MACHINE LEARNING IN GALAXY GROUPS DETECTION

by

RAFEE TARIQ IBRAHEM

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
May 2017

Abstract

The detection of galaxy groups and clusters is of great importance in the field of astrophysics. In particular astrophysicists are interested in the evolution and formation of these systems, as well as the interactions that occur within galaxy groups and clusters. In this thesis, we developed a probabilistic model capable of detecting galaxy groups and clusters based on the Hough transform. We called this approach probabilistic Hough transform based on adaptive local kernel (PHTALK). PHTALK was tested on a 3D realistic galaxy and mass assembly (GAMA) mock data catalogue (at close redshift $z < 0.1$)(mock data: contains information related to galaxies' position, redshift and other properties). We compared the performance of our PHTALK method with the performance of two versions of the standard friends-of-friends (FoF) method. As a performance measures, we used the precision versus recall curve. Furthermore, to test the efficiency of recovering the galaxy groups' and clusters' properties, we also used completeness and reliability, fragmentation and merging, velocity and mass estimation of the detected groups. The new PHTALK method outperformed the FoF methods in terms of reducing the detection of spurious agglomerations (false positives (FPs)). This smaller sensitivity to the false positive (FP) is mainly due to the clear description of the galaxy groups' model based on astrophysical prior knowledge; in particular, the fingers of god (FoG) pattern (a pattern formed by the projected velocity dispersion of galaxies, inside a galaxy group, along the line of sight). However, the FoF methods seem to outperform the PHTALK in terms of detecting galaxy groups or clusters that do not follow the FoG pattern. The main advantage of our probabilistic model is its flexibility to incorporate any prior knowledge expressed in terms of a galaxy group model.

ACKNOWLEDGEMENTS

First of all, thanks Allah for his mercy and grace which allow me to keep my motivation and accomplish my thesis. I would like to express my sincere thanks and respect to Guru and supervisor Professor Peter Tino for his helpful and tactful supervision. My genuine thanks to Professor Trevor Ponman, Professor Arif Babul, Professor Gary Mamon and Dr Richard Pearson for their valuable time, notes, suggestions, and critics. I would like to thank The Higher Committee for Education Development in Iraq (HCED) for offering this golden opportunity and all the support to continue my research. I would like to thank the local committee in the School of Computer Science for offering useful guidance. I deeply appreciate the moral support of my family and friends.

This thesis was copy-edited for the conventions of language, spelling and grammar by Janet's Proofreading Service.

CONTENTS

LIST OF ABBREVIATIONS

AGN active galactic nuclei. 32

BCG brightest cluster galaxy. 24

CHT circle Hough transform. 56, 58

Com completeness. 103

Dec declination. 4, 18, 67

FHT frequency Hough transform. 56

FN false negative. 64, 65, 102, 125, 126

FoF friends-of-friends. i, iii, xx, xxi, 4, 7, 30, 33, 34, 37–43, 45, 94, 95, 104, 106–110, 114–117, 121

FoG fingers of god. i, xviii, 5, 38, 60, 108

FP false positive. i, 7, 64, 65, 102, 125, 126

FPs false positives. i, 5

GAMA galaxy and mass assembly. i, iii, xviii–xxi, 5, 7, 29, 30, 75–77, 79, 94, 97, 107–109, 117

GL gravitational lensing. 31

HMF halo mass function. 26, 47, 81

HT Hough transform. iii, 48, 49

ICM intra-cluster medium. 32

LoS line of sight. xix, xx, 5, 38, 42, 46, 61, 67, 68, 70, 76, 77, 79–81, 90, 94, 99, 100, 104, 106

MF matched filter. 36, 37

ML maximum likelihood. 52

Mpc mega-parsec. xix, 10, 12, 15, 22–24, 27, 32, 80, 83

PDF probability density function. 53, 55

PFoF probabilistic friends of friends. 37, 40–43

PG predicted group. 103

PHT probabilistic Hough transform. 48, 53

PHTALK probabilistic Hough transform based on adaptive local kernel. i, iv, xx, xxi, 7, 75, 87, 94, 95, 97, 98, 104–110, 114–117, 120–122

PHTM probabilistic Hough transform model. 7, 60, 64, 67, 70, 75

PPV positive-predictive-value. 125, 126

Pr precision. 66, 102, 108, 125

RA right ascension. 4, 18, 67

Re recall. 66, 102, 108, 125

Rel reliability. 103

SFR star formation rate. 21

SHT standard Hough transform. 52–54

sr steradians. 18

SSS SuperCOSMOS Sky Survey. 55

SZ Sunyaev Zeldovich. 32

TG true group. 103

TID time intensity diagram. 55

TP true positive. 102, 125, 126

TPR true-positive rate. 125, 126

TPs true positives. 5, 96

VDM Voronoi-Delaunay method. 34

VGCF Voronoi galaxy cluster finder model. 34

WMAP Wilkinson Microwave Anisotropy Probe. 81

LIST OF SYMBOLS

B background level parameter. 44

D distance. 10, 12

D_H Hubble distance. 15

D_L luminosity distance. 16

D_a angular diameter distance. 16

D_c comoving distance. 15

D_m transverse comoving distance. 16

$D_{\oplus}$ the obtained dilated matrix. 72

E evolution scaling parameter. 14

F flux. 16

G gravitational constant. 13

G_k a grid point or galaxy group suspect. 63

H a Hough transform value. 64

L luminosity. 16

L_{max} a maximum perpendicular linking length. 38

M the absolute magnitude. 17

M_h a halo mass. 26

M_{500} the mass w.r.t. over-density equivalent to 500 times the critical density of the Universe. 28

M_{vr} an estimated mass based on virial theorem. 101

N_G the number of groups/grid points. 63

N_M the number of mass bands. 89

N_{gal} the number of galaxies. 64

N_{grp} the number of galaxies in a group. 61, 82

N_{gf} the number of galaxies in all groups after the flux limit effect. 86

N_{bg} the generated number of background galaxies in each z band. 87

N_{grp_f} the number of galaxies in a group after the flux limit effect. 83

Q a distribution over galaxy count. 100

R the 3D radius. 27

R_f the ratio between perpendicular and parallel linking lengths. 38

Rot_{3D} the rotation matrix. 68

V_c the comoving volume. 20

Z_k the redshift of suspected galaxy group centre. 90

ΔN the number of galaxies in each bin of radius r around the group centre. 77

$\Delta\lambda$ the amount of change in the wavelength. 11

Ω solid angle. 18

Ω_K the Universe curvature parameter. 13

Ω_M density parameter. 13

Ω_R radiation parameter. 13

Ω_Λ cosmological parameter. 13

Θ an angular separation. 39

$\bar{n}$ a comoving number density. xiii, 38

β declination. 18

$\mathcal{C}$ the galaxy count inside a cylindrical volume $\mathcal{V}$. 99

$\mathcal{R}$ group responsibility. 45

$\mathcal{V}$ a cylindrical volume oriented along LoS, based on mass and redshift. 99

$\mathcal{J}$ the redshift bins. 87

χ the physical size of an object. 16

δN the group mass function. 82

ℓ the mass index. 89

$\ell_\perp$ a perpendicular linking length. 38

ϵ_k the prolonged shape factor. 61

γ an angle of positioning a galaxy g around the centre of the group. 85

$\hat{M}$ the mean estimation of a group mass. 100

λ_{em} the emitted wavelength. 11

λ_{ob} the observed wavelength. 11

$\oplus$ dilation process. 128

ϕ Schechter luminosity function. 25

ψ a modulating parameter. 93

ρ_0 the mean density. 26

ρ_s the characteristic density. 27

ρ_{crit} critical density. 13

$\mathrm{e_h}$ an intercept. 49

$\mathrm{s_h}$ a slope. 49

$\mathcal{Y}_\mathrm{h}$ the angle between x-axis and the distance from the origin ρ_h. 51

σ_{mv} the rms variance. 26

τ a detection threshold. 65

θ right ascension. 18

$\tilde{M}$ the final mass estimation after the merging processes. 104

$\tilde{N}$ a normalization term. 68

φ the angular size of an object. 16

b a linking length based on the comoving number density of galaxies $\bar{n}$. 38

c the speed of light. 11

c_{200} a concentration parameter w.r.t. over-density equivalent to 200 times the critical density of the Universe. 28

c_{500} a concentration parameter w.r.t. over-density equivalent to 500 times the critical density of the Universe. 28

d the total galaxy density. 86

f a fitting function. 26

g a galaxy. 63

m the apparent magnitude. 17

r a 2D radius. 26

r_s the scale radius. 27

r_{200} the characteristic over-density radius w.r.t overdensity equivalent to 200 times the critical density of the Universe. 28

r_{500} the characteristic over-density radius w.r.t overdensity equivalent to 500 times the critical density of the Universe. 28

v a velocity. 10

z a redshift. 5, 6, 11, 15

erg the unit of energy. 32

F a cumulative distribution function. 68

H a Hough space. 49

H_0 Hubble constant. 10

$M_\odot$ solar mass unit. 6

Ns the intensity of generated noise/points/galaxies. 61

S an image space. 49

T a multiplicative factor. 61

ASTROPHYSICAL TERMS

active galactic nuclei (AGN) Regions of extremely high luminosity at the centre of some galaxies. 31

AGN feedback An emitted radiation due to accretion onto a supermassive black hole. 2

angular size An angular measurement describes how large a celestial object appears from a given point of view. 16

baryonic matter All observed space components. 13

comoving distance The distance between two nearby objects moving according to Hubble flow and it remains constant with epoch. 14

comoving number density The number of objects (eg. galaxies) per comoving volume. 26

comoving volume The volume measure of the number densities of objects as if they are non-evolving and being entrapped into the Hubble flow with constant redshift. 14

cosmic microwave background (CMB) An electromagnetic radiation left over from an early stage of the Universe in Big Bang cosmology. 32

dark energy A theoretical form of energy postulated to act in opposition to gravity. 13

dark matter A non-luminous material postulated to exist in space. 13

filaments Narrow dispersions of galaxies which appear as threads throughout space and distribute orthogonally to the line of sight. 3

galactic winds Streams of high-speed charged particles often observed blowing out of galaxies. 31

Gravitational lensing A distribution of matter between a distant light source and an observer, which is capable of bending the light from the source as the light travels towards the observer. 31

halo It has different meanings. It means the spherical matter around a galaxy or a galaxy group/cluster. Also, it may represent more than one galaxy group/cluster or it can be used to describe a group of mock galaxies. 20

halo mass The mass of an assumed galaxy group or a halo . 26

halo occupation distribution (HOD) Provides a view of how galactic matter is distributed within each of the dark matter clumps. 29

Hubble flow All far galaxies appear to move away from us according to the expansion of the Universe. 10

light curve A graph showing the variation in the light received over a period of time from a variable star or other varying celestial object. 55

luminosity The total energy radiated from an object per second. 15

nebulae An interstellar cloud of dust and other ionized gases. Originally, nebula was a name for any diffuse astronomical object, including galaxies beyond the Milky Way. xviii, 11

redshift The Doppler effect in terms of light variations based on how far the celestial object is from the observer. 3, 4, 11

solid angle An angular area on the celestial sphere. It is expressed regarding two angular displacements. It is used to find the comoving volume. 18

star masking An obscuring of the light from celestial objects by a huge bright star. 40

the critical density The average density of matter required for the Universe to halt its expansion. 13

transverse comoving distance The comoving distance based on assumed shape of the Universe whether it is flat, hyperbolic or a sphere. 16

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND MOTIVATION

The basic active element of cosmic structure is the galaxy. The nature of gravity impels the galaxies to aggregate into groups. The galaxies orbit around an attracting centre (usually the most massive luminous galaxy). The formation of galaxy groups or clusters is supposed to happen according to the theory of inflation and hierarchical structure formation, beginning as over-dense fluctuations in the Universe and increasing and merging matter during cosmic history to become current huge structures (Wardlow 2010).

Astrophysicists are interested in detecting galaxy groups and clusters to study and analyse the evolution and formation of these kind of systems and to try to investigate more about their activities. In this chapter, we will mention some of the key points of detecting galaxy groups/clusters, identify the difficulties of detecting the pattern of interest, and present a brief overview of the following chapters.

1.1 The Importance of Detecting Galaxy Groups and Clusters

In general, galaxies tend to expand away from one another. However, in some parts of space, there can be an over-density of galaxies, which means that the gravitational field in these areas is sufficiently strong to prevent these galaxies from escaping from one another; and therefore, they remain bound and interact together, forming groups or clusters of galaxies.

Studying the corresponding interactions of the evolution and the formation of galaxies (such as star formation, stellar nucleosynthesis and AGN feedback) is paramount in understanding and obtaining more consistent descriptions of the cosmic structure and the environmental impact of these interactions on galactic and extragalactic levels (Liang et al. 2016). Galaxy groups and clusters play a significant role in explaining the evolution of the Universe and measuring its baryonic content. Moreover, they can signify the gravitational lenses and contribute to the estimation of cosmological parameters such as the variation in the density field in fixed physical scales. Also, the clusters act as laboratories to study the evolution of cluster properties and contents such as gases, shapes, colours and the star formation history of the member galaxy (Tyson et al. 1984). Furthermore, studying galaxy groups and clusters is instrumental in probing the history of the structure and the formation of galaxies; since clusters retain an imprint of how they were formed as well as providing a history of nucleosynthesis in the Universe (Mushotzky 2004).

The detection of galaxy groups is known as a very complex ill-posed problem. These groups spread within clouds of gases which prevents the distinction of the amount and the boundaries of these clusters; furthermore, it is difficult to assess if they are stable or unstable groups. There are some problems in detecting galaxy groups' boundaries, especially if they have overlapped with filaments (some scattered galaxies spread as threads in the field). To identify galaxy groups and clusters, astrophysicists generate mock data (a mock survey) simulating reality, due to the incompleteness of the real data that they can obtain from the telescopes. In the mock data, they know exactly the components of their simulations and the contents of the celestial systems (e.g. stars, galaxies, galaxy groups). In addition, they can produce as much data as necessary to analyse some phenomena.

Given the amounts of available survey data, automated discovery of galaxy groups and clusters with a degree of uncertainty is of utmost interest to astrophysicists. Furthermore, the introduction of ever more powerful computing techniques has enabled applications of machine learning which can make many contributions to astronomical dataset analysis (Kramer et al. 2013). If we compare the current clusters' identification techniques with Abell's technique (i.e. using the early photographic plate surveys, Abell applied some criteria [1] to obtain a homogeneous catalogue), these modern ways appear more reasonable, since they are automated and objective (Gal et al. 2003). Also, they have a logical selection function and suppose former minimal limitations on the features of the system to be identified (Koester et al. 2007).

[1] These criteria include counting the galaxy in a specific fixed physical area, selecting a minimum number of galaxies within 2 magnitude (see section 2.4) of the third brightest galaxy in a cluster, specifying a minimum and maximum redshift to the clusters.

Many large galaxy surveys have been conducted to identify galaxy positions in the sky and their recession (line-of-sight) velocities. However, there are difficulties in classifying galaxies into groups and clusters due to observational problems such as redshift distortion (especially in photometric observations), edge effects and bright stars masking regions (Duarte 2014). Some approaches have achieved success; however, they are very complex in terms of computational and statistical processing. Other heuristic methods, such as FoF (a standard method used to detect galaxy groups/clusters, see section 2.14), tend to be used to obtain fast outcomes but with a high false positive rate. Thus, the necessity to propose new probabilistic models appears. A rationale for building a probabilistic model is to give some uncertainty in a principled way, considering prior knowledge discerned from the local density and the theoretical notions of the galaxy groups' distribution as much as possible; then coming up with probabilistic answers. Furthermore, the benefit of applying a probabilistic approach is to have more flexibility in updating the model based on the new physical simulations and findings.

The main questions of this research can be presented as: "Can we identify galaxy agglomerations in a probabilistic way?"; "What are the pros and cons for this kind of detection?"; and "How is the performance compared with the modified FoF method?".

1.2 The Major Challenges in Galaxy Group Detection

The observer on Earth surveys the Universe over a certain patch of the sky specified through two angles - right ascension (RA) and declination (Dec) (θ, β thereafter).

Besides the spatial position in the sky (θ, β), the velocity of the object along the line of sight (LoS) can be deduced from the redshift (z) (see Eq.2.2). A typical example of the form of a galaxy survey (GAMA mocks: see section 2.12 for more details) is shown (as a 2-D slice) in Figure 1.1. The patterns of interest (galaxy groups/clusters) are weak signals because they are swamped inside a huge background/foreground environment of galaxies. They (i.e. galaxy groups/clusters) form prolonged features, due to the projection of their galaxy velocities, along the LoS. Thus, they are called the "fingers of god" (FoG). As an example, the mean positions of FoG patterns are marked by red points in Figure 1.1. The challenge is to detect patterns corresponding to the real galaxy groups' true positives (TPs), while reducing the detection of similar patterns formed by the fore/background and chance superposition (FPs).

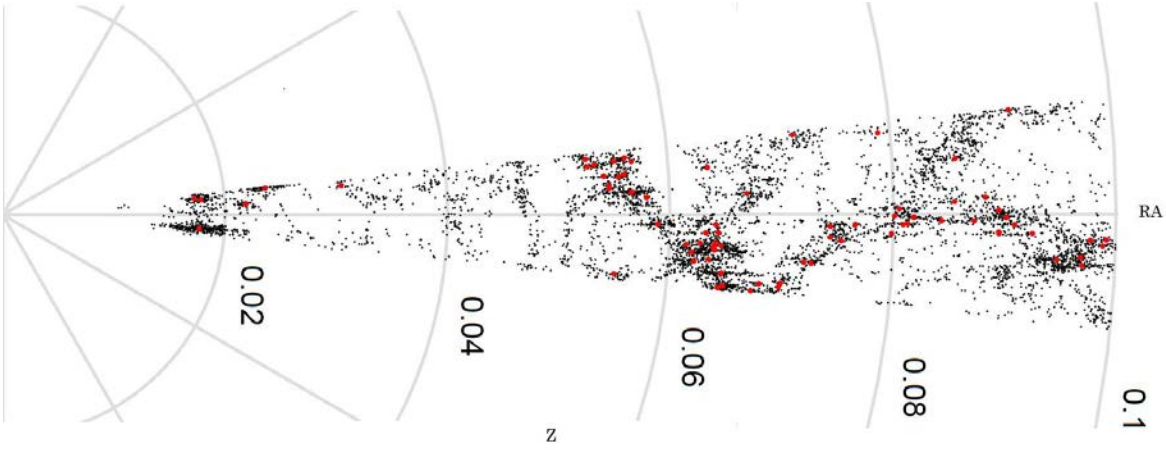


Figure 1.1: 2-D slice from a volume of GAMA mock data: RA vs Z. Red points signify the centre of FoG patterns

A new hybrid, theoretical plus data driven, model for detecting galaxy groups is needed due to the large amount of false detections in the previous techniques. In general, the existing galaxy group finders have many free parameters that need to be carefully set before applying the analysis. This raises issues regarding the gen-

erality of the results and the stability of the calibration process. A mere mixture distributions model will not be useful in this scenario because of the noisy foreground/background galaxies. This noisy environment makes the detection process of the pattern of interest infeasible. The main contributions of this work are building an adaptive probabilistic model to generalize the possibility of detecting groups and clusters of galaxies, learning from previous knowledge based on the expertise of astrophysicists. The input of the probabilistic model is the coordinate of each galaxy: the two angles (θ,β), redshift z. The output from the probabilistic model is a density mixture of voted galaxies belonging to a given galaxy group/cluster position as will be illustrated in detail in Chapters 4, 5 and 6.

Our research has been limited to finding galaxy systems (groups/clusters) within the close redshift ($0.01 < z \leq 0.1$) and mass (size) (10^{12} $M_{\odot}$ - 10^{15} $M_{\odot}$) .

1.3 Thesis Outline

The rest of the thesis consists of six chapters:

In Chapter 2, some astronomical concepts and terminologies related to the cosmology of the Universe and the generation of a mock galaxy survey are presented. Also, we explain some related works concerning detecting galaxy groups in a probabilistic way, in addition to other techniques that have been used in detecting galaxy groups and clusters.

In Chapter 3, we demonstrate some machine learning concepts relevant to the Hough transform and its ability to detect objects in a noisy environment. Examples

of the utilisation of the Hough transform in astronomy and in detecting spherical shapes are presented.

Chapter 4 explains the basic formulation of the probabilistic Hough transform model (PHTM) and the preliminary data generations and modifications of the method in simple 2D data and a 2D model experiment. Furthermore, we illustrate the basic 3D mock data and the preliminary 3D model.

In Chapter 5, the generation of complete and well-defined realistic 3D data is discussed, by including the flux limit effect and applying the 3D probabilistic Hough transform based on adaptive local kernel (PHTALK); which is the updated version of the PHTM.

Chapter 6 presents the application of the modified PHTALK on the most sophisticated realistic GAMA mock data and compares the results with two versions of the FoF method. In addition, we illustrate the problem of FP and demonstrate some suggestions to reduce them. Furthermore, we investigate the recovered groups and their properties such as galaxy group members and galaxy groups' velocity and mass estimation.

Chapter 7 delivers the conclusion and suggests some future work.

1.4 Publications

A part of Chapters 4 and 5 contains a paper (Ibrahem et al. 2015) which was submitted to the ICONIP-2015 conference and was published by Springer in Neural

Information Processing, volume 9491 of the series Lecture Notes in Computer Science, pp 323-331 with the title "Automated Detection of Galaxy Groups Through Probabilistic Hough Transform". We were awarded 'best paper' in the College of Engineering and Physical Sciences at the University of Birmingham in December 2015 for this work. Chapter 6 is a paper which is to be submitted to the MNRAS Journal (in prep.). My own contributions are around 80% of the papers.

CHAPTER 2

ASTRONOMY BACKGROUND AND TERMINOLOGY

Isaac Newton and Richard Bentley were pioneers in trying to describe and understand the Universe, based on physical laws. However, Newton realised that his gravitational law could not fully describe a homogeneous, static and isotropic Universe on a cosmological scale (Janiak 2009). In 1916, Albert Einstein proposed a theory of the Universe based on the general theory of relativity, but he could not find a static solution for the Universe either(Einstein 1916, Figueiro Spinelli 2011). Hence, he introduced the cosmological constant Λ as a force that acts against gravity. Later, Alexander Friedmann found a solution for Einstein's field equation that described the expansion of the Universe. Then Hubble's observations, in 1928, of the distance and recession velocity of galaxies confirmed Friedmann's findings (Hubble 1929). This evidence of the expansion of the Universe was a spark that motivated new studies of the evolution and structure of the Universe, based on observations and particle physics.

The aims of this chapter are to outline some key concepts in cosmology; to

illustrate some properties of large celestial systems (galaxies, groups, clusters); and to present a brief overview of some works on detecting galaxy groups and clusters. These cosmology concepts will be employed in subsquent chapters for generating galaxy mock data surveys and generating a new probabilistic model to improve the detection process.

2.1 The Expansion of the Universe

In 1928, Edward Hubble identified a linear relation between the velocity v of distant galaxies, which can be measured using the Doppler shifts of spectral lines, and their distance D from Earth can be calculated as

$$v = \mathrm{H}_0 D, \tag{2.1}$$

where v is measured in km/sec; D in Mpc; and H_0=69.7 which is the Hubble constant in $\mathrm{km\,s^{-1}Mpc^{-1}}$; this is a measure of the slope of the line through the distance versus recession velocity data. Eq. (2.1) is utilised to estimate the age and the size of the Universe. Also, it can be used to estimate the mass and the intrinsic brightness of the stars in the nearby galaxies (i.e. local Universe). The zero in the Hubble constant refers to the current time, because H_0 changes with time (Hogg 1999).

In Figure 2.1, the line passes through the origin point, which represents the Earth at zero distance and zero speed. Hubble concluded that all far galaxies appear to move away from us according to the expansion of the Universe, which is called **'Hubble flow'**. By this movement, the wavelengths will stretch according

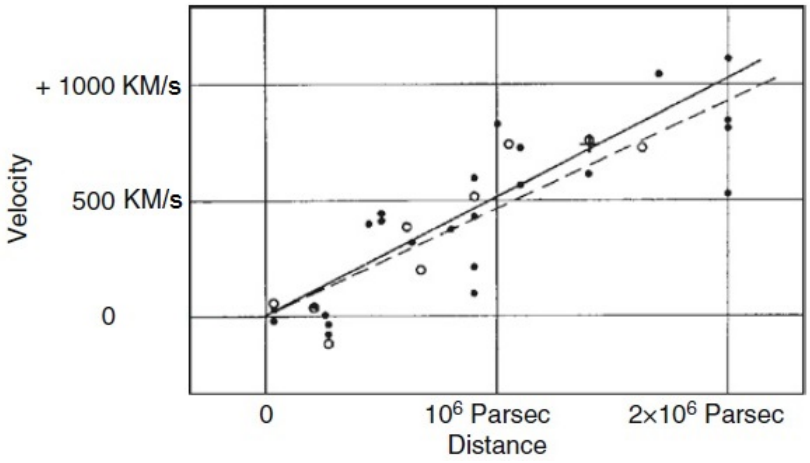


Figure 2.1: The Hubble diagram shows the radial velocity of galaxies vs. their distance. "*The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.*" (Hubble 1929).

to the Doppler effect and reach the earth.

The change in the wavelength between the observed λ_{ob} and emitted λ_{em} wavelengths is annotated as $\Delta\lambda$ or λ_{ob}-λ_{em}. The difference between the emitted and observed wavelength object photons is called **'redshift'** z,

$$z \approx \frac{\Delta\lambda}{\lambda_{em}} \approx \frac{\lambda_{ob} - \lambda_{em}}{\lambda_{em}} \approx \frac{\upsilon}{c}, \tag{2.2}$$

where c is the speed of light, equal to 299792.458 km s^{-1}. As shown in Figure 2.2, when the object (the emitter) is moving away from the observer, the emission and the absorption features of that object's wavelength will appear shifted toward the red end of the spectrum (Hogg 1999).

The redshift of a galaxy is considered as a proxy for the third spatial dimension

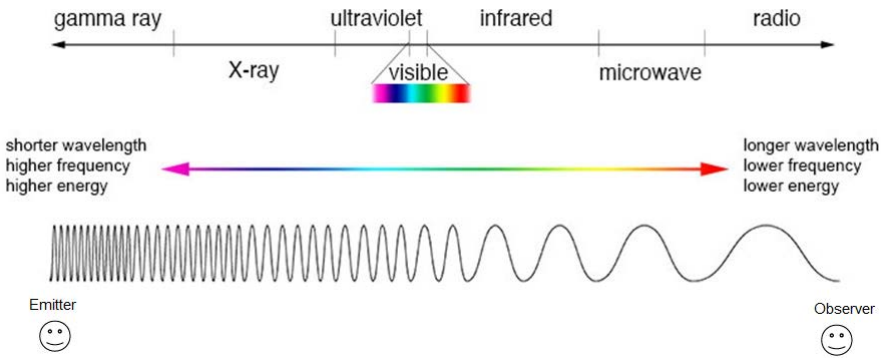


Figure 2.2: The electromagnetic spectrum and the redshift phenomena.

D which results from the combination of two velocities v: the recessional velocity (due to the expansion of the Universe), and the local peculiar velocity (motion of a galaxy relative to the galaxy group/cluster frame) of the galaxy (Miller 2012).

Hubble also compared the recessional velocities of the galaxies with their apparent magnitudes (i.e. brightness) and found that the fainter, smaller galaxies have higher redshifts (Hubble 1929).

2.2 The Current Cosmology Paradigm

The current theoretical paradigm of hierarchical structure formation in the Universe is the Lambda cold dark matter(ΛCDM) model which depends on two assumptions: the evolution of the matter in the Universe can be interpreted by the general relativity theory; and over a large scale (>100 Mpc) the Universe is isotropic and homogeneous (Kurek & Szydlowski 2008, Ostriker & Steinhardt 1995, Peebles 1980, Turner 1997). ΛCDM, along with the theory of cosmic inflation, explains the initial conditions of structure formation and predicts its hierarchical

nature due to gravitational instability. Observations have confirmed the hierarchies of the formation of the Universe (Tegmark et al. 2006). As small objects crash into one another, the hierarchies begin to form self-gravitated hot gases, which condense into large gravitationally bound systems such as galaxies, then galaxies agglomerate to create groups and clusters as a result of cosmic processes (Tegmark et al. 2004). This hierarchical nature is the core aspect of all cosmic formation of structures and their evolution (Press & Schechter 1974). In ΛCDM, the Universe is flat and it is composed of around 4% baryonic matter, 20% dark matter, and 76% dark energy.

The necessary amount of matter for the Universe to be flat can be computed from Friedmann's equation (Friedman 1922):

$$\rho_{crit} = \frac{3\mathrm{H(z)}^2}{8\pi G}, \tag{2.3}$$

which is called the **'critical density'**, where H(z) is the Hubble constant at redshift z and G is the gravitational constant. The density ratio of the Universe at redshift z to the critical density $\rho_{crit}(z)$:

$$\Omega_{\mathrm{M}} = \frac{\rho(z)}{\rho_{\mathrm{crit}}(z)} = \frac{8\pi G\rho_{\mathrm{M}}}{3\mathrm{H}_0^2},$$

which is a dimensionless quantity known as the **'density parameter'**. Through multiple contributors (related to the evolution of the Universe), astrophysicists determine the geometry of the Universe whether it is homogeneous, isotropic and matter dominated. Including the matter parameter (ordinary mass plus dark matter, Ω_M), the other contributors are , radiation (Ω_R), curvature (the flatness of the Universe, Ω_K), and the cosmological constant (Ω_Λ), which includes the effective

mass density of dark energy. They are defined by the following equations (Hogg 1999, Serjeant 2010):

$$\Omega_R = \frac{8\pi G\rho_R}{3\mathrm{H}_0^2},$$

$$\Omega_\Lambda = \frac{\Lambda c^2}{3\mathrm{H}_0^2}, \text{and}$$

$$\Omega_K = -\frac{Kc^2}{a^2\mathrm{H}_0^2},$$

where $\rho = \rho_M + \rho_R$ is the current mass density. The total density:

$$\Omega_{total} = \Omega_R + \Omega_M + \Omega_k + \Omega_\Lambda = 1.$$

The Friedmann equation can be written in terms of density parameters as

$$\frac{\mathrm{H}(z)}{\mathrm{H}_0} = \sqrt{\Omega_R(z)^4 + \Omega_M(z)^3 + \Omega_k(z)^2 + \Omega_\Lambda},$$

where $(1 + z)$ is the redshift at the present day (it is the ratio of the size of the Universe today to its size at redshift z). Supposing a flat Universe, both the curvature Ω_K and the radiation Ω_R parameters become negligible. Thus, the scaling parameter of the evolution E will be

$$E(z) = \frac{\mathrm{H}(z)}{\mathrm{H}_0} = \sqrt{\Omega_M(z)^3 + \Omega_\Lambda}, \tag{2.4}$$

which will be used in sections (2.3.1 and 2.6) to find the comoving distance and comoving volume.

