncertainties. The probability distributions in Equation 5.6.20 would now attempt to estimate the line of sight position distribution of a galaxy in real space in a statistical manner. One way to do this would be to convolve the redshift space pdf, $p_{\rm measured}(z)$, with a distribution that statistically represents our lack of knowledge of the galaxy velocity, p_v. This would define $p(z)$ as

$$p(z) = \int dz' p_{\rm measured}(z') p_v(z' - z)\,. \tag{5.6.22}$$

This means the connection criterion can now be isotropic, but would require using the real space correlation function in Equation 5.6.20. The real space auto/cross correlation function of two samples can be estimated by deprojecting the projected auto/cross correlation function (Arnalte-Mur et al., 2009). A sensible choice for p_v would be one derived from a Gaussian velocity distribution, with velocity dispersion σ_v. σ_v could be a constant free parameter for all galaxies or depend on the local density or type of galaxy. It could also change depending on the mass of halo a galaxy lives in, which would require an iterative approach similar to the one in Yang et al. (2005).

One complication we have not addressed is the calculation of the local density. We have seen section 5.5.2 that the local density can be used to modify the linking length and significantly improve group finding. One could try to extend this scheme for the pairwise density calculations or simply extend the density kernel along the line of sight as was done in Eke et al. (2004) and Robotham et al. (2011).

5.7 Conclusion

This chapter explains that the well known FOF algorithm is a subset of the more general MCL graph clustering algorithm (Van Dongen, 2000). MCL has one free parameter, inflation (Γ), which when set to 1 produces the same results as for an FOF algorithm. Working in real space we use MCL to detect galaxy groups in a realistic galaxy mock catalogue constructed from an N-body simulation using the `GALFORM` semi analytic model.

We use the variation of information (VI) (Meilă, 2003) to compare group catalogues to the real halos. A smaller value of VI produces a better clustering. We validate this choice by showing that the minimum value of VI for a simple FOF approach is found at linking lengths that agree with previous best fit parameters, such as those found in Eke et al. (2004).

When we allow Γ to vary away from 1, we find that for a simple constant linking length the FOF algorithm produces the best group catalogue. This is because the FOF algorithm produces too many spurious small groups and too few large groups and increasing inflation only acts to make this discrepancy worse.

We vary the linking length as a function of the local density of the two galaxies in a pair to try to address the multiplicity dependency of the results. This local density enhancement is normalised in such a way that can be measured from the real data and requires no free parameter. This scheme significantly improves the results of both the FOF and MCL approaches. Using this scheme the MCL algorithm produces the catalogue with the minimum value of VI.

The MCL group catalogue with density enhancement is shown to be have better completeness and purity than the comparable FOF catalogues, and in particular a completeness and purity that is more constant as a function of multiplicity. The MCL catalogue also best estimates the number of groups at a given multiplicity. Compared to the best FOF approach, it significantly improves the purity of, and the estimate of the number of, high multiplicity groups. This is most likely because it helps address the problem of bridges linking large structures together in FOF approaches.

MCL allows pairwise connection amplitudes that are not just ones and zeros,

which may prove very useful in catalogues with mixed redshift measurement precision, such as will be produced by the PAU Survey. Even in real space, where pairwise connections are not probabilistic, we have shown MCL can produce better group catalogues than an FOF approach. We have proposed a scheme to extend this work to catalgoues with uncertain galaxy positions in a way that is driven by the data. To do this we use the two point correlation function to replace lost line of sight information about a galaxy pair with average information about that type of pair of galaxies. We have shown this scheme produces sensible results when considering a single pair of galaxies with Gaussian position uncertainties. Future work will test this scheme on an appropriate mock catalogue.

Chapter 6

Conclusions and future work

This chapter summarises the work presented in this thesis, provides insight into ongoing work, and explains the avenues open for future work.

6.1 Point processes and Euclid

Chapter 2 uses point processes to provide catalogues that can be used to validate the clustering pipelines of the Euclid survey (Laureijs et al., 2011). In particular, it provides a means to build catalogues with analytically known higher order moments of the two point function using two common point processes, the segment Cox process (Stoyan et al., 1995) and the Thomas process (Thomas, 1949). We also provide predictions for the three point correlation function of the isotropic Thomas process.

The three point function predictions for the Thomas process are yet to be validated and could be explored in future work. Future work could also extend the work done here to look at the results for higher order correlation functions of a Neyman Scott process with a more realistic cluster profile for use in interpreting the one halo term results of HOD modelling. Furthermore, spherically symmetric cluster profiles fail to boost the likelihood of aligned triangles as is seen in the three point function in the real Universe. Using cylindrically symmetric cluster profiles rather than spherically symmetric ones could provide an interesting avenue for exploration into modeling the three point function analytically.

The Euclid two point correlation function pipeline has passed the strict validation tests using the segment Cox process and is now undergoing validation using realistic mock catalgoues, but the three point function has yet to be validated at all. This work has provided a viable method for the three point correlation function pipeline to be tested against an analytic prediction.

6.2 Galaxy clustering measurements, 2PCF, and DESI

Chapter 3 introduced a publicly available two point correlation function code, 2PCF, written in C++, that is fast, flexible and contains the features needed for modern galaxy redshift survey clustering statistics. The code is similar in scope and approach to CUTE (Alonso, 2012), but adds flexible binning, on the fly jackknife resampling calculations, more flexible IO and the pairwise upweighting scheme of Bianchi & Percival (2017). An extension of this pair upweighting scheme to account for shifting the survey in each run is also explained and implemented. We tested the performance scaling of this code. Under the right circumstances, it scales linearly with volume, quadratically with density, and close to ideally with increasing numbers of CPU cores.

The code has been used to investigate constraints on f(R) modified gravity models using marked correlation functions in Hernández-Aguayo et al. (2018). The use of the code in Smith et al. (2018) is summarised, which shows that the implementation of the pair upweighting scheme here is sufficient to correct the clustering statistics of the DESI BGS (DESI Collaboration et al., 2016) for the complicated target selection effects.

