Master's thesis
Astronomy

Your Title Here

Anni Järvenpää

April 4, 2018

Tutor: Dr. Till Sawala
Professor Peter Johansson

Censors: prof. Smith
doc. Smythe

UNIVERSITY OF HELSINKI
Department of Physics

PL 64 (Gustaf Hällströmin katu 2a)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Physics

Tekijä — Författare — Author

Anni Järvenpää

Työn nimi — Arbetets titel — Title

Your Title Here

Oppiaine — Läroämne — Subject

Astronomy

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	April 4, 2018	0 pages

Tiivistelmä — Referat — Abstract

Abstract goes here.

Avainsanat — Nyckelord — Keywords

Your keywords here

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1. Introduction

1.1 TL;DR version of prerequisite information

1. galaxies form
    - Why?
    - When?
    - How?
    - Where?
2. galaxies form in groups
3. our local group is one of these
4. something about large scale distribution of galaxies

1.2 History of Local Group Research

LG objects visible with naked eye -> realization they are something outside our galaxy -> realization they are something very much like our galaxy

First determining distance was difficult, now mass is more interesting question

1.3 Aim of This Thesis

Whatever the main results end up being, presented in somewhat coherent manner and hopefully sugar-coated enough to sound Important and Exciting.

2. Theoretical Background

Cosmology determines the properties of the Universe, including its origin, the rules by which it evolves and the structures that arise within it (Mo et al., 2010). Thus many fields of astronomy and astrophysics, including the study of galaxies and galaxy groups such as the Local Group and its members, are tightly connected to the study of cosmology. This section gives a brief explanation of the current cosmological understanding, its relevant implications and how the Local Group is currently viewed in the cosmological context.

2.1 Basics of Cosmology

Understanding processes that happen on large scales or take long times requires understanding some basic cosmology. The following sections will cover the most basic concepts of cosmology, the evolution of the Universe on both large and small scales, and the ΛCDM model which is currently the cosmological model that best matches the observations on multiple scales (Mo et al., 2010). Sections 2.1.1 and 2.1.2 cover large scales at which the cosmological principle applies and section 2.1.3 covers smaller scales of single dark matter halo.

2.1.1 Evolution of the Universe

Current cosmological understanding is based on a subset of the general theory of relativity together with simple hypotheses such as the cosmological principle, which states that on large scales the Universe is spatially homogeneous and isotropic (Mo et al., 2010). At a given location, an observer might not see this, as is easy to understand when considering two observers located at the same point but moving relative to each other. In this situation, it is clear that at least one of the observers will see a dipole in the surrounding velocities and thus will not observe the universe to be isotropic. Nonetheless, in an isotropic universe, for every point in space we can define a so-called fundamental observer as the observer who sees the universe as isotropic (Mo et al., 2010). These fundamental observers correspond to a cosmological rest frame, which can be determined at a given location e.g. by observing the cosmic microwave background and subtracting the velocity corresponding to the observed dipole component (Mo et al., 2010).

The existence of such fundamental observers has interesting consequences. The existence of any large-scale flows is clearly prohibited, as these would violate the isotropy. In a three-dimensional universe any curl of the velocity field around any fundamental observer is also forbidden. This can be easily seen by considering the hairy ball theorem, which states that a sphere cannot have a nowhere-zero tangent vector field (Renteln, 2013). Thus if there was a curl in the surrounding velocity field, there would always be at least one direction in which the tangential velocity is zero and thus the field would not be isotropic. This means that there cannot be any tangential motion and thus the only allowed motion happens in the radial direction: the Universe can either expand or contract.

The expansion of the Universe can be parametrized using the dimensionless scale factor a, whose time evolution is governed by the Friedmann equations

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3\frac{P}{c^2}\right) + \frac{\Lambda c^2}{3} \tag{2.1}$$

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3}, \tag{2.2}$$

where G is the gravitational constant, P is pressure, ρ is energy density and K and Λ are cosmology-related parameters (Mo et al., 2010). K specifies the curvature of the Universe, determined by the overall density of the Universe (Mo et al., 2010). The allowed values for K are -1 corresponding to a hyperbolic universe, 0 to a flat one and 1 to a spherical universe (Mo et al., 2010). Modern measurements suggest that the Universe is flat within the measurement error, but small deviations from the $K = 0$ cannot be ruled out (Planck Collaboration, 2016). Λ is the cosmological constant driving the expansion of the Universe, often described as dark energy or vacuum energy (Mo et al., 2010).

The concept of scale factor a is crucial for this thesis as it both affects the distances and velocities measured and governs the structure formation. It increases or decreases as the Universe expands or contracts, and thus it can be used to relate distances at different times (Mo et al., 2010). For a universe that is monotonically expanding, the scale factor can also be used as an alternative time coordinate, as it has a one-to-one correspondence to time. It is often convenient to measure not the proper distance l between a pair of objects, but instead their so-called comoving distance r: (Mo et al., 2010)

$$r = \frac{l}{a} \tag{2.3}$$

For a pair of objects with no relative motion, the comoving distance will remain constant as the size of the Universe changes. A comoving coordinate system is often used in cosmological simulations, as the expansion of the universe is included in the scale factor instead of all of the coordinates having to be recalculated as the size of the universe varies (Griebel et al., 2007).