2.3 Cosmological Distances

Distance measures are used to obtain the distance between two objects or events in the Universe. They are utilised to infer some non-directly observable quantities from those which are observable. There are different ways to find the distance between two celestial objects in the Universe. Astrophysicists specify which to use based on the available parameters of the object of interest, such as its physical size or intrinsic luminosity. Some of the distance measures that were used in our simulation are listed as follows,

2.3.1 Comoving Distance

The distance between two nearby objects in the Universe remains constant with epoch if both objects are moving according to the Hubble flow; this is called the **'comoving distance'**.

The comoving distance D_c from the Earth to a distant object is found by integrating small εD_c contributions between them along the radial ray from $z = 0$ (our position) to the z of the object's position (Peebles 1993), expressed as:

$$D_c = D_H \int_0^z \frac{dz'}{E(z')}, \tag{2.5}$$

where D_H is the Hubble distance,

$$D_H = \frac{c}{\mathrm{H}_0} = 3000 h^{-1}\mathrm{Mpc}; \tag{2.6}$$

E is the evolution scaling parameter Eq. (2.4); and $\frac{dz}{E(z)}$ is proportional to the

time-of-flight of a photon travelling through the redshift interval dz divided by the scale factor at that time. The comoving distance is the proper distance divided by the scale factor, because the speed of light is constant (Hogg 1999).

2.3.2 The Angular Diameter Distance

If the actual physical size χ of the object of interest and its angular size φ are known, then astrophysicists tend to use the angular diameter distance D_a:

$$D_a(z) = \frac{\chi}{\varphi}, \tag{2.7}$$

to find the object distance; D_a has an inverse relation with angular size φ; D_a increases until $z \sim 1.5$, then turns over and decreases as $z \to \infty$. Thus, the objects with the lowest angular size are the most distant. This is also related to the transverse comoving distance D_m:

$$D_a = \frac{D_m}{1+z}. \tag{2.8}$$

Assuming the flatness of the Universe $\Omega_K = 0$, then $D_m = D_c$ (Hogg 1999).

2.3.3 Luminosity Distance

Another way to find the distance from one object to another is to use the luminosity distance (D_L), which is how far away the object is in Euclidean space. The D_L is calculable if astrophysicists can measure the total amount of flux (F) emitted by an object in energy per area per time, and its luminosity (L) in energy per time

(Hogg 1999) using

$$D_L = \sqrt{\frac{L}{4\pi F}}.$$

D_L is also related to the comoving transverse distance D_m, through Etherington's reciprocity relation, and related to the radial comoving distance D_c:

$$D_L(z) = (1+z)D_m = (1+z)D_c. \tag{2.9}$$

2.4 Absolute and Apparent Magnitude

The apparent magnitude m is the brightness measurement of an object as seen by an observer; the brighter an object, the lower its magnitude. The scale is backward and logarithmic. The value of the apparent magnitude is adjusted by considering the absence of the Earth's atmosphere. An object's brightness differs based on its distance from the observer; an extremely bright object looks dim if it is far away. The term brightness is another way to say the flux of light, in Watts per square metre, coming towards us and varies inversely with the square distance. If two objects have the same apparent magnitude, there are two possibilities:
1- The objects are the same distance from the earth.
2- They are of different distances with a different value of luminosity (i.e. the highly luminous object is located far away from the earth, while the less luminous is located close to the earth) as shown in Figure 2.3.

The absolute magnitude or the real magnitude M is the apparent magnitude

that the object would have if it is positioned at a distance of 10 parsecs (32.6 light-years) from the Earth. To compare the intrinsic brightness of celestial objects regardless of their distances, astrophysicists tend to convert apparent magnitude to absolute magnitude. This conversion is based on the distance from the object to the observer using the following relation:

$$m - M = 5 \log_{10}(D_L(pc)/10(pc)) \tag{2.10}$$

where $m - M$ known as the **'distance modulus'** and D_L is the luminosity distance(Bradt 2007, Schneider 2014).

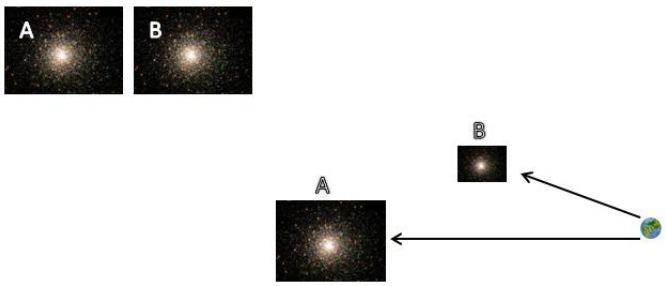


Figure 2.3: Left: group A and B have the same apparent magnitude; Right: however, group A is more luminous and a farther distance than B from the Earth.

2.5 The Solid Angle

The solid angle Ω is an angular area on the celestial sphere. It is expressed regarding two angular displacements: the RA or θ and Dec or β. The hatched area in Figure 2.4 represents the solid angle and can be declared in 'square degrees' or 'square radians' units; the latter are called steradians (sr). The solid angle will be used in Chapter 5 to generate the data in a specific comoving volume (see section

2.6).

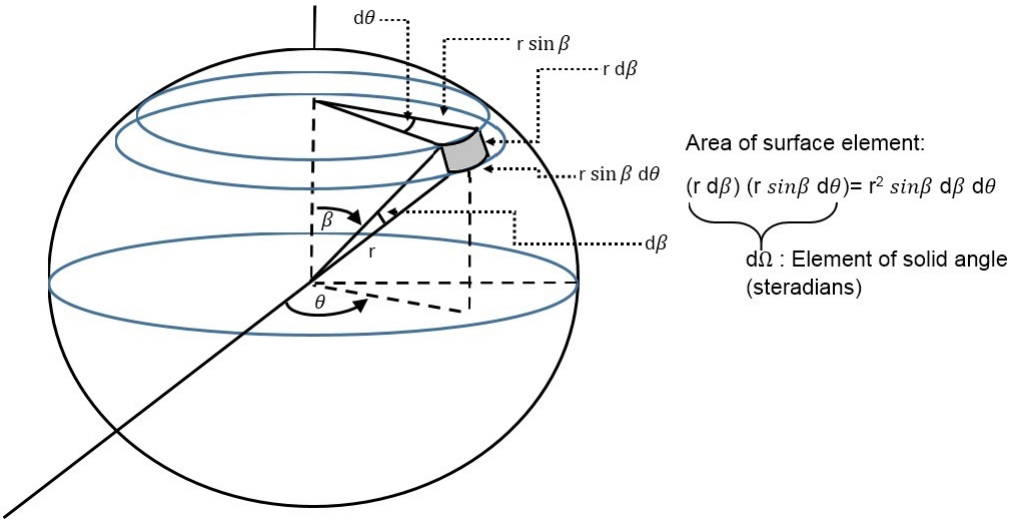


Figure 2.4: The concept of solid angle,Ω, which is the 'angular' area on a sphere. The element of solid angle dΩ can be declared in terms of steradians (sr) which is equivalent to an element of surface area on a sphere of unit radius (Bradt 2007).

The unit of the solid angle (sr) is a dimensionless quantity of magnitude 1rad × 1rad where 1 rad= $\frac{360}{2\pi} = 57.3°$. The equivalent number in square degrees is:

$$1sr = \frac{360}{2\pi} \cdot \frac{360}{2\pi} = (57.3)^2 = 3282\,\mathrm{deg}^2$$

To obtain the correct solid angle, integration over the area of the sphere should be conducted. However, for small solid angles ($\leq$ 100 deg2 = 0.03 sr), the relative solid angle in steradians is gained through calculating the area of the sky, in (sr), as if the piece of sky was a flat piece of paper (Bradt 2007).

2.6 Comoving Volume

The comoving volume(V_c) is the volume measure of the number densities of objects as if they are non-evolving and being trapped into the Hubble flow with constant redshift. The V_c element per solid angle $d\Omega$ element and redshift interval dz is:

$$dV_c = \frac{D_H \cdot (1+z)^2 \cdot D_a^2 \cdot d\Omega \cdot dz}{E(z)} \tag{2.11}$$

and the total comoving volume (V) is the integral of dV_c from z_{min} to z_{max} (Hogg 1999):

$$V = \int_{z_{min}}^{z_{max}} dV_c(z)\ dz \tag{2.12}$$

2.7 What is a Galaxy?

A galaxy is a dynamically bounded system[1] of stars, stellar remnants, planets, gases, and dark matter. It has a different number of stars, ranging from 10^7 stars as in dwarfs, to 10^{14} stars as in massive galaxies. The schematic structure of a galaxy is shown in Figure 2.5; which illustrates what a galaxy comprises: the disk which contains the spiral arms, the halo, and the nucleus or central bulge. Also, it contains at least three other components that are "invisible": the galactic magnetic field; charged particles trapped in the galactic magnetic field; and a halo of **dark matter** that is of unknown composition but that makes itself felt through its gravitational influence on the visible matter (Amores 2011, Robin et al. 2003). Most likely this kind of galaxy is shaped as a spiral, or barred spiral which

[1]The change of patterns, of galaxies or galaxy groups, will take much long time because these systems are far away from us. Thus, the detection process is relatively stable and is not been affected.

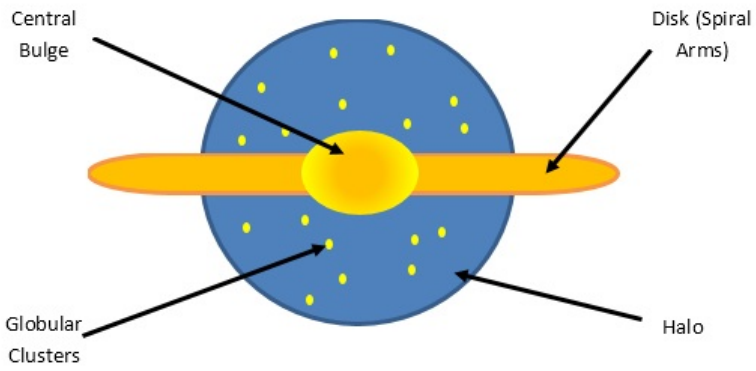


Figure 2.5: Schematic side view of Milky Way galaxy (Amores 2011).

is younger and has a more recent star formation than other types. There are different galaxy morphologies: such as the elliptical shape galaxy, which is the oldest passive type and it has no clear substructure; the irregular, which has a weak or no structure; and the lenticular, which is a transition between the spiral and elliptical morphology (Dressler 1984, Stott 2007).

It has been shown that the older galaxies occur in a denser environment compared to the younger galaxies; this is based on the colour-density relation, which leads to a stronger clustering of older galaxies. Also, the galaxies with a high star formation rate (SFR) tend to be more clustered than those with a lower SFR. In both elliptical and spiral galaxies, the degree of clustering among them depends on their size, and how many types of galaxies are gathered together(Frost 2010).

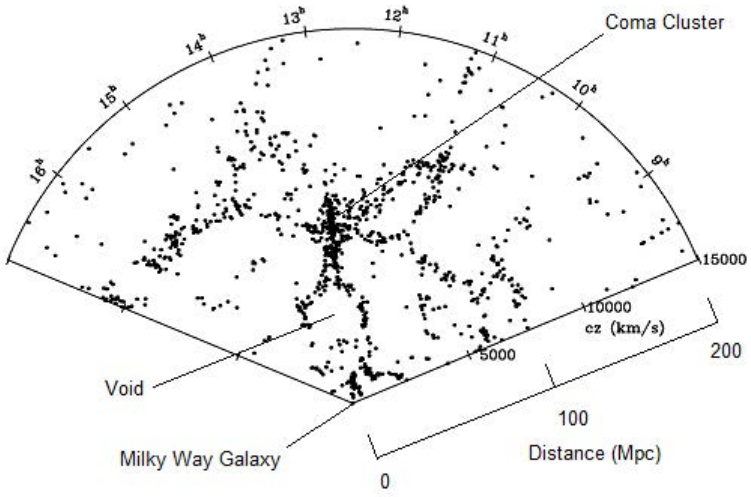


Figure 2.6: The First Slice Universe Survey - contains 1057 galaxies out to approximately 200 Mpc distance. It illustrates non-random distributions of groups and clusters throughout the field; rather, they form a filamentary structure around nearly empty voids. The distances shown assume $H_0 = 75$ km/s/Mpc (Huchra 1988).

2.8 What are Galaxy Groups/Clusters and Their Properties?

The basic active element of cosmic structure is the galaxy. The nature of gravity impels the galaxies to aggregate into groups. The galaxies orbit around an attracting centre (usually the most massive luminous galaxy). Figure 2.6 (Huchra 1988) illustrates the distribution of galaxies for the closest redshift $Z < 0.05$; they are not randomly distributed throughout the space, but tend to aggregate along filaments in clusters and groups (Snaith 2011).

The smallest virialized system of aggregation in the cosmic hierarchy is called a galaxy group. Galaxy groups usually contain around 2-50 galaxies and have a mass of around 10^{12}- 10^{14} $M_\odot$ (Huchra & Geller 1982). The majority of galaxies

(around 70%) in the local Universe are placed in groups (O'Sullivan et al. 2014). Groups normally have a radius r between 0.5 and 1.5 Mpc and a velocity dispersion σ_v <500 km s^{-1} (Saviane et al. 2007).

Galaxy groups can be found in different structural states(i.e. virialized and non-virialized systems). Poor systems with few numbers of galaxies can be identified by optical detection methods; whereas, some galaxy groups are X-ray luminous systems. This kind of group can be detected in X-ray observations if they have enough galaxies with hot intra-group medium. Galaxy groups can be detected in different redshift bands. Some galaxy groups can contain early and late types of galaxies. Due to the low velocity dispersion of galaxy groups, the interactions between galaxy group members creates a huge impact through galaxy evolution processes, such as merging and transformation.

A substantial part of the baryons in galaxy groups contains hot, diffused gas and it is supposed that galactic interactions would have a significant impact on this gas (Liang et al. 2016). Many studies have found that galaxy groups are more likely to be in rich environments. Thus, astrophysicists are more interested in analysing the relations between the galaxies and their neighbourhood systems (Dressler 1984, Erfanianfar 2014, Martinez & Muriel 2006).

Clusters are the largest gravitationally bounded system in the Universe (Malin 2001). As in some types of galaxy groups, clusters contain three components: galaxies, gas and dark matter. The galaxies contribute as a very small proportion $(1 - 2\%)$ to the cluster mass. The rest is a diffused hot gas $(5 - 18\%)$ called intergalactic medium (IGM) and unseen component (dark matter 80%) (Sarazin

1986). Galaxy clusters typically contain between 50-1000 galaxies and have a mass range of 10^{14}-10^{15} $M_{\odot}$. The core of the cluster, the centre, is dominated by a huge elliptical galaxy, which is called the brightest cluster galaxy (BCG) (Santos, J. S. et al. 2009). Clusters are usually defined with a radius r range of (1-3) Mpc and a velocity dispersion σ_v range of (800-1400) $\mathrm{km}s^{-1}$ (Murdin 2001).

The detection of galaxy groups is more difficult and they look fainter than galaxy clusters, especially in the high redshift. Compared to galaxy clusters, galaxy groups are lower in density contrast against galaxies in the field (Knobel 2011).

Astrophysicists have further classified groups and clusters of galaxies into virialized and non-virialized. Virialized clusters are characterized by richness, as well as having a symmetrical spherical shape. Galaxies are more condensed in the core, which comprises of large elliptical galaxies. Mostly, these kinds of clusters contain elliptical and lenticular-shaped galaxies. Meanwhile, non-virialized clusters are more likely to be scattered and disorganized and furthermore, lack the condensed core of the virialized type. Although they contain smaller numbers of galaxies, they are characterized by the diversity of the galaxies morphologies. However, the spiral-shaped galaxies are more dominant.

2.9 Luminosity Function

The luminosity function is an observational quantity that specifies a member distribution of a class of objects based on their luminosity (Schneider 2014, Stahler & Palla 2008). Moreover, it is used to estimate the luminosity for each particular object (e.g., star, galaxy). It is used to optain information related to primor-

dial density fluctuation; processes that change one type of galaxy into another; processes that create or destroy galaxies; and processes that transform mass into light. Also, the approximated distribution of a galaxy can be obtained using the Schechter luminosity function.

2.9.1 Schechter Function

In the early 1970s Press and Schechter calculated the mass distribution of groups, later Schechter applied this function to fit the luminosity distribution of galaxies in Abell clusters. There are two versions of the luminosity function: $\phi(L)$ per dL and $\phi(M)$ per dM, where L is the luminosity and M is the absolute magnitude of the celestial object, as shown in Figure 2.7 (Schechter 1976). The Schechter function in terms of luminosity can be written as:

$$\phi(L)dL = \phi^*(\frac{L}{L^*})^\alpha \exp(-\frac{L}{L^*})d(\frac{L}{L^*}),$$

where L^* is the characteristic luminosity that separates the low and high luminosity parts; α is the faint end slope of the luminosity function for small L, and ϕ^* is the number density (the normalization of the distribution). At low luminosity ($L < L^*$), there is a power law ($\phi \propto L^\alpha$) where $\alpha \sim$ -0.5 $-$ -1.5 (flat to steep); while at higher luminosity ($L > L^*$) there is an exponential cut-off ($\phi \propto e^{-L}$) that means very luminous galaxies are very rare. The previous equation can be written in terms of magnitude, $\frac{L}{L^*} = 10^{0.4(M^*-M)}$, where M^* is the characteristic magnitude.

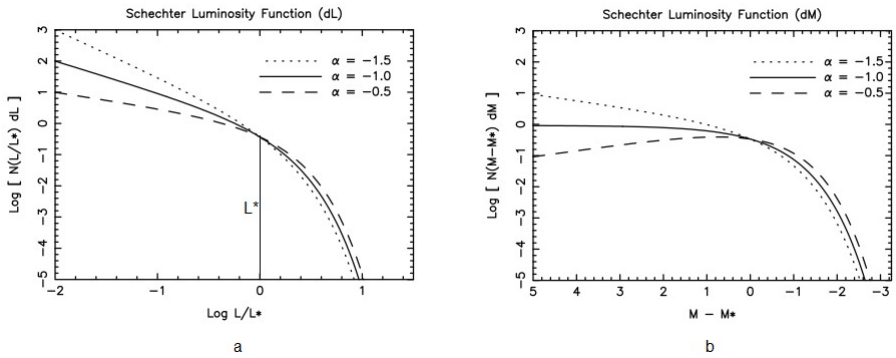


Figure 2.7: Sketch of the Schechter luminosity function.

2.10 The Halo Mass Function (HMF)

The theories of the cosmological structure formation assume that the dark matter forms into a massive gravitationally bound structure called halos. The halo mass function (HMF) quantifies the number of the halos per unit comoving volume of the Universe as a function of their mass. The HMF is affected by the cosmological parameters Ω_M, Ω_Λ, in addition to the nature of the dark matter. The comoving number density of halos per unit logarithm of the halo mass M_h was introduced by Press & Schechter (1974):

$$\frac{dn}{dlnM_h} = \frac{\rho_0}{M_h}.f(\sigma_{mv}).\left|\frac{dln\sigma_{mv}}{dlnM_h}\right| \tag{2.13}$$

where ρ_0 is the cosmology dependent mean density of the Universe, σ_{mv} is the rms variance of the mass within a sphere of radius r containing mass M_h and $f(\sigma_{mv})$, where f represents the functional form that defines a particular HMF fit (fitting function). Many researchers have refined the form of f to produce a better match to cosmological simulations. An on-line web-based tool (HMFcalc) has been used to calculate $\frac{dn}{dlnM}$ (Murray et al. 2013) [1].

[1] http://hmf.icrar.org/

For generating our mock data in Chapter 5, Reed's fitting function $f(\sigma_{mv})$ has been chosen, which has been improved and modified from the Sheth-Tormen (S-T) mass function (Reed et al. 2007, Sheth & Tormen 1999) with mass range (10^{12}-10^{15}) $M_{\odot}$; mass bin width 0.05; redshift range 0.01- 0.1; and cosmological parameters Ω_M=0.28, Ω_Λ=0.72, Ω_K=0 and H=69.7 km/s/Mpc.

2.11 The Radial Distribution (The Surface Mass Density)

Navarro, Frenk and White (NFW) found that the density profile of dark matter halos can be described by a universal two - parameter function over a wide range of halo masses (Navarro et al. 1996):

$$\rho(R) = \frac{\rho_s}{\xi(1+\xi)^2}, \text{ where } \xi = \frac{R}{r_s},$$

and ρ_s is the characteristic density; r_s is a scale radius; and both are mass-dependent scaling parameters. R is the 3D radius from the centre of an expected group. The projected NFW profile derived by (Bartelmann 1996) is used to describe the surface density of galaxy groups and clusters:

$$\Sigma(r) = \frac{2\rho_s r_s}{\left(\frac{r}{r_s}\right)^2 - 1} \; f\left(\frac{r}{r_s}\right), \tag{2.14}$$

where r is the 2D radius and by using $x = r/r_s$:

$$f(x) = \begin{cases} 1 - \dfrac{2}{\sqrt{x^2-1}} \tan^{-1}\left(\sqrt{\dfrac{x-1}{x+1}}\right) & : x > 1 \\ 1 - \dfrac{2}{\sqrt{1-x^2}} \tanh^{-1}\left(\sqrt{\dfrac{1-x}{1+x}}\right) & : x < 1 \\ 0 & : x = 1 \end{cases}, \tag{2.15}$$

and

$$\rho_s = \frac{200}{3} \frac{c_{200}^3}{[\ln(1+c_{200}) - c_{200}/(1+c_{200})]} \tag{2.16}$$

with $c_{200} = r_{200}/r_s$, $r_{200} = r_{500}/0.67$ (Liang et al. 2016),

$$r_s = \frac{r_{500}}{c_{500}}. \tag{2.17}$$

Here, c_{200} and c_{500} are the concentration parameters. c_{500} can be calculated:

$$c_{500} = 29.1 M_{500}^{-0.091} (1+z_k)^{-0.44} \tag{2.18}$$

and r_{200} and r_{500} are the characteristic over-density radii, r_{500} can be defined as:

$$r_{500} = \sqrt[3]{\frac{M_{500}}{500 \cdot \frac{4}{3}\pi \cdot \rho_{crit(z_k)}}}, \tag{2.19}$$

where $\rho_{crit(z_k)}$ is the critical density of the Universe at redshift (z_k) as shown in Eq. (2.3). M_{500} is the mass w.r.t. over-density equivalent to 500 times the critical density of the Universe.

2.12 Galaxy and Mass Assembly (GAMA) Mock Survey

The GAMA mock survey style is one of the major data product for the GAMA project which was constructed by Peder Norberg and various collaborators in Durham. It is a 3D more realistic simulation constructed from a large numerical simulation of cosmic structure formation (cosmological N-body simulation). In the simulation the galaxies are attached to the dark halos according to a halo occupation distribution (HOD). The outcome from the simulation then can be utilized as a mock survey, which is comparable with the real observations. Thus, the distribution of dark halos is known (Liu et al. 2008).

The aims of generating GAMA mock survey are to study the evolution and formation of galaxy systems, test galaxy systems' grouping quality and furthermore, due to the need for mock data to cover multi-wavelength bands (i.e. ultraviolet, infrared, and radio frequencies), this makes GAMA an important data mock survey for combining a varied collection of galaxy properties.

The GAMA survey spans across three equatorial fields measuring 12×5 deg^2 centred at RA=9h, Dec=0.5° (G09), RA=12h, Dec=-0.5° (G12) and RA=14.5h, Dec=-0.5° (G15); spectroscopic coverage is m=19.8 magnitude, which helps to discover more galaxies at high redshift(z). A sample of GAMA data is presented in Figure 2.8. The characteristics of the GAMA survey are given in Driver et al. (2011); while the survey input catalogue was described in Baldry et al. (2010) and the spectroscopic tiling algorithm in Robotham et al. (2010) (Robotham et al.

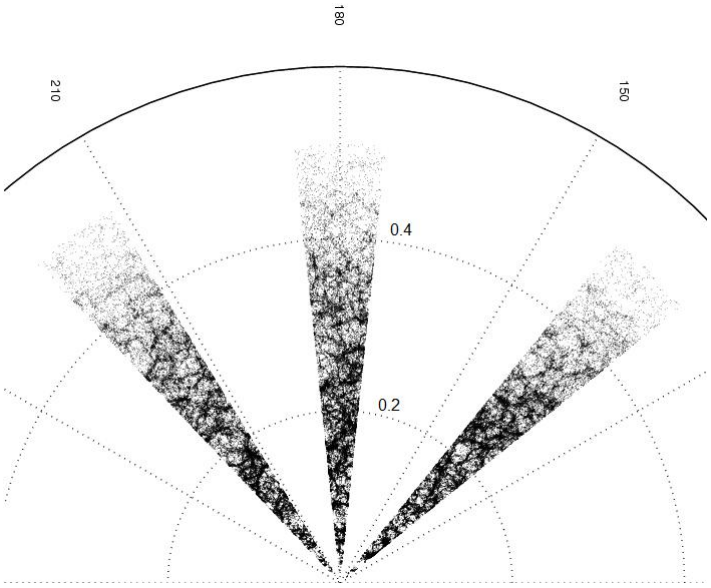


Figure 2.8: The right ascension vs. redshift (z) of volume no.1 of GAMA mock survey; the galaxies spread till z=0.5.

2011).

In Chapter 6, we test our model on the GAMA data and compare the outcome (i.e. predicted galaxy groups) with the galaxy groups identified via two FoF versions by Eke et al. (2004) and Robotham et al. (2011) [1]. We have nine volumes of GAMA. Each volume was divided into three cones for simplicity with redshift (z) cut-off till (0.1); as a result, we have 27 cones in total.

2.13 Identifying Galaxy Groups/Clusters

Detecting galaxy groups and clusters helps astrophysicists to explain the origin of the Universe and to understand how galaxies can be formed and evolve over time. Galaxy groups have a direct environmental impact on galaxies and are the direct

[1]The FoF methods have been implemented by astrophysicists.

feedback receiver of the galaxies' gravitational potential, such as galactic winds, active galactic nuclei (AGN) and radiation. Furthermore, finding galaxy groups and clusters helps to study and measure cosmological relations such as mass to light ratios. Thus, mock data surveys are also necessary to validate cosmological simulations and estimate cosmological parameters (Tempel et al. 2016).

There are different methods of detecting galaxy groups/clusters, which can be categorized according to the data type used; there are four main categories.

2.13.1 Gravitational Lensing (GL)

Gravitational lensing represents the appearance of galaxies, which are located in the background of a giant galaxy cluster as an arc pattern. In other words, galaxy cluster systems can affect and bend the light of a background celestial source, due to their strong gravitational effect in which some copies are produced from the same source in the background as an arc like pattern. This gravitational lensing (GL) is mainly used to detect huge systems such as galaxy clusters, especially in the large redshift range. Gladders et al. (2003) used strong lensing to detect a few systems. Weak lensing also has been applied to detect galaxy clusters and their masses at high redshift; only huge clusters with a mass $>10^{14}$ $M_{\odot}$ can be identified due to the cross section for the gravitational lensing falling consecutively with z. Weak lensing provides a way to check the shape and the profile of the system at relatively large radii. Some researchers have concluded that the utilization of weak lensing and the strong galaxy lens observations close by the groups could help in estimating the total mass and gain accurate information about matter distribution in the groups (Moller et al. 2001, Oguri et al. 2009).

2.13.2 Sunyaev-Zeldovich (SZ) Effect

The Sunyaev Zeldovich (SZ) effect is the deformation of cosmic microwave background (CMB) radiation photons due to the hot gas intra-cluster medium (ICM). The SZ effect is an efficient method for identifying clusters at high redshift; since it varies the magnitude based on the mass of the cluster, regardless of the redshift. However, the SZ signal is limited to higher mass clusters due to confusion of the SZ signal with the background, which may affect the brightness and reduce the possibility of identifying the mass of galaxy groups (Connelly 2012, Holder et al. 2007, Johansson 2011, Reichardt et al. 2013, Sunyaev & Zeldovich 1970, Vikhlinin et al. 2014).

2.13.3 The Radiation of X-ray from Hot Intra-group/cluster Medium (IGM/ICM)

The signifying of X-ray emissions from galaxy groups/clusters was one of the important discoveries of the UHURU X-ray satellite (Schneider 2014). Recently, many groups/clusters of galaxies have been observed in X-ray through satellites' telescopes such as ROSAT and Chandra (Jones et al. 2014, Rosati et al. 2002). In addition to the active galactic nuclei (AGN), galaxy clusters are the most luminous X-ray sources. The luminosity of galaxy clusters is between 10^{43} and 10^{45} erg/s; while galaxy groups' luminosity is between 10^{41} and 10^{43} erg/s. However, there is a possibility that the detection is biased towards galaxy groups with rich IGM; it may not be efficient to detect a wide spectrum of fainter galaxy groups which are dominant in the Universe. The X-ray photons are detected for a region of a size around 1 Mpc (i.e. the X-ray photon from a single galaxy cannot be detected)

(Mirkazemi 2014).

2.13.4 Redshift Surveys

Groups and clusters of galaxies can be detected as overdensity systems in redshift surveys (Berlind et al. 2006, Duarte & Mamon 2015). The redshift surveys allow searching for systems in two (RA and Dec) plus redshift (related to the distance) dimensions. These surveys are more accurate spectroscopy but cheaper and faster to obtain photometry (Miller 2012). They can be built from the ground and so are easier to obtain than X-ray surveys. There are three main classes of methods for detecting celestial systems based on the redshift surveys characteristics (Ascaso et al. 2012):

Geometrical methods

One of the common approaches for galaxy groups/clusters' detection is friends of friends (FoF) method (Eke et al. 2004, Huchra & Geller 1982, Robotham et al. 2011, Tempel et al. 2016). Galaxies are considered within the same group and linked together if they are within a particular linking length. The linking length is specified according to the typical overdensity of the galaxies within groups. Several approaches have been proposed to determine the linking length and to measure local galaxy densities (Berlind et al. 2006, Ramella 1989). Also, FoF has been utilized in finding dark matter halos within N-body simulation (Knebe et al. 2011).

Another geometrical approach is the Voronoi tessellation which can be defined

as a group of convex polygon cells and their locations can be identified by groups of points (galaxies) in the plane. Each cell contains all points close to its location to some extent. The cell size is inversely proportional to the local point density (Miller 2012). The distribution of galaxies over space has been modelled using the natural partitioning of the Voronoi tessellation. M. Ramella et al. (2001) developed a Voronoi galaxy cluster finder model (VGCF) model which uses galaxy positions and magnitudes to find clusters and determine their key features, such as size, richness, and contrast above the background (Zaninetti 2006). By conducting an adaptive search on galaxy redshift to find high-density regions, Marinoni et al. (2002) developed the Voronoi-Delaunay method (VDM) to identify and reconstruct galaxy groups. The VDM works based on calculating the 3D Voronoi tessellation and its Delaunay triangulations. However, the calculation of the Delaunay mesh and Voronoi tessellation is not easy in non Euclidean space. Furthermore, at the edges of the galaxy sample the border effects appear from the formation of the infinite size of the Voronoi cells; and the cell size close to the edges might not reflect the local density because the galaxy distribution is unclear beyond the sample size(Duarte 2014).

The FoF method has many parameters to tweak based on the data surveys. Also, both FoF and Voronoi techniques seem to be arbitrary and heuristic methods, and they are not related to physical group/cluster profiles/properties (i.e. no assumption has specified). Despite the ability to detect an irregular type of galaxy group/cluster (non-virilized system or non-relaxed type), they cannot avoid the contamination of many interlopers and consider the interlopers as members of actual galaxy groups/clusters.