Immediate future developments will focus on including cross correlation function calculations, as this may prove particularly useful for the redshift space pairwise connection scheme proposed in Chapter 5 to help with galaxy group detection. Providing support for periodic boxes has also been a highly requested feature, so will be high on the list of priorities.

The survey strategy of DESI BGS is still under revision, partly due to uncer-

tainty surrounding the assumptions of what will be possible observationally in bright time. It is possible that a revised Moon model will reduce the capability of the survey. Smith et al. (2018) provided results that will be very useful in modifying survey strategy in this circumstance, as it has shown that large scale cosmology measurements can be made with similar precision with only one pass as with the currently planned three passes. This will have to be weighed against small scale and environment science cases which may prefer a drop in area rather than a drop in completeness.

The natural extension of the work of Smith et al. (2018) is to test the recovery of the three point correlation function in DESI BGS using the scheme of Bianchi & Percival (2017). Testing the three point function will be significantly more difficult than for the two point. While the pair upweighting scheme of Bianchi & Percival (2017) extends naturally to triplet counts, it is difficult to extend to work with algorithms that allow faster three point calculations such as the one presented in Slepian & Eisenstein (2015).

6.3 PAUS

Chapter 4 presents a mock catalogue for the Physics of the Accelerating Universe Survey (PAUS) (Castander et al., 2012) built from the N-body MR7 simulation (Guo et al., 2013) using the `GALFORM` semi analytic model presented in Gonzalez-Perez et al. (2014b). We use it to quantify the competitiveness of the narrow band imaging for measuring spectral features and galaxy clustering. This mock catalogue agrees well with observed number counts and redshift distributions. We demonstrate the importance of including emission lines in the narrow band fluxes. We show that PAUCam has sufficient resolution to measure the strength of the 4000Å break to the nominal PAUS depth. We predict the evolution of a narrow band luminosity function and show how this can be affected by the OII emission line. We use new rest frame broad bands (UV and blue) along with D4000 and redshift to define galaxy samples and provide predictions for galaxy clustering measurements. We show that systematic errors in the recovery of the projected clustering due to photometric

redshift errors in PAUS are significantly smaller than the expected statistical errors. The galaxy clustering on two halo scales can be recovered quantitatively without correction, and all qualitative trends seen in the one halo term are recovered. In this analysis mixing between samples reduces the expected contrast between the one halo clustering of red and blue galaxies and demonstrates the importance of a mock catalogue for interpreting galaxy clustering results.

There are two points to be addressed in the next version of the catalogue. The first is that the catalogue cannot currently be used for photometric redshift code validation as the interpolation scheme gives rise to a discreteness in the observed redshifts. Figure 6.1 shows this effect. It shows the lightcone redshift label z_{spec} vs the inferred redshift using the PAUS photoz pipeline z_{photo}. z_{spec} is continuous but the z_{photo} shows significant discreteness that lines up with the snapshots of the N body simulation used to build the mock catalogue. This is because the flux in a band is found by interpolating the values in that band from the adjacent snapshots. In the case of broad bands this change is small, so provides a good approximation, but this is not the case with narrow bands. This is because the spacing between snapshots moves the observer frame by more than the width of a single filter.

The second point to address is one of resolution. The MR7 simulation has a minimum halo mass of $2{\times}10^{10}M_{\odot}$, which translates to a minimum flux in a particular band for which the catalogue will be highly complete. For a magnitude limited galaxy survey this sets a minimum redshift where the mock is highly complete. For a PAUS mock catalogue limited by $i < 23$ this will lie somewhere between $0.1 < z < 0.2$. This means that a mock catalogue that will be used to calibrate a galaxy group finder should use an N body simulation with higher resolution.

A mock catalogue built using the pMillennium N body simulation (Baugh et al., 2018), which has 4x the number of snapshots and around an order of magnitude better mass resolution, should address both of these points.