Observations have confirmed that the Universe is indeed expanding, first observed by Hubble (1929). The rate of the expansion is denoted with the Hubble parameter $H(t)$, which is defined using the proper distance and the rate of change of the proper distance between a pair of fundamental observers:

$$H(t)\, l = \frac{\mathrm{d}l}{\mathrm{d}t} \tag{2.4}$$

The Hubble parameter is also closely related to the scale factor, as (Mo et al., 2010)

$$H(t) = \frac{\dot{a}(t)}{a(t)}. \tag{2.5}$$

The first measurements trying to determine the current value of the Hubble parameter, often called the Hubble constant, were prone to error as the distance estimates to extragalactic objects were inaccurate. Later estimates offer reasonably accurate results such as the Planck Collaboration (2014) value of 67.77 ± 0.77 km/s/Mpc for the current expansion speed, though the results of different experiments have considerable scatter.

Another factor affecting measurements based on extragalactic objects is that the observed proper velocities contain not only the expansion of the Universe but also peculiar motions of the objects, which is why measurements based on other probes such as the cosmic microwave background in case of Planck Collaboration (2016) are valuable. Information contained in the peculiar motions can still be interesting. In this thesis, the mass of the Local Group is estimated using radial velocity measurements within a few megaparsecs of a number of simulated Local Universe analogues. At scales this small, the expansion of the Universe is greatly affected by local gravity fields, and thus expansion measurements can be used to infer the mass enclosed within the Local Group.

While the scale factor is one possible way of expressing time, it is not the only one. As the universe expands and we observe objects receding, the light emitted from them is shifted to longer wavelengths. The further away the emitter is, the more the space between the observer and emitter will expand making the effect stronger. The relative change in the wavelength is called redshift z, defined as

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} \tag{2.6}$$

where λ_o is the observed and λ_e the emitted wavelength (Mo et al., 2010). If the effect of peculiar motions is ignored, redshift is directly related not only to the distance to the emitter but also to the scale factor at the time of the emission.

For arbitrary scale factors at the time of the emission and observation, the relation between z and a is (Mo et al., 2010)

$$1 + z = \frac{a(t_o)}{a(t_e)} \tag{2.7}$$

2.1.2 Composition of the Universe

Using the Friedmann equations 2.1 and 2.2 to model the evolution of the Universe requires not only the equations and cosmological parameters introduced in section 2.1.1 but also the composition of the Universe to be known in order to determine the density, pressure, cosmological constant and ultimately even the curvature of the Universe. According to the current understanding, the Universe is made of three components: non-relativistic matter, relativistic matter and dark energy (Mo et al., 2010). At present time, about 69 % of the energy density ρ in the Universe is dark energy and 31 % is non-relativistic matter, including both baryonic and dark components (Planck Collaboration, 2016). Most of the non-relativistic matter, around 84 %, is dark matter, with baryonic matter only contributing around 16 % of total matter energy density (Planck Collaboration, 2016). The energy density of relativistic matter, i.e. photons and standard-model neutrinos, in the present-day universe is negligible (Mo et al., 2010).

This has not always been true, as these ratios change over time as the Universe evolves. It is easy to see that, as the Universe expands, the energy density of matter behaves as a^{-3}, as the volume of the Universe increases as a^3 and in an adiabatic system no matter is created or disappears. Radiation is diluted similarly to matter, but in addition to the effect of increasing volume, the growing universe also causes the wavelength of the radiation to increase, resulting in the energy density decreasing as a^{-4}. The dark energy, as the name suggests, can be thought as arising from the space itself, and thus a change in a does not affect the energy density of the component (Mo et al., 2010). The change of energy density of a component may

Component	ρ	P	T
Matter	a^{-3}	a^{-5}	a^{-2}
Radiation	a^{-4}	a^{-4}	a^{-1}
Dark energy	a^{0}	a^{0}	

Table 2.1: Evolution of energy density, pressure and temperature of the Universe as the scale factor varies. Adapted from Mo et al. (2010).

also correspond to a change in the pressure or temperature of the component (Mo et al., 2010). These effects are shown in table 2.1.

The different time evolutions of the components also mean that the dominant component of the Universe changes as the scale factor grows. At very early times the Universe was radiation dominated, but as the radiation energy density decreases faster than the energy densities of the two other components, a matter dominated era followed. As dark energy is the only component with constant energy density, it will be the final dominant energy component. This transition from matter to dark energy dominated era has happened in the recent past (Mo et al., 2010).

The geometry of the Universe is also determined by its contents. A density threshold known as critical density and defined as

$$\rho_{crit,0} \equiv \frac{3H_0^2}{8\pi G} \tag{2.8}$$

acts as a threshold value that separates the different geometries (Mo et al., 2010). Subscript zero comes from $z = 0$ and denotes present-day values, but all introduced quantities can also be determined at any other time. The overall density of the Universe can be parametrized using this critical density:

$$\Omega_0 \equiv \frac{\bar{\rho}_0}{\rho_{crit,0}} \tag{2.9}$$

where $\bar{\rho}_0$ is the mean density of the Universe and Ω_0 is known as the density parameter (Mo et al., 2010). Different values of Ω_0 correspond to different geometries: for values of $\Omega_0 < 1$ the Universe is hyperbolic, for $\Omega_0 = 1$ flat and for $\Omega_0 > 1$ spherical (Mo et al., 2010).