Furthermore, other techniques have been used to create catalogues of galaxy groups and clusters. For example, the hierarchical group finding scheme (Materne 1978) and the percolation technique (Klypin & Shandarin 1993)

Colour/red sequence clustering methods

This type of detection is based on galaxy colours in addition to their close spatial locations. In Miller (2012) divided this type of grouping into two divisions. First, it can be defined as a model of specific colour and luminosity of the galaxy groups and clusters and its distance from such centres, such as the maxBCG method (Koester et al. 2007) based on the red sequence of galaxies. The red sequence method relies on the fact that galaxy clusters contain well defined, highly regular lenticular and elliptical galaxies, which are known as a red sequence. These types of galaxies are based on their colour relation with their magnitude, employed as an indicator of the existence of galaxy clusters (Gladders & Yee 2000). If we plot the relation between the colour and magnitude of galaxies in a cluster, the galaxies will show almost a linear relation. Thus, in this method, the projection effect is no longer a problem, because if galaxies are located in different redshifts they will not produce a consistent red sequence. However, this kind of tracing requires a considerable number of red galaxies which leads to bias in finding old groups or relaxed dynamical states of the galaxy clusters (rich in red early galaxies), rather than identifying the typical kind of groups (Dariush et al. 2010, Koester et al. 2007). Also, this method depends on the quality of the optical data and the accuracy of redshift estimation of the systems, based on the filters that have been used to gain the colour of the red sequence.

Secondly it can be defined as a non-determined model; but the groups and

clusters are considered to form as a set of galaxies, located according to a position as well as colour, and without specified centres. The C4 method is an example based on galaxy colours (Miller et al. 2005). In the maxBCG method, the likelihood has been calculated of each galaxy being the brightest depending on two factors: the magnitude, along with the degree to which other galaxies are grouped around it in magnitude; and the position. While in the C4 algorithm, four-dimensional colours will be assigned a 'probability box' around each galaxy; and to find out the number of galaxies which have similar gradients of colour box a sample of galaxies is selected. The galaxies are called a cluster or group if they have the same gradient of colour in the colour box (Nichol et al. 2001).

However, this kind of galaxy groups' finder suffers from having many requirements, such as galaxy surveys that need to have an accurate measure of colour and brightness. While in reality, with some galaxies at high redshift, it is difficult to identify their colours. In addition, some galaxy groups contain mostly spiral blue galaxies, especially at high redshift, which makes them difficult to detect (Oemler 1974). Furthermore, colour clustering methods are a suitable identifier for some scenarios, but they may be biased towards a specific type of galaxy group/cluster; which leads to inefficient detection and less diverse types of galaxy groups (Donahue et al. 2002), than the geometrical methods. Also, colour clustering can have fewer interloper galaxies from the back/foreground than the geometrical methods.

Model based methods

This kind of detection incorporates many galaxy group/cluster properties, distribution and luminosity to model them; such as the matched filter (MF), which convolve the data with a shaped filter based on the pattern signal of interest.

The peaks are obtained by matching the proposed filter and the observed galaxies which will be considered as potential galaxy groups/clusters. Groups/clusters can be detected by specifying the local maxima within a moving window using a specified size and centring on each element from the galaxy map array (Kepner et al. 1999). If the centre of the window is a local maximum, then the group/cluster will be registered. Many different approaches of MF have been suggested, such as the adaptive MF (Dong et al. 2008, Kepner et al. 1999, Lobo et al. 2000), 3D MF (Milkeraitis et al. 2010) and multi-scale wavelets techniques (Moretti et al. 2004). These kind of methods are able to detect galaxy groups/clusters with a high completeness, little contamination and less false-positive rates. The main drawback of these methods is that they may fail to detect some groups/clusters which are not symmetrical, or which may not be similar to the suggested profile (M. Ramella et al. 2001).

Other approaches of galaxy groups/clusters' identification are based on probabilistic formulations and they can be a combination from all the aforementioned (i.e. geometrical, colour and model based) methods; such as extending FoF to probabilistic friends of friends (PFoF) (Liu et al. 2008); or including model-based analysis (Ascaso et al. 2012, Dominguez Romero et al. 2012, Duarte & Mamon 2015, Yang et al. 2005). Our method in Chapters 5 and 6 can be classified under this kind of formulation to identify galaxy groups.

The reasons behind the trend for developing probabilistic based galaxy groups/clusters' finders are: avoiding the unnatural heuristic methods (which needs to be tuned based on each mock galaxy survey), reducing the free parameters, developing more natural and sufficient methods with a strong theoretical background,

having a flexibility of including cosmological prior knowledge and adding physically related galaxy group/cluster characteristics to help improve the detection process. The probabilistic framework enables one to deal consistently with issues such as redshift distortion. All the previous methods, apart from the probabilistic method are decisive (e.g. taking the decision that a galaxy belongs to a particular group/cluster of galaxies). While in reality, there is a degree of uncertainty and this needs to be considered. The drawback is that some systems can not be detected if they are small and do not follow the same pattern as the object of interest. Thus, our method has been developed to tackle some of these scenarios.

In the following sections, we describe the widely used FoF method by Eke et al. (2004) and some of the related probabilistic existing models of galaxy group finders.

2.14 Friends of Friends (FoF)

FoF uses spatial and velocity information of galaxies to locate galaxy groups/clusters as overdensity regions. It has three free parameters to be set in regards to mock catalogue characteristic: a linking length b based on the comoving number density of galaxies $\bar{n}$, which is specified for each galaxy; an upper limit of the perpendicular linking length L_{max} depending on the nature of the galaxy survey; and the ratio R_f between the perpendicular to ($\ell_\perp$) and along the LoS ($\ell_\parallel$) linking lengths. The R_f guarantees that $\ell_\parallel$ linking length is larger than $\ell_\perp$ considering the elongation along the LoS due to the impact of the FoG (i.e. to avoid the redshift distortion). Galaxies are linked together after comparing their separation along and perpendicular to the LoS with their linking lengths. If the separation between

two galaxies is less than the linking criteria regarding linking lengths, this pair of galaxies will be linked; otherwise, they will be classified as field galaxies. The linked pair will be called friends and groups of friends (friends of friends) will form a galaxy group/cluster. The comoving linking lengths for a particular galaxy are:

$$\ell_{\perp} = \min\left[L_{max}(1+z), \frac{b}{\bar{n}^{\frac{1}{3}}}\right], \tag{2.20}$$

$$\ell_{\parallel} = R_f \cdot \ell_{\perp}. \tag{2.21}$$

Two galaxies a and g, at comoving distances D_{c_a} and D_{c_g}, respectively with angular separation $\Theta_{a,g}$, are linked together if they satisfy two conditions:

$$\Theta_{a,g} \leq \frac{1}{2}\left(\frac{\ell_{\perp,a}}{D_{c_a}} + \frac{\ell_{\perp,g}}{D_{c_g}}\right), \tag{2.22}$$

where

$$\Theta_{a,g} = \frac{180}{\pi}\tan^{-1}\left(\frac{\sqrt{\cos^2\beta_g \sin^2(\theta_g - \theta_a) + [\cos\beta_a \sin\beta_g - \sin\beta_a \cos\beta_g \cos(\theta_g - \theta_a)]^2}}{\sin\beta_a \sin\beta_g + \cos\beta_a \cos\beta_g \cos(\theta_g - \theta_a)}\right), \tag{2.23}$$

and

$$|D_{c_a} - D_{c_g}| \leq \frac{\ell_{\parallel,a} + \ell_{\parallel,g}}{2}. \tag{2.24}$$

The performance of FoF is controlled through scaling the linking length depending on the number of galaxies that are identified as a function of redshift. The optimizing of the linking length and calibrating the related parameters are based on the mock surveys onto which FoF has been conducted [e.g. (Berlind et al. 2006, Eke et al. 2004, Robotham et al. 2011)]. Choosing a small linking

length leads to many incomplete, fragmented groups/clusters, increases the possibility of obtaining false positive groups by linking separate galaxies via "bridges" (Knebe et al. 2011). While setting a large linking length implies reducing the number of false positive groups/clusters, but at the same time will increase the contamination of the expected galaxy groups/clusters by combining field galaxies as members. Furthermore, two factors will affect the linking length indirectly by affecting the comoving number density $\bar{n}$, the low density of galaxies at high redshifts in the flux limited catalogues causes decreasing in $\bar{n}$ and increasing in the mean inter- galaxy separation. Secondly, the completeness variation of the mock survey also leads to a change in $\bar{n}$. The FoF method is easy to interpret and it does not need assumptions related to the shape of the pattern of interest; however, this can also count as a disadvantage because it can assume that many dense regions are galaxy groups/clusters. Moreover, as structure formation is hierarchical, to find substructures, FoF requires different linking lengths (Knebe et al. 2011).

2.15 Some of the Existing Probabilistic Approaches

Due to some observational issues related to galaxy surveys such as edge effects, bright star masking and poor measurements of redshift in photometric surveys, some probabilistic approaches to finding galaxy groups depending on photometric and spectroscopic data have been developed to address these problems (Duarte & Mamon 2015).

- **Probability Friends of Friends (PFoF) Methods (Jian et al. 2014, Li & Yee 2008, Liu et al. 2008)**

There are two PFoF approaches applied and developed which are based on photometric mock surveys and consider poor precision of redshift in this kind of survey.

The first PFoF method was developed by Li & Yee (2008), taking into account the photometric redshift probability density of each galaxy and galaxy groups/clusters, and conducting extensive modification on the conventional FoF algorithm, with an updated definition related to the linking length. Each galaxy in the survey is considered as a seed of a group. For each seed and depending on a comparison in probability between the redshift probability density function(PDF) of both the group seed and galaxy, the probability membership ratio in redshift space has been calculated:

$$P_{\rm ratio} = \frac{\int_0^\infty P_{\rm gal}(z) \cdot P_{\rm group}(z) dz}{maxP}$$

where z is the photometric redshift; $P_{\rm gal}(z)$ is the PDF of the galaxy; $P_{\rm group}(z)$ is the PDF of the group; and $maxP$ is the normalization term (the maximum value of the numerator). The $P_{\rm ratio}$ finds the amount of overlapping between the galaxy PDF and group PDF. For each galaxy to be added to the group, it has to pass the friendship criteria in redshift space $P_{\rm ratio,crit}$. If $P_{\rm ratio} \geq P_{\rm ratio,crit}$, the new galaxy will be a member in this particular group. Then the group PDF will be recalculated again considering the current group members (which is the likelihood for all these members to occur at the same redshift):

$$P_{\rm group}(z) = \frac{\prod_{\rm nm} P_{\rm nm}(z)}{\int_{z_{\rm min}}^{z_{\rm max}} \prod_{\rm nm} P_{\rm nm}(z') dz'},$$

where nm is the number of galaxies in the group and $P_{\rm nm}$ is the PDF of the individual galaxy currently within the group. The process will be repeated until no additional galaxies can be added or removed from the group and then continued for all galaxies in the survey to be considered as a seed group galaxy and add its members of galaxies. The resulting groups will be checked for duplications (merging those with most likely memberships and removing the less significant). This approach has been tested on Virgo Consortium Millennium Simulation mock

catalogues (Springel et al. 2005) and on the CNOC2 group catalogue and it was confirmed that it recovered more than 80% for mock groups of at least $2{\times}10^{13}M_{\odot}$ with 10% false detection rate of galaxy groups containing ≥ 8 galaxies.

The second PFoF method was developed by Liu et al. (2008) to overcome the disadvantages of the traditional FoF and EXT- FoF methods (Botzler et al. 2004, Eke et al. 2004) when detecting galaxy groups in a galaxy catalogue with large error dispersion in redshift (to deal with photometric redshift uncertainty). Two criteria have been tested to check two galaxies are physically linked, one related to the perpendicular to the LoS direction and another along the LoS direction. The first criteria is the separation distance between two galaxies $D_{i,j}$ has to be less than the comoving linking length $\ell_{\perp}$ (i.e. $D_{i,j} \leq \ell_{\perp}$). While for the second criteria, they have calculated the probability of the distance between any two galaxies along LoS to be less than the parallel linking length $\ell_{\parallel}$:

$$P(|z_2 - z_1| \leq \ell_{\parallel}) \equiv \int_0^{\infty} \mathrm{P}_1(z)dz \cdot \int_{z-\ell_{\parallel}}^{z+\ell_{\parallel}} \mathrm{P}_2(z')dz',$$

where P_1 and P_2 are the probability distribution functions for both galaxies in the LoS direction. If the probability P is larger than an artificial threshold P_{th} (i.e. $P(|z_2 - z_1| \leq \ell_{\parallel}) > P_{\mathrm{th}}$), then the galaxies are associated. The PFoF here has aimed to measure the probability of two galaxies being associated and not just the intersection of the two distribution functions. It considered the probability amplitude besides the error distribution width. The DEEP2 mock catalogues and additional simulated photometric redshift error have been used to measure the performance of the PFoF method. In addition, PFoF was compared with the conventional FoF method. It was confirmed that the outcome of the method is better than the outcome from the traditional FoF if applied on given data with

the same redshift uncertainty. Recently, further tests and optimizations on PFoF have been carried out (Jian et al. 2014).

We could not use the PFoF methods (Jian et al. 2014, Li & Yee 2008, Liu et al. 2008) to compare with our method because both PFoF methods are based on the redshift probability density functions (PDF) of both the group seed and individual galaxies. These PDFs are used in PFoF to check the pertinence between the galaxy and the galaxy group/cluster. In our case, the galaxies are represented only through the spatial information without PDFs. As no such PDFs are available from the GAMA mock survey. In addition, PFoF methods were designed to be applied on photometric mock surveys which are known to yield a poor redshift estimate.

- **Yang et al. (Yang et al. 2007, Yang et al. 2005) Method**

In Yang et al.'s method, the potential galaxy groups were found using the traditional FoF algorithm with small linking lengths. For each potential group, the luminosity $L_{19.5}$ of the group was estimated based on its tentative members with absolute magnitude $\leq$ -19.5 for a group within redshift z $\leq$ 0.09. Then the halo mass M_h was estimated based on the mass to light relation $M_h/L_{19.5}$ and also the size and the velocity dispersion. The galaxy memberships of these groups is updated based on density contrast $p(r, \Delta z)$ using the estimated information (halo mass, size and velocity dispersion). They assumed that the distribution of galaxies inside groups are following the same distribution of dark matter particles in cosmological simulations (NFW profile); while the velocity distribution of galaxies $p(\Delta z)$ in terms of the differences between galaxies' and groups' redshifts is distributed normally:

$$p(r, \Delta z) = \frac{\mathrm{H}_0 \ \Sigma(r) \ p(\Delta z)}{c\tilde{\rho}},$$

where $\tilde{\rho}$ is the mean density of the Universe. For a given group, if a galaxy has $p(r, \Delta z) \geq B$ where B is the chosen background level, then this galaxy will be assigned to the group, otherwise it will be dismissed. If a galaxy can be assigned to multiple groups based on this condition, it is assigned to the group where the galaxy has obtained the highest $p(r, \Delta z)$ value. The iteration process continues to reassess the centres of the galaxy groups and their luminosities $L_{19.5}$ after the new updates in the galaxy memberships until the membership convergence is reached, then the mass - luminosity relation will be recalculated $L_{19.5}$, M_h, size and velocity dispersion until the relation $M_h/L_{19.5}$ - $L_{19.5}$ will be converged. They used mock catalogues constructed for the SDSS DR4 to check the performance of the group finder in terms of completeness of true members, contamination by interlopers, and accuracy of the assigned masses. They used the density profile in assessing the galaxy membership to each galaxy group. While in our probabilistic approach, we used the density profile to signify the potential galaxy group positions through the Hough Transform.

- **Romero et al. (Dominguez Romero et al. 2012) Method**

Romero et al. improved Yang et al.'s (Yang et al. 2007, Yang et al. 2005) method as they realised that Yang's method is similar to the well-known "k-means" clustering method. The main improvement was in the galaxy assignment process. Inspired by the "soft k-means" algorithm (MacKay 2003), they introduced a soft assignment to each galaxy, making it belong to each cluster rather than a hard, equal degree of assignment to a given group. These are called "responsibilities", which represent that a galaxy belongs partially to k galaxy groups. Romero et al. started by taking all galaxies as centre of galaxy groups. They used the same estimations of the system's (i.e. group/ cluster) properties as in Yang's method. However, instead of using merely the luminosity of the members to find the estimated luminosity of

the whole group, they used a weighted luminosity of the members by incorporating their responsibilities. The responsibility of group k for each galaxy i is:

$$\mathcal{R}_{(k)}^{(i)} = \frac{\pi_{(k)} p_{(k)}(r, \Delta z)}{\sum_{(k')} \pi_{(k')} p_{(k')}(r, \Delta z)}.$$

Afterwards, they followed Yang's method and assigned the satellite galaxies to the potential group candidates based on comparing their density contrast $p(r, \Delta z)$ with the background threshold value B. They allowed a lower background level value so that the likelihood of a galaxy being in a given galaxy group is higher than that of the central galaxy of a nearby halo. They used the likelihood value to classify the galaxies into central group galaxy or satellite. If the likelihood of relevance to another halo is over the background level, the galaxy is classified as a satellite. The method continues by updating the central galaxy group (i.e. the mean position of the responsible galaxies) and calculating the groups' weighting parameters π_k based on $\mathcal{R}_{(k)}^{(i)}$ values. The method iterates the assignment and the update procedures until the convergence in the membership occurs. It was checked by a mock catalogue generated by using the Millennium Simulations, in addition to applying the method on the NYU-DR7 galaxy catalogue.

- **MAGGIE (Duarte & Mamon 2015)**

Duarte & Mamon (2015) created the MAGGIE method which is a probabilistic abundance matching grouping algorithm. MAGGIE consists of the combination of previously measured universal distribution of halo interlopers in projected phase space and knowledge of NFW halos with realistic internal kinematics. The MAGGIE method was performed on two kinds of orientation: group luminosities and stellar masses. They compared MAGGIE's performance with an optimized version of FoF for detecting groups of at least three galaxies extracted from two sub-

samples that are complete in distance and luminosity within a mock, flux-limited, SDSS Legacy redshift survey.

The method assigned a probability to each galaxy of being in a given group. The galaxy can be assigned to many groups. The MAGGIE method is based on an iterative process assuming that the galaxies' basic information is known: Ra, Da, redshift, stellar masses (galaxy masses), luminosities, absolute magnitudes and their apparent magnitudes. Then it start with a seed of significant positions to be considered as galaxy groups, assuming either the most massive galaxies or the most luminous are the potential group centres. For each group, the virial radius was estimated by assuming the halo mass corresponds to the virial mass. The halo mass can be estimated at the first iteration based on the relation between the halo mass and the central stellar mass of galaxy from (Behroozi et al. 2010) and the ratio relation for luminosity; with learning from the previous iterations for the next iterations. Then, all galaxies are assigned within an angular separation corresponding to the virial radius to become members of the group. They compute the probabilities of galaxies to become members of a given group:

$$P(R, v_z) = \frac{g_h(R, v_z)}{g_h(R, v_z) + g_i(R, v_z)}$$

where g_h is the density profile inside the group (a multiplicative combination between Gaussian profile along the LoS and NFW profile orthogonal to the LoS); and g_i is the interloper density profile. Then, galaxy group members will be filtered (i.e. whether they contribute in estimating the mass and luminosity of the group through weighted multiplicity or are dismissed) based on comparing their $P(R, v_z)$ values and a free parameter threshold $\rho_{\rm mem}$, with an additional condition that these members of a given group should have absolute magnitude values less

than the absolute magnitude of the sample. Then, MAGGIE uses the abundance matching technique which assumes a one to one relation between the stellar mass of the group central galaxy and the halo mass of the group. By comparing the cumulative distribution function (CDF) of the two quantities, it appears that the number of groups above a specific central galaxy stellar mass is equal to the number of groups above a specific corresponding halo mass. Thus, for a certain HMF, MAGGIE can predict the halo mass of a group with a given galaxy central stellar mass by comparing the CDF with that predicted by the HMF. The MAGGIE method reiterates the computation of groups with the halo mass - central stellar mass relation until it reaches a convergence in the number of groups.

Our approach in Chapters 5 and 6 differs from Yang et al. (2007), Yang et al. (2005) and MAGGIE (Duarte & Mamon 2015) in the way the distribution of galaxies within galaxy groups is utilised. In our case this distribution forms a core of a probabilistic Hough transform targeted towards primarily finding the galaxy group positions, rather than the group membership. The group memberships are then later inferred based on the detected positions and estimated group mass.

CHAPTER 3

HOUGH TRANSFORM BACKGROUND

The aims of this chapter are to define the HT and the probabilistic Hough transform (PHT) principles, along with their ability to detect objects in a noisy environment. Examples of the utilisation of the Hough transform in astronomy and in detecting spherical shapes are presented.

3.1 Introduction

The idea behind the HT is to recognize patterns (eg. lines, circles) which are essential parts of computer vision and digital image processing. These geometric features of interest could be regular or irregular in shape and could be embedded within a noisy background. A fine description of the pattern of interest should be determined if the utilization of the HT concept needs to be considered. Paul Hough originally devised the HT to identify the intricate patterns in a picture and to recognize particle tracks in pictures derived from a bubble chamber based on a voting procedure (Hough 1962). The HT can be utilized for the recognition of any shape that can be described in parametric form. The basic principle is to

convert the image point (data point) from its space (S) into Hough (parameter) space (H) and then identify the peaks in the H-space. Each peak (intersection) in the H-space represents a pattern in S-space. Besides, the location of this peak in the H-space leads to the determination of the position of this pattern in the S-space.

The main advantages of the HT are its tolerance of discontinuities in shape boundary and robustness against the noise. Thus it can determine the patterns more accurately, and the results mostly do not require post processing. Also, there is a trade-off between work in image space and parameter space by handling inaccurate edge locations. The classical HT was concerned with the identification of lines in the image, but later the HT was extended to identifying positions of irregular shapes. The disadvantages of the HT are associated with its large storage and computational requirements (Kesidis & Papamarkos 2000). For these reasons, many approaches have been proposed in the literature, regarding the reduction of computation time and memory requirements (Chiu et al. 2010, Guo et al. 2009, McLaughlin 1998). For more details, the HT surveys and the references therein, such as (Antolovic 2008, Hassanein et al. 2015, Illingworth & Kittler 1987, Mukhopadhyay & Chaudhuri 2015) can be consulted.

3.2 Straight Line Detection using HT

Using a simple slope s_h - intercept e_h parametric representation of a line, every point (x_i, y_i) in image space can correspond to many lines in parametric space

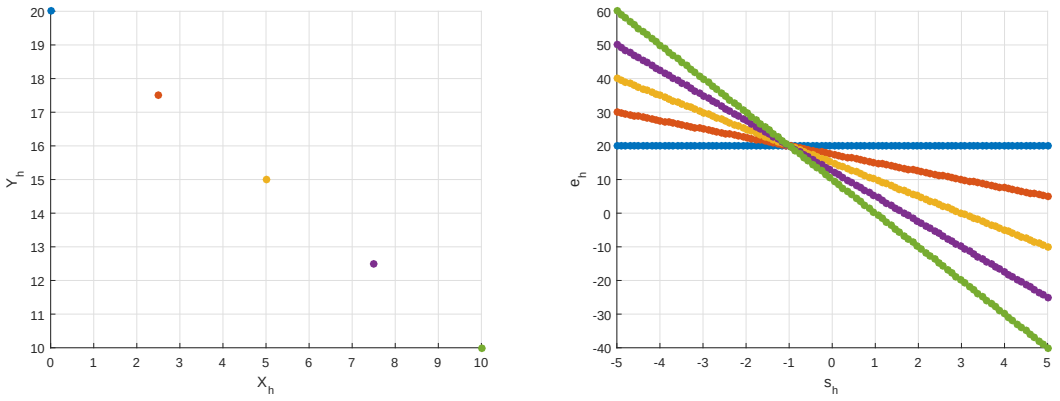
(H-space) (s_h, e_h) passing through it,

$$y_i = s_h \times x_i + e_h. \tag{3.1}$$

By partitioning the parameter space into a number of cells in a grid, every point (x_i, y_i) now "casts a vote" as to what line could be passing through it in the parametric space. This is simply done by associating a counter (accumulator : initialized to 0) with each cell in the parameter space and then taking the points (x_i, y_i) in the image one-by-one, incrementing a cell's counter if the parameter line $e_h = -x_i \times s_h + y_i$ passes through it. This way the cells with highest counter values correspond to lines in the image with the strongest support. Hence the line detection is cast as peak detection in the parameter space. Random points in an image (e.g. a noisy environment surrounding patterns) are unlikely to contribute coherently to one bin of the accumulator and therefore produce only a very low-level background of counts in the H-space (Ballester 1996, Laschinsky 2012, M. & Muthukrishnan 2015).

As shown in Figure 3.1a the five points (0,20), (2.5,16), (5,15), (7.5,12.5) and (10,10) in the image space are transformed into five lines in the Hough space with the slope s_h=-1 and intercept e_h=20 as shown in Figure 3.1b.

The important point in the HT is the ability to detect each point or pattern in the image space in accordance with the parametric constraints. This leads in some cases to an unfavourable situation if there is a high noise background, when noise points or spurious patterns may satisfy the constraints, which will then lead to a high level of false positive results. In our case, we utilize a peak detection



(a) Image space (b) Hough space

Figure 3.1: Line detection

process to overcome these false positives as much as possible. There are no specific generalization rules for achieving good outcomes because the method will always rely on the nature of the properties of the data which are observed.

The parametric representation of a line in Eq. (3.1) has an infinite slope in the case of vertical lines. To deal with this a solution was presented by Duda & Hart (1971), is called a *singularity-free normal parameterization* and is depicted in Eq. (3.2).

$$\rho_{\mathrm{h}} = \mathrm{x_i} \cos\left(\mathcal{Y}_{\mathrm{h}}\right) + \mathrm{y_i} \sin\left(\mathcal{Y}_{\mathrm{h}}\right) \tag{3.2}$$

where ρ_{h} is the typical form of the line representing the distance from the origin; $\mathcal{Y}_{\mathrm{h}}$ is an angle between the x-axis and the ρ_{h}; thus the parameter space here changes from ($\mathrm{s_h}$,$\mathrm{e_h}$) (slope,intercept) into (ρ_{h}, $\mathcal{Y}_{\mathrm{h}}$) (length, angle). The cost of this change is the trigonometric computing of these functions. If there are multiple points in the image that are collinear then their sinusoidal shapes in the Hough space will interpolate as shown in Figure 3.2.

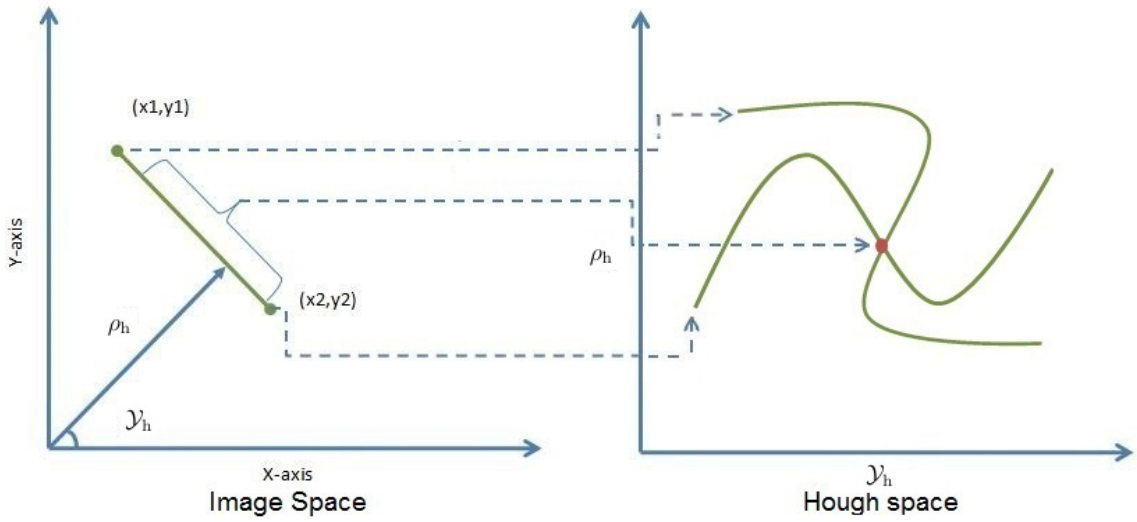


Figure 3.2: Normal form - sinusoidal shapes

This form seems more efficient if there is a similarity coincidence between a line and noise that can be detected more precisely by the angles of the patterns, to distinguish the genuine from the spurious (Laschinsky 2012). The same example above leads to the sinusoidal lines in Hough Space as in Figure 3.3.

In the case of galaxy group patterns' detection, our approach will consider the mixture of posterior distributions as an accumulator of galaxies' votes for a given galaxy group/cluster position, as will be illustrated in the following chapters.

3.3 Probabilistic Hough Transform

Probabilistic Hough transform was defined by Stephens (1991), who noticed that there is a relationship between the standard Hough transform (SHT) and maximum likelihood (ML) method due to the voting procedures in the SHT are similar to

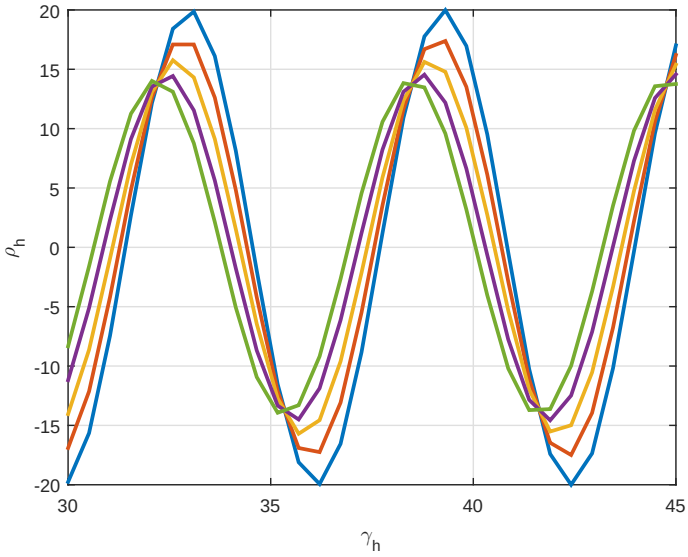


Figure 3.3: Line detection: Hough space

the highest probability regions in the probability density function (PDF).

By converting the detection of the patterns into a parameter estimation problem and supposing that each set of parameters represents a model, the challenge will be how to fit that model into a specific pattern. The analysis of the ML leads to the definition of the PHT.

If the assumption is reasonable about the input feature error properties of such a system, then the outcomes of the PHT are equal to those of the SHT. However, if the assumption is too far from the correct properties, the SHT will not work well and that will require some adjustments to the model of the input feature errors during the detection process.

The PHT adds previous knowledge to the Hough space, but it is a more com-

putationally expensive procedure compared to the SHT. Besides, it is dependent on floating point arithmetic and therefore, does not work well with low dimensional features but it is more resistant than the SHT when the dimensionality is increased. The PHT is defined as a continuous function; while, the SHT is defined as a discrete model. Hence, the PHT can be applied more reliably to any model.