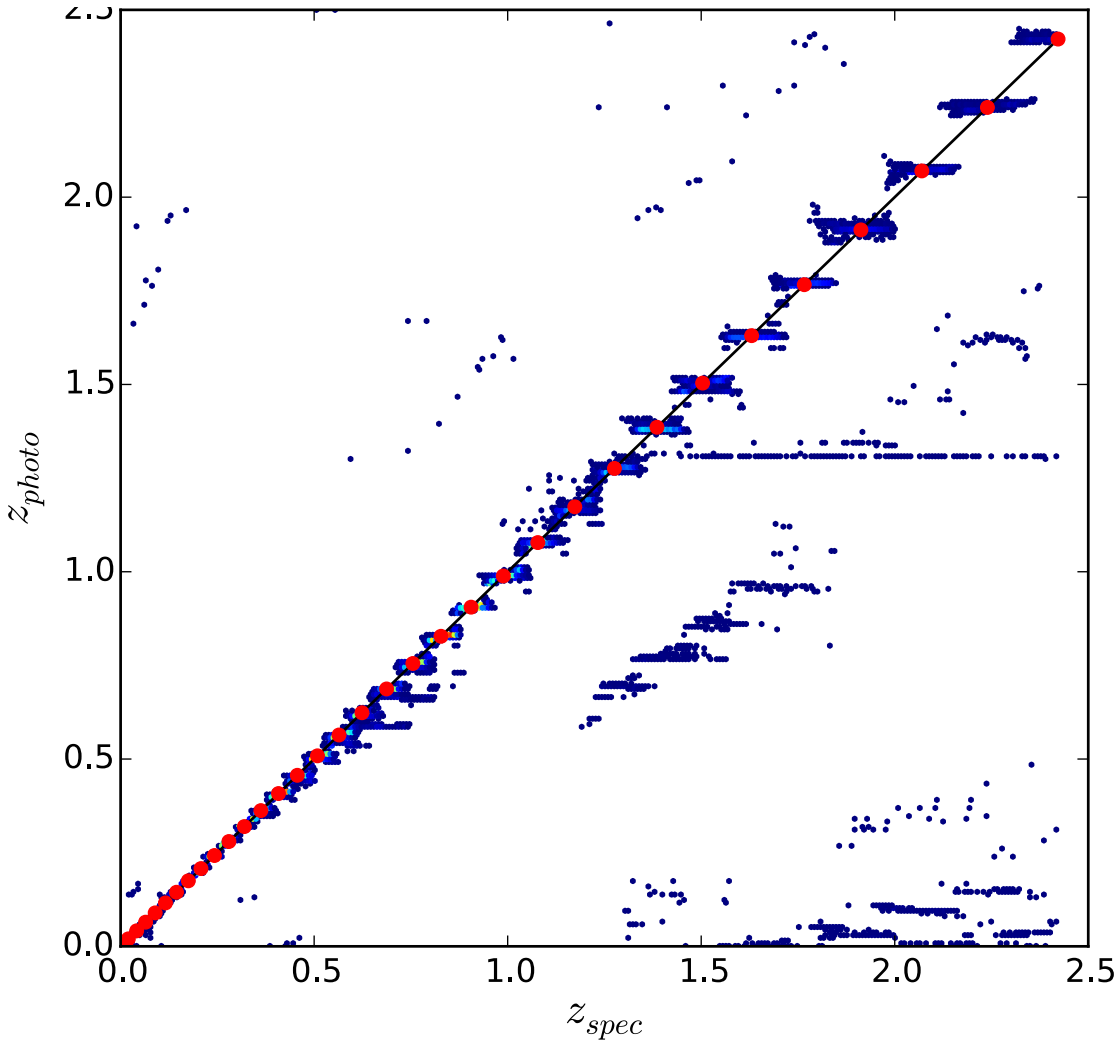

Figure 6.1: Scatter plot of redshift as labeled in the PAUS lightcone, z_{spec}, and measured photometric redshift using the 40 simulated PAUS narrow bands and the PAUS photo z pipeline, z_{photo}. The blue dots show the galaxies and the large red dots show the redshift locations of the snapshots of the MR7 N-body simulation used to build the mock catalogue. Plot provided by Alex Alarcon.

6.4 Galaxy groups and MCL

Chapter 5 introduces the Markov clustering algorithm MCL (Van Dongen, 2000) as a viable galaxy group finding algorithm. We show that the well known FOF algorithm is a subset of the MCL graph clustering algorithm. We test multiple models and each time optimise any free parameters by minimising the variation of information measure (Meilă, 2003). The variation of information measure is a single statistic that we use to find the best catalogue with a reasonable balance between purity and completeness. We show that in real space MCL produces a catalogue with better purity and completeness than the comparable FOF catalogue, and a more accurate cumulative multiplicity function. MCL allows probabilistic pairwise connections which may be useful in surveys with mixed redshift precision such as PAUS. We suggest a possible method of extending this to redshift space that uses the two point correlation function to replace lost line of sight information.

Work is ongoing to test this scheme of finding connection probabilities in a catalogue with line of sight position uncertainties. It uses the same mock catalogue as the real space work but in redshift space with Gaussian errors representing roughly the photometric redshift errors of galaxies from the PAU Survey. There are some computational challenges that are introduced with this. First, the code currently being used to produce the pairwise connection probabilities was built using the two point correlation function code `2PCF`. This code uses a local cell search to speed up calculations. In real space, the cell size does not need to be particularly large to encompass all realistically connected galaxies, but once galaxies have large position errors the size of cells must be significantly increased, and the number of considered pairs increases dramatically, increasing the runtime. Further, each pair of galaxies considered now requires a double integral over the two position pdfs. A Gaussian truncated at 3σ may stretch over more than 60 h^{-1}Mpc in a PAUS-like mock catalogue, so many points are required to resolve sub-Mpc pairwise separations, and the code scales with the resolution squared. Nevertheless, early results trying to use the scheme in a way to help recover the local density have proved promising.

Future work will complete this testing for a sample with PAUS-like uncertainties. This work can be used to help produce a galaxy group catalogue using the PAU

Survey data. Such a group catalogue would allow us to probe further down the halo mass function than ever before and provide insight into how galaxy groups have evolved over redshift. Such a group catalogue would provide the best picture of the galaxy-halo connection at low halo masses and medium redshifts (z $\sim$0.5) until 4MOST WAVES-Deep (Driver et al., 2016) produces a spectroscopic catalogue of similar scope to PAUS, starting in 2023. WAVES-Wide will produce a large solid angle survey of similar depth but will use a photometric redshift cut to select galaxies with $z < 0.2$. Connections to higher redshifts ($z \sim 1$) will be possible with the MOONS survey in 2021 (Cirasuolo & MOONS Consortium, 2016).

This author believes communities who use a FOF scheme for galaxy group detection should strongly consider switching to the MCL algorithm for future work. The extra free parameter, inflation, Γ, provides additional flexibility that can be used to better tune a group finder to a specific use case, or simply improve results, as has been evidenced in specific circumstances considered in Chapter 5.

Bibliography

Alam S., et al., 2015, ApJS, 219, 12

Alonso D., 2012, preprint, (arXiv:1210.1833)

Anderson L., et al., 2012, MNRAS, 427, 3435

Anderson L., et al., 2014a, MNRAS, 439, 83

Anderson L., et al., 2014b, MNRAS, 441, 24

Arnalte-Mur P., Fernández-Soto A., Martínez V. J., Saar E., Heinämäki P., Suhhonenko I., 2009, MNRAS, 394, 1631

Arnalte-Mur P., et al., 2014, MNRAS, 441, 1783

Balogh M. L., Morris S. L., Yee H. K. C., Carlberg R. G., Ellingson E., 1999, ApJ, 527, 54

Barsanti S., et al., 2018, ApJ, 857, 71

Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101

Baugh C. M., Gaztanaga E., Efstathiou G., 1995, MNRAS, 274, 1049

Baugh C. M., Benson A. J., Cole S., Frenk C. S., Lacey C. G., 1999, MNRAS, 305, L21

Baugh C. M., et al., 2018, preprint, (arXiv:1808.08276)