As the contents of the Universe can be divided into the three categories of dark energy and relativistic and non-relativistic matter, the total density parameter is also a sum of three density parameters:

$$\Omega_0 = \Omega_{m,0} + \Omega_{r,0} + \Omega_{\Lambda,0} \tag{2.10}$$

where m, r and Λ stand for non-relativistic matter ("matter"), relativistic matter ("radiation") and dark energy (Mo et al., 2010). Each of these can be calculated similarly to the equation 2.9 but replacing the overall density with density of the corresponding component (Mo et al., 2010). As in the Universe the value of Ω_0 is very close to unity, the different density parameters conveniently correspond to the relative densities of the components.

2.1.3 Structure Formation

All structures in the Universe arise from the small perturbations in the matter density in the early universe and the physics that govern their evolution. These primordial density fluctuations that later develop into the structures such as galaxy clusters and voids separating them can still be seen in the cosmic microwave background (CMB) (Planck Collaboration, 2016). Starting at the end of inflation, the density contrast of such perturbations in the dark matter begins to grow. While dark matter is collisionless, the pressure experienced by baryons slows their structural evolution down. Consequently, their evolution is delayed relative to the dark matter, and baryons sink into potential wells already formed from the dark matter.

At $z \approx 1100$, the time of recombination and origin of the CMB photons, these fluctuations are still very small, having $\Delta\rho/\rho \approx 10^{-4}$, but sufficiently overdense regions have already collapsed (Mo et al., 2010). The gas inside these structures seen as hot spots in the CMB has just reached a density maximum and is shock-heated, while gas inside smaller fluctuations has already begun expanding again, and larger scales have not yet reached their maximum density at the time of the CMB.

At first, the evolution of overdense regions differs from the evolution of the surroundings only by the speed of the expansion of the Universe in that region: expansion is slower in denser regions. Later some of these volumes will start to collapse. This requires the volume to be sufficiently dense to allow gravity to overcome the expansion of the universe as described by the Friedmann equations.

To understand the evolution of density perturbations at early times when the perturbations are still small and evolve linearly, let us consider an ideal fluid of density ρ, moving at proper velocity $\mathbf{v}$ and experiencing the gravitational field with potential ϕ. Growth of a perturbation in this medium is governed by three equations: the equation of continuity describing the conservation of mass, the Euler equation governing the motions in the fluid and the Poisson equation describing the gravitational field, or

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} + \rho \nabla_{\mathbf{x}} \cdot \mathbf{v} = 0\,, \tag{2.11}$$

$$\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = -\frac{\nabla_{\mathbf{x}} P}{\rho} - \nabla_{\mathbf{x}}\phi \tag{2.12}$$

and

$$\nabla^2_{\mathbf{x}}\phi = 4\pi G\rho \tag{2.13}$$

respectively (Mo et al., 2010). Here $\mathbf{x}$ denotes proper coordinates and $\frac{\mathrm{D}}{\mathrm{D}t}$ is the convective time derivative, defined as

$$\frac{\mathrm{D}}{\mathrm{D}t} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} \tag{2.14}$$

and describing the time derivative when moving with the fluid (Mo et al., 2010).

Next let us follow Longair (2008) and introduce a small perturbation by replacing $\mathbf{v}$, ρ, P and ϕ with $\mathbf{v_0}+\delta\mathbf{v}$, $\rho_0+\delta\rho$, $P_0+\delta P$ and $\phi_0+\delta\phi$ respectively, having subscript zero represent the properties of the unperturbed medium. Now equations 2.11–2.13 are

$$\frac{\mathrm{D}\rho_0}{\mathrm{D}t}+\frac{\mathrm{D}\delta\rho}{\mathrm{D}t}+\rho_0\nabla_{\mathbf{x}}\cdot\mathbf{v}_0+\rho_0\nabla_{\mathbf{x}}\cdot\delta\mathbf{v}+\delta\rho\nabla_{\mathbf{x}}\cdot\mathbf{v}_0+\delta\rho\nabla_{\mathbf{x}}\cdot\delta\mathbf{v}=0\,, \tag{2.15}$$

$$\frac{\mathrm{D}\mathbf{v}_0}{\mathrm{D}t}+\frac{\mathrm{D}\delta\mathbf{v}}{\mathrm{D}t}=-\frac{\nabla_{\mathbf{x}}(P_0+\delta P)}{\rho_0+\delta\rho}-\nabla_{\mathbf{x}}\phi_0-\nabla_{\mathbf{x}}\delta\phi \tag{2.16}$$

and

$$\nabla^2_{\mathbf{x}}\phi_0+\nabla^2_{\mathbf{x}}\delta\phi=4\pi G\rho_0+4\pi G\delta\rho. \tag{2.17}$$