Furthermore, the PHT is defined as a mathematical form of the SHT; which is represented in the log of the PDF of the output parameters given all input patterns (Stephens 1991).

$$\mathrm{PHT(y)} = \ln \mathrm{P(y|X_n)} \tag{3.3}$$

where $\mathrm{X_n}$ refers to a set of random variables $(x_1, x_2, x_3, ..., x_n)$ (features) in image space; Y is a set of patterns collected from points $(y_1, y_2, y_3, ..., y_m)$ in Hough space. From Bayes' theorem :

$$\mathrm{PHT(y)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \ln \mathrm{P(x_i|y)} + \ln \mathrm{P(y)} + \mathrm{C} \tag{3.4}$$

The purpose of developing the PHT was to improve the performance of the conventional HT in solving problems that occur when there are many unknown parameters (i.e. the existence of uncertainty).

Our approach will rely on finding the mixture of local posterior distributions

instead of the log of the PDF of the output parameters, as will be illustrated in subsequent chapters.

3.4 HT in Astronomy

The HT is utilised in many applications in astronomy; for example, a modified HT was applied by Ragazzoni & Barbieri (1994) to obtain the correct cycle number for each observation. This procedure determined the durations of some of the periodic events during a specific period. Also, the HT can be used to analyse astronomical light curve times. The light curve of an eclipsing binary star has been identified with reasonable accuracy and confirmed through these observations (Ragazzoni & Barbieri 1996). A variant HT was adapted by Llebaria et al.(Llebaria & Lamy 1999) to understand the temporal evolution of radial structures on the solar corona 'polar plumes', by tracking the coherent trajectories of the time intensity diagram (TID) on a set of images from the LASCO/C2 coronagraph. Another utilisation of HT was to clean the SuperCOSMOS Sky Survey (SSS) by Storkey et al. (2004). They employed their approach to clean some types of records in the SSS: linear phenomena can appear on the plate such as scratches fibres and satellite tracks, circular halos around bright stars and diffraction spikes close to bright stars. Additionally, they developed a probabilistic technique combining the HT, renewal processes and hidden Markov models and applied it to the SSS data to develop a dataset of spurious object detections, and confidence measures, which allow this unwanted data to be removed from consideration (Storkey et al. 2014).

HT has been employed to identify gravitational waves' signals, such as developing an incoherent method by Krishnan et al. (2004). The HT was utilised to dis-

tinguish patterns in the time-frequency plane of data collected via an earth-based gravitational wave detector. Two HT search methods were applied, depending on the type of data used with the HT: the Fourier transforms of the detector data, or the output of a coherent, matched filtering type search. Another version of HT called frequency Hough transform (FHT) was developed by Astone et al. (2014) for scanning the sky to identify continuous gravitational wave signals via hierarchical data analysis. Starting with a coarse grid of the parameter space, the area around the revealed candidates from applying FHT was processed with a refined analysis.

Furthermore, the HT has been extended to many varieties such as arc-line, circular and elliptical shaped detections in the astronomy context. For instance, Ballester (1994) developed two methods based on the detection of straight lines and parabolas for a spectral data reduction domain. The HT was utilised for fast detection of echelle orders and automated arc-line identification by cross-matching arc spectra and line catalogues. Hollitt & Johnston-Hollitt (2012) developed a circle Hough transform (CHT) and explored its response in detecting the circular or arc-like forms of cosmological objects. Images containing noise alone, as well as images containing point sources, were examined. The CHT was applied to different images and the extent of the filtering was investigated as well as the robustness of the presence of noise. It was found that the CHT had the effect of identifying the circular structures. However, the CHT is computationally challenging, in terms of both computational effort and memory consumption, making it quite time-consuming. Lastly, another automated procedure by Massone et al. (2014) was developed to identify curves and elliptical shapes in medical and astronomical images. A set of classes of curves were defined to utilise in detecting patterns. Their recognition method was applied to astronomical images provided by NASA's solar dynamic observatory satellites to identify the front ends of solar eruptions. The

main limitations of their approach are its dependency on the defined profiles of curves in the catalogue, and the known computational requirements of the HT for the optimisation problem.

3.5 HT for Detecting Spherical Shapes

There have been few attempts to use the HT in 3D space due to its high level of time and memory consumption. Tsuji & Matsumoto (1978) presented an adjusted HT that could be used to detect a 3D ellipsoidal shape with five parameters. They identified the approximate positions of the centres of the ellipses and all points inside these ellipses. After checking whether the candidates are inside or not, the true points can be used to calculate the five parameters of these ellipses. To scale down both the time and space of the ellipsoid identification, Hsu & Huang (1990) proposed a partitioned method based on the HT, splitting the original parameter space into many small parameter spaces, to reduce the dimension of detecting the ellipsoidal shapes using the independent properties of ellipsoid parameters. In addition, they utilised information derived from a given image to reduce the search space range of each parameter. They were able to detect the ellipsoidal shapes with noise no more than $\mathcal{N}(0,0.01)$. Moreover, They could identify multiple ellipsoids inside each other, but it was important that the ellipsoid centre was not occluded due to the identification of the ellipsoid depending on the information from its central part. Taylor (1990) developed a methodology for recomposing parametrised surfaces, using a parameter set decomposed into multiple-subsets. The full parametrisation of the surface is revealed through the detection of the conjunctions of the individual parameter into these subsets. These recognitions are obtained through a multi-window parameter estimation method; multi-resolution k-tree parameter subspace searching and voting; and a conflict filtration process

to avoid the false parameter hypothesis and find a unique parameter set for each surface region. This methodology has been shown to provide consistent results in the detecting of multiple spherical and cylindrical shapes.

The circular HT (CHT) has been used to detect the ball in soccer game images to verify goals (D'Orazio et al. 2004). In this application, many problems must be faced; such as occlusions, shadows, objects similar to the ball, and real-time processing. The implementation of this method has to solve this issue through a visual framework. The identification of the ball should be fast, in terms of time for processing, and robust about reducing the rate of false positives. D'Orazio et al. applied two sequential steps: first they used a modified directional CHT to recognise the most significant candidate regions that might contain the ball; and second, neural back-propagation was applied to determine whether the selected region included the ball or whether there was just a false positive in that region. Some experiments have been conducted to prove that this proposed method obtains a fair detection score.

In the medical sector, for successful joint replacement surgery, it is necessary to identify the joint's geometric centre; Glas et al. developed a technique to automatically determine the sub-voxel position and size of a sphere in unsegmented 3D images generated by CT and MRI scans, using the direction and strength of the gradient.Their technique is stable to size and robust to noise. Just a quarter of a sphere is needed to detect the centre of the humeral head (van der Glas et al. 2002).

In the following chapters, the developing of a new probabilistic model based on

the Hough transform to detect galaxy group and cluster patterns will be illustrated in detail. Also, the model will be tested on simple and realistic mock data and its performance will be evaluated.

CHAPTER 4

PRELIMINARY GALAXY GROUP DETECTION BASED ON PROBABILISTIC HOUGH TRANSFORM

The aim of this chapter is to construct the basic PHTM for detecting elongated patterns (i.e. imitating the FoG patterns). These patterns are imposed and generated in a simple 2D flat area and a 3D cone shaped area[2] within a noisy environment. The PHTM is based on the mixture of posterior distributions as an accumulator of galaxies' votes being within a given galaxy group/cluster position (as will be illustrated in section 4.1.2).

For each data type in this and the following chapters, we describe the generation of the data and illustrate the implementation of the PHTM and its modifications. At the same time, the results for each scenario are discussed.

[2] imitating the real 3D cone as constructed by astrophysicists when observing the sky.

4.1 2D Experiment: Testing the Concept

Before the methodology is demonstrated on realistic 3-D data in the next chapter, we will first test our method in a large set of controlled experiments in 2-D, where we can control the amount of background noise. This section shows preliminary proof of the efficiency of utilizing the Hough transform concept in detecting 2D simple prolonged patterns in a noisy environment.

4.1.1 Preliminary 2D Data Generation: Flat Area

In the 2-D flat setting, the LoS direction was assumed as the y-axis and the galaxy groups are represented by points (galaxies) generated from Gaussian distributions elongated along the y-axis. In each group we generate 10–25 points from such Gaussian distributions. Six groups are created with different fixed means and the same covariance Cov:

$$Cov = \begin{bmatrix} \epsilon_k \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}, \tag{4.1}$$

where $\sigma = 0.5$, $\epsilon_{kk} = 0.025$ to form the elongated shape of galaxy groups as shown in Figure 4.1.

The background is generated from uniform distribution to make the scene more sophisticated and imitate reality. The number of background points Ns is determined as TN_g, where N_{grp} is the number of galaxies in galaxy groups and T is a multiplicative factor in the range (5- 30), as presented in Figure 4.2. In each setting there are six galaxy groups at fixed positions as shown in Figure 4.1.

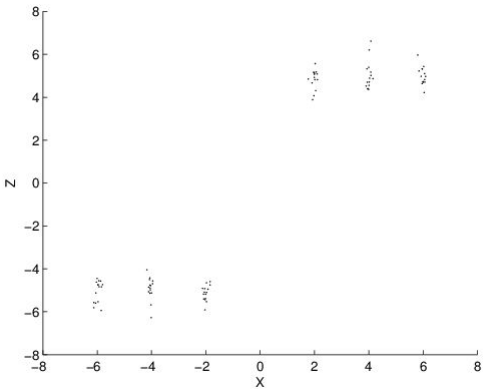
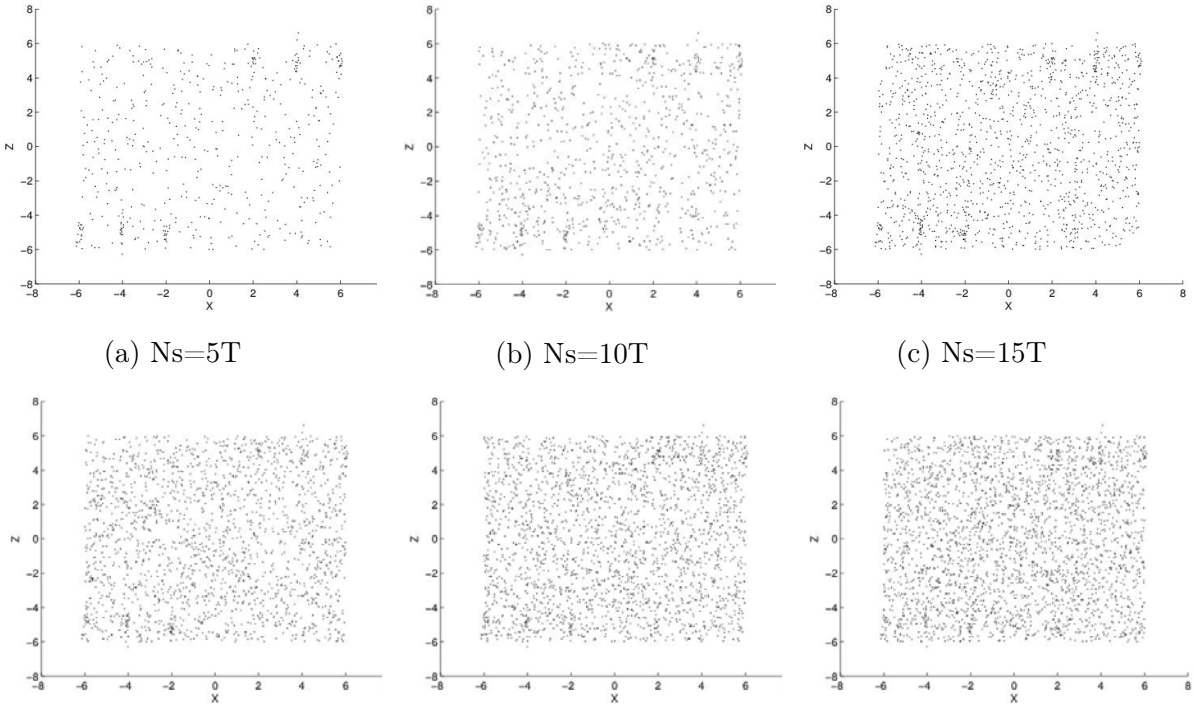


Figure 4.1: Synthetic data before adding noise



(a) Ns=5T (b) Ns=10T (c) Ns=15T

(d) Ns=20T (e) Ns=25T (f) Ns=30T

Figure 4.2: Synthetic data after adding the uniform noise

4.1.2 Preliminary 2D PHTM Model: Flat Area

The search space is covered by a regular structure of N_G grid points. On each grid point, we position a noise model representing a possible galaxy group position and ask all observed galaxies (points) to ascertain whether they are likely to have come from that group. Formally, for the k-th grid point (x_k, y_k) we have a Gaussian noise model centred at $\mu_k = (x_k, y_k)$ with axis-aligned (diagonal) covariance matrix $Cov = diag(\epsilon_k\sigma^2, \sigma^2)$ with variance along the y-axis σ^2 and variance along the x-axis $1/\epsilon_k$ times smaller, i.e. $\epsilon_k\sigma^2$ (we used $\sigma = 0.5$ and $\epsilon_k = 0.025$). The likelihood model for the G_k-th grid point is thus a multivariate Gaussian with mean μ_k and covariance Cov:

$$p(g_q|G_k(\mu_k, Cov)) = \frac{1}{\sqrt{2\pi|Cov|}} e^{-0.5(g_q-\mu_k)Cov^{-1}(g_q-\mu_k)^T} \tag{4.2}$$

Given a galaxy g_q, $q = 1, 2, ..., N_{\text{gal}}$, the degree to which it belongs to the possible group centred at the k-th grid point μ_k is quantified through posterior:

$$P(G_k|g_q) = \frac{p(g_q|\mu_k, Cov)P(G_k)}{P(g)}, \tag{4.3}$$

where $P(g)$ is the normalization term:

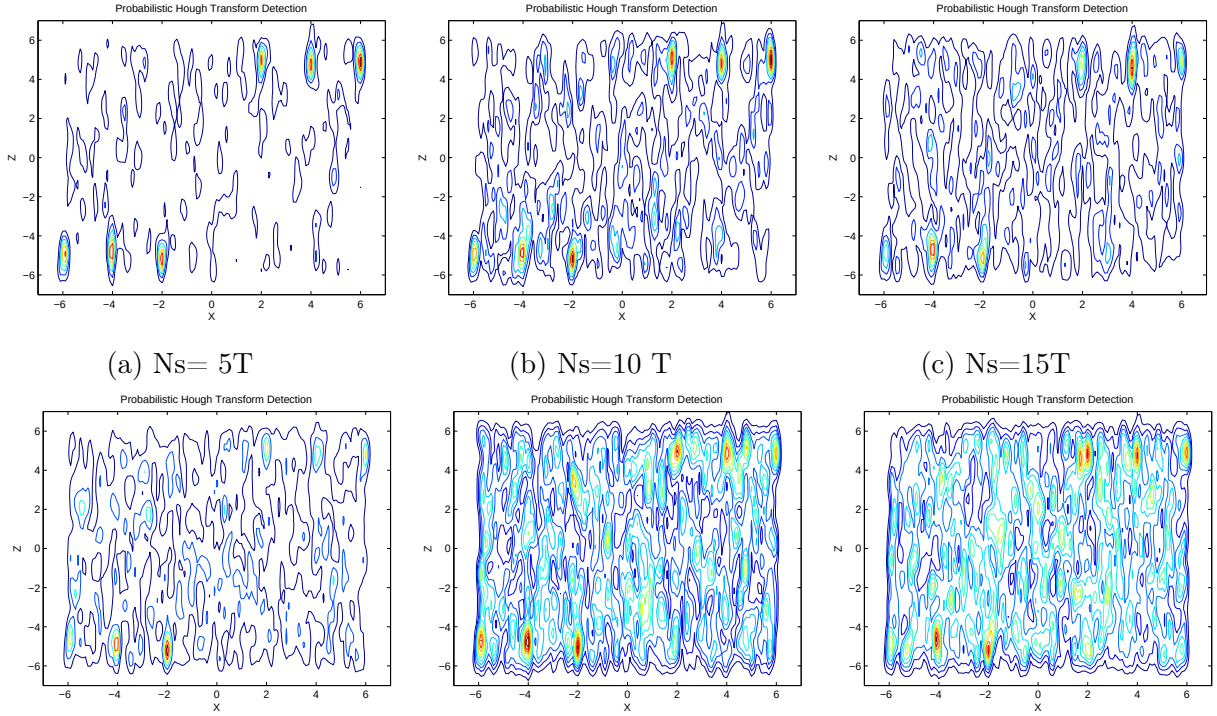
$$P(g) = \sum_{j=1}^{N_G} P(g_q|\mu_j, Cov)P(j). \tag{4.4}$$

We assume no preferred positions for galaxy groups, i.e. flat prior $P(G_k) = 1/N_G$. The posterior can be interpreted as a 'soft' vote of the q-th galaxy for the possible galaxy group at position μ_k. The overall vote for the presence of a galaxy group at μ_k is then obtained as a flat mixture of posteriors H given by the observed galaxies N_{gal} (which is called PHTM):

$$H(x_k, y_k) = \frac{1}{N_{\text{gal}}} \sum_{q=1}^{N_{\text{gal}}} P(G_k|g_q) \tag{4.5}$$

4.1.3 Results and Discussion: 2D Flat Area

The $H(x, y)$ values of the true peaks (true galaxy groups) were affected by the intensity of the noise as shown in Figure 4.3. As long as the intensity of the noise increases, the difficulty of finding the true galaxy groups increases, due to the increasing of spurious peaks (i.e. FP). In addition, the number of true group members (galaxies) is not enough to keep the centre position of the galaxy group candidate prominent (i.e. sometimes the high intensity of the galaxies (noise) close to the true group will form $H(x, y)$ values approximately equal to the true group $H(x, y)$ values). For example, the three upper true groups in Figure 4.3d have almost disappeared, which means the false negative (FN) increase because we can not see the true first three groups clearly. Whereas, when the number of background galaxies has increased further, the appearance of the upper three groups has recovered but in kind of blurry noisy way, as shown in Figure 4.3e and Figure 4.3f.



(a) Ns= 5T (b) Ns=10 T (c) Ns=15T

(d) Ns= 20T (e) Ns=25T (f) Ns= 30T

Figure 4.3: The results of applying the probabilistic Hough transform method for all noise scenarios: the contour formed to any prolonged pattern with enough characteristics to be a group candidate. The contours with the highest peaks with their centres in a red colour, means that they have the highest value positions as group candidates

Given a detection threshold $\tau > 0$, the possible galaxy groups are detected as peaks above τ in the $H(x, y)$ landscape. Note that high values of τ will produce over-cautious conservative detections with a significant number of undetected true galaxy groups (FN). On the other hand, low τ will lead to insignificant low peaks declared as group candidates (FP).

Noticeably, in Figure 4.3f due to the high noise, the first upper group from the left has dispersed into two group candidates and the first lower group from the

left has nearly disappeared. To obtain the true groups correctly with decreasing or avoiding the detection of false peaks, we have applied a simple peak detection based on the convolution technique of moving a 3 by 3 sliding window through all grid cells to find the highest peaks by comparing the $H(x, y)$ values within the window and then choosing the highest as a representative for all surrounding grid cells.

4.1.4 Precision vs. Recall Test

The performance of the detector has been evaluated using a precision versus recall curve (see appendix A). To evaluate the model, each intensity of noise scenario has been replicated 30 times. Thus, each precision (Pr) vs. recall (Re) curve in Figure 4.4 represents the average of each intensity scenario (5T–30T). We can easily notice the curves tend to go down as the intensity of the noise starts to increase, because we obtain more false positive peaks when the amount of noise increases.

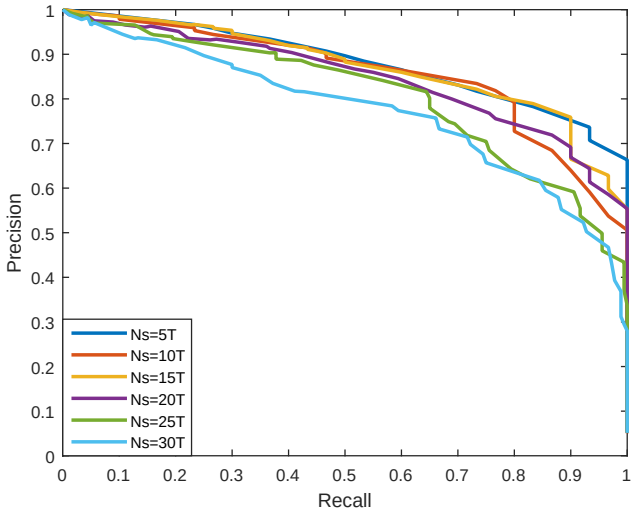


Figure 4.4: Precision versus recall for all intensity scenarios

However, we are able to detect all the true groups correctly as the recall can indicate reaching to value one exactly. On the other hand, the precision will be affected by the false positive peaks and go down gradually through increasing the intensity of the noise.

4.2 3D Data : Preliminary Experiment

We can describe the generation of the preliminary 3D mock data in radian space; assuming galaxies in galaxy groups have distributed normally as a prolonged ellipsoidal pattern along the LoS. While field galaxies (fore/back-ground) have distributed uniformly with different intensities of noise (5T- 30T); we apply the PHTM model to the 3D data after some adjustments.

4.2.1 Preliminary 3D Data Generation

In the case of 3D cone mock data (θ,β and z), where θ and β denote to the RA and Dec, respectively and z is the redshift, we generate the fore/back- ground galaxies as uniform with θ_x, β_x and z generated as:

$$\theta_x = (\theta_{\max} - \theta_{\min}) \cdot \mathcal{U}(0,1) + \theta_{\min} \tag{4.6}$$

By finding the cumulative distribution of β F:

$$\begin{aligned} \mathrm{F}(\beta) &= \frac{1}{\tilde{N}} \cdot \int_{\beta_{\min}}^{\beta_x} \sin(\beta)\mathrm{d}\beta \\ &= \frac{\cos\beta_{\min} - \cos\beta_x}{\mathcal{U}(0,1)\tilde{N}} \\ \beta_x &= \cos^{-1}(\cos\beta_{\min} - \mathcal{U}(0,1)\tilde{N}) \end{aligned} \tag{4.7}$$

where $\tilde{N}$= $\cos\beta_{\min} - \cos\beta_{\max}$. In the same way the redshift z is calculated as:

$$z_x = \sqrt[3]{\mathcal{U}(0,1) \cdot (z_{\max}^3 - z_{\min}^3) + z_{\min}^3} \tag{4.8}$$

While galaxy groups have been generated as 3D Gaussian shapes, elongated, and oriented along the LoS by employing the rotated covariance matrix:

$$Cov_{\mathrm{Rot3D}} = Rot_{\mathrm{3D}} \cdot \begin{bmatrix} \epsilon_k\sigma^2 & 0 & 0 \\ 0 & \epsilon_k\sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \cdot Rot_{\mathrm{3D}}^T \tag{4.9}$$

where Rot_{3D} is the rotation matrix for each group using the perpendicularity of three vectors as shown in Figure 4.5. Given the LoS direction v=(v_x,v_y,v_z) in the Cartesian system, the rotation matrix Rot_{3D} can be derived by considering the local frame $u = (u_x, u_y, u_z)$, $s = (s_x, s_y, s_z)$ and v. We impose: $u \perp$v, $s \perp$ v and $u \perp s$. In other words, the dot products $\mathrm{v}^T u$, $\mathrm{v}^T s$ and $u^T s$ vanish. This leads to an undetermined system. By imposing $u = (0, v_z, -\mathrm{v}_y)$ we automatically satisfy

$\mathrm{v}^T u = 0$. Substituting u in $u^T s = 0$, we obtain:

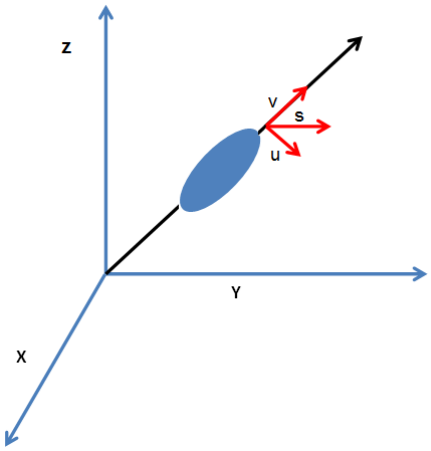


Figure 4.5: Rotation vectors

$$\mathrm{v}_z s_y - \mathrm{v}_y s_z = 0, \quad \frac{\mathrm{v}_y s_z}{\mathrm{v}_z} = s_y. \tag{4.10}$$

Using $\mathrm{v}^T s = 0$, we get:

$$\mathrm{v}_x s_x + \frac{\mathrm{v}_y^2}{\mathrm{v}_z} s_z + \mathrm{v}_z s_z = 0, \tag{4.11}$$

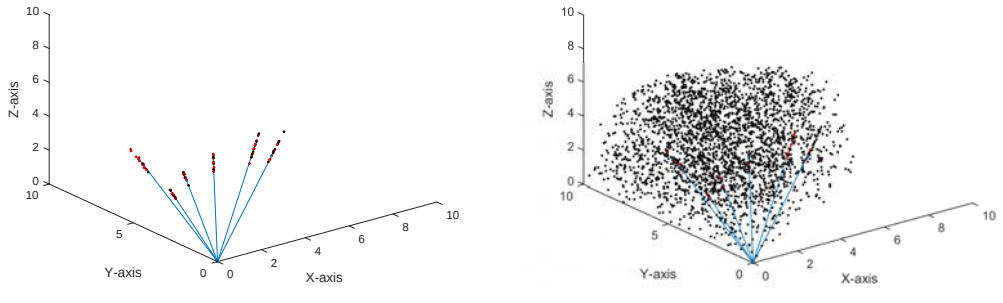
yielding:

$$s_x = \frac{-s_z(\mathrm{v}_y^2 + \mathrm{v}_z^2)}{\mathrm{v}_x \mathrm{v}_z}. \tag{4.12}$$

we are left with one free parameter, s_z, that can be assigned an arbitrary value (we used $s_z = 1$). After normalization of u, s and v into unit vectors, the rotation matrix is formed as follows:

$$Rot_{3\mathrm{D}} = \begin{bmatrix} u_x & s_x & \mathrm{v}_x \\ u_y & s_y & \mathrm{v}_y \\ u_z & s_z & \mathrm{v}_z \end{bmatrix}. \tag{4.13}$$

The generated 3D cone is shown in Figure 4.6 with no noise, and with noise=30T.



(a) (b)

Figure 4.6: (a) 3D cone shape with no noise: six groups have been generated along LoS assuming the origin point is the position of the observer. (b) 3D cone shape with uniformly distributed noise=30T.

4.2.2 Preliminary 3D PHTM Model

After translating from the spherical system (θ, β, z) to the Cartesian $x(\theta, \beta, z)$, $y(\theta, \beta, z)$, z(β, z), the noise model will be a 3-D Gaussian formulated in the corresponding Cartesian coordinate system (x,y,z) and elongated along the LoS. The noise model at the k-th grid point takes the form $p(g_q|(x_k, y_k, \mathrm{z_k}), Cov_{\mathrm{Rot3D}}) = \mathcal{N}(\mu_k, Cov_{\mathrm{Rot3D}})$. To apply the PHTM, the accumulators' set has initiated as an equally likely uniform 3D cone mesh-grid as in Figure 4.7.

Algorithm (1) illustrates the main points of PHTM method :

Algorithm 1 PHTM Algorithm

Input:

RA(θ), Dec(β) and redshift(z): Spatial information consists of galaxy groups and field galaxies.

Output:

H: A landscape of mixture local posterior distribution of galaxies for the expected group positions.

Method:

1- Construct the accumulator as a 3D cone regular grid mapping on the spatial location of the data as in Figure 4.7.

2- For each galaxy g and galaxy group G_k, find the posterior probability of the galaxy group/cluster given the galaxy as shown in Eq.(4.3).

3- Sum the posteriors together to form the landscape see Eq.(4.5).

4- Identify the peaks of the landscape, which are the patterns of interest.

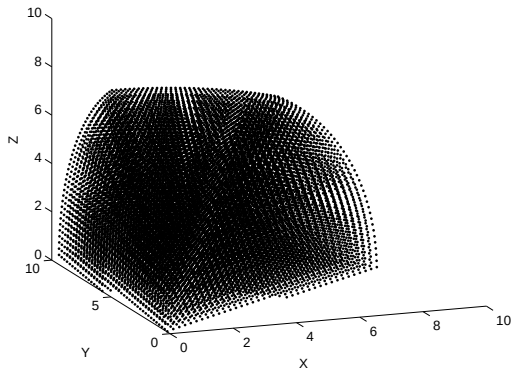


Figure 4.7: A uniform 3D cone meshgrid

4.2.3 Results and Discussion

To find the precision vs. recall curves, we should apply a peak detection method. We detect the local maxima peaks for all scenarios (5T-30T) utilizing the image dilation process, which is a fast and reasonably accurate method that has been used in image processing and detection. A simple example of the dilation process is shown in Appendix(B). The mask form (4.14) has been applied on the 3D $H(x,y,z)$ values. The outcome data from the dilation process ($D_{\oplus}$) is compared with the 3D $H(x,y,z)$ values. If the value of $D_{\oplus}$ is less than the corresponding H value, then the specific (x,y,z) coordinate of the H data point will be considered as a peak.

$$msk_{d_1,d_3} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad msk_{d_2} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{4.14}$$

The final precision vs. recall curves (each curve represents the average of 20 replications) for all 3D scenarios (5T-30T) for 6 groups are illustrated in Figure 4.8. As in the 2D flat area, the detection process has degraded when the intensity of noise has been increased. In reality, due to the limited sensitivity of observational devices, more distant galaxies are less likely to be detected than those comparable at closer redshift; this is called the flux limit effect. The model developed so far will not work in real cosmology since it does not account for the flux limit effect. In the next chapter, we will discuss in detail the generation of the mock galaxy groups and galaxies in the field with the flux limit effect; and furthermore, how we adjust the model to compensate for the incompleteness of groups along the

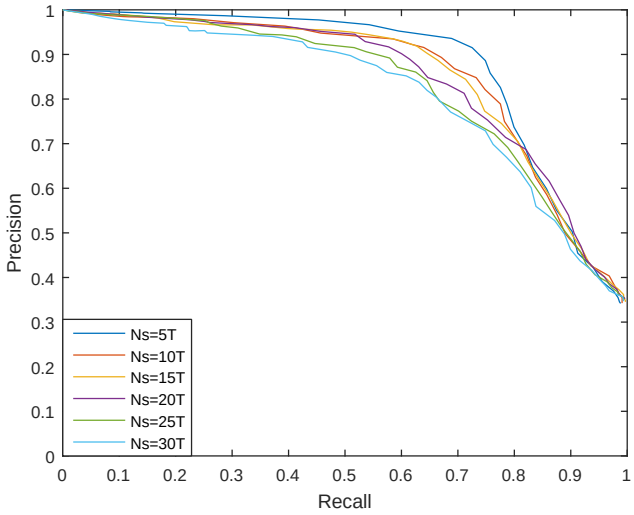


Figure 4.8: Precision vs. recall curves for 3D cone-shape scenarios

redshift bands.