Benítez N., 2000, ApJ, 536, 571

Benson A. J., 2001, MNRAS, 325, 1039

Benson A. J., 2010, Phys. Rep., 495, 33

Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587

Berlind A. A., Kazin E., Blanton M. R., Pueblas S., Scoccimarro R., Hogg D. W., 2006a, ArXiv Astrophysics e-prints,

Berlind A. A., et al., 2006b, ApJS, 167, 1

Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1

Bianchi D., Percival W. J., 2017, MNRAS, 472, 1106

Blanton M. R., et al., 2003, ApJ, 592, 819

Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476

Bray A. D., et al., 2015, ApJ, 811, 90

Bruzual G., 1983, ApJ, 273, 105

Carretero J., et al., 2017, PoS, EPS-HEP2017, 488

Castander F. J., et al., 2012, in Ground-based and Airborne Instrumentation for Astronomy IV. p. 84466D, doi:10.1117/12.926234

Cirasuolo M., MOONS Consortium 2016, in Skillen I., Balcells M., Trager S., eds, Astronomical Society of the Pacific Conference Series Vol. 507, Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields. p. 109

Coil A. L., et al., 2011, ApJ, 741, 8

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Cole S., et al., 2005, MNRAS, 362, 505

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

Contreras S., Baugh C. M., Norberg P., Padilla N., 2013, MNRAS, 432, 2717

Cooray A., Sheth R., 2002, Phys. Rep., 372, 1

Coupon J., Arnouts S. a., 2015, MNRAS, 449, 1352

Cowley W., Baugh C., Cole S., Frenk C., Lacey C., 2017, preprint, (arXiv:1702.02146)

Cox D. R., 1955, Some Statistical Methods Connected with Series of Events"

Crain R. A., et al., 2015, MNRAS, 450, 1937

DESI Collaboration et al., 2016, preprint, (arXiv:1611.00036)

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

Driver S. P., et al., 2011, MNRAS, 413, 971

Driver S. P., Davies L. J., Meyer M., Power C., Robotham A. S. G., Baldry I. K., Liske J., Norberg P., 2016, The Universe of Digital Sky Surveys, 42, 205

Eisenstein D., 2005, New Astronomy Reviews, 49, 360

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

Eke V. R., et al., 2004, MNRAS, 348, 866

Farrow D. J., et al., 2015a, MNRAS, 454, 2120

Farrow D. J., et al., 2015b, MNRAS, 454, 2120

Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23

Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, L66

Gaztañaga E., Norberg P., Baugh C. M., Croton D. J., 2005, MNRAS, 364, 620

Gerke B. F., et al., 2005, ApJ, 625, 6

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Frenk C. S., Wilkins S. M., 2013, MNRAS, 429, 1609

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014a, MNRAS, 439, 264

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014b, MNRAS, 439, 264

Gonzalez-Perez V., et al., 2017, preprint, (arXiv:1708.07628)

Guo Q., White S., Angulo R. E., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351

Guzzo L., et al., 2014, A&A, 566, A108

Hagen G., Schweder T., 1995, in Blix A. S., Walle L., yvind Ulltang eds, Developments in Marine Biology, Vol. 4, Whales, seals, fish and man. Elsevier Science, pp 27 – 33, doi:https://doi.org/10.1016/S0163-6995(06)80006-5, http://www.sciencedirect.com/science/article/pii/S0163699506800065

Hamilton A. J. S., 1992, ApJ, 385, L5

Hartuv E., Shamir R., 2000, Information Processing Letters, 76, 175

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

Hellwing W. A., Koyama K., Bose B., Zhao G.-B., 2017, Phys. Rev. D, 96, 023515

Hernández-Aguayo C., Baugh C. M., Li B., 2018, MNRAS, 479, 4824

Heymans C., et al., 2012, MNRAS, 427, 146

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

Ilbert O., 2012, CFHT T0007 photo-z catalogue http://iapix.iap.fr/~hudelot/CFHTLS/CFHTLS-zphot-T0007/cfhtls_wide_T007_v1.2_Oct2012.pdf

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

Jarvis M., Bernstein G., Jain B., 2004, MNRAS, 352, 338

Jian H.-Y., et al., 2014a, ApJ, 788, 109

Jian H.-Y., et al., 2014b, ApJ, 788, 109

Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, MNRAS, 440, 2115

Kaiser N., 1987, MNRAS, 227, 1

Kauffmann G., et al., 2003, MNRAS, 341, 33

Kim H.-S., Baugh C. M., Cole S., Frenk C. S., Benson A. J., 2009, MNRAS, 400, 1527

Knobel C., et al., 2009, ApJ, 697, 1842

Knobel C., et al., 2012, ApJ, 753, 121

Koyama K., 2016, Reports on Progress in Physics, 79, 046902

Kriek M., van Dokkum P. G., Whitaker K. E., Labbé I., Franx M., Brammer G. B., 2011, ApJ, 743, 168

Kuijken K., et al., 2015, MNRAS, 454, 3500

Lacey C. G., et al., 2016, MNRAS, 462, 3854

Lagos C. D. P., Lacey C. G., Baugh C. M., Bower R. G., Benson A. J., 2011a, MNRAS, 416, 1566

Lagos C. D. P., Lacey C. G., Baugh C. M., Bower R. G., Benson A. J., 2011b, MNRAS, 416, 1566

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

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

Laureijs R., et al., 2011, preprint, (`arXiv:1110.3193`)

Lilly S. J., Le Fevre O., Hammer F., Crampton D., 1996, ApJ, 460, L1

Lilly S. J., et al., 2007, ApJS, 172, 70

Linde A., 2014, preprint, (`arXiv:1402.0526`)

Liu H. B., Hsieh B. C., Ho P. T. P., Lin L., Yan R., 2008, ApJ, 681, 1046

Liu R., Feng S., Shi R., Guo W., 2014, Procedia Computer Science, 31, 85

Loveday J., et al., 2012, MNRAS, 420, 1239

Loveday J., et al., 2018, MNRAS, 474, 3435

Madau P., Ferguson H. C., Dickinson M. E., Giavalisco M., Steidel C. C., Fruchter A., 1996, MNRAS, 283, 1388