These equations can be greatly simplified by subtracting the unperturbed versions from each and assuming that the initial state is homogeneous and isotropic i.e. $\nabla P_0=0$ and $\nabla\rho_0=0$. Using the knowledge that in the linear regime the perturbations are small and thus discarding second order terms, the equations take the following forms:

$$\frac{\mathrm{D}}{\mathrm{D}t}\frac{\delta\rho}{\rho_0}=-\nabla_{\mathbf{x}}\cdot\delta\mathbf{v} \tag{2.18}$$

$$\frac{\partial\delta\mathbf{v}}{\partial t}+(\delta\mathbf{v}\cdot\nabla_{\mathbf{x}})\mathbf{v}_0=-\frac{\nabla_{\mathbf{x}}\delta P}{\rho_0}-\nabla_{\mathbf{x}}\delta\phi \tag{2.19}$$

$$\nabla^2_{\mathbf{x}}\delta\phi=4\pi G\delta\rho \tag{2.20}$$

As the Universe is expanding, it is natural to transit to comoving coordinates, defined in equation 2.3 and denoted by $\mathbf{r}$. This also affects the velocities, and using the dot to denote time derivative we can write

$$\mathbf{v}=\dot{\mathbf{x}}=\dot{a}(t)\mathbf{r}+a(t)\dot{\mathbf{r}}=\dot{a}(t)\mathbf{r}+\mathbf{u}, \tag{2.21}$$

where $\mathbf{u}$ denotes the perturbed comoving velocity. From this it follows that $\delta\mathbf{v}=a\mathbf{u}$. In comoving coordinates the operator $\nabla_{\mathbf{x}}$ is also replaced with $\frac{1}{a}\nabla_{\mathbf{r}}$. Using these and $(a\mathbf{u}\cdot\nabla_{\mathbf{x}})\dot{a}\mathbf{r} = \mathbf{u}\dot{a}$, the equations can be written as

$$\frac{\mathrm{D}}{\mathrm{D}t}\frac{\delta\rho}{\rho_0} = -\nabla_{\mathbf{r}}\cdot\mathbf{u} \tag{2.22}$$

$$\frac{\partial\mathbf{u}}{\partial t} + 2\frac{\dot{a}}{a}\mathbf{u} = -\frac{\nabla_{\mathbf{r}}\delta P}{\rho_0 a^2} - \frac{1}{a^2}\nabla_{\mathbf{r}}\delta\phi. \tag{2.23}$$

$$\nabla_{\mathbf{r}}^2\delta\phi = 4\pi G\delta a^2\rho \tag{2.24}$$

Taking the comoving divergence, equation 2.23 gives

$$\nabla_{\mathbf{r}}\cdot\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\nabla_{\mathbf{r}}\cdot\mathbf{u} = -\frac{\nabla_{\mathbf{r}}^2\delta P}{\rho_0 a^2} - \frac{1}{a^2}\nabla_{\mathbf{r}}^2\delta\phi. \tag{2.25}$$

There the last term on the right side contains the left side of equation 2.24, the first term on the left can be found in equation 2.22 after taking the convective time derivative and the second term on the left already contains the right side of 2.22. Inserting these yields

$$\frac{\mathrm{D}^2}{\mathrm{D}^2 t}\frac{\delta\rho}{\rho_0} + 2\frac{\dot{a}}{a}\frac{\mathrm{D}}{\mathrm{D}t}\frac{\delta\rho}{\rho_0} = \frac{\nabla_{\mathbf{r}}^2\delta P}{\rho_0 a^2} + 4\pi G\delta\rho. \tag{2.26}$$

The overdensities in the Universe are often denoted using density contrast $\Delta = \delta\rho/\bar{\rho}$, where the bar denotes the universal mean (Mo et al., 2010). In this case the unperturbed density ρ is the mean density, so the left side of equation 2.27 can be expressed using Δ. If the density perturbations are assumed to be adiabatic, the adiabatic sound speed c_s relates perturbations in density and pressure as $\delta P/\delta\rho = c_s^2$. Inserting these yields

$$\frac{\mathrm{D}^2\Delta}{\mathrm{D}^2 t} + 2\frac{\dot{a}}{a}\frac{\mathrm{D}\Delta}{\mathrm{D}t} = \frac{c_s^2\nabla_{\mathbf{r}}^2\Delta\rho_0}{a^2\rho_0} + 4\pi G\Delta\rho_0. \tag{2.27}$$

Now a trial solution of $\Delta \propto e^{i(\mathbf{k}_r\cdot\mathbf{r}-\omega t)}$ corresponding to a wave with comoving wavevector $\mathbf{k}_r$ yields a wave equation

$$\frac{\mathrm{D}^2\Delta}{\mathrm{D}^2 t} + 2\frac{\dot{a}}{a}\frac{\mathrm{D}\Delta}{\mathrm{D}t} = \Delta(4\pi G\rho_0 - k^2 c_s^2). \tag{2.28}$$

Here $\mathbf{k}_r$ has been replaced by $a\mathbf{k}$ to transform it to proper coordinates. The wave equation is a linear second-order partial differential equation that describes the evolution of perturbations.

Whether the waves described by equation 2.28 are oscillatory or unstable depends on the sign of the right side of the equation. If $c_s^2 k^2 > 4\pi G \rho_0$ the perturbations are oscillating sound waves supported by the internal pressure of the denser regions, but if $c_s^2 k^2 < 4\pi G \rho_0$ the wave is unstable and the modes grow (Longair, 2008). In a static universe these density perturbations grow exponentially but in an expanding universe the growth is slower: e.g. in simple Einstein-de Sitter universe with $\Omega_0 = 1$ and $\Omega_\Lambda = 0$ the growth is only algebraic (Longair, 2008).

The equation 2.28 represents the perturbations well if all matter behaves as non-relativistic fluid so that Newtonian physics suffice to describe the perturbations, the perturbations are assumed to be spatially small compared to the observable universe and the density contrast is smaller than unity (Mo et al., 2010). In the present-day Universe it is clear that many structures with $\Delta \gg 1$ exist so the linear model alone is not sufficient to explain the evolution of structures. In the general case the evolution of these non-linear perturbations cannot be predicted analytically and simulations are often used instead to gain insight into their dynamics (Mo et al., 2010).