CHAPTER 5

REALISTIC 3D DATA GENERATION AND MODEL MODIFICATIONS

To identify any phenomenon, we need to have an accurate description of it, whether it happens in Space or on Earth. Also, to devise a reasonable approach to detecting aggregated galaxies (galaxy groups and clusters), we cannot just assume any randomness of data patterns. Astrophysicists have discovered that each kind of celestial object has its properties (movement direction, distribution, and whether influenced by gravitational effects from its neighbourhood objects). It is impossible that we can find such objects or groups of objects without studying their relations and aggregations, in addition to what kind of possible patterns they can form.

Due to the limitations of the telescopes' observations of the real data, astrophysicists tend to generate mock data to simulate reality. Furthermore, in the actual data, there are some difficulties in suggesting the ground truths of such phenomena such as galaxy groups and clusters. Whereas the ground truths are more likely to be known of these phenomena in the generated data, which helps in

the study of the possible densities of celestial objects and finding a way to distinguish between them. Therefore, mock galaxy surveys tend to simulate very detailed processes (phenomena: e.g. redshift distortion, flux limit factors and gravitational lensing, celestial objects); and that leads to hugely sophisticated data.

Mock galaxy surveys are effective to test and evaluate different cosmological models such as galaxy groups' detectors and to analyse statistical properties: such as luminosity functions, the distribution of the particles in the group, the behaviour of the gases, the interactions between the galaxies during the collapse process and the star formation process.

In this chapter, the distributions of galaxies inside galaxy groups and clusters in GAMA mock surveys are confirmed, via comparing to the theoretical profiles that have been provided by astrophysicists, the generation of complete and well-defined realistic 3D data, by including the flux limit effect, is discussed and the 3D probabilistic Hough transform based on the adaptive local kernel (PHTALK); which is the updated version of the PHTM is applied.

5.1 Galaxy Distributions Inside Galaxy Groups and Clusters

To perform theoretical based modelling, we would need to know how galaxies are distributed in a group - on average; but there is no analytical model describing this. In reality, this model is needed as an input and the only way to determine what to input is via numerical simulations. Galaxy formation and evolution is

not amenable to analytical modelling. The evolution of galaxies is complex and the only way to understand their distribution is to look at numerical simulations. Ideally, one would like to take some very different numerical simulations, determine galaxy distributions (spatial and velocity) in those and then use that template for different mock data. However, in this case we have one large simulation and so we can use a few volumes of the GAMA mock survey (see section 2.12) to determine the distributions of galaxies in their systems (i.e. groups and clusters). We have analysed the distributions towards two directions: one related to galaxy redshift distributions (velocity dispersion) along the LoS and the other related to galaxy radial distributions orthogonal to the LoS.

5.1.1 Galaxy Redshift Distributions

Empirically, we collected true groups from the GAMA data based on their mass (we collected the groups within two mass bands (10^{12}-10^{13} $M_{\odot}$) and (10^{13}-10^{14} $M_{\odot}$), since most galaxy groups are concentrated within these bands). For each mass band we assumed five redshift bands [$0.01 < z \leq 0.1$].

To plot redshift dispersion, we take group galaxies and subtract from their redshifts the group's redshift (which is the x-axis in Figure 5.1) and find their histogram. We compare the empirical curve (blue colour) with the theoretical curve (red colour). We establish Gaussian distributed z projected onto the LoS of the galaxy group, with redshift dispersion σ_z, that is calculated as:

$$\sigma_z(M_{500}) = (1+z)\frac{\sigma_v(M_{500})}{c}, \tag{5.1}$$

where $\sigma_v(M_{500})$ is the velocity dispersion of a galaxy group distributed as (Pearson

et al. 2015):

$$3\log_{10}\left(\frac{\sigma_v}{537.2}\text{km/s}\right) = 0.94\log_{10}\left(\frac{M_{500}E(z)}{10^{14}}\right) - 0.0403 + \mathcal{N}(0, 0.26) \qquad (5.2)$$

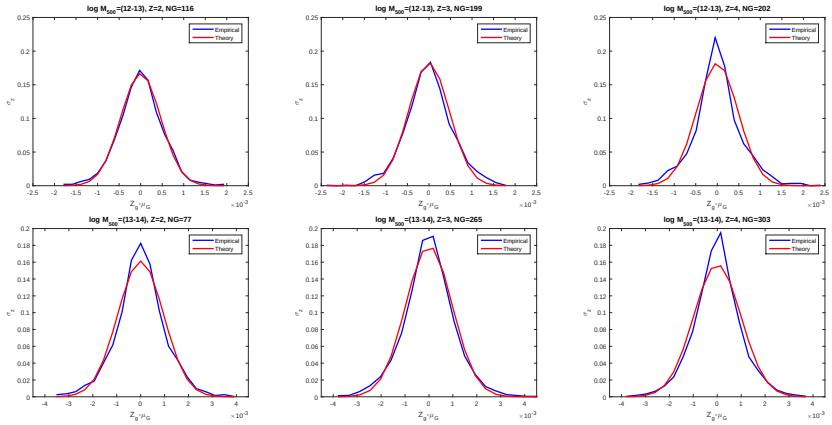


Figure 5.1: Redshift galaxy distributions along LoS: the blue curve is the empirical redshift dispersion and the red curves are the theoretical curves with σ_z, calculated as shown in Eq. (5.1)

We obtained an empirical distribution, similar to the theoretical profile, which resembles a Gaussian profile.

5.1.2 Galaxy Radial Distributions

For the empirical radial galaxy distributions, we again collect true groups from GAMA data based on two bands of mass ($10^{12} - 10^{13}$ $M_\odot$) and ($10^{13} - 10^{14}$ $M_\odot$).

For a given group we conduct concentric circles around the group centre (as in Figure 5.2) with r_1 (radius of the first circle) being bin_1, dr_1=r_2-r_1 being bin_2 (annulus) and so on. We count the number of galaxies ΔN in each bin. Then we

calculate the number of galaxies per square distance, for bin_1 as

$$\Sigma(\mathrm{r}) = \frac{\Delta N_1}{\pi \mathrm{r}_1^2}, \tag{5.3}$$

while the rest of the bins are calculated as:

$$\Sigma(\mathrm{r}) = \frac{\Delta N_a}{2\pi \mathrm{r}_a \mathrm{dr}_a}, \tag{5.4}$$

where $\mathrm{dr}_a = \mathrm{r}_{a+1} - \mathrm{r}_a$.

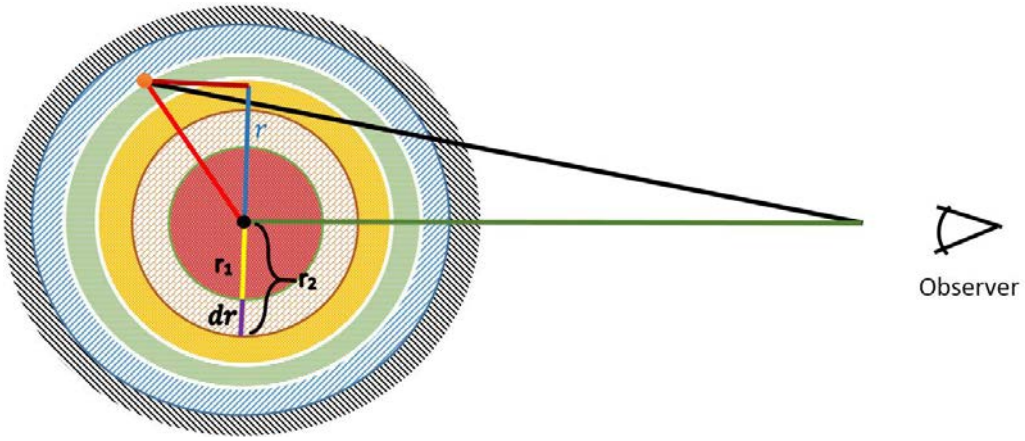


Figure 5.2: A schematic plot of computing the parameters to check the radial galaxy distributions.

For the theoretical section, we find the projected radii (r, the blue line in Figure 5.2) for all galaxies belonging to each group. We find the corresponding NFW values, $\Sigma(r)$ as mentioned in section 2.11, for each projected radius r of each galaxy per group.

We look at the stacked galaxy profiles (empirical profiles) and compare the similarity in shape to the NFW profile (theoretical profiles) by plotting the em-

pirical and theoretical curves as shown in Figure 5.3. We have for each band of mass the empirical (starred blue curve) and theoretical (dashed black curve) for all z bands.

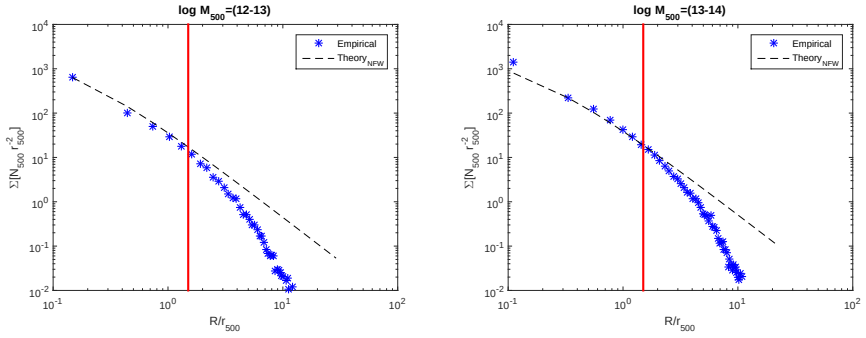


Figure 5.3: The radial galaxy distributions orthogonal to LoS for two bands of mass: the blue curves are the empirical radial dispersions and the black curves are the theoretical (NFW profile) curves calculated as shown in section (2.11). The area of interest is between 10^{-1} (x-axis) and the red vertical line based on the radius reasonable cut-off ($1.5r_{500}$).

In both cases, the shape is consistent with the sampled NFW and sampled Gaussian. In our detailed control experiments, we employ two sets of data: our generation of the mock data including the flux limit effect (in this chapter) and Galaxy And Mass Assembly (GAMA) mock data (in the next chapter). On both mock data, we use redshift cut-off till (0.1) with apparent magnitude limit $m =$ 19.8 and mass range (10^{10}M$_\odot$-10^{15}M$_\odot$).

5.2 Restricted Mock Data

Before we apply our model on GAMA mock (more realistic data), we need to test and improve our approach based on restricted data at close redshifts $z \leq 0.1$ with specific ranges of (θ=RA and β=Dec) angles as a control experiment. We generate 3-D realistic data consisting of two parts: galaxy groups' generation and

fore/background galaxies' generation. The groups are generated from a joint distribution, consisting of a Gaussian distribution of dispersed, projected velocities along the LoS and the radial distribution in the orthogonal complement of the LoS, formulated using the Navarro, Frenk and White (NFW) density profiles (Navarro et al. 1996). In addition, our limited mock data have been generated with different intensities of foreground/background noise, which is (5, 10 or 15) times the intensity of galaxies in all generated galaxy groups, to test the efficiency of our approach.

To achieve a large set of controlled experiments, we have tried to be more concise in generating data in a reasonable way through simulating the distribution of galaxy groups and the dispersion of the galaxies inside. The galaxy groups in our data are similar to the real galaxy groups; however, the foreground/background is uniformly distributed, as we have assumed the test will be within the closest redshift cosmology; RA, Dec 5×5 deg^2 have been specified.

The flux limit effect in the generation of the synthesised galaxy groups has been included in the next section and in the fore/background objects' intensities based on the degree of brightness using the Schechter luminosity function, as illustrated in section (5.2.2).

We considered a flat ΛCDM, $\Omega_k = 0$ cosmology and cosmological parameters Ω_M=0.28, Ω_Λ=0.72 and h=$H_0/100$ (kms^{-1} Mpc^{-1})=0.697.

5.2.1 Generating Galaxy Groups

The steps for generating simulated galaxy groups oriented along the LoS with overdensity equivalent to 500 times the critical density of the Universe are:

1- Specifying a small solid angle Ω by assuming both (Ra=θ) and (Dec=β) angles from -2.5 to 2.5 (i.e. Δ for θ, $\beta = 5$):

$$\Omega = [\sin\beta]_{\beta_1}^{\beta_2} d(\beta) d(\theta) \tag{5.5}$$

2- Finding the total comoving volume, for a patch of the sky of solid angle Ω (in sr); the comoving volume element dV_c within a redshift range dz (centred at Z band) is given by Eq.(2.11). The total comoving volume (V_c) can be found by Eq. (2.12)

3- Using the online tool (HMFcalc)[1] by (Murray et al. 2013) to find the halo mass function (HMF) $dn/d\ln(M_h)$ as shown in Eq.(2.13), which quantifies the number of halos per unit comoving volume of the Universe as a function of their mass, with respect to the following characteristics: transfer function Wilkinson Microwave Anisotropy Probe (WMAP); $\Delta_{\rm halo} = 500$ on critical density. The Reed fitting function f has been chosen, which has been improved and modified from the Sheth-Tormen(S-T) mass function (Reed et al. 2007, Sheth & Tormen 1999), mass range (10^{10}M$_\odot$-10^{15}M$_\odot$), bin width 0.05, and, redshift range (0.01- 0.1).

[1] http://hmf.icrar.org/

4- Calculating the group mass function δN, which is the mean of the Poisson distribution:

$$\delta N = \frac{dn}{d\ln M_h} d\ln mV \tag{5.6}$$

where $\frac{dn}{d\ln M_h}$ is acquired through Eq. (2.13); $d\ln m$ represents the log mass bin width (i.e. the offset among the masses' intervals); V is the total comoving volume at redshift range dz centred at z.

5- Applying the Poisson random number generator by considering δN values as mean values to find the number of groups is required to generate per Z interval per comoving volume of that interval.

6- Finding the richness of the galaxy group from the relation that is scaled to a lower absolute magnitude limit $M = -16.5$, which is independent of redshift;

$$\log_{10}\left(\frac{N_{\rm grp}}{224}\right) = 0.97\log_{10}\left[\frac{M_{500}\cdot\mathcal{U}(0,1)}{10^{14}}\right] - 0.0411 + \mathcal{N}(0,0.2) \tag{5.7}$$

This relates the $\log_{10}$ of mass M_{500} to the $\log_{10}$ of the number of galaxies $N_{\rm grp}$ within a group and has a scatter in $\log_{10}$ space of 0.2 - i.e. the distribution in log richness at a given M_{500} can be drawn from a Gaussian $\mathcal{N}(\mu_i, 0.2)$.

7- Including the effect of the flux limit on the number of galaxies in each group. We calculate the ratio and the number density of the luminosity function of each group at its redshift z_c. The calculation is achieved by integrating over the galaxy luminosity function $\Phi(M)$ annotated by $\Phi(M)_{\rm gal}$ in Figure 5.5.

The $N_{\rm grp}$ value is within an absolute magnitude limit (-25 – -16.5); whereas to add redshift dependence, we should find $N_{\rm grp}$ within the limit (-25 – $M(z)$) for a particular z_c value of galaxy group;

$$f_{\rm scale}(z_c) = \frac{\int_{-25}^{M(z_c)} \Phi(M) dM}{\int_{-25}^{-16.5} \Phi(M) dM} \tag{5.8}$$

where $M(z_c)$ is the conversion from the apparent magnitude m to the absolute magnitude M at a redshift z_c as depicted in Eq. (2.10), $S(M)$ is the luminosity Schechter function (Schechter 1976);

$$\Phi(M) = \frac{\ln(10)}{2.5} \cdot \phi^* \cdot \left(10^{\frac{M^*-M}{2.5}}\right)^{(\alpha+1)} \cdot \exp\left\{-10^{\frac{M^*-M}{2.5}}\right\}, \tag{5.9}$$

with parameters from the r-band cluster luminosity function of (Pearson et al. 2015, Popesso et al. 2005); $\Phi^* = 1.49 \times 10^{-2} h^3\,{\rm Mpc}^3$ is the number density; $M^* = -21.35 + 5\log_{10} h$ is characteristic magnitude; and $\alpha = -1.3$ is the faint end slope.

The final number of galaxies $N_{{\rm grp}_f}$ for a certain galaxy group considering the flux limit effect at redshift z_c is:

$$N_{{\rm grp}_f} = N_{\rm grp} f_{\rm scale}(z_c). \tag{5.10}$$

8- For each group, we specify an arbitrary group centre in the spherical coordinate and within a specified range of (z, Ra=θ, Dec=β) as mentioned in Eq. (4.6, 4.7, 4.8).

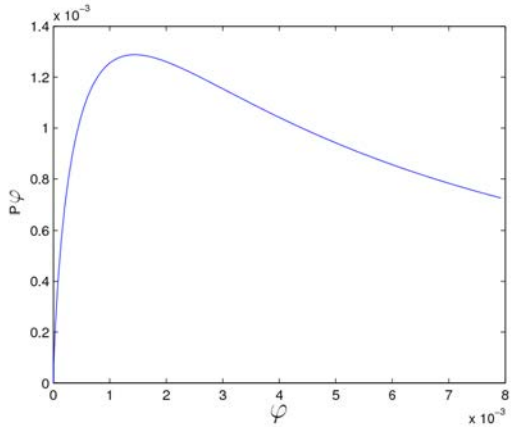


Figure 5.4: The probability of projected NFW profile for a galaxy group where φ in arcmin unit and the area under the curve equal to unity.

9- Finding the radial distribution - the surface mass densities using the projected NFW profile as depicted in Eq. (2.14). However, before applying $\Sigma_{\rm NFW}(r)$, we have to change the actual size of the group radius r into an angular size (φ) through dividing by D_a Eq. (2.7). Instead of r and r_s, we have $\varphi = \frac{r}{D_a(z_c)}$, φ in arcmin unit, and $\varphi_s = \frac{\rm r_s}{D_a(z_c)}$. r values are located within ($r_{\rm min}$-$r_{\rm max}$) range; where $r_{\rm max}$=1.5 r_{500}, is the maximum radius of the group.

The probability of observing a galaxy g at angular size φ_g, as shown in Figure 5.4, is:

$$P(\varphi_g) = \frac{2\pi\varphi_g\Sigma_{\rm NFW}(\varphi_g)}{\int_{\varphi_{\rm min}}^{\varphi_{\rm max}} 2\pi\varphi\Sigma_{\rm NFW}(\varphi)d\varphi} \tag{5.11}$$

10- Specifying an arbitrary angular distance $\varphi_{\rm win}$ within the range $(\varphi_{\rm min}, \varphi_{\rm max})$ for each galaxy from its particular group centre using the inverse transform method,

a- generate a random number $\mathcal{U}(0,1)$ and deliver the random variable $\varphi_{\rm win}$:

$$\varphi_{\rm win} = \varphi_i, \text{ iff } \mathrm{F}(\varphi_{i-1}) < \mathcal{U} \leq \mathrm{F}(\varphi_i) \tag{5.12}$$

where F is the cumulative distribution function.

b- generate the angle γ, $\mathcal{U}(0, 2\pi)$, which represents the azimuthal, angle of positioning galaxy g around the centre of the group.

$$\theta_g = \theta_c + \varphi_{\rm win}\cos(\gamma) \tag{5.13}$$

$$\beta_g = \beta_c + \varphi_{\rm win}\sin(\gamma) \tag{5.14}$$

11- Find the velocity distribution $\boldsymbol{v}$, which represents how the galaxies distribute in velocity space (Pearson et al. 2015) as in Eq 5.2.

$$\Delta v_g = \mathcal{N}(0, \sigma_v) \tag{5.15}$$

The approximate redshift values of galaxies within a particular group after translating them into the correct positions are:

$$\boldsymbol{z_f} = z_c + \frac{(1+z_c)\Delta v_g}{c} \tag{5.16}$$

then translate from the spherical space (θ_g, β_g, z_f) to the Cartesian space $x(\theta_g, \beta_g, z_f)$, $y(\theta_g, \beta_g, z_f)$, $\mathrm{z}(\beta_g, z_f)$.

The generation process and the generated galaxy groups have been checked

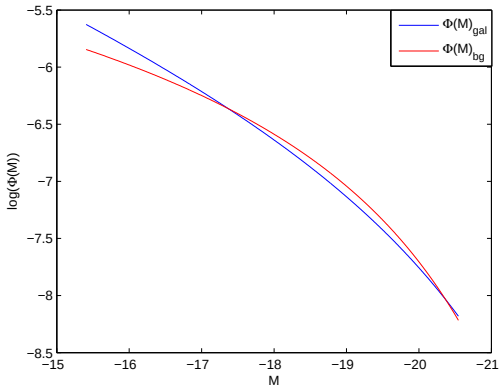


Figure 5.5: The profile of the luminosity function vs. the magnitude range for both galaxies in groups and in background respectively

and verified by astrophysicists.

5.2.2 Generating Fore/Back-ground galaxies

The fore/back-ground galaxies have been generated and distributed uniformly with different degrees of faint based on the redshift and the comoving volume as follows:

a- finding the density of galaxies in the entire region $d = TN_{gf}$; where N_{gf}, $N_{gf} = \sum N_{\mathrm{grp}_f}$, is the number of galaxies in flux limited galaxy groups and $T = (5, 10, 15)$ is a multiplicative factor to imitate the galaxy intensity surveys.

b- following the same relation of generating the galaxies of the galaxy groups with the degree of faint (section 5.2.1, step 7); we have changed the parameter values $M^* = -20.44 + 5\log_{10}(h)$ and $\alpha = -1.05$ in order to find $\Phi(M)$ for the background based on the redshift as annotated by $\Phi(M)_{bg}$ in Figure 5.5.

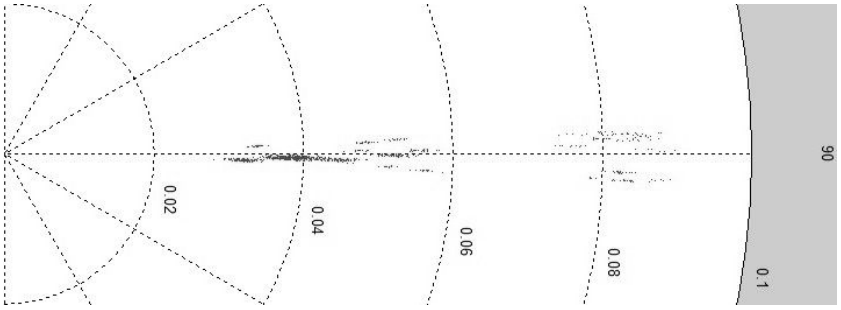
c- supposing we have $\mathcal{J}$ redshift bins, we find the comoving volume at each redshift bin $V_c(z_b)$. The number of background galaxies $N_{\rm bg}$ to generate at redshift bin z_b is:

$$N_{\rm bg_b} = d \cdot \frac{V_c(z_b) f_{\rm scale}(z_b)}{\sum_{i=1}^{\mathcal{J}} V_c(z_i) f_{\rm scale}(z_i)} \tag{5.17}$$

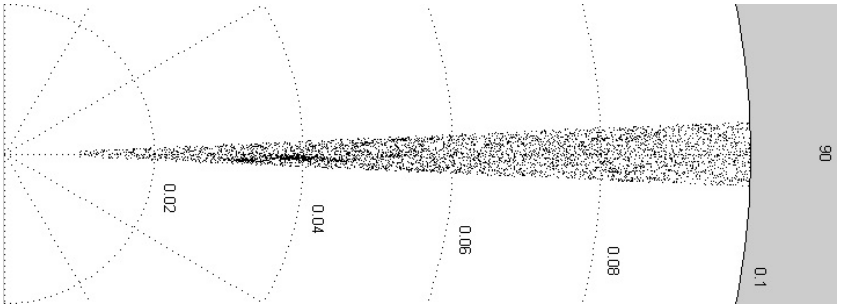
A 2D projection, RA versus redshift, slice of 26 galaxy groups of the generated mock data is depicted in Figure 5.6a. Background/foreground galaxies are generated with an intensity equivalent to 5 times the number of galaxies in all flux limited groups as shown in Figure 5.6b. The intensity of the fore/back-ground galaxies has been increased , $T = (5, 10, 15)$, purely to demonstrate the power of the HT and verify the idea from the machine learning point of view. It is worth mentioning that the intensity of fore/back-ground galaxies with $T = 5$ could correspond to a realistic situation (observation). However, the density model of the fore/back-ground is completely different; thus, these two situations cannot be compared.

5.3 The Groups Finder: Probabilistic Hough Transform Based on Adaptive Local Kernel (PHTALK)

To reduce the false positive peaks without affecting the detection of the actual group peaks, the model has been improved further to deal with the flux limit effect (a galaxy group/cluster of a specific mass at high redshift appears fainter with less number of galaxies comparing to a galaxy group/cluster has the same mass located at closest redshift). To accelerate the model computation time, the calculations have confined to only the galaxy positions to be considered as probable potential galaxy group candidates.



(a) RA vs. Z: galaxy groups with no noise



(b) RA vs. Z: galaxy groups with the flux limit intensity noise= 5T

Figure 5.6: One projected 2D slice of the new mock data in a polar system

The key object in our probabilistic Hough transform is a probability distribution characterizing a galaxy group of mass M_{500} centred at G, assigning to each possible galaxy position g in the cone a density value $P(g|G, M_{500})$. Given a prior $P(G|M_{500})$ specifying how likely a-priori it is to find a galaxy group of mass M_{500} at location G, we can then calculate the posterior probability of the galaxy group at a particular position G, given an observed galaxy g, $P(G|g, M_{500}) \propto P(g|G, M_{500})P(G|M_{500})$. In particular, given a set of observed galaxies g_q, $q = 1, 2, .., N_{\text{gal}}$, and a set of candidate group positions G_k, $k = 1, 2, ..., N_G$, we have

$$P(G_k|g_q, M_{500}) = \frac{P(g_q|G_k, M_{500})P(G_k|M_{500})}{P(g_q|M_{500})}, \tag{5.18}$$

where:

$$P(g_q|M_{500}) = \sum_{j=1}^{N_G} P(g_q|G_j, M_{500})P(G_j|M_{500}). \tag{5.19}$$

The posterior $P(G_k|g_q, M_{500})$ can be interpreted as a probabilistic vote of a galaxy g_q for the group of mass M_{500} positioned at G_k. The Hough landscape for galaxy groups of mass M_{500} is then obtained by accumulating votes from all galaxies for each group position G_k in a flat mixture:

$$H(G_k|M_{500}) = \frac{1}{N_{\rm gal}} \sum_{q=1}^{N_{\rm gal}} P(G_k|g_q, M_{500}). \tag{5.20}$$

However, galaxies at higher redshift are more difficult to observe and we compensate for this flux limit effect by weighting the galaxy contributions with weights $w(g_q)$ depending on their redshift, instead of simply giving each galaxy equal vote weight $1/N_{\rm gal}$:

$$H(G_k|M_{500}) = \sum_{q=1}^{N_{\rm gal}} w(g_q).P(G_k|g_q, M_{500}) \tag{5.21}$$

where $w(g_q) \geq 0$, $\sum_{q=1}^{N_{\rm gal}} w(g_q) = 1$.

In general, the mass of the expected group at position G_k is unknown and we express this uncertainty through a distribution over possible mass bands $\ell = 1, 2, ..., N_M$, $P(M_{500}^{\ell}|G_k)$. We therefore calculate the expected vote for G_k with respect to the mass distribution around G_k:

$$\begin{aligned} H(G_k) &= \mathbf{E}_{P(M_{500}|G_k)}[H(G_k|M_{500})] \\ &= \sum_{\ell=1}^{N_M} P(M_{500}^{\ell}|G_k) H(G_k|M_{500}^{\ell}) \end{aligned} \tag{5.22}$$

In the following sections the basic building blocks of the general model introduced above will be expanded in greater detail.

5.3.1 The Likelihood Model $P(g_q|G_k, M_{500})$

Consider a group position G_k in a 3D cone; we denote the redshift component of G_k by Z_k. Given a galaxy g_q with redshift z_q and angle φ_q^k from the LoS going through G_k, the $P(g_q|G_k, M_{500})$ is modelled as a joint distribution,

$$P(g_q|G_k, M_{500}) = p(z_q|G_k, M_{500})p(\varphi_q^k|G_k, M_{500}), \tag{5.23}$$

of a Gaussian distribution of LoS velocities and the radial distribution in the orthogonal complement of the LoS formulated using the projected Navarro, Frenk and White(NFW) density profile (Bartelmann 1996, Navarro et al. 1996).

The projected z_q along the LoS, relative to the group position G_k can be obtained as $\Delta\tilde{z}_q^k = \Delta z_q^k \cos(\varphi_q^k)$, where $\Delta z_q = z_q - Z_k$, $\varphi_q^k = \arccos(\mathbf{v}_k \cdot \mathbf{u}_q)$, $\mathbf{v}_k$ and $\mathbf{u}_q$ are unit vectors along the LoS of the galaxy group G_k and galaxy g_q respectively, as depicted in the schematic of Figure 5.7.

We assume Gaussian distributed z projected onto the LoS of the galaxy group, with dispersion $\tilde{\sigma}_z^k$,

$$\tilde{\sigma}_z^k(M_{500}) = (1+z)\frac{\tilde{\sigma}_v^k(M_{500})}{c}, \tag{5.24}$$

where $\tilde{\sigma}_v^k(M_{500})$ is a random quantity distributed as shown in Eq.(5.2). This equa-

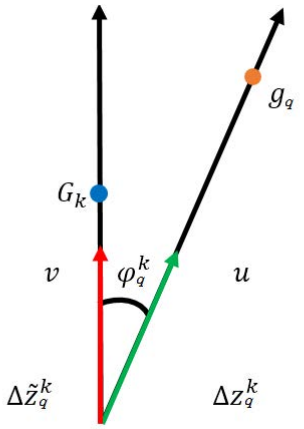


Figure 5.7: A schematic graph of finding the required parameters of the likelihood model

tion can be rewritten as:

$$(\sigma_v^k)^3 = \left(\frac{1.5510^8}{10^{0.0403}}\right)\left(\frac{M_{500}E(Z_k)}{10^{14}}\right)^{0.94} 10^{\epsilon}. \tag{5.25}$$

Denoting

$$A = \left(\frac{1.5510^8}{10^{0.0403}}\right)\left(\frac{M_{500}E(Z_k)}{10^{14}}\right)^{0.94}$$

we get

$$\sigma_v^k = A^{1/3}\mathbf{E}[10^{\epsilon_1}], \tag{5.26}$$

where $\epsilon_1 \sim \mathcal{N}(0, \sigma_\epsilon/3)$. We have $10^{\epsilon_1} = \exp\{\epsilon_1 \ln 10\}$ therefore,

$$\sigma_v^k = A^{1/3}\mathbf{E}\big[\exp\{\epsilon_2\}\big] \tag{5.27}$$

where $\epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2})$, and $\sigma_{\epsilon_2} = (\sigma_\epsilon \ln 10)/3$.