Martí P., Miquel R., Castander F. J., Gaztañaga E., Eriksen M., Sánchez C., 2014a, MNRAS, 442, 92

Martí P., Miquel R., Castander F. J., Gaztañaga E., Eriksen M., Sánchez C., 2014b, MNRAS, 442, 92

Marulli F., et al., 2013, A&A, 557, A17

Marulli F., Veropalumbo A., Moresco M., 2016, Astronomy and Computing, 14, 35

McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Newman J. A., Scoccimarro R., Zehavi I., Schneider D. P., 2011, ApJ, 726, 13

McCullagh N., Neyrinck M., Norberg P., Cole S., 2016, MNRAS, 457, 3652

Meilă M., 2003, in Schölkopf B., Warmuth M. K., eds, Learning Theory and Kernel Machines. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 173–187

Merson A. I., et al., 2013, MNRAS, 429, 556

Mitchell P. D., Lacey C. G., Baugh C. M., Cole S., 2013, MNRAS, 435, 87

Moles M., et al., 2008, AJ, 136, 1325

Molino A., et al., 2014, MNRAS, 441, 2891

Moller J., Waagepetersen R. P., 2004, Statistical Inference and Simulation for Spatial Point Processes

Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563

Neyman J., Scott E. L., 1958, J. Roy. Statist. Soc.

Nichol R. C., et al., 2006, MNRAS, 368, 1507

Norberg P., et al., 2002, MNRAS, 332, 827

Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009a, MNRAS, 396, 19

Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009b, MNRAS, 396, 19

Padilla C., et al., 2016, in Ground-based and Airborne Instrumentation for Astronomy VI. p. 99080Z, doi:10.1117/12.2231884

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

Planck Collaboration et al., 2018, preprint, (`arXiv:1807.06209`)

Pozzetti L., et al., 2016, A&A, 590, A3

Pujol A., et al., 2017, MNRAS, 469, 749

Riess A. G., et al., 1998, AJ, 116, 1009

Robotham A. S. G., et al., 2011, MNRAS, 416, 2640

Sánchez C., et al., 2014, MNRAS, 445, 1482

Schaeffer S. E., 2007, Comput. Sci. Rev., 1, 27

Schaye J., et al., 2015, MNRAS, 446, 521

Schechter P., 1976, ApJ, 203, 297

Schneider M. D., et al., 2013, MNRAS, 433, 2727

Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20

Scodeggio M., et al., 2018, A&A, 609, A84

Sheth R. K., Hui L., Diaferio A., Scoccimarro R., 2001, MNRAS, 325, 1288

Skibba R. A., et al., 2014, ApJ, 784, 128

Skibba R. A., et al., 2015, ApJ, 807, 152

Slepian Z., Eisenstein D. J., 2015, MNRAS, 454, 4142

Smee S. A., et al., 2013, AJ, 146, 32

Smith A., Cole S., Baugh C., Zheng Z., Angulo R., Norberg P., Zehavi I., 2017, MNRAS, 470, 4646

Smith A., Cole S., Baugh C., Norberg P., Stothert L., 2018

Snethlage M., Martínez V. J., Stoyan D., Saar E., 2002, A&A, 388, 758

Soneira R. M., Peebles P. J. E., 1978, AJ, 83, 845

Springel V., et al., 2005, Nature, 435, 629

Stasińska G., 1990, A&AS, 83, 501

Stothert L., et al., 2018, preprint, (`arXiv:1807.03260`)

Stoyan D., 1994, statistics, 25, 267

Stoyan D., Kendall W., Mecke J., 1995, Stochastic Geometry and its Applications

Tempel E., Kruuse M., Kipper R., Tuvikene T., Sorce J. G., Stoica R. S., 2018, preprint, (`arXiv:1806.04469`)

Thomas M., 1949, A generalization of Poissons binomial limit for use in ecology

Treyer M., et al., 2018, MNRAS, 477, 2684

Tulin S., Yu H.-B., 2018, Phys. Rep., 730, 1

Van Dongen S., 2000, PhD thesis, University of Utrecht

Viel M., Becker G. D., Bolton J. S., Haehnelt M. G., 2013, Phys. Rev. D, 88, 043502

Vlasblom J., Wodak S. J., 2009, BMC Bioinformatics, 10, 99

Wagner S., Wagner D., 2007, Technical Report 4, Comparing Clusterings - An Overview. Karlsruhe

Wang Y., Yang X., Mo H. J., van den Bosch F. C., Weinmann S. M., Chu Y., 2008, ApJ, 687, 919

Wang L., Weinmann S. M., De Lucia G., Yang X., 2013, MNRAS, 433, 515

White S. D. M., 1979, MNRAS, 186, 145

Wolf C., et al., 2004, A&A, 421, 913

Wu J., Xiong H., Chen J., 2009, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09. ACM, New York, NY, USA, pp 877–886, doi:10.1145/1557019.1557115, `http://doi.acm.org/10.1145/1557019.1557115`

Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005, MNRAS, 356, 1293

Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, ApJ, 671, 153

York D. G., et al., 2000, AJ, 120, 1579

Zehavi I., et al., 2002, ApJ, 571, 172

Zehavi I., et al., 2005, ApJ, 630, 1

Zehavi I., et al., 2011, ApJ, 736, 59

de la Torre S., et al., 2013, A&A, 557, A54

van Uitert E., et al., 2017, MNRAS, 467, 4131

Appendix A

Appendix to chapter 4

A.1 Galaxy clustering statistics and code

We calculate galaxy clustering using the appropriately normalised Landay-Szalay estimator (Landy & Szalay, 1993b)

$$\xi(r_{\rm p},\pi) = \frac{DD(r_{\rm p},\pi) - 2DR(r_{\rm p},\pi) + RR(r_{\rm p},\pi)}{RR(r_{\rm p},\pi)}, \tag{A.1.1}$$