One special case in which an analytical solution can be presented is spherical top-hat collapse in which there is a uniform spherical density perturbation with no angular momentum inside a uniform density field (Longair, 2008). The evolution of such a perturbation resembles the evolution of an individual universe with $K = 1$: initially the perturbation expands with the surrounding Universe, but its expansion slows down relative to its surroundings and eventually a sufficiently dense perturbation reaches a point after which it starts to contract (Longair, 2008). As the size of the perturbation as a function of time is a cycloidic function, it is easiest to express in parametric form

$$a_p = \frac{\Omega_0}{2(\Omega_0 - 1)}(1 - \cos\theta) \tag{2.29}$$

$$t = \frac{\Omega_0}{2H_0(\Omega_0 - 1)^{3/2}}(\theta - \sin\theta) \tag{2.30}$$

where a_p is the relative size of the perturbation and t is time (Longair, 2008). Parameter θ evolves from 0 to 2π. From this equation one can see that a_p reaches its maximum at $\theta = \pi$. This maximum size is known as the turnaround radius of the perturbation and the corresponding time as the turnaround time (Mo et al., 2010).

At $\theta = 0$ and $\theta = 2\pi$ the model predicts a radius of zero and therefore an infinite density for the perturbation. Thus it is clear that in addition to the often unrealistic assumption of a perfectly uniform and spherical density perturbation the model has other limitations as well. In reality, smaller perturbations are present within the perturbation to cause the larger perturbation to fragment and the perturbation feels tidal forces from other perturbations that surround it, which apply torques (Longair, 2008). Different structures can also collide and merge.

The evolution of these fragments depends on their composition. Collisionless matter such as cold dark matter will simply experience relaxation and end up as virialized structures, but structures containing baryons have more variation (Mo et al., 2010). In gas, the non-linear evolution of the perturbation is more complicated as the gas feels pressure and shocks can occur when the gas compresses (Mo et al., 2010). Unless the gas is able to radiate away energy by effective cooling, the relaxation ends when the structure is in hydrostatic equilibrium (Mo et al., 2010).

In this work, only dark matter is studied. These collapsed objects made of dark matter are called dark matter haloes (Mo et al., 2010). The exact definition for when an object is dense enough to be considered a halo and where the edge of a halo lies vary somewhat depending on the source, but for halo catalogues analyzed in this thesis a halo is defined to extend to the radius at which the mean density of a spherical volume drops below 200 times the critical density of the Universe. The mass and radius of such a halo are denoted M_{200} and r_{200} where the used definition is ambiguous.

As the cosmological model and its parameters affect the structure formation, analyzing the observations of structures at different stages of their evolution provides a way to compare cosmological models (Mo et al., 2010). Currently the standard model is the so-called ΛCDM model, according to which the Universe consists mostly of dark energy and cold dark matter in ratios given in section 2.1.2, with baryonic matter making up only a small fraction of total mass (Mo et al., 2010). Observations at the scale of the Local Group and its surroundings are not well suited for constraining the nature of dark energy, but the existence of dark matter can already be seen and its properties studied at scales of an individual galaxy (Mo et al., 2010).

For example the mass of possible warm dark matter particle can be greatly restricted (**?**) and hot dark matter made of standard model neutrinos can be excluded as a sole dark matter component based on the fact that non-linear structure has been able to form in the distribution of galaxies by the current age of the Universe (White et al., 1984). This is due to the properties of hot dark matter particles, i.e. low-mass dark matter particles that would have decoupled from the radiation while still relativistic and their thermal motions would allow them to escape from gravity wells and smooth out structures smaller than some tens of Mpc (Mo et al., 2010). The same connection between the particle mass and speed of thermal motions and thus size of smoothed-out structures also sets the lower limit for a warm dark matter particle mass.

2.2 The Local Group and Its Mass

Galaxy groups are systems of galaxies defined by the number of galaxies within a volume. Exact definitions vary, but typically they are required to have at least three galaxies and a volume with numerical overdensity of the order of 20 (Mo et al., 2010). The upper limit of the size of a galaxy group is set by the least massive galaxy clusters: again, the different definitions exist but typically if a group would have over 50 members with apparent magnitudes $m < m_3 + 2$, m_3 denoting the magnitude of the third brightest member, it is classified as a galaxy cluster instead (Mo et al., 2010).

The Local Group, the galaxy group containing the Milky Way, is the best-known of these galaxy groups as it offers a great opportunity for precise observations: much fainter dwarf galaxies can be observed in it than in any other galaxy group and objects subtend large areas on the sky allowing smaller details to be observed than in more distant galaxies (Mo et al., 2010; **?**). This makes it an appealing target for testing astrophysical and cosmological models (**?**). The Local Group has two main members: the Milky Way and the Andromeda Galaxy (M31). In total more than hundred galaxies are known to exist within 3 Mpc of the Sun, more than half of these being satellites for either of the primaries (**?**). As the distance range is quite wide, some of the more distant galaxies are likely not bound to the Local Group, but on the other hand there are likely numerous faint dwarf galaxies that remain unseen (**?**). The number of these unknown galaxies within the Local Group could be as high as several hundred (**?**).