Denoting $O = 1/(\sqrt{2\pi}\sigma_{\epsilon_2})$, we find:

$$\begin{aligned}
\mathbf{E}\big[\exp\{\epsilon_2\}\big] &= O\int_{-\infty}^{\infty}\exp\left\{\frac{-u^2}{2\sigma_{\epsilon_2}^2}\right\}\exp\{u\}du \\
&= O\int_{-\infty}^{\infty}\exp\left\{\frac{-u^2+2u\sigma_{\epsilon_2}^2}{2\sigma_{\epsilon_2}^2}\right\}du \\
&= O\int_{-\infty}^{\infty}\exp\left\{\frac{-(u-\sigma_{\epsilon_2}^2)^2}{2\sigma_{\epsilon_2}^2}+\frac{\sigma_{\epsilon_2}^2}{2}\right\}du \\
&= \exp\left\{\frac{\sigma_{\epsilon_2}^2}{2}\right\}O\int_{-\infty}^{\infty}\exp\left\{\frac{-(u-\sigma_{\epsilon_2}^2)^2}{2\sigma_{\epsilon_2}^2}\right\}du \\
&= \exp\left\{\frac{\sigma_{\epsilon_2}^2}{2}\right\}.
\end{aligned} \tag{5.28}$$

Hence,

$$\sigma_v^k = A^{1/3}\exp\left\{\frac{\sigma_{\epsilon_2}^2}{2}\right\}, \tag{5.29}$$

then

$$\sigma_z^k = (1+z)(\sigma_v^k/c), \tag{5.30}$$

and

$$p(z_i|G_k, M_{500}) = \frac{1}{\sqrt{2\pi}\sigma_z^k}\exp\left(-0.5\frac{(\Delta\tilde{z}_i^k)^2}{(\sigma_z^k)^2}\right). \tag{5.31}$$

The angle φ_q^k is used to find the projected radius $r_q^k = \varphi_q^k D_a(Z_k)$, where $D_a(Z_k)$ is obtained via Eq. (2.8).

The surface mass density of the group, $p(\varphi_q^k|G_k, M_{500})$, can be obtained using

the projected normalized NFW profile:

$$p(\varphi_q^k|G_k, M_{500}) = \frac{r_q^k \Sigma(r_q^k)}{\int_0^{r_{max}} r\Sigma(r)\ dr} \tag{5.32}$$

where $r_{max} = 1.5r_{500}$ is the maximum radius. The projected NFW profile $\Sigma(r_q^k)$ is described in section 2.11.

5.3.2 Considering the Degree of Faintness

In Eq. (5.21) a weight is given to the vote of each galaxy g_i based on its redshift z_i and absolute magnitude M_i to compensate for the shortfall in the number of galaxies during the voting procedure especially at high redshift. In other words, we are more likely to observe galaxies of the same magnitude close by (at smaller redshift) than at high redshift. The weights should sum to 1 and need to be inversely related to the (Schechter 1976) luminosity function Eq.5.9.

The weight for galaxy g_i is then

$$w(g_i) = \frac{\mathcal{S}_i^{\psi}}{\sum_{q=1}^{N_{\rm gal}} \mathcal{S}_q^{\psi}}, \tag{5.33}$$

with $-1.5 \leq \psi \leq 1.5$ modulating the influence of $\mathcal{S}_i$ on the weight profile[1] and,

$$\mathcal{S}_i = \int_{-25}^{M_i} \Phi(M)\ dM, \tag{5.34}$$

where M_i is the conversion from the apparent magnitude m to the absolute mag-

[1] $\psi = 0$ corresponds to the equal weight setting Eq. (5.20), large values of ψ concentrate weights on galaxies with largest $\mathcal{S}_i$.

nitude at redshift z_i as depicted in Eq.(2.10).

For the generated data in this chapter, we have created three different cones, in order to check the value of beta and that gave better results. We found $\psi = 0.5$ to work robustly on the generated mock data. The same process was carried out for the GAMA mock data in the next chapter; where we selected three different cones and checked their outcomes by applying different ψ values. Then we found $\psi = 1.3$ to work robustly on the GAMA mock data.

5.4 Testing PHTALK on the Newly Generated Mock Data

We apply both the PHTALK method and FoF method by Eke et al. (2004) on the generated data and compare the outcomes of both methods using precision versus recall for each intensity of noise scenario. We assume a uniform prior for each of $P(G_k|M_{500})$ in Eq. (5.18) and $P(M_{500}|G_K)$ in Eq. (5.22); then search for any detected galaxy group within a tolerance boundary of (1.5 virial radii) perpendicular to the LoS and ($2\sigma_z$) along the LoS around the ground truth galaxy group centres, within tolerance (boundary) based on the true halo mass limit, to evaluate the performance of the methods (see Appendix C). We generate 30 examples for each intensity of fore/background with fixed galaxy group positions and the average Pr versus Re curves of them; depicted in Figure 5.8 for one stripe from the mock data cones, consisting of 25 galaxy groups and with three different intensities of noise background, T=(5,10,15). We compare the mean positions of the expected galaxy groups' candidate for both FoF and PHTALK, with the mean positions of the actual galaxy groups of the simulated data.

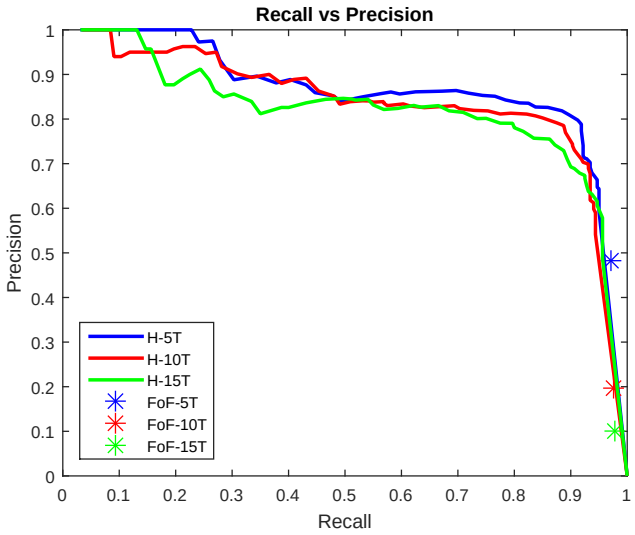


Figure 5.8: Precision vs. recall for the mock data example: the curves represent the PHTALK detection outcome, which is denoted as (H); while the coloured stars represent the FoF by the Eke et al. (2004) method outcome.

Note that while it is very natural to create precision versus recall curves from PHTALK (by varying τ), this turned out to be cumbersome for FoF (modifying free parameters can lead to abrupt changes in performance). Therefore, we report a single value (star) of (precision versus recall) obtained with the parameter setting recommended in Eke et al. (2004). The star points represent the performance of the FoF method; while the curves represent the performance of PHTALK. As noticed in Figure 5.8, the FoF indicators (stars) decrease (along y-axis) means the FP increases as long as the intensity of noise increases. While on the other hand, the PHTALK curves remain approximately consistent through all intensity of noise scenarios T=(5,10,15). It is shown clearly that PHTALK has not been affected dramatically by increasing the noise fore/back-ground, in contrast to the FoF case. In conclusion, PHTALK is more robust for identifying the patterns in a highly noisy environment.

The reason for using realistic mock data (i.e. 3D generated data and GAMA mock data (in Chapter 6)) instead of using real data is related to the testing and validation procedures. The TPs are not clearly known (prominent) in the real data, thus testing and validation procedures related to the efficiency of the detection cannot be conducted.

CHAPTER 6

GAMA MOCK DATA AND FURTHER PHTALK MODEL UPDATES

Due to the existence of filament patterns in the real and realistic mock surveys (such as GAMA mocks), PHTALK model needs to be updated to overcome the spurious galaxy group patterns and focus on detecting the actual groups. We suggest and implement some procedures to suppress and refute FP peaks (spurious galaxy groups candidates). As well as improve the probabilistic model accuracy in signifying galaxy groups positions.

The generated mock data in the previous chapter does not have the filament patterns. Thus, we did not suffer severely from the FP and simple filtration based on 3D image dilation to find the local maxima does work very well. While using the realistic GAMA mock data, we reduced the false positives based on three different factors of a given galaxy: a prior knowledge from the local density of the galaxy, its luminosity, and its H value.

6.1 PHTALK Updates

6.1.1 The prior $P(G_k|M_{500})$

The prior $P(G_k|M_{500})$ specifies how likely it is to find a galaxy group of mass M_{500} at position G_k. We determine $P(G_k|M_{500})$ in two ways: (i) based on theoretical expectation of finding a galaxy group of mass M_{500} at the given Z_k of G_k. There is no local density information of galaxies included in this measure. The prior calculated in this way expresses a bias towards groups with smaller mass at a given z and reflects the evolving mass distribution in groups; (ii) based on local density of galaxies, we produce an estimate of the mass of the galaxy group, should such a group indeed exist at position[1] G_k.

The prior $P(G_k|M_{500})$ based on Z_k

The prior of a group position G_k given a M_{500}^{ℓ} band needed in Eq. (5.18) can be determined as:

$$P(G_k|M_{500}^{\ell}) = \frac{P(G_k)P(M_{500}^{\ell}|G_k)}{\sum_{j=1}^{N_G} P(G_j)P(M_{500}^{\ell}|G_j)}, \tag{6.1}$$

where, assuming no a-priori preference for some grid points over the others, $P(G_k) = 1/N_G$. As explained above, the mass likelihood $P(M_{500}|G_k)$ expresses the notion that some group masses will be more likely than others at redshift Z_k. To ex-

[1] Note that this is strictly speaking not a prior, as we use the observed galaxies to express the level of confidence that at position G_k there is a group of mass M_{500}. However, it seems natural to use the local galaxy density for this purpose, as relying purely on Z_k can be insufficient. Indeed, there can be two group candidate positions with the same z but very different local galaxy densities. This cannot be easily resolved through likelihood formulation, as the likelihood term in our case specifies how likely it is that a certain galaxy is a member of a group at G_k, *assuming that indeed there is a galaxy group at* G_k.

press explicitly that in this case the conditioning on the grid position G_k is really conditioning on the Z_k of G_k, we write $P(M_{500}|G_k)$ as $P(M_{500}|Z_k)$.

To obtain $P(M_{500}|Z_k)$, we calculate the group mass function δN from the halo mass function (HMF) (using the online tool (HMFCalc) developed by (Murray et al. 2013)) Eq.5.6.

δN is the mean $\lambda(M_{500}, Z_k)$ of Poisson distribution over galaxy group number N_m at Z_k, giving the expected number of galaxy groups of mass M_{500} at redshift band centred at Z_k, per unit comoving volume. Binning the group mass into N_M bins and using the mean of the (normalized) galaxy group number, we formulate $P(M^{\ell}_{500}|Z_k)$, $\ell = 1, 2, ..., N_M$, as,

$$P(M^{\ell}_{500}|Z_k) = \frac{\lambda(M^{\ell}_{500}, Z_k)}{\sum_{l=1}^{N_M} \lambda(M^{l}_{500} Z_k)}. \tag{6.2}$$

This expression is used in Eq.6.1 to determine the local mass distribution and in Eq.5.22 to marginalize the Hough landscape over group mass.

The prior $P(G_k|M_{500})$ based on local galaxy density

Inspired by (Smith et al. 2012) we also formulated an alternative way of calculating $P(G_k|M_{500})$ according to the local galaxy density.

Given a position G_k of redshift Z_k, we investigate the distribution of galaxy count $\mathcal{C}$ inside a cylindrical volume $\mathcal{V}$ (within $2\sigma_z$ along LoS and $1.5r_{500}$ per-

pendicular to LoS, given a group mass M_{500}) in the local z-band $Z_k \pm \delta_z$[1]. To that end we estimate distribution of galaxy counts inside volume $\mathcal{V}(g_i, M_{500})$ centred around "suspected" galaxy group centres g_i within the local z-band $Z_k \pm \delta_z$. To declare a galaxy g_i a suspected centre of a group we first check the richness within $\mathcal{V}(g_i, 10^{12})$ and reject galaxies with richness ≤ 3. Next, for each mass band $\ell = 1, 2, ...N_M$, we check the normality of projected velocities along LoS of the galaxies inside $\mathcal{V}(g_i, M_{500}^{\ell})$ using Shapiro-Wilk test ($p = 0.1$, code obtained from (BenSaida 2014)). We record the richness

$$\mathcal{C}(g_i, M_{500}^{\ell}) = |\{g_j \in \mathcal{V}(g_i, M_{500}^{\ell}), j = 1, 2, ..., N_{\rm gal}\}|, \tag{6.3}$$

for all volumes $\mathcal{V}(g_i, M_{500}^{\ell})$ that pass the normality test. For each mass band ℓ, the counts $\mathcal{C}(g_i, M_{500}^{\ell})$ are then used to construct a distribution[2] $Q(\mathcal{C}|Z_k, M_{500}^{\ell})$ over galaxy counts in groups of mass within M_{500}^{ℓ} at redshift $Z_k \pm \delta_z$.

Given a position G_k, we obtain the galaxy counts $\mathcal{C}(G_k, M_{500}^{\ell})$ for every mass band $\ell = 1, 2, ..., N_M$, and then estimate the probability of a particular group mass band M_{500}^{ℓ} at G_k through

$$P(M_{500}^{\ell}|G_k) = \frac{Q(\mathcal{C}(G_k, M_{500}^{\ell})|Z_k, M_{500}^{\ell})}{\sum_{\ell'} Q(\mathcal{C}(G_k, M_{500}^{\ell'})|Z_k, M_{500}^{\ell'})}. \tag{6.4}$$

Finally, we estimate the group mass $\hat{M}$ for a given position G_k as the mean of

[1] δ_z=0.001.

[2] We used smoothed normalized histogram estimation using Matlab smoothing function (Local regression using weighted linear least squares and a 2nd degree polynomial model with assigns lower weight to outliers in the regression and zero weight to data outside six mean absolute deviations).

mass estimates ${M_{\rm vr}}^{\ell}$ obtained through virial theorem,

$$\hat{M}(G_k) = \sum_{\ell=1}^{N_M} M_{vr}^{\ell}\ P(M_{500}^{\ell}|G_k), \tag{6.5}$$

where

$$M_{\rm vr}^{\ell} = \mathcal{F}\frac{(\sigma_p^{\ell})^2\ r^{\ell}}{G}, \tag{6.6}$$

with $\mathcal{F}$=3 (Barschel 2007), σ_p^{ℓ} is the estimated velocity dispersion at G_k (estimated using *gapper* estimator Eq.(6.7) (Beers et al. 1990, Wainer & Thissen 1976)) and r^{ℓ} is the projected radius of the group at G_k estimated as the average of the projected radii of its members.

For a group of galaxies with count $\mathcal{C}(G_k, M_{500}^{\ell})$ and velocities $v_1 \leq v_2 \leq ... \leq v_{\mathcal{C}}$, we evaluated the gaps $gp_q = v_{q+1} - v_q$, $q = 1, ..., \mathcal{C}-1$. Each gp_q is associated with a weight related to its position in the ordered list, $w_q = q(\mathcal{C} - q)$. The estimator is defined as

$$\sigma_p = \frac{\sqrt{\pi}}{\mathcal{C}(\mathcal{C}-1)} \sum_{q=1}^{\mathcal{C}-1} w_q gp_q. \tag{6.7}$$

Besides the mass estimation described above, the mass distribution Eq.6.4 is also used in Eq.5.22. The probability of a group position G_k given a M_{500}^{ℓ} band needed in Eq. (5.18) can be calculated as

$$P(G_k|M_{500}^{\ell}) = \frac{Q(\mathcal{C}(G_k, M_{500}^{\ell})|Z_k, M_{500}^{\ell})}{\sum\limits_{k'} Q(\mathcal{C}(G_{k'}, M_{500}^{\ell})|Z_{k'}, M_{500}^{\ell})}. \tag{6.8}$$

6.2 Experiments

The performance of group finders were compared using two families of measures. One evaluates the methods by viewing them as detectors of groups (group detection measures), the other one quantifies how closely are the properties of true groups matched by those of the detected groups (group properties measures).

6.2.1 Group Detection Measures

Precision (Pr), also known as group reliability (purity), is the percentage of the detected groups that are true groups,

$$\mathrm{Pr} = \frac{\mathrm{truepositive(TP)}}{\mathrm{TP} + \mathrm{FP}}, \tag{6.9}$$

where TP is the number of detected groups that are true groups and FP is the number of detected groups that are not true groups.

Recall (Re), sensitivity or group completeness, is the percentage of true groups that have been detected,

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6.10}$$

where FN is the number of true groups that are not detected.

6.2.2 Group Properties Measures

Completeness represents the fraction of galaxies in true group (TG) that have been recovered in the predicted group (PG)

$$\text{completeness(Com)} = \frac{\text{TG} \cap \text{PG}}{\text{TG}}, \tag{6.11}$$

while *reliability* is the fraction of galaxies in the PG that belong to the actual group (TG)

$$\text{reliability(Rel)} = \frac{\text{TG} \cap \text{PG}}{\text{PG}} \tag{6.12}$$

Before applying the completeness and reliability, we follow the strategy of (Duarte & Mamon 2015) to link the predicted galaxy groups PG with the true galaxy groups TG based on the central galaxy of the true groups (i.e. the brightest galaxy in the group).

Also, for each predicted group P_i linked to a true group, we calculate its *fragmentation rate*

$$\text{Frag}_i = \frac{D_i}{|P_i|}, \tag{6.13}$$

where $|P_i|$ is the number of galaxies in P_i and D_i is the number of galaxies from the linked true group that were not detected.

Merging is the fraction of true groups that were (incorrectly) merged into single predicted groups.

Velocity dispersion and mass estimation of the predicted groups

To stabilize velocity dispersion estimation for detected groups, we first excluded suspected "groups" with galaxies not distributed normally (using Shapiro-Wilk test) due to the fact that the projected velocity of galaxies inside real galaxy groups/clusters distribute normally along LoS. The velocity dispersion σ_p of each detected group was estimated using *gapper* estimator (see Eq. 6.7).

To estimate the mass $\tilde{M}_i$ (in Figures 6.9 and 6.10 annotated as M_P) for each predicted group P_i linked to a true group with mass T_i (in Figures 6.9 and 6.10 annotated as M_T), we use the virial theorem Eq.6.6. The estimation quality is quantified through bias $\log_{10}(\tilde{M}_i/T_i)$. We also assess the match between velocity dispersions/masses of true groups and the corresponding predicted groups through scatter plots and correlation coefficients.

6.2.3 Experimental Results

Pre/Post-processing

Before applying the PHTALK method, in order to increase efficiency and reduce false positives, we positioned the grid points on the observed galaxies that have a potential to be galaxy group centres. The potential is evaluated by positioning a cylindrical volume[1] defined by the mass[2] $10^{12}\mathrm{M}_{\odot}$, within $1.5r_{500}$ perpendicular to LoS and $2\sigma_z$ along LoS. We only consider as grid points that contain more than 3

[1] In FoF method by Eke, the cylindrical shape was used instead of ellipsoidal volume, as it improved the group detection performance (Eke et al. 2004).

[2] reasonable lower bound on mass for groups we would like to find

galaxies within this volume (see Appendix C).

After the application of PHTALK, the potential group centres correspond to local peaks in the Hough landscape $H(G_k)$ Eq. 5.22. Recall that the grid positions correspond to observed galaxy positions and hence can be associated with the corresponding galaxy luminosities. We sort the grid points in descending order according their luminosities. We then process the grid points corresponding to the peaks of the Hough landscape in this order. It may be that some of the close-by peaks can be merged into a single group. Given a peak grid point G_k, we first estimate the potential group mass $\hat{M}(G_k)$ around it as the expected value over mass bands Eq.6.5 and collect all peaks within that volume. Finally, the merged group centre will be the peak grid point G_j with the highest hough value $H(G_j)$. Processing the peak grid points in descending order of their luminosity ensures that the more luminous positions act as merger seeds that will absorb less luminous smaller groups (the first merging process based on the $H(G_k)$ peak values).

Finally, we applied a group merging operation based on the estimated group membership. For each grid point G_k corresponding to a predicted galaxy group centre with estimated mass $\hat{M}(G_k)$ (see section - 6.1.1, Eq.6.5) we collect galaxies within the cylindrical boundary $\mathcal{V}(G_k, \hat{M}(G_k))$ (see section 6.1.1). If two close-by groups overlap in galaxy membership in more than one half of the smaller group members, the two groups are merged. As before, the centre of the new merged group will be the peak grid point G_j with the highest Hough value $H(G_j)$ and the mass $\tilde{M}$ will be estimated.

Precision vs. Recall results

We use precision versus recall curve (to compare the performance of detecting galaxy group centre positions by PHTALK and two FoF versions (Eke et al. (2004) and Robotham et al. (2011)). For the FoF methods we calculate the galaxy group centres as the mean positions of the corresponding predicted groups, while in PHTALK we identify the group centres as the dominant peaks of the Hough landscape. We search for any detected galaxy group within a tolerance boundary of 1.5 virial radii perpendicular to LoS, and $2\sigma_z$ along LoS around the ground truth galaxy group centres as annotated in GAMMA mock data.

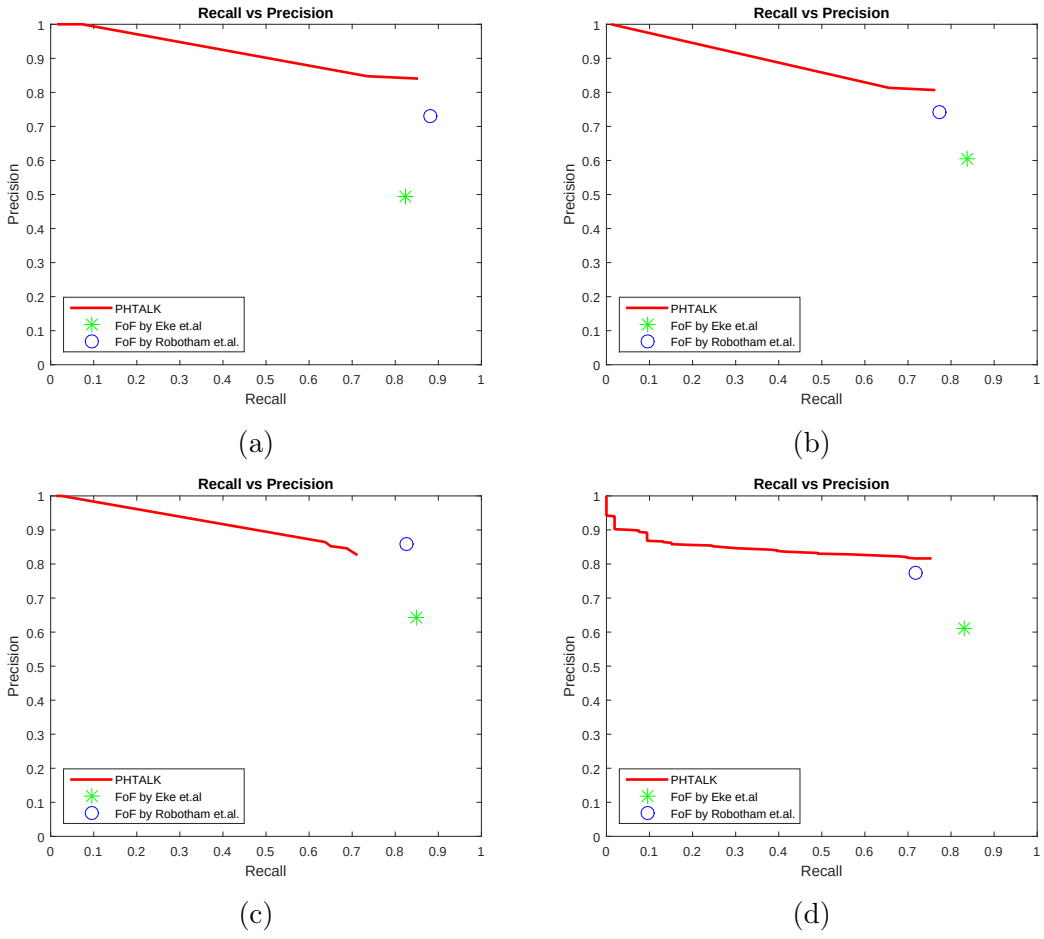
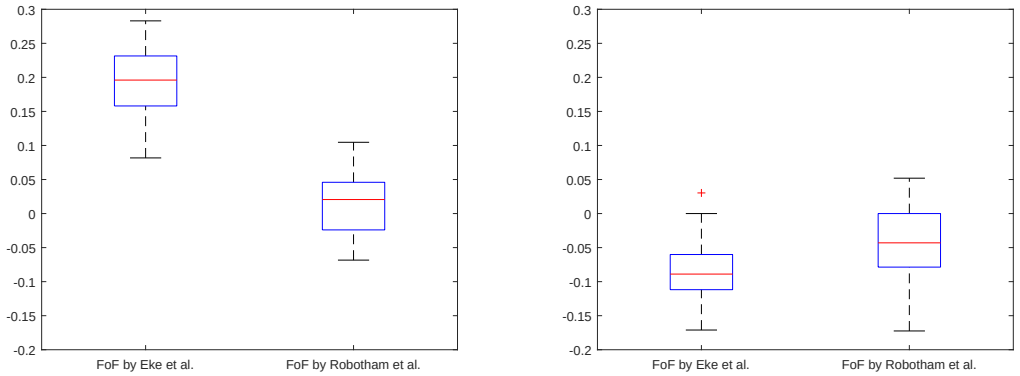


Figure 6.1: Precision vs. Recall for four example cones of GAMA mock: (a) volume 1 - cone 1, (b) volume 6 - cone 3, (c) volume 7 - cone 3, and (d) volume 9 - cone 3.

Figure 6.1 shows examples of Pr vs. Re results for few cones taken from different GAMA volumes. Pr vs. Re values of FoF by Eke and FoF by Robotham are presented as stars and circles, respectively. The settings of the FoF methods were taken from Eke et al. (2004) and Robotham et al. (2011). For PHTALK full Pr vs. Re curves can be obtained naturally by manipulating the peak detection threshold on the Hough landscape. The curves are shown as solid lines.

To quantitatively compare the galaxy group finders across the GAMA survey, for each cone and for a given FoF method, we evaluate the difference between Pr values of PHTALK and FoF, i.e. $Pr_{(\mathrm{PHTALK})} - Pr_{(\mathrm{FoF})}$. Box plots of the differences across the cones are presented in figure 6.2a. Analogously, box plots of the differences in recall values, $Rec_{(\mathrm{PHTALK})} - Rec_{(\mathrm{FoF})}$ are shown in figure 6.2b. Overall, PHTALK has superior Precision performance over the FoF methods (less false positives), at the cost of inferior Recall values (more false negatives). This tendency is more pronounced for FoF by Eke. Note, however, that in this case the positive values of precision differences are approximately twice the (absolute value of negative) values of recall differences. So the overall gain in precision at the expense of recall is more favourable for PHTALK. For FoF by Robotham, the situation is less distinct, but when compared to PHTALK, the balance between precision and recall is slightly favourable for FoF by Robotham. To evaluate statistical significance of these results, we performed Wilcoxon Signed-rank test (see Appendix D) at 5% significance level. For recall, PHTALK is significantly better than both FoF by Eke and FoF by Robotham. When the precision values are compared, PHTALK is significantly better than FoF by Eke, but no significant difference between PHTALK and FoF by Robotham has been found.

The results confirm observations based on Figure 6.1: Compared with both FoF methods, PHTALK suffers from less false positives (better Pr results). However, PHTALK misses some of the galaxy groups correctly detected by the FoF method (worse Re results). Those are predominantly groups with weak FoG signatures, that is not well-formed groups, or groups with small number of galaxies (e.g. <5). Our model based method will obviously suffer in such situations. Also, FoF by Robotham et al. tends to fragment high mass groups into many smaller groups (see Figures 6.7 and 6.10).



(a) (b)

Figure 6.2: Box plots represent the values of difference between PHTALK and FoF versions across 27 cones for precision 6.2a and recall 6.2b respectively

Finally, we note that in GAMA mocks ($z \leq 0.1$) there are 1924 true galaxy groups with ≥ 5 galaxy members. FoF by Eke et al., FoF by Robotham et al. and PHTALK were able to detect $\approx 95\%$, 85% and 87% of them, respectively.

Completeness vs. Reliability Results

To compare the capabilities of identification of group members by PHTALK and the two FoF methods considered in this study, we evaluate the Completeness (Com) and Reliability (Rel) measures for all detected groups across the 27 cones. For the PHTALK method a threshold value τ on the Hough landscape defining which peaks to consider needs to be specified. For all cones τ is set to a small value close to minimum Hough count $H_{min} = \min_k H(G_k)$. This generous threshold setting is possible thanks to the post-processing steps described above. In particular, $\tau = H_{min} + \frac{H_{max} - H_{min}}{100}$, where $H_{max} = \max_k H(G_k)$.

In Figure 6.3 we show the mean Com and Rel values for each of the 27 cones.

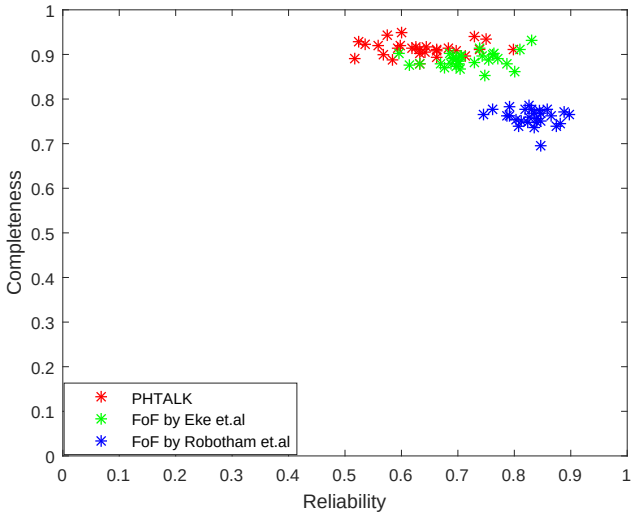


Figure 6.3: Mean Reliability and Completeness values of the studied methods for all cones.

PHTALK is better or comparable to FoF by Eke et al. and superior to FoF by Robotham et al. in terms of completeness. On the other hand, reliability performance of FoF by Robotham et al. is superior to both PHTALK and FoF by Eke et al.

To extend this analysis according to group mass, we categorized the predicted galaxy groups into four mass bands, (11.5 - 12.375, 12.375 - 13.25, 13.25 - 14.125, 14.125 - 15) $M_{\odot}$ (log scale), based on the masses of the linked true galaxy groups. Figure 6.4 shows the mean Com and Rel values along with std dev bars for each of the four mass bands. PHTALK tends to overestimate group membership of small groups by including more interlopers (Figure 6.4(a)), whereas FoF by Robotham et al. misses greater portion of galaxies from large groups (Figure 6.4(d)). This finding is confirmed by more detailed study of the distribution of reliability and completeness values through CDF curves in Figure 6.5 and Figure 6.6, respectively.