DD, DR and RR are normalised Data-Data, Data-Random and Random-Random pair counts. The number of randoms set is always ten times the number of galaxies in a sample, and they are uniformly distributed in the comoving volumes of the samples. $r_{\rm p}$ and π are, respectively, the galaxy pair separations transverse and parallel to the line of sight. These separations are defined in terms of the pair of galaxy vectors $\underline{x}_1$ and $\underline{x}_2$

$$\pi = \left| \frac{(\underline{x}_1 - \underline{x}_2).(\underline{x}_1 + \underline{x}_2)}{|\underline{x}_1 + \underline{x}_2|} \right|, \tag{A.1.2}$$

$$r_{\rm p} = \sqrt{(\underline{x}_1 - \underline{x}_2)^2 - \pi^2}. \tag{A.1.3}$$

In this analysis we consider only projected galaxy clustering to minimise the impact of the PAUS redshift error. The projected correlation function is given by

$$w_{\rm p}(r_{\rm p}) = 2\int_0^{\pi_{\rm max}} \xi(r_{\rm p},\pi){\rm d}\pi, \tag{A.1.4}$$

where the value of $\pi_{\rm max}$ is a parameter to be set.

Fig. A.1 shows the systematic loss of signal in the projected galaxy clustering for samples with different values of photometric redshift errors relevant to PAUS for two different values of $\pi_{\rm max}$. The sample used was (-19.5 $< M_B^h <$ −19.0) in the redshift range $0.5 < z < 0.63$. The real PAUS data will have a distribution of photometric redshift errors rather than the single Gaussian error assumed here so this plot can inform us on the systematic errors we may introduce for different error distributions. The larger value of $\pi_{\rm max}$ recovers more of the signal but at the cost of increasing the statistical noise. The difference in spectroscopic result between $\pi_{\rm max}$ = 50 and 100h^{-1}Mpc is less than 2%. A value of $\pi_{\rm max}$ of 100 h^{-1}Mpc would allow us to use galaxies in the sample with three times the nominal PAUS redshift error and recover the projected clustering within the statistical errors. See Arnalte-Mur et al. 2009 and Arnalte-Mur et al. 2014 for further discussion on projected correlation recovery in photometric redshift surveys.

All clustering results are calculated using a two point clustering code which is publicly available on github [1]. This is an OpenMP accelerated code which has the ability to calculate monopole and 2D decompositions of the correlation function with flexible linear or logarithmic binning, multiple input/output types and on-the-fly jackknife errors at the expense of very little extra computing time.

The galaxy pairs were binned logarithmically in both $r_{\rm p}$ and π, which can help reduce the increase in statistical error for large values of π_{max}.

A.2 Clustering samples

Fig. A.2 shows the volume limited cuts used to create galaxy clustering samples. The faint limit in M_B^h at each redshift was chosen such that the faintest samples were over 99% complete in a catalogue i ¡ 23 without errors. The scatter in the colour term between the observed i-band and M_B^h is responsible for any small amount of incompleteness. The high completeness of the samples can be seen from Fig. A.2 by noting that the bottom right corners of the faintest boxes do not overlap with

[1] `https://github.com/lstothert/two_pcf`

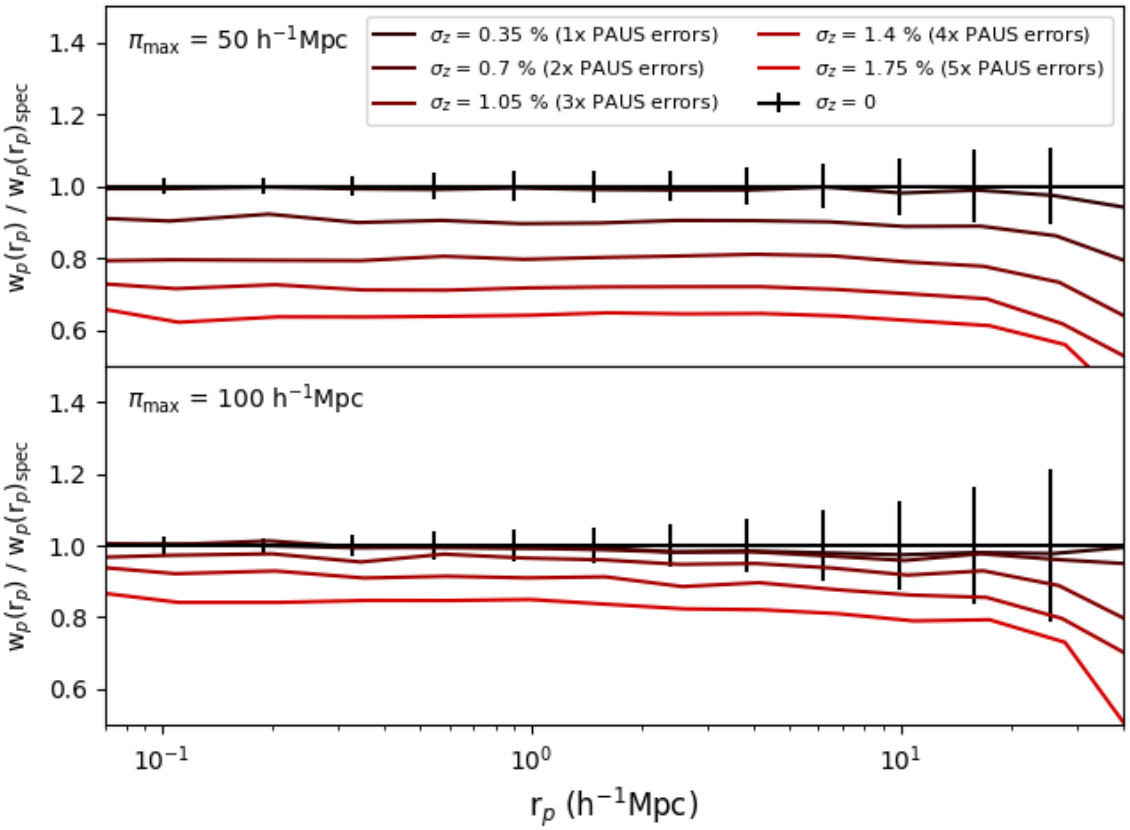

Figure A.1: The recovery of the projected galaxy clustering for samples of different Gaussian photometric redshift errors and different values of $\pi_{\rm max}$. Each curve is normalised by the spectroscopic result integrated to the same $\pi_{\rm max}$. The error bars represent the jackknife errors on the spectroscopic result.

galaxies with mean i band magnitude of 23. These samples are therefore the samples we would choose if we had perfect photometry, and we then deduce the recoverability of the results when realistic errors are included. The cuts at lower redshift must have a more conservative limit in M_B^h than at higher redshift as the scatter between PAUS Blue and the apparent i band magnitude is larger at lower redshift. This is because at the lowest redshift the wavelength difference between the two bands is maximised in the PAUS redshift range so the colour term, and the corresponding colour scatter, is the largest.