The number of galaxiesis is not the only uncertainty regarding the Local Group: also the total mass of the system and the individual masses of the Milky Way and M31 galaxies are still highly uncertain with different estimates easily differing by a factor of 2–3 (**??**). This is problematic as in addition to constraining dark matter models as mentioned in section 2.1.3, the Local Group has a key role in testing the currently dominant ΛCDM model. The model explains the large-scale structure of the Universe where it is able to make accurate predictions, but at distance scales of a single galaxy group the quality of the predictions is more questionable (**?**).

Two of the possible discrepancies between observations and ΛCDM predictions are the missing satellites problem and the too-big-to-fail problem. The first arises from the fact that both analytical calculations and cosmological simulations produce more low-mass satellites in systems similar to the Milky Way or M31 than can be observed (**?**). To some extent the number of luminous galaxies and dark matter haloes can be expected to differ as stars do not form in all haloes due to reionization and feedback processes such as supernova feedback (**??**). The problem is also somewhat alleviated by the high estimated number of satellite galaxies that are luminous but faint and yet unseen. It is still uncertain whether the missing satellites problem actually exists, but many modern simulations such as (**?**) agree well with the observed number of faint satellite galaxies.

Regardless of the existence of the missing satellites problem, the most massive dark matter haloes in dark matter only simulations also seem to be more numerous than observed galaxies in the the Local Group are. For example the Milky Way has only three satellite galaxies with central densities high enough to allow maximum circular velocities of more than 30 km/s, but dark matter only simulations tend to produce double or even triple this number of similar mass dark matter haloes (**?**). In this case the central masses of the satellites can be reduced as a result of satellites interacting with the primary and e.g. tidal stripping and ram pressure stripping lowering the central densities (**?**). As with the missing satellites problem, the difference between the simulations and observations decreases as baryons and baryonic effects are included in the simulations (**?**).

Both these possible problems are sensitive to the mass of the Local Group: a massive halo is expected to have more subhaloes than a less massive one has. Thus the mass of the Local Group and its members is an intriquing question not only for building up knowledge about our surroundings but also for testing cosmological models. Numerous studies have been conducted to find out its mass, and the following sections outline some of the methods that have been used.

2.2.1 Timing Argument

One possible way of estimating the lower end of the range of possible Local Group masses is to use the timing argument, first introduced by **?**. It is based on the mutual kinematics of the Milky Way and M31 galaxies and the assumption that they have formed close to each other and are now on orbiting their mutual mass centre, approaching each other for the first time. This corresponds to the structure formation model presented in section 2.1.3: initially the expansion of the Universe drives the pair further away from each other, but eventually the mass of the system is able to overcome the expansion and the galaxies start to approach each other.

For a zero angular momentum system, **?** obtain an estimate of minimum total mass of the system using Kepler's third law

$$P^2 = \frac{4\pi}{GM} a \tag{2.31}$$

and the fact that energy is conserved:

$$\frac{GM}{2a} = \frac{GM}{D} - E_k. \tag{2.32}$$

Many of the quantities that appear in the equations are either known or can be approximated: the current distance D between the Milky Way and the center of mass of the system can be estimated using the distance to the M31 galaxy, an upper limit for period P can be obtained using the current age of the Universe and kinetic energy, E_k, only depends on the velocity of the galaxy, easily obtained using radial velocity measurements. Thus what remains to be solved are semimajor axis, a, and the effective mass at the center of gravity, M. The lower limit for the mass of the system derived by **?** was $1.8 \times 10^{12}\ M_{\odot}$. This is clearly less than current estimates that tend to favour masses around $5 \times 10^{12}\ M_{\odot}$ (**??**), but still significantly larger than the observed baryonic mass content of the two galaxies (**?**).

The estimate can be improved by using the Kepler's laws in parametric form:

$$r = a(1 - \cos(E)) \tag{2.33}$$

and

$$t = \left(\frac{a^3}{\mu}\right)^{1/2} (E - \sin(E)) \tag{2.34}$$

where E is the eccentric anomaly and μ is gravitational parameter (**?**). These two equations can be combined to determine the value of $\frac{\mathrm{d}r}{\mathrm{d}t}$ as

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}E}\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}E}\left(\frac{\mathrm{d}t}{\mathrm{d}E}\right)^{-1}. \tag{2.35}$$

This yields the following equation:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \left(\frac{\mu}{a}\right)^{1/2} \frac{\sin(E)}{1 - \cos(E)}. \tag{2.36}$$

Solving for a and $\sqrt{\mu}$ from equations **??** and **??** yields

$$a = \frac{r}{1 - \cos(E)} \tag{2.37}$$

and

$$\sqrt{\mu} = \frac{a^{\frac{3}{2}}}{t}(E - \sin(E)) \tag{2.38}$$

respectively and inserting them into equation **??** gives

$$\frac{vt}{r} = \frac{\sin(E)\,(E - \sin(E))}{(1 - \cos(E))^2}, \tag{2.39}$$

where velocity v and distance r can be deduced from observations and if the Milky Way and M31 are on their first orbit then t is the age of the Universe. Now E can be solved numerically. Inserting equations **??** and the solved value of E into **??**, a value for μ and the combined mass of the pair is acquired.

Getting an estimate is simple, but the results are burdened with the assumptions that are made about the system. Modern values for the velocity and distance yield a virial mass of $(4.23 \pm 0.45) \times 10^{12}\ M_{\odot}$ for the Local Group (**?**). This is a considerably higher mass than calculations of the Milky Way and M31 masses using kinematic tracers of the gravitational field suggest (**?**), but at least some of the effect might be explained by the fact that the timing argument is sensitive to a larger volume of mass compared to satellite galaxies and other kinematic tracers (**?**).