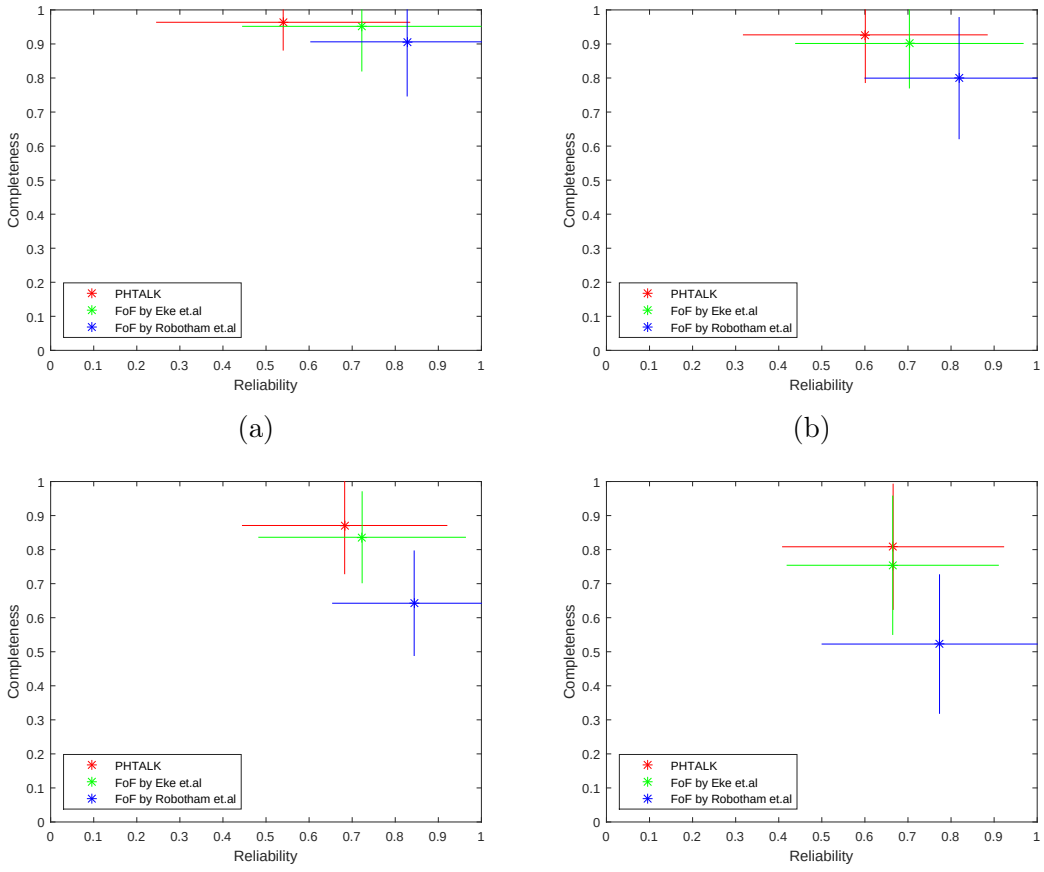


Figure 6.4: Completeness versus Reliability in the mass band 11.5-12.375$M_{\odot}$ (a), 12.375-13.25$M_{\odot}$ (b), 13.25-14.125$M_{\odot}$ (c) and14.125-15$M_{\odot}$ (d).

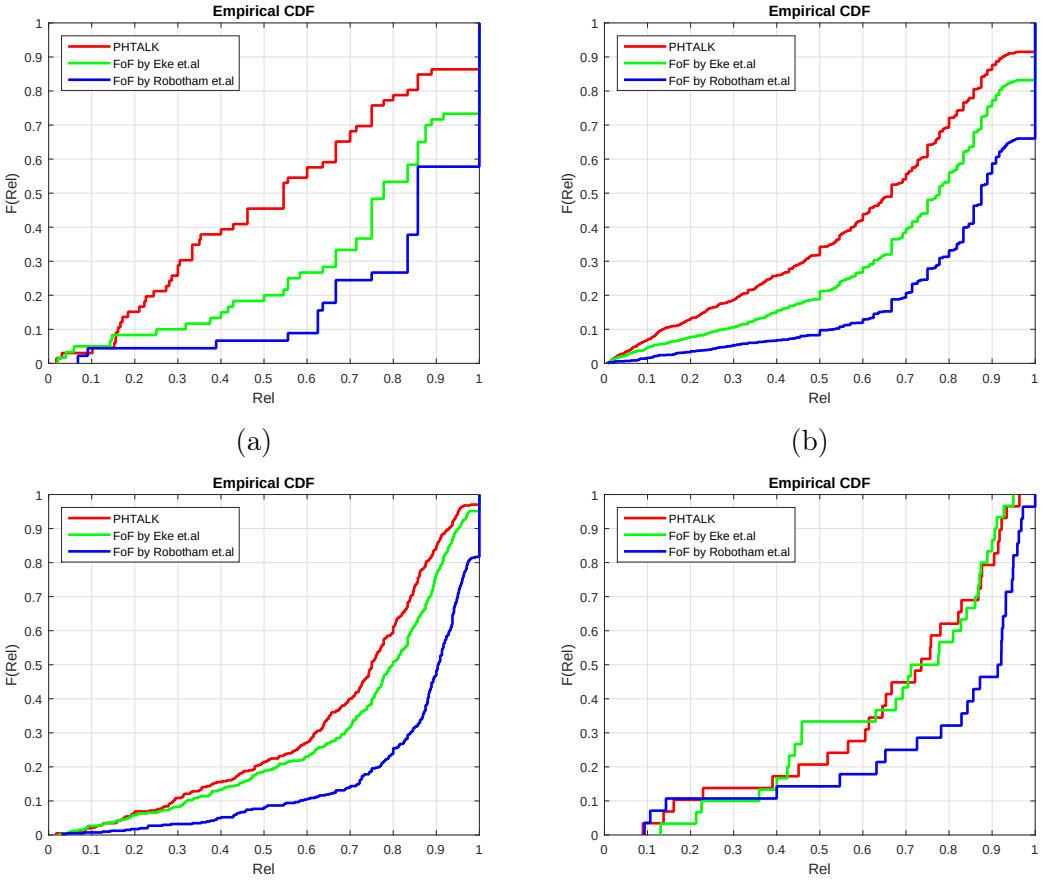


Figure 6.5: Reliability CDF in the mass band 11.5-12.375$M_{\odot}$ (a), 12.375-13.25$M_{\odot}$ (b), 13.25-14.125$M_{\odot}$ (c) and14.125-15$M_{\odot}$ (d).

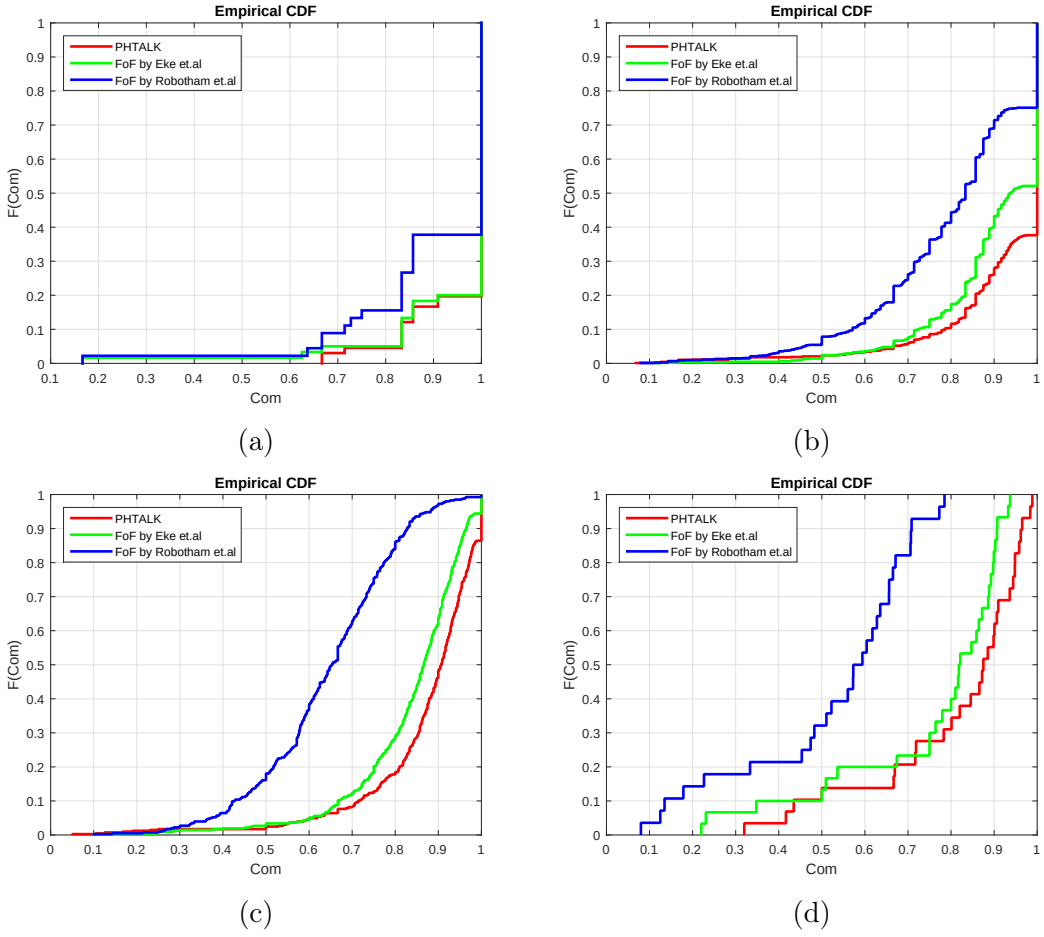


Figure 6.6: Completeness CDF in the mass band 11.5-12.375$M_{\odot}$ (a), 12.375-13.25$M_{\odot}$ (b), 13.25-14.125$M_{\odot}$ (c) and14.125-15$M_{\odot}$ (d).

Fragmentation and merging results

Fragmentation rates shown in Figure 6.7 for the four mass bands confirm the completeness results. For larger groups PHTALK has consistently the lowest rates, whereas FoF by Robotham et al. exhibits the largest fragmentation.

Merging of two true groups occured only once in a predicted group by PHTALK at redshift ≈ 0.075 within the second mass band $(12.375 - 13.25 M_{\odot})$. No merging has been detected for the FoF methods.

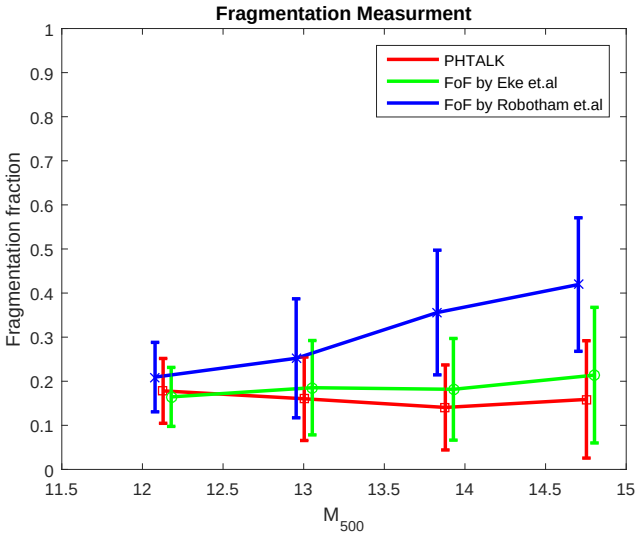


Figure 6.7: Fragmentation Measurement: the x-axis represents the mass bands M_{500} in log scale, the FoF methods' results have shifted slightly for illustration purposes, y-axis represent the fragmentation rate.

Velocity dispersion and mass estimation results

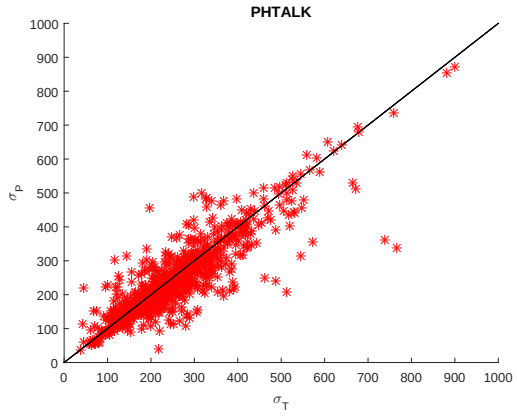
Scatter plots of predicted versus true velocity dispersions for pairs of linked estimated and true groups are presented in Figure 6.8. All methods have comparable

performance with correlation coefficient for PHTALK, FoF by Eke et al., and FoF by Robotham et al. equal to 87.57%, 86.59%, and 87.36%, respectively. Scatter plots of predicted versus true group masses are shown in Figure 6.9. PHTALK and FoF by Eke et al. are comparable and superior to FoF by Robotham et al. - correlation coefficient values 73.22%, 74.7% and 69.25%.

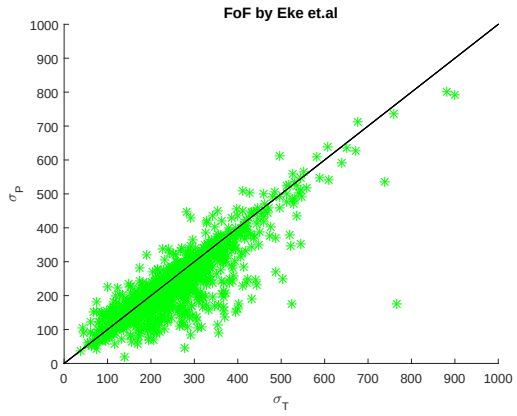
The accuracy of estimating the mass is evaluated through bias measure discussed in section 6.2.2 and presented in Figure 6.10 for the four mass bands. Overall, PHTALK and FoF by Eke et al. tend to overestimate and underestimate, respectively, the mass of smaller groups, while FoF by Robotham et al. underestimates the mass of both small and large groups.

6.2.4 Enhancing the estimated mass

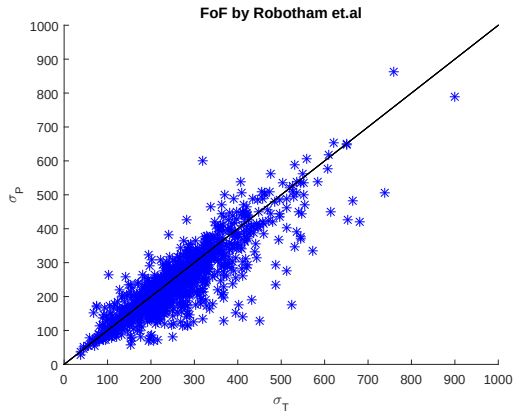
We attempted to improve the mass estimation of the predicted galaxy groups as follow, for each predicted galaxy group and depending on its current mass estimation and its current members, we recalculate the mean position of the group and consider the closest galaxy to that mean position as a centre of the group. Also, we use the cylindrical volume (see Appendix D) by repositioning it centred around the new centre position to recollect the members of the group and re-estimate the mass based on the current group members using the virial theorem. For each predicted group, the process will be iterated until the convergence between the current estimation of its mean position and the previous estimation of its mean position occurs. Figure 6.11 shows the final estimation of the mass after the iterated process. In Figure 6.11b some galaxy groups have gained better mass estimate, while others have deviated towards lower mass estimates than their true



(a)

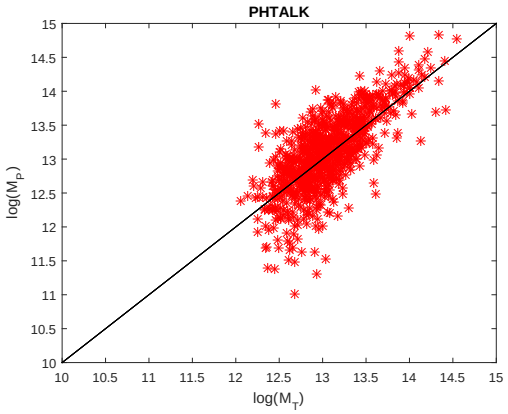


(b)

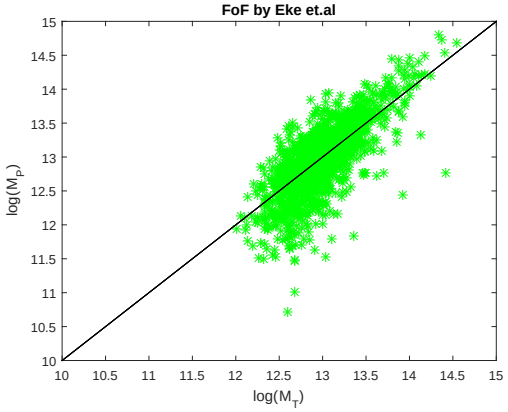


(c)

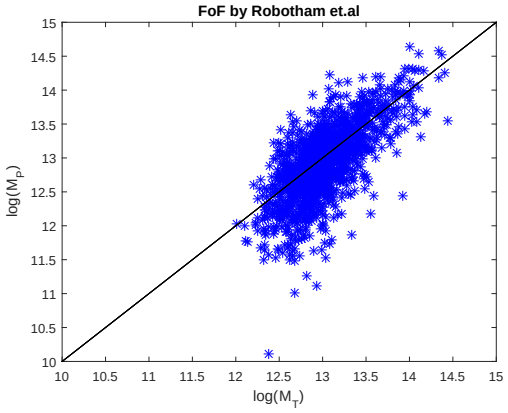
Figure 6.8: Predicted velocity dispersion σ_P vs. actual velocity dispersion σ_T in km/s of (a) PHTALK, (b) FoF by Eke, and (c) FoF by Robotham et al. methods.



(a)



(b)



(c)

Figure 6.9: Predicted mass $Mass_p$ vs. actual mass $Mass_o$ in log scale of the total predicted galaxy groups from GAMA mock with reliability ≥ 0.5; (a) PHTALK, (b) FoF by Eke, and (c) FoF by Robotham et al. methods

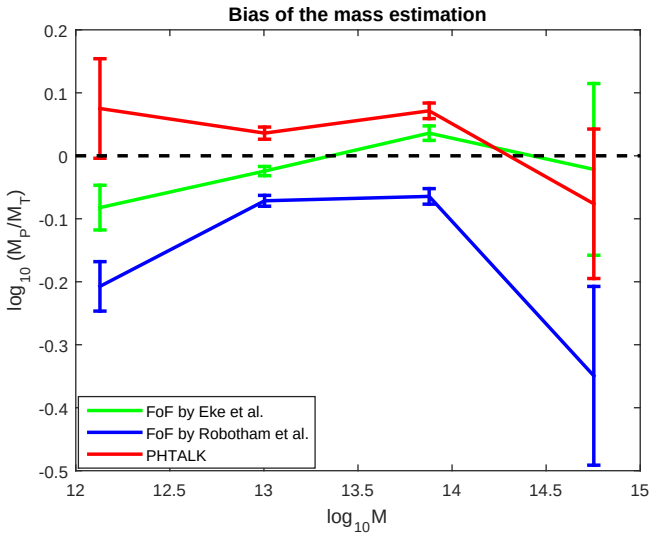
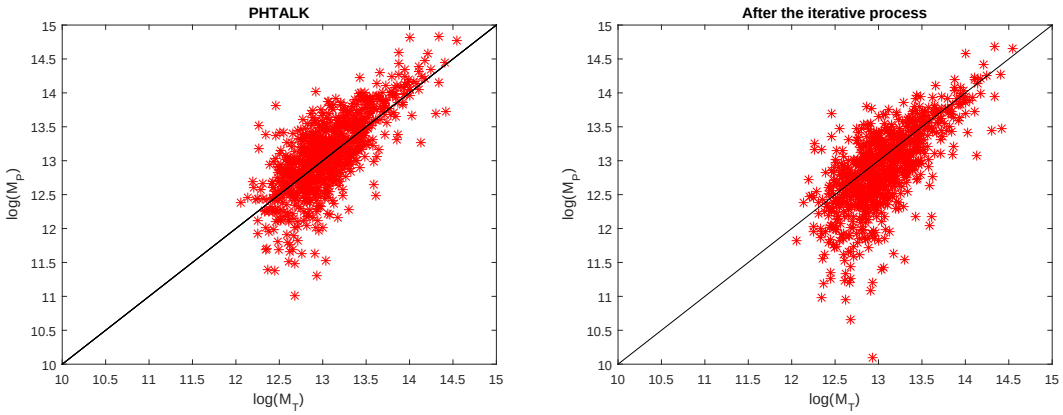


Figure 6.10: Mass estimation bias: the x-axis represents the mean of the mass bands, y-axis is the bias value $\log_{10}(M_P/M_T)$, the error bars correspond to standard errors

masses.



(a) (b)

Figure 6.11: Mass estimation, (a) the estimation of the mass before the iteration process. (b) mass estimation after iteration process.

We did not rely on the luminosity of the members (i.e considering the high luminous galaxy as a centre of a group) because we have used this technique on the

previous mass estimation as shown in Figure 6.11a in addition to the issue related to overusing the luminosities of the current galaxy members in the iterative process may lead to a huge deviation in estimating the true centre position, members and mass of a galaxy group. Due to the possibility of combining field galaxies that have higher luminosities than the true galaxy members during the iterative process. In this case, field galaxies may dominate the groups and lead to miss the true group members and their masses if these high luminous field galaxies considered as centre positions mistakenly.

Although astronomers use the repetitive approach till converge occur in order to estimate some properties, in our case it did not work due to the overdoing of mass estimation (i.e. for a given galaxy group G_k, the mass has been estimated first ($\hat{M}(G_k)$) as illustrated in Eq.6.5 and then re-estimated ($\tilde{M}(G_k)$) after the merging processes as mentioned in sub-section (Pre/Post-processing - 6.2.3))

CHAPTER 7

CONCLUSIONS AND FUTURE WORKS

In general, the existing galaxy group finders have many free parameters that need to be carefully tuned before applying analysis. This raises issues regarding generality of the results and stability of the calibration process. Hough transform based models have been shown to be effective in the detection of patterns of interest in cluttered scenes.

Probabilistic Hough transform formulation enables us to include explicitly prior expectations of the shape of interest through the likelihood model and to treat the background noise consistently. In addition, we can include a particular model for describing the interloper galaxies in order to filter out field galaxies and obtain more purity in terms of galaxy group members.

The PHTALK is a principled approach which effectively incorporates a type of prior knowledge on what is the expected appearance of a group/cluster. As such, it could naturally be modified to seek out groups/clusters with distinct characteristics

- a property which FoF methods do not share. The PHTALK approach is adaptable and flexible to include more astrophysics' information. It is able to detect galaxy groups with more than five galaxies in our simulation and in the GAMA mock data. It has to be said, however, that PHTALK is time-consuming due to the costly voting process for each grid point, in the accumulator, by all galaxies in the cone.

Thus, we have used parallel computing to speed up the performance and have positioned grid points only on observed galaxies that were suspected to be galaxy group centres. The PHTALK approach has less false positives in detecting galaxy groups' positions than both FoF methods. However, both FoF methods can detect galaxy groups which are not well-formed, or where their members follow a different kind of distribution profile. In such special cases our model PHTALK obviously cannot be effective.

We can summarize some reasons for such limited cases of missing true groups as follows:

- The true groups suffer from a lack of intensity of galaxies (i.e. not enough prominent galaxies) around the centre.
- Galaxy groups do not have the prolonged shape 'finger of god' very well formed.
- Some galaxy groups are close to each other; hence they can be detected as a single group.

Possible reasons for obtaining false positive groups (peaks) include:

- True groups with high mass are identified as several smaller groups, due to the fragmentation factor (i.e. if the true group mass has been underestimated).
- Some fore/back-ground galaxies may follow the same distribution by incidence of the pattern of interest (i.e. true galaxy groups).

This work includes several model decisions that helped to find the true estimation of a galaxy group's centre and the mass estimation of a galaxy group. For example: we estimated the mass based on the local density; incorporated that estimation in calculating PHTALK; merged the fragmented peaks that were obtained from PHTALK based on their locations, luminosities and H values, using virial theorem in the final mass estimation.

The main contributions of this work are as follows:

- Building an adaptive probabilistic model to generalize the possibility of detecting groups and clusters of galaxies with fewer free parameters and based on the coordinates of the galaxies.
- A well-grounded principled framework based on probability theory.
- A very natural framework to incorporate prior knowledge.
- It is the first time to propose the Hough transform for galaxy groups/clusters detection.

- It has provided a principled way, to deal with uncertainty with respect to the mass of the galaxy groups/clusters; as it is a probabilistic formulation, we can integrate it out over the masses.

Future works may consist of collecting the galaxies for each significant grid position (i.e. group) with a probabilistic technique, after filtering out the field galaxies. This can be done by including an interloper profile in the collecting process. It would be interesting to include an appropriate description of filaments in the process of finding the galaxy groups. This should obviously be helpful because most potential galaxy groups are condensed in the intersection regions of the filaments. Analysing the distribution of galaxies that are scattered within the filaments outside known galaxy groups is more likely to improve the detection outcomes.

Appendices

APPENDIX A

PRECISION VERSUS RECALL (PR V. RE)

To evaluate the performance of the detector, positive − predictive − value(PPV) otherwise called precision (Pr), which is the percentage of the detected groups that have been identified as actual groups, and the true − positiverate(TPR) otherwise called recall (Re), which is the percentage of real groups that have been detected, have been calculated; as shown in Eq. (A.1) and Eq. (A.2) respectively. We inspected the true positive(TP), which means the ground truths (the mean positions of the true groups) that are detected. Moreover, the false negative (FN) means the ground truths that are not detected; while the false positive(FP) indicates the detected peaks that are not true peaks.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{A.1}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{A.2}$$

To obtain a more balanced and fair trade-off between the FP and FN, the measurement of the PPV and TPR have been combined into a single curve. The curve is well known in the machine learning field as the precision versus recall curve. In our detection case, each point in the curve will represent the precision versus recall at a particular threshold τ; where τ will take 100 $H(x, y)$ values within the range of minimum and maximum $H(x, y)$ values increasing gradually. The detected peaks above τ will be tested to check if they are FP or TP.

APPENDIX B

SIMPLE PEAK DETECTION USING DILATION

Dilation is the one of fundamental morphological image processing operations. It is used interchangeably with another morphological operation called erosion to noise suppression and image smoothing (i.e. opening and closing operation) (Maragos 2005). These operations are used in peak detection by convolving the image with a mask (i.e. structuring element). As an example, for peak detection let us assume we have a matrix with pixel values

$$A = \begin{bmatrix} 34 & 24 & 433 & 123 & 123 & 654 \\ 234 & 21 & 32 & 65 & 78 & 34 \\ 23 & 454 & 54 & 96 & 24 & 2 \end{bmatrix}, \tag{B.1}$$

and mask

$$msk = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \tag{B.2}$$

find the dilation $\oplus$ between the A and msk

$$D_{\oplus} = A \oplus msk. \tag{B.3}$$

The obtained dilated matrix is

$$D_{\oplus} = \begin{bmatrix} 234 & 433 & 123 & 433 & 654 & 123 \\ 454 & 454 & 454 & 433 & 654 & 654 \\ 454 & 234 & 454 & 78 & 96 & 78 \end{bmatrix}. \tag{B.4}$$

By comparing the original image values with the outcome values of the dilated process, we can identify the original image positions that have values larger than the dilated positions as prominent peaks. The identified peak positions in A matrix are the positions with the values of '1' as shown bellow:

$$A > D_{\oplus} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{B.5}$$

APPENDIX C

CHECKING THE CYLINDRICAL REGION AROUND THE POTENTIAL PEAKS

To check whether a galaxy g_i with position $\mathbf{g}_i$ is within the cylindrical volume centred around a given peak G_k positioned at $\mathbf{G}_k$ as illustrated in Figure C.1, we calculate the difference vector: $\mathbf{v}_i = \mathbf{G}_k - \mathbf{g}_i$, then the unit vector in the direction of $\mathbf{G}_k$ has been found denoting by $\mathbf{u}_k$; then the dot product $(\cdot)$ between $\mathbf{u}_k$ and $\mathbf{v}_i$ has been calculated to get the projected length $(\mathbf{b}_i)$ w.r.t $\mathbf{G}_k$ along the LoS:

$$\mathbf{b}_i = \mathbf{u}_k \cdot \mathbf{v}_i \text{ , where } -\alpha \cdot \sigma_z <= \mathbf{b}_i <= \alpha \cdot \sigma_z, \tag{C.1}$$

the length of $\mathbf{b}_i$ should not exceed the interval $[-\alpha.\sigma_z, \alpha.\sigma_z]$ where σ_z is the redshift dispersion estimated using Eq. (5.30) and $\alpha = 2$ is the quantile factor.

The perpendicular distance (X) of g_i from the LoS should not exceed the varial radius $(1.5r_{500})$ calculated using the NFW profile (section 2.11), depending on a

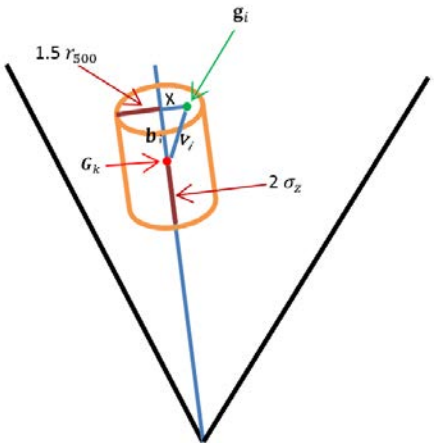


Figure C.1: Schematic illustration of the cylindrical volume around the suspected group position.

particular mass and current central galaxy group redshift $z(G_k)$.

$$X = \|\mathbf{v}_i - \mathbf{b}_i \cdot \mathbf{u}_k\| \ , \ \text{and} \ X <= 1.5 \cdot r_{500} \tag{C.2}$$

APPENDIX D

WILCOXON SIGNED-RANK TEST

Most common statistical methods depend on assumptions related to the distribution of the data. In the testing of the mean, it is assumed that the distribution is normal. However, in practice, if there is a doubt about the normality of the population, especially when we have small sample, there are some inference methods which do not require a particular distribution of the data; these are called **'non-parametric methods'**. For the inference of the mean difference in data of matched pairs, we use the Wilcoxon signed-rank test (Neuhauser 2011, SEGRE n.d.). For matched pairs' data, the absolute value of the differences (magnitude with no sign) has been compared. We discard any zero differences from the list then assign ranks for all differences after sorting them in increasing order, keeping track of which rank attached to positive difference values. We sum all the ranks that attached to the positive sign difference (S^+). If the null hypothesis is true and there is no difference between the distributions of the matched pairs' data, (S^+) has the mean:

$$\mu_{S^+} = \frac{s^{\pm}(s^{\pm}+1)}{4}, \tag{D.1}$$

and standard deviation

$$\sigma_{S^+} = \sqrt{\frac{s^{\pm}(s^{\pm}+1)(2 \cdot s^{\pm}+1)}{24}}, \tag{D.2}$$

where $s^{\pm}$ is the total sample size (i.e. positive (S^+) and negative (S^-) differences). Otherwise, the hypothesis (i.e. no differences in distributions) will be rejected, if (S^+) is far from its mean. Then we find the one or two-sided P-value of (S^+) from special tables.

REFERENCES

Amores, E. B. d. (2011), Data analysis and simulations of the large data sets in the galactic astronomy, *in* 'Numerical Analysis - Theory and Application', InTech.

Antolovic, D. (2008), 'Review of the hough transform method, with an implementation of the fast hough variant for line detection', *Department of Computer Science, Indiana University* .

Ascaso, B., Wittman, D. & Benítez, N. (2012), 'Bayesian cluster finder: clusters in the CFHTLS Archive Research Survey', *MNRAS* **420**, 1167–1182.

Astone, P., Colla, A., D'Antonio, S., Frasca, S. & Palomba, C. (2014), 'Method for all-sky searches of continuous gravitational wave signals using the frequency-hough transform', *Phys. Rev. D* **90**, 042002.