All samples selected along with their completeness and purity once errors are included are listed in Tables A.1 and A.2. The definitions of the completeness and purity in those tables can be written as follows. Define N_{ij} as the number of galaxies that lie in sample i in the catalogue without errors and in sample j in the catalogue including errors. Define N_{i*} as the number of galaxies in sample i in the catalogue without errors. Define N_{*j} as the number of galaxies in sample j in the catalogue with errors. The completeness of sample i can now be defined as N_{ii} / N_{i*} and the purity as N_{ii} / N_{*i}. Satellite fraction and median halo mass are galaxy weighted quantities. A halo with many satellites may therefore make multiple contributions to the number of satellites and halo masses in a sample. The samples here were split in uniform redshift steps but future work may choose to make the lower redshift bins larger than the higher redshift bins to match the sizes of the volumes probed.

There is high completeness and purity amongst samples split only by redshift and M_B^h seen in table A.1, which drops when samples are further split by colour in table A.2. This shows that the driving source of sample mixing in this work is the colour split. In a fixed luminosity bin the completeness and purity falls with redshift as the photometry errors are larger for apparently fainter samples. This also holds once galaxies are split by colour.

The number density of the brightest galaxies increases with increasing redshift as the star formation rate of the universe, and therefore the amplitude of the M_B^h luminosity function, increases with redshift. These trends are also seen in fainter samples but aren't as clear once errors are included. Brighter galaxies live in larger halos and this trend is particularly strong for red galaxies. These red galaxies also

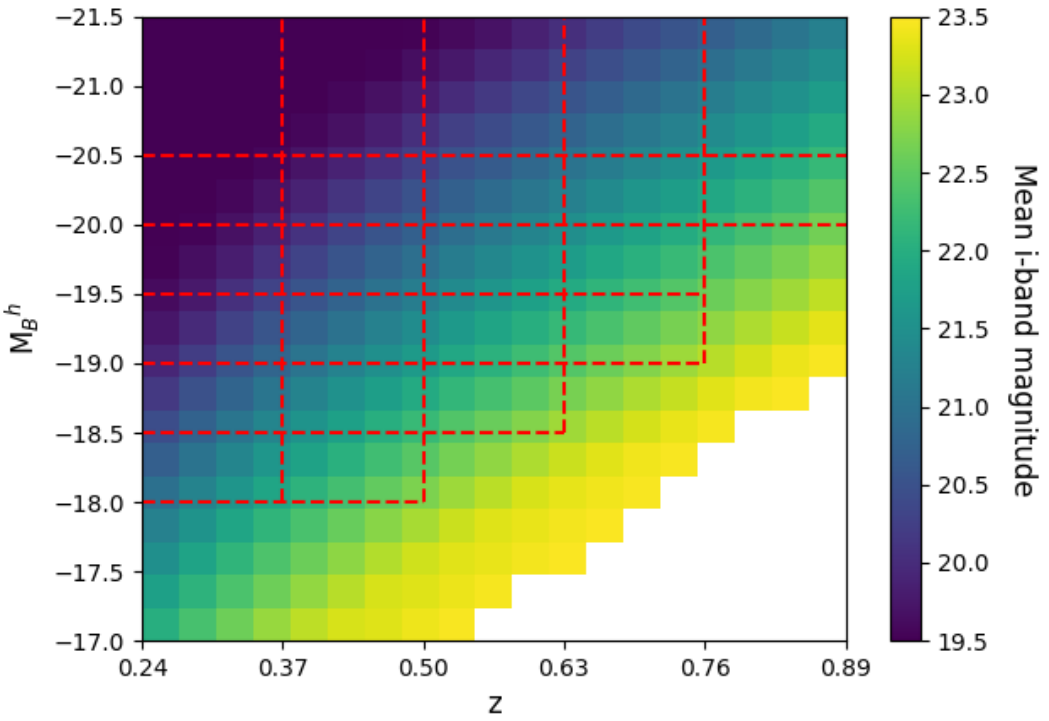

Figure A.2: Rest-frame M_B^h vs redshift, colour coded by mean i-band magnitude for a PAUS mock built to i ¡ 25 without including redshift or photometry errors. The lightcone was built deeper than nominal PAUS depth so as to be certain about the completeness values of the samples. The plot stops at i¡ 23.5 so the colour gradient through the boxes is more obvious to the reader. The boxes show the sample limits used in the galaxy clustering analysis, chosen to be 99% complete to $i < 23$ in this lightcone. Note the boxes do not touch the $i = 23$ coloured squares.

on average live in significantly larger halos than their blue counterparts with the same luminosity and redshift. At fixed colour and luminosity the median halo mass increases with decreasing redshift as the dark matter growth rate is large on small non-linear scales over this redshift range.