The assumption of the galaxies being on their first approach is likely to be valid as a higher number of completed orbits would result in a mass so high that it does not seem physically realistic. Applying Kepler's laws also requires the masses to be approximated as point masses. If the dark matter haloes of the galaxies are assumed to be spherically symmetrical, this is not a problem if the haloes are sufficiently far from each other, but at times when the center of one halo is encompassed in the other halo, the accuracy of the model is compromised. Indeed the mass estimates acquired using timing argument seem to be best consistent with the mass enclosed in two spheres surrounding the two galaxies with radii of half the distance between the galaxies (**?**).

The assumption of radial orbit is likely fairly good as the estimated tangential velocities are small. For example **?** give a tangential velocity of 17 km/s for M31 with $v_t < 34$ km/s at 1 σ confidence, which is small compared to their radial velocity of -109.3 ± 4.4 km/s. It is also possible to carry out the calculations above for elliptical orbits as is done by **?**. A non-zero tangential velocity increases the resulting mass: for example **?** give a mass of $4.27 \pm 0.53 \times 10^{12}\ M_\odot$ when the tangential velocity is included. Even if external forces felt by the galaxies do not result in significant tangential velocities for the galaxies, they can affect the system in other ways. Large-scale structure surrounding the Local Group can apply forces and torques and smaller members of the Local Group interact with the primary members as was noted by **?**.

2.2.2 Other Mass Estimation Methods

Other structures can also be used to infer the masses of Milky Way and M31 galaxies. For example **?** use Leo I, a Milky Way satellite with high galactocentric velocity, to estimate the mass of the Milky Way by a calculation similar to the timing argument. This is done by assuming that the satellite is bound to the Milky Way which is the only body gravitationally influencing it and that the satellite is on a radial orbit, having passed its periapsis once and now moving away from the Milky Way and towards the apoapsis. This yielded a lower limit of $1.3 \times 10^{12}\ M_{\odot}$ for the mass of the Milky Way (**?**).

Another option is to use the Large and Small Magellanic Clouds, a pair of well-studied Milky Way satellites. For example **?** use circular velocities within the Magellanic clouds and their distances and velocities relative to the Milky Way to estimate its mass. This is done by constructing the probability density functions (PDFs, see section **??**) of these three measurements from a simulation halo catalogue and then using Bayes' theorem to find the Milky Way mass range that is most probable given the observations.

Naturally this type of analysis can suffer from effects not included in the model. If for example the system happens to be peculiar in some sense, its characteristics might cause effects that are not reflected in the PDFs of the selected variables. This is studied by e.g. **?** who measure the effects of local galaxy density and nearby clusters and the fact that the Magellanic Clouds are a fairly close pair and thus rare. They find that subhalo pairs similar to the Magellanic Clouds are rare in haloes resembling the Milky Way and including the subhalo pair criterion in analysis used in **?** brings the estimated mass down considerably, reducing the most probable mass by a factor of two (**?**).

Including more constraints for the sample from which the PDF is determined naturally brings it closer to the one that would represent the Milky Way, but the range of allowed values has to be large enough to cover the uncertainty of measurements from the real system. Tightening the criteria also requires having a larger simulation to find a statistically representative sample of haloes. Naturally the Milky Way system can also be located in either of the tails of the probability distribution, which is true for all mass estimation methods based on statistics instead of physics.

Larger number of objects can also be used in analysis. For example the properties of satellite galaxies can be utilized in mass estimation as is done by e.g. **?** who use the velocity dispersion of the satellites as an indicator of the virial velocity of the host halo and **?** who use the ellipticity distribution of Milky Way satellites to constrain the range of likely masses. Other useful objects include individual high-velocity stars from which the galactic escape speed can be estimated (**?**) and the orbits of stellar streams (**?**) among others.

The mass can also be determined based on a larger volume than is occupied by the two main galaxies of the Local Group and their satellites: for example **?** and **?** estimate the mass of the Local Group by stydying the local Hubble flow. This is possible as the gravitational interaction of the surrounding galaxies with the Milky Way and M31 galaxies cause the surrounding galaxies to recede slower than ones surrounding empty regions of the Universe, and the strength of this effect can be translated to a mass estimate using e.g. Bayesian analysis. The velocity dispersion around the Hubble flow can also be used as is done by for example **?**. The properties of the Hubble flow are also used in this thesis as one estimator of the Local Group mass.