Baldry, I., Robotham, A., Hill, D., Driver, S., Liske, J., Norberg, P., Bamford, S., Hopkins, A., Loveday, J., Peacock, J. et al. (2010), 'Galaxy and mass assembly (gama): the input catalogue and star–galaxy separation', *MNRAS* **404**(1), 86–100.

Ballester, P. (1994), 'Hough transform for robust regression and automated detection', *AAP* **286**, 1011–1018.

Ballester, P. (1996), 'Hough transform and astronomical data analysis', *Vistas in Astronomy* **40**(4), 479–485.

Barschel, C. (2007), Structure formation in the universe, Technical report, Aachen University of Technology - RWTH I. Physics Institute B Sommerfeldstr. 14 D-52074 Aachen.

Bartelmann, M. (1996), 'Arcs from a universal dark matter halo profile', *Astron.Astrophys.* **313**, 697–702.

Beers, T. C., Flynn, K. & Gebhardt, K. (1990), 'Measures of location and scale for velocities in clusters of galaxies - A robust approach', *AJ* **100**, 32–46.

Behroozi, P. S., Conroy, C. & Wechsler, R. H. (2010), 'A comprehensive analysis of uncertainties affecting the stellar mass-halo mass relation for 0< z< 4', *The Astrophysical Journal* **717**(1), 379.

BenSaida, A. (2014), 'Shapiro-wilk and shapiro-francia normality tests.'. Online; accessed 09 Jan 2017.
URL: *http://uk.mathworks.com/matlabcentral/fileexchange/13964-shapiro-wilk-and-shapiro-francia-normality-tests*

Berlind, A. A., Frieman, J. A., Weinberg, D. H., Blanton, M. R., Warren, M. S., Abazajian, K., Scranton, R., Hogg, D. W., Scoccimarro, R., Bahcall, N. A., Brinkmann, J., Gott, J.Richard, I., Kleinman, S., Krzesinski, J., Lee, B. C., Miller, C. J., Nitta, A., Schneider, D. P., Tucker, D. L. & Zehavi, I. (2006), 'Percolation galaxy groups and clusters in the sdss redshift survey: Identification, catalogs, and the multiplicity function', *Astrophys.J.Suppl.* **167**, 1–25.

Botzler, C. S., Snigula, J., Bender, R. & Hopp, U. (2004), 'Finding structures in photometric redshift galaxy surveys: an extended friends-of-friends algorithm',

MNRAS **349**(2), 425–439.
URL: *http://dx.doi.org/10.1111/j.1365-2966.2004.07468.x*

Bradt, H. (2007), *Astronomy methods: A physical approach to astronomical observations*, Cambridge, UK: Cambridge Univ. Pr.

Chiu, S.-H., Liaw, J.-J. & Lin, K.-H. (2010), 'A fast randomized hough transform for circle/circular arc recognition', *International Journal of Pattern Recognition and Artificial Intelligence* **24**(03), 457–474.

Connelly, J. L. (2012), Optically and x-ray selected galaxy groups at intermediate redshift, PhD thesis, LMU.

Dariush, A. A., Raychaudhury, S., Ponman, T. J., Khosroshahi, H. G., Benson, A. J., Bower, R. G. & Pearce, F. (2010), 'The mass assembly of galaxy groups and the evolution of the magnitude gap', *MNRAS* **405**, 1873–1887.

Dominguez Romero, M. J. d. L., Garcia Lambas, D. & Muriel, H. (2012), 'An improved method for the identification of galaxy systems: measuring the gravitational redshift by dark matter haloes', *MNRAS: Letters* **427**(1), L6–L10.
URL: + *http://dx.doi.org/10.1111/j.1745-3933.2012.01326.x*

Donahue, M., Scharf, C. A., Mack, J., Lee, Y. P., Postman, M., Rosati, P., Dickinson, M., Voit, G. M. & Stocke, J. T. (2002), 'Distant cluster hunting. ii. a comparison of x-ray and optical cluster detection techniques and catalogs from the rosat optical x-ray survey', *AJ* **569**(2), 689.
URL: *http://stacks.iop.org/0004-637X/569/i=2/a=689*

Dong, F., Pierpaoli, E., Gunn, J. E. & Wechsler, R. H. (2008), 'Optical cluster finding with an adaptive matched-filter technique: Algorithm and comparison with simulations', *The Astrophysical Journal* **676**(2), 868.

D'Orazio, T., Guaragnella, C., Leo, M. & Distante, A. (2004), 'A new algorithm for ball recognition using circle hough transform and neural classifier', *Pattern recognition* **37**(3), 393–408.

Dressler, A. (1984), 'The evolution of galaxies in clusters', *ARAA* **22**, 185–222.

Driver, S., Hill, D., Kelvin, L., Robotham, A., Liske, J., Norberg, P., Baldry, I., Bamford, S., Hopkins, A., Loveday, J. et al. (2011), 'Galaxy and mass assembly (gama): survey diagnostics and core data release', *MNRAS* **413**(2), 971–995.

Duarte, M. (2014), Toward a new level of modeling of environmental effects on galaxies, PhD thesis, Institut d Astrophysique de Paris.
URL: *http://www.theses.fr/2014OBSP0250*

Duarte, M. & Mamon, G. A. (2015), 'Maggie: Models and algorithms for galaxy groups, interlopers and environment', *MNRAS* **453**(4), 3848–3874.

Duda, R. & Hart, P. (1971), Use of the hough transformation to detect lines and curves in pictures, Technical Report 36, AI Center, SRI International.

Einstein, A. (1916), 'Die Grundlage der allgemeinen Relativitätstheorie', *Annalen der Physik* **354**, 769–822.

Eke, V. R. et al. (2004), 'Galaxy groups in the 2dfgrs: The group - finding algorithm and the 2pigg catalog', *MNRAS* **348**, 866.

Erfanianfar, G. (2014), The group galaxy population through the cosmic time, PhD thesis, Ludwig-Maximilians-Universität München.

Figueiro Spinelli, P. (2011), Weak Lensing Analysis of Galaxy Groups, PhD thesis, LMU.

Friedman, A. (1922), 'Über die krummung des raumes', *Zeitschrift für Physik A Hadrons and Nuclei* **10**(1), 377–386.

Frost, M. I. (2010), The Clustering of Galaxies in the SWIRE Survey, PhD thesis, University of Sussex.
URL: *http://sro.sussex.ac.uk/2363/*

Gal, R. R., Carvalho, R. R. d., Lopes, P. A. A., Djorgovski, S. G., Brunner, R. J., Mahabal, A. & Odewahn, S. C. (2003), 'The northern sky optical cluster survey. ii. an objective cluster catalog for 5800 square degrees', *The Astronomical Journal* **125**(4), 2064.

Gladders, M. D., Hoekstra, H., Yee, H. K. C., Hall, P. B. & Barrientos, L. F. (2003), 'The incidence of strong - lensing clusters in the Red - Sequence Cluster Survey', *AJ* **593**, 48.

Gladders, M. D. & Yee, H. K. C. (2000), 'A new method for galaxy cluster detection. 1. The algorithm', *AJ* **120**, 2148.

Guo, S., Pridmore, T., Kong, Y. & Zhang, X. (2009), 'An improved hough transform voting scheme utilizing surround suppression', *Pattern Recogn. Lett.* **30**(13), 1241–1252.
URL: *http://dx.doi.org/10.1016/j.patrec.2009.05.003*

Hassanein, A. S., Mohammad, S., Sameer, M. & Ragab, M. E. (2015), 'A survey on hough transform, theory, techniques and applications', *arXiv preprint arXiv:1502.02160* .

Hogg, D. W. (1999), Distance measures in cosmology. Unpublished manuscript.

Holder, G. P., McCarthy, I. G. & Babul, A. (2007), 'The sunyaev-zeldovich background', *MNRAS* **382**(4), 1697–1706.
URL: *http://mnras.oxfordjournals.org/content/382/4/1697.abstract*

Hollitt, C. & Johnston-Hollitt, M. (2012), 'Feature detection in radio astronomy using the circle hough transform', *PASA* **29**, 309–317.

Hough, P. V. C. (1962), 'A method and means for recognizing complex patterns', US Patent: 3,069,654.

Hsu, C.-C. & Huang, J. S. (1990), 'Partitioned hough transform for ellipsoid detection', *Pattern recognition* **23**(3-4), 275–282.

Hubble, E. (1929), 'A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae', *Proceedings of the National Academy of Science* **15**, 168–173.

Huchra, J. P. (1988), Redshift surveys and the description of clustering in the universe, *in* J. M. Dickey, ed., 'The Minnesota lectures on Clusters of Galaxies and Large-Scale Structure', Vol. 5 of *Astronomical Society of the Pacific Conference Series*, pp. 41–70.

Huchra, J. P. & Geller, M. J. (1982), 'Groups of galaxies. i - nearby groups', *Astrophysical Journal* **257**(June 15), 423–437.

Ibrahem, R. T., Tino, P., Pearson, R. J., Ponman, T. J. & Babul, A. (2015), Automatic Detection of Galaxy Groups by Probabilistic Hough Transform, *in* 'Neural Information Processing Volume 9491, 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings Part III, Eds Arik, S., Huang, T., Lai, W.K., Liu, Q., p.323-331', pp. 323–331.

Illingworth, J. & Kittler, J. (1987), A survey of efficient hough transform methods., *in* 'Alvey Vision Conference', pp. 1–8.

Janiak, A. (2009), *The Stanford Encyclopedia of Philosophy: Newton's Philosophy*, The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115.

Jian, H.-Y., Lin, L., Chiueh, T., Lin, K.-Y., Liu, H. B., Merson, A., Baugh, C., Huang, J.-S., Chen, C.-W., Foucaud, S. et al. (2014), 'Probability friends-of-friends (pfof) group finder: Performance study and observational data applications on photometric surveys', *The Astrophysical Journal* **788**(2), 109.

Johansson, D. (2011), Observations of submillimeter galaxies and of the Sunyaev-Zeldovich effect toward clusters of galaxies, PhD thesis, Department of Earth and Space Sciences, Chalmers University of Technology.

Jones, C., Andrade-Santos, F., Forman, W. R., Murray, S. S. & Churazov, E. (2014), Characterizing Planck-detected Clusters of Galaxies with Chandra, *in* 'AAS/High Energy Astrophysics Division', Vol. 14 of *AAS/High Energy Astrophysics Division*, p. 111.09.

Kepner, J., Fan, X., Bahcall, N., Gunn, J., Lupton, R. & Xu, G. (1999), 'An automated cluster finder: The adaptive matched filter', *The Astrophysical Journal* **517**(1), 78.

Kesidis, A. & Papamarkos, N. (2000), 'A window-based inverse hough transform', *Pattern Recognition* **33**(6), 1105 – 1117.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320399001673*

Klypin, A. & Shandarin, S. (1993), 'Percolation technique for galaxy clustering', *apj* **413**, 48–58.

Knebe, A., Knollmann, S. R., Muldrew, S. I., Pearce, F. R., Aragon-Calvo, M. A., Ascasibar, Y., Behroozi, P. S., Ceverino, D., Colombi, S., Diemand, J., Dolag, K., Falck, B. L., Fasel, P., Gardner, J., Gottlöber, S., Hsu, C.-H., Iannuzzi, F., Klypin, A., Lukić, Z., Maciejewski, M., McBride, C., Neyrinck, M. C., Planelles, S., Potter, D., Quilis, V., Rasera, Y., Read, J. I., Ricker, P. M., Roy, F., Springel, V., Stadel, J., Stinson, G., Sutter, P. M., Turchaninov, V., Tweed, D., Yepes, G. & Zemp, M. (2011), 'Haloes gone MAD: The Halo-Finder Comparison Project', *MNRAS* **415**, 2293–2318.

Knobel, C. (2011), Galaxy groups with zCOSMOS, PhD thesis, ETH ZURICH.

Koester, B. P., McKay, T. A., Annis, J., Wechsler, R. H., Evrard, A. E., Rozo, E., Bleem, L., Sheldon, E. S. & Johnston, D. (2007), 'Maxbcg: A red-sequence galaxy cluster finder', *The Astrophysical Journal* **660**(1), 221.

Kramer, O., Gieseke, F. & Polsterer, K. L. (2013), 'Learning morphological maps of galaxies with unsupervised regression', *Expert Systems with Applications* **40**(8), 2841–2844.

Krishnan, B., Sintes, A. M., Papa, M. A., Schutz, B. F., Frasca, S. & Palomba, C. (2004), 'Hough transform search for continuous gravitational waves', *Physical Review D* **70**(8), 082001. PRD.

Kurek, A. & Szydlowski, M. (2008), 'The Lambda-CDM model on the lead: A Bayesian cosmological models comparison', *AJ* **675**, 1–7.

Laschinsky, H. (2012), The Hough Transform as Event Eilter for the ANTARES Neutrino Telescope - A Simulation Study, PhD thesis, The Faculty of Science, The Friedrich Alexander University.

Li, I.-h. & Yee, H. K. C. (2008), 'Finding Galaxy Groups In Photometric Redshift Space: the Probability Friends-of-Friends (pFoF) Algorithm', *AJ* **135**, 809.

Liang, L., Durier, F., Babul, A., Dave, R., Oppenheimer, B. D., Katz, N., Fardal, M. & Quinn, T. (2016), 'The growth and enrichment of intragroup gas', *MNRAS* **456**(4), 4266–4290.
URL: *http://mnras.oxfordjournals.org/content/456/4/4266.abstract*

Liu, H. B., Hsieh, B., Ho, P. T., Lin, L. & Yan, R. (2008), 'A new galaxy group finding algorithm: Probability friends-of-friends', *The Astrophysical Journal* **681**(2), 1046.

Llebaria, A. & Lamy, P. (1999), Time domain analysis of solar coronal structures through hough transform techniques, *in* D. M. Mehringer, R. L. Plante & D. A. Roberts, eds, 'Astronomical Data Analysis Software and Systems VIII', Vol. 172 of *Astronomical Society of the Pacific Conference Series*, p. 46.

Lobo, C., Iovino, A., Lazzati, D. & Chincarini, G. (2000), 'Eisily looking for distant clusters of galaxies - a new algorithm and its application to the eis-wide data', *AAP* **360**, 896–910.

M., R. & Muthukrishnan, R. (2015), Contributions to a study on robust statistics and its applications in computer vision, PhD thesis, Department of statistics, Bharathiar University.

M. Ramella, W. Boschin, D. Fadda & M. Nonino (2001), 'Finding galaxy clusters using voronoi tessellations', *A&A* **368**(3), 776–786.
URL: *http://dx.doi.org/10.1051/0004-6361:20010071*

MacKay, D. J. (2003), *Information theory, inference and learning algorithms*, Cambridge university press.

Malin, D. (2001), 'The hydra cluster of galaxies'. Online; accessed 21 June 2016.
URL: *http://apod.nasa.gov/apod/ap010416.html*

Maragos, P. (2005), 'Morphological filtering for image enhancement and feature detection', *Analysis* **19**, 18.

Marinoni, C., Davis, M., Newman, J. A. & Coil, A. L. (2002), 'Three-dimensional Identification and Reconstruction of Galaxy Systems within Flux-limited Redshift Surveys', *APJ* **580**, 122–143.

Martinez, H. J. & Muriel, H. (2006), 'Groups of galaxies: relationship between environment and galaxy properties', *MNRAS* **370**, 1003–1007.

Massone, A. M., Perasso, A., Campi, C. & Beltrametti, M. C. (2014), 'Profile detection in medical and astronomical images by means of the hough transform of special classes of curves', *Journal of Mathematical Imaging and Vision* pp. 1–15.

Materne, J. (1978), 'The structure of nearby clusters of galaxies - Hierarchical clustering and an application to the Leo region', *AAP* **63**, 401–409.

McLaughlin, R. A. (1998), 'Randomized hough transform: improved ellipse detection with comparison', *Pattern Recognition Letters* **19**(3), 299–305.

Milkeraitis, M., Van Waerbeke, L., Heymans, C., Hildebrandt, H., Dietrich, J. P. & Erben, T. (2010), '3d-matched-filter galaxy cluster finder- i. selection functions and cfhtls deep clusters', *MNRAS* **406**(1), 673.
URL: + *http://dx.doi.org/10.1111/j.1365-2966.2010.16720.x*

Miller, C. J. (2012), Galaxy clusters, *in* M. Way, J. Scargle, K. Ali & A. Srivastava, eds, 'Advances in Machine Learning and Data Mining for Astronomy', Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Taylor & Francis, chapter 16, pp. 337–348.

Miller, C. J., Nichol, R., Reichart, D., Wechsler, R. H., Evrard, A. et al. (2005), 'The c4 clustering algorithm: Clusters of galaxies in the sloan digital sky survey', *AJ* **130**, 968.

Mirkazemi, S. M. (2014), Cluster through the cosmic time, PhD thesis, LMU Munchen: Faculty of Physics.
URL: *http://nbn-resolving.de/urn:nbn:de:bvb:19-175085*

Moller, O., Natarajan, P., Kneib, J. P. & Blain, A. W. (2001), 'Probing the mass distribution in groups of galaxies using gravitational lensing', *AJ* .

Moretti, A., Guzzo, L., Campana, S., Lazzati, D., Panzera, M. et al. (2004), 'The brera multi-scale wavelet hri cluster survey. 1. selection of the sample and number counts', *A&A* **428**, 21–37.

Mukhopadhyay, P. & Chaudhuri, B. B. (2015), 'A survey of hough transform', *Pattern Recognition* **48**(3), 993–1010.

Murdin, P. (2001), *Encyclopedia of Astronomy & Astrophysics*, Taylor & Francis.
URL: *https://books.google.co.uk/books?id=W8zLQgAACAAJ*

Murray, S. G., Power, C. & Robotham, A. S. G. (2013), 'Hmfcalc: An online tool for calculating dark matter halo mass functions', *Astronomy and Computing* **3**, 23–34.

Mushotzky, R. F. (2004), 'Clusters of galaxies: An x-ray perspective', *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution* p. 123.

Navarro, J., C.S., F. & White, S. (1996), 'The structure of cold dark matter halos', *APJ* **462**, 563.

Neuhauser, M. (2011), *Nonparametric statistical tests: A computational approach*, CRC Press.

Nichol, R. C., Miller, C., Connolly, A. J., Chong, S.-S., Genovese, C., Moore, A. W., Reichart, D., Schneider, J., Wasserman, L., Annis, J., Brinkman, J., Böhringer, H., Castander, F., McKay, T., Postman, M., Sheldon, E., Szapudi, I., Romer, K. & Voges, W. (2001), *SDSS-RASS: Next Generation of Cluster-Finding Algorithms*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 613–623.

Oemler, Jr., A. (1974), 'The systematic properties of clusters of galaxies. photometry of 15 clusters', *APJ* **194**, 1–20.

Oguri, M., Hennawi, J. F., Gladders, M. D., Dahle, H., Natarajan, P., Dalal, N., Koester, B. P., Sharon, K. & Bayliss, M. (2009), 'Subaru Weak Lensing Measurements of Four Strong Lensing Clusters: Are Lensing Clusters Overconcentrated?', *APJ* **699**, 1038–1052.

Ostriker, J. P. & Steinhardt, P. J. (1995), 'Cosmic concordance', *arXiv preprint astro-ph/9505066* .

O'Sullivan, E., Kolokythas, K., Raychaudhury, S., Vrtilek, J. M. & Kantharia, N. (2014), First results from the complete local-volume groups sample, *in* 'Astronomical Society of India Conference Series', Vol. 13 of *Astronomical Society of India Conference Series*, pp. 259–262.
URL: *http://inspirehep.net/record/1281817/files/arXiv:1402.4676.pdf*

Pearson, R., Ponman, T., Norberg, P., Robotham, A. & Farr, W. (2015), 'On optical mass estimation methods for galaxy groups', *MNRAS* **449**, 3082–3106.

Peebles, P. J. E. (1980), *The large-scale structure of the universe*, Princeton University Press.

Peebles, P. J. E. (1993), *Principles of Physical Cosmology*, Princeton University Press.

Popesso, P., Böhringer, H., Romaniello, M. & Voges, W. (2005), 'Rass-sdss galaxy cluster survey. ii. a unified picture of the cluster luminosity function', *AAP* **433**, 415–429.

Press, W. H. & Schechter, P. (1974), 'Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation', *APJ* **187**, 425–438.

Ragazzoni, R. & Barbieri, C. (1994), 'Cycle-number determination via the hough transform: The technique and an application to gw cephei', *ASP* **106**, 683–687.

Ragazzoni, R. & Barbieri, C. (1996), 'Confirmation of the period of gw cep found by hough transform', *Information Bulletin on Variable Stars* **4293**, 1.

Ramella, Massimo; Geller, M. J. . H. J. P. (1989), 'Groups of galaxies in the center for astrophysics redshift survey', *Astrophysical Journal, Part 1* **344**(Sept. 1, 1989), 57–74.

Reed, D., Bower, R., Frenk, C., Jenkins, A. & Theuns, T. (2007), 'The halo mass function from the dark ages through the present day', *MNRAS* **374**, 2–15.

Reichardt, C. L., Stalder, B., Bleem, L. E., Montroy, T. E., Aird, K. A., Andersson, K., Armstrong, R., Ashby, M. L. N., Bautz, M., Bayliss, M., Bazin, G., Benson, B. A., Brodwin, M., Carlstrom, J. E., Chang, C. L., Cho, H. M., Clocchiatti, A., Crawford, T. M., Crites, A. T., de Haan, T., Desai, S., Dobbs, M. A., Dudley, J. P., Foley, R. J., Forman, W. R., George, E. M., Gladders, M. D., Gonzalez, A. H., Halverson, N. W., Harrington, N. L., High, F. W., Holder, G. P., Holzapfel, W. L., Hoover, S., Hrubes, J. D., Jones, C., Joy, M., Keisler, R., Knox, L., Lee, A. T., Leitch, E. M., Liu, J., Lueker, M., Luong-Van, D., Mantz, A., Marrone, D. P., McDonald, M., McMahon, J. J., Mehl, J., Meyer, S. S., Mocanu, L., Mohr, J. J., Murray, S. S., Natoli, T., Padin, S., Plagge, T., Pryke, C., Rest, A., Ruel, J., Ruhl, J. E., Saliwanchik, B. R., Saro, A.,

Sayre, J. T., Schaffer, K. K., Shaw, L., Shirokoff, E., Song, J., Spieler, H. G., Staniszewski, Z., Stark, A. A., Story, K., Stubbs, C. W., ØěÂăuhada, R., van Engelen, A., Vanderlinde, K., Vieira, J. D., Vikhlinin, A., Williamson, R., Zahn, O. & Zenteno, A. (2013), 'Galaxy clusters discovered via the sunyaev-zel'dovich effect in the first 720 square degrees of the south pole telescope survey', *The Astrophysical Journal* **763**(2), 127.
URL: *http://stacks.iop.org/0004-637X/763/i=2/a=127*

Robin, A. C., Reylé, C., Derrière, S. & Picaud, S. (2003), 'A synthetic view on structure and evolution of the Milky Way', *AAP* **409**, 523–540.

Robotham, A., Driver, S., Norberg, P., Baldry, I., Bamford, S., Hopkins, A., Liske, J., Loveday, J., Peacock, J., Cameron, E. et al. (2010), 'Galaxy and mass assembly (gama): optimal tiling of dense surveys with a multi-object spectrograph', *Publications of the Astronomical Society of Australia* **27**(1), 76–90.

Robotham, A. S. G., Norberg, P., Driver, S. P., Baldry, I. K., Bamford, S. P., Hopkins, A. M., Liske, J., Loveday, J., Merson, A., Peacock, J. A., Brough, S., Cameron, E., Conselice, C. J., Croom, S. M., Frenk, C. S., Gunawardhana, M., Hill, D. T., Jones, D. H., Kelvin, L. S., Kuijken, K., Nichol, R. C., Parkinson, H. R., Pimbblet, K. A., Phillipps, S., Popescu, C. C., Prescott, M., Sharp, R. G., Sutherland, W. J., Taylor, E. N., Thomas, D., Tuffs, R. J., van Kampen, E. & Wijesinghe, D. (2011), 'Galaxy and mass assembly (gama): the gama galaxy group catalogue (g3cv1)', *MNRAS* **416**(4), 2640–2668.

Rosati, P., Borgani, S. & Norman, C. (2002), 'The evolution of x-ray clusters of galaxies', *Ann. Rev. Astron. Astrophys.* **40**, 539–577.

Santos, J. S., Rosati, P., Gobat, R., Lidman, C., Dawson, K., Perlmutter, S., Bohringer, H., Balestra, I., Mullis, C. R., Fassbender, R., Kohnert, J., Lamer,

G., Rettura, A., Rite, C. & Schwope, A. (2009), 'Multiwavelength observations of a rich galaxy cluster at z 1 - the hst/acs colour-magnitude diagram', *A&A* **501**(1), 49–60.
URL: *https://doi.org/10.1051/0004-6361/200811546*

Sarazin, C. L. (1986), 'X-ray emission from clusters of galaxies', *Rev. Mod. Phys.* **58**, 1–115.
URL: *http://link.aps.org/doi/10.1103/RevModPhys.58.1*

Saviane, I., Ivanov, V. D. & Borissova, J. (2007), *Groups of Galaxies in the Nearby Universe: Proceedings of the ESO Workshop Held at Santiago de Chile, December 5-9, 2005*, Springer Science & Business Media.

Schechter, P. (1976), 'An analytic expression for the luminosity function for galaxies.', *APJ* **203**, 297–306.

Schneider, P. (2014), *Extragalactic astronomy and cosmology: an introduction*, Springer.

SEGRE, A. (n.d.), *Nonparametric Tests*, WHFreeman, chapter 16.
URL: *bcs.whfreeman.com/webpub/statistics/PSBE4e/companion_chapters/Moore_4e_CH16.pdf*

Serjeant, S. (2010), *Observational Cosmology*, Cambridge University Press.

Sheth, R. K. & Tormen, G. (1999), 'Large scale bias and the peak background split', *MNRAS* **308**, 119.

Smith, A., Hopkins, A., Hunstead, R. & Pimbblet, K. (2012), 'Multiscale probability mapping: groups, clusters and an algorithmic search for filaments in sdss', *MNRAS* **422**, 25–43.

Snaith, O. (2011), The environment of galaxies and groups of galaxies, PhD thesis, University of Central Lancashire.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J. & Pearce, F. (2005), 'Simulations of the formation, evolution and clustering of galaxies and quasars', *NAT* **435**, 629–636.

Stahler, S. W. & Palla, F. (2008), *The Formation of Stars*, Wiley-VCH Verlag GmbH.
URL: *http://dx.doi.org/10.1002/9783527618675*

Stephens, R. (1991), 'A probabilistic approach to the hough transform', *Image and Vision Computing* **9**(1), 66 – 71. The first {BMVC} 1990.

Storkey, A. J., Hambly, N. C., Williams, C. K. I. & Mann, R. G. (2004), 'Cleaning sky survey data bases using hough transform and renewal string approaches', *MNRAS* **347**(1), 36–51.

Storkey, A. J., Hambly, N. C., Williams, C. K. I. & Mann, R. G. (2014), 'Renewal Strings for Cleaning Astronomical Databases', *ArXiv e-prints* .

Stott, J. P. (2007), The evolution of galaxies in massive clusters, PhD thesis, Durham University.

Sunyaev, R. & Zeldovich, Y. (1970), 'The Spectrum of Primordial Radiation, its Distortions and their Significance', *Comments on Astrophysics and Space Physics* **2**, 66.

Taylor, R. W. (1990), An efficient implementation of decomposable parameter spaces, *in* 'Pattern Recognition, 1990. Proceedings., 10th International Conference on', Vol. 1, IEEE, pp. 613–619.

Tegmark, M., Blanton, M. R., Strauss, M. A., Hoyle, F., Schlegel, D., Scoccimarro, R., Vogeley, M. S., Weinberg, D. H., Zehavi, I., Berlind, A., Budavari, T., Connolly, A., Eisenstein, D. J., Finkbeiner, D., Frieman, J. A., Gunn, J. E., Hamilton, A. J. S., Hui, L., Jain, B., Johnston, D., Kent, S., Lin, H., Nakajima, R., Nichol, R. C., Ostriker, J. P., Pope, A., Scranton, R., Seljak, U., Sheth, R. K., Stebbins, A., Szalay, A. S., Szapudi, I., Verde, L., Xu, Y., Annis, J., Bahcall, N. A., Brinkmann, J., Burles, S., Castander, F. J., Csabai, I., Loveday, J., Doi, M., Fukugita, M., Gott, III, J. R., Hennessy, G., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T. A., Kunszt, P., Munn, J. A., O'Connell, L., Peoples, J., Pier, J. R., Richmond, M., Rockosi, C., Schneider, D. P., Stoughton, C., Tucker, D. L., Vanden Berk, D. E., Yanny, B., York, D. G. & SDSS Collaboration (2004), 'The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey', *APJ* **606**, 702–740.

Tegmark, M., Eisenstein, D. J., Strauss, M. A., Weinberg, D. H., Blanton, M. R., Frieman, J. A., Fukugita, M., Gunn, J. E., Hamilton, A. J., Knapp, G. R. et al. (2006), 'Cosmological constraints from the sdss luminous red galaxies', *Physical Review D* **74**(12), 123507.

Tempel, E., Kipper, R., Tamm, A., Gramann, M., Einasto, M., Sepp, T. & Tuvikene, T. (2016), 'Friends-of-friends galaxy group finder with membership refinement-application to the local universe', *A&A* **588**, A14.

Tsuji, S. & Matsumoto, F. (1978), 'Detection of ellipses by a modified hough transformation', *Computers, IEEE Transactions on* **C-27**(8), 777–781.

Turner, M. S. (1997), 'The case for lambdacdm', *arXiv preprint astro-ph/9703161* .

Tyson, J. A., Valdes, F., Jarvis, J. F. & Mills, A. P. (1984), 'Galaxy mass-distribution from gravitational light deflection', *Astrophysical Journal* **281**(2), L59–L62. Ta270 Times Cited:106 Cited References Count:29.

van der Glas, M., Vos, F. M., Botha, C. P. & Vossepoel, A. M. (2002), Determination of position and radius of ball joints, *in* 'Medical Imaging 2002', International Society for Optics and Photonics, pp. 1571–1577.

Vikhlinin, A. A., Kravtsov, A. V., Markevich, M. L., Sunyaev, R. A. & Churazov, E. M. (2014), 'Clusters of galaxies', *Physics-Uspekhi* **57**(4), 317.
URL: *http://stacks.iop.org/1063-7869/57/i=4/a=317*

Wainer, H. & Thissen, D. (1976), 'Three steps towards robust regression', *Psychometrika* **41**(1), 9–34.

Wardlow, J. L. (2010), The Role of Obscured Activity in Galaxy Formation, PhD thesis, Durham University.

Yang, X., Mo, H. J., van den Bosch, F. C., Pasquali, A., Li, C. & Barden, M. (2007), 'Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties', *APJ* **671**, 153–170.

Yang, X., Mo, H., van den Bosch, F. C. & Jing, Y. (2005), 'A halo-based galaxy group finder: calibration and application to the 2dfgrs', *MNRAS* **356**, 1293–1307.

Zaninetti, L. (2006), 'On the large-scale structure of the universe as given by the voronoi diagrams', *Chinese Journal of Astronomy and Astrophysics* **6**(4), 387.