z-min Volume	z-max ($10^6 h^{-3}$Mpc^{-3})	M_B^h bright	M_B^h faint	Comp (%)	Purity (%)	$\overline{n}$ ($10^{-3} h^{-3}$Mpc^{-3})	Sat frac	Median M_{halo} ($10^{11} h^{-1} M_\odot$)
0.24	0.37	-18.5	-18.0	89.6	88.6	7.51	0.273	2.43
4.626		-19.0	-18.5	92.4	91.9	6.22	0.259	3.58
		-19.5	-19.0	94.6	93.5	4.89	0.221	4.58
		-20.0	-19.5	95.4	94.7	3.41	0.153	5.48
		-20.5	-20.0	96.2	95.4	1.8	0.1	6.22
		None	-20.5	97.3	96.2	1.02	0.073	8.66
0.37	0.5	-18.5	-18.0	81.9	81.7	8.14	0.291	2.34
8.262		-19.0	-18.5	87.5	86.9	6.73	0.275	3.53
		-19.5	-19.0	90.8	90.8	5.44	0.242	4.6
		-20.0	-19.5	93.2	92.9	3.88	0.179	5.36
		-20.5	-20.0	94.1	94.2	2.1	0.113	6.05
		None	-20.5	96.1	96.1	1.22	0.079	8.79
0.5	0.63	-19.0	-18.5	81.2	82.0	6.22	0.273	3.31
12.22		-19.5	-19.0	87.0	86.2	5.31	0.234	4.28
		-20.0	-19.5	90.2	89.9	3.91	0.174	5.03
		-20.5	-20.0	92.2	91.7	2.15	0.117	5.78

		None	-20.5	95.2	95.0	1.25	0.076	8.26
0.63	0.76	-19.5	-19.0	82.2	83.7	4.96	0.232	4.09
16.177		-20.0	-19.5	86.9	86.7	3.91	0.177	4.85
		-20.5	-20.0	89.9	89.6	2.23	0.118	5.5
		None	-20.5	94.1	93.9	1.3	0.08	8.08
0.76	0.89	-20.5	-20.0	87.9	87.4	2.62	0.129	5.42
19.922		None	-20.5	93.1	92.9	1.53	0.083	7.99

Table A.1: Table of galaxy clustering samples used in this analysis. Completeness, purity and satellite fraction are defined in the text. $\overline{n}$ is the number density of the sample.

z-min	z-max	Colour	M_B^h bright	M_B^h faint	Comp (%)	Purity (%)	$\overline{n}$	Sat frac	Median M_{halo}
Volume	$(10^6 h^{-3}\mathrm{Mpc}^{-3})$						$(10^{-3}h^{-3}\mathrm{Mpc}^{-3})$		$(10^{11}h^{-1}\mathrm{M}_\odot)$
0.24	0.37	red	-18.5	-18.0	73.9	63.9	3.43	0.481	13.3
4.626			-19.0	-18.5	85.8	80.2	3.11	0.47	19.4
			-19.5	-19.0	92.2	88.2	2.52	0.401	20.4
			-20.0	-19.5	93.5	91.6	1.68	0.286	20.9
			-20.5	-20.0	94.9	92.8	0.788	0.193	23.2

			None	-20.5	96.6	94.6	0.454	0.13	64.7
		blue	-18.5	-18.0	71.7	78.3	4.08	0.097	1.91
			-19.0	-18.5	81.8	86.2	3.11	0.048	2.36
			-19.5	-19.0	88.9	91.0	2.37	0.029	2.89
			-20.0	-19.5	92.2	92.6	1.73	0.023	3.63
			-20.5	-20.0	94.0	94.3	1.01	0.026	4.69
			None	-20.5	95.9	95.5	0.571	0.028	6.37
0.37	0.5	red	-18.5	-18.0	56.3	46.3	3.94	0.412	6.61
8.262			-19.0	-18.5	69.5	63.6	3.4	0.426	12.2
			-19.5	-19.0	81.1	76.9	2.88	0.401	16.7
			-20.0	-19.5	88.2	84.0	2	0.31	17.9
			-20.5	-20.0	91.2	88.5	0.964	0.212	22.2
			None	-20.5	94.1	93.9	0.571	0.136	50.7
		blue	-18.5	-18.0	55.2	64.2	4.2	0.178	1.95
			-19.0	-18.5	66.3	71.1	3.33	0.12	2.4
			-19.5	-19.0	76.3	81.0	2.55	0.062	2.88
			-20.0	-19.5	84.5	88.3	1.88	0.04	3.51
			-20.5	-20.0	89.3	91.7	1.14	0.029	4.41
			None	-20.5	94.2	94.4	0.651	0.028	6.02

0.5	0.63	red	-19.0	-18.5	64.2	61.0	2.94	0.448	12.7
12.22			-19.5	-19.0	74.9	73.5	2.58	0.4	14.9
			-20.0	-19.5	83.9	82.8	1.86	0.316	16.6
			-20.5	-20.0	88.8	87.7	0.94	0.219	18.1
			None	-20.5	93.4	93.5	0.579	0.129	36.4
		blue	-19.0	-18.5	65.2	69.5	3.27	0.114	2.29
			-19.5	-19.0	75.6	75.7	2.73	0.076	2.8
			-20.0	-19.5	84.1	84.8	2.05	0.044	3.35
			-20.5	-20.0	89.5	89.5	1.21	0.038	4.17
			None	-20.5	94.1	93.7	0.67	0.03	5.78
0.63	0.76	red	-19.5	-19.0	66.5	65.2	2.42	0.378	11.6
16.177			-20.0	-19.5	76.1	73.6	1.88	0.302	12.9
			-20.5	-20.0	83.4	81.4	0.989	0.214	14.8
			None	-20.5	91.3	91.3	0.61	0.132	27.7
		blue	-19.5	-19.0	67.0	70.6	2.54	0.093	2.77
			-20.0	-19.5	76.2	78.3	2.02	0.06	3.32
			-20.5	-20.0	84.4	85.5	1.25	0.041	4.01
			None	-20.5	91.7	91.5	0.686	0.034	5.56
0.76	0.89	red	-20.5	-20.0	77.8	74.7	1.22	0.214	12.1

19.922		None	-20.5	88.7	87.9	0.735	0.129	22.6
	blue	-20.5	-20.0	78.5	80.4	1.41	0.054	3.88
		None	-20.5	88.5	88.9	0.795	0.041	5.39

Table A.2: Table of galaxy clustering samples used in this analysis including colour splits. Completeness, purity and satellite fraction are defined in the text. $\overline{n}$ is the number density of the sample.