Different papers present varying values for the masses of the Local Group and its galaxies, both due to the strengths and limitations of the methods used and differences in the used observed and simulated data. A collection of mass estimates for Milky Way is shown in figure **??** by **?**. The masses are calculated by different authors using various mass estimation methods, some of which have been briefly

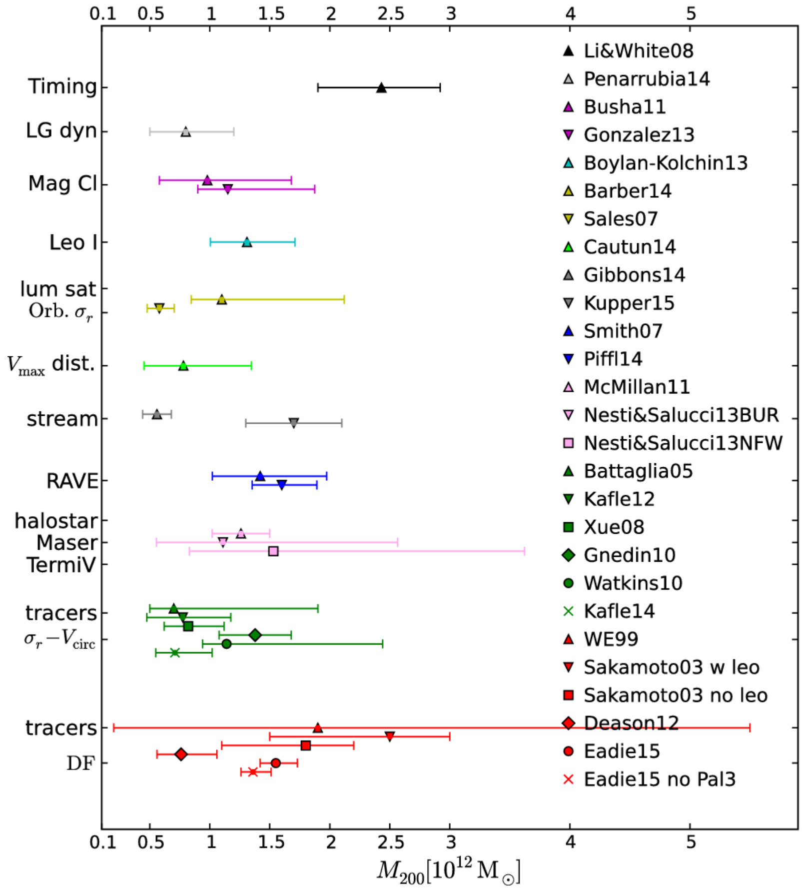

Figure 2.1: M_{200} masses of the Milky Way as obtained by different authors using various mass estimation methods, collected and plotted by **?**. Error bars show the 1σ range where assuming Gaussian distribution was approppriate. Masses derived using similar methods are plotted in same colour.

introduced in this section. The plot clearly illustrates the fact that currently the masses of the Local Group galaxies have errors of over a factor of two.

Especially the timing argument stands out from the other methods. This is in agreement with a study of **?** where timing argument masses were found to overestimate the true mass of a Local Group like system with small tangential velocity and low local overdensity by a mean factor of 1.6. This is due to the timing argument not being sensitive to the tangential velocity or environment of the system.

The other mass estimates also have considerable scatter, sometimes even within the same mass estimation method. At least some of this scatter can be a result of some estimates being more than ten years old and thus their methods and data possibly outdated, but the scatter also reflects the fact that the exact mass of the Local Group is currently a question without a definitive answer. Thus exploring new ways of estimating the mass and refining the old results is important.

3. Simulations and Simulation Codes

Simulations are a valuable tool in astrophysical research as they bypass many major restrictions characteristic to observational astronomy such as not being able to observe the evolution of a single object with the exception of the most rapid events such as supernovae, and the observations being constrained to a single view plane. The data used in this master's thesis also originates from a cosmological N-body simulation. This section shortly introduces first some fundamental operational principles of N-body simulations and then specifics of the [insert name or other identification here] simulations from which the data used here originates from.

3.1 N-body simulations

N-body simulations are a type of computer simulations that follow a number of particles interacting with each other, often used in computational astrophysics (Binney and Tremaine, 2008). Their concept is simple: the current states of the particles are known and a physical model is used to calculate how the states of the particles should advance over a small period of time known as the time step (Binney and Tremaine, 2008). Simple dark matter only simulations might only handle gravitational interactions of the particles, but many simulation codes such as GADGET-2 (Springel, 2005) (see section 3.2.1 for more about GADGET family simulation codes) and Enzo (Norman et al., 2007) are also able to handle hydrodynamics of baryonic matter and can include star formation, feedback and other baryonic processes.

There are multiple ways to handle both the force calculations and calculating new positions for the particles (Binney and Tremaine, 2008). For the force calculations most popular algorithms are based on either hierarchical trees or particle meshes, but for calculating the positions, a wide variety of integrators have been developed for a range of different needs (Binney and Tremaine, 2008). The [insert identification here] simulations are run using a modified GADGET-3 and thus use the TreePM method for force calculations accompanied by a leapfrog integrator. TreePM is a mix of a hierarchical tree for close range forces and a particle mesh for distant forces, the former of which is more interesting. Therefore the Barnes and Hut (1986) hierarchical tree algorithm is introduced in the following subsection, followed by a short description of the leapfrog integrator in the next one.

3.1.1 Hierarchical Tree Algorithm

In many applications of computational astrophysics the desired number of particles in a simulation is too great to allow calculating interparticle forces by direct summation as its time complexity is $\mathcal{O}(n^2)$. One of the alternatives is organizing the particles into a tree data structure, which allows distant particles residing close to each other to be approximated as single more massive particle. This approach was first introduced by Appel (1985) and Barnes and Hut (1986), the latter of which will be followed here.

Constructing the tree starts with setting the full simulation box as the root of the tree. This root cube is then subdivided into eight equally sized sub-cubes called cells. These eight cells are children of the root node. This starts a recursive process where each new cell is again divided into eight subcells recursively until each of the cells contains either no particles, one particle or eight subcells. When a new cell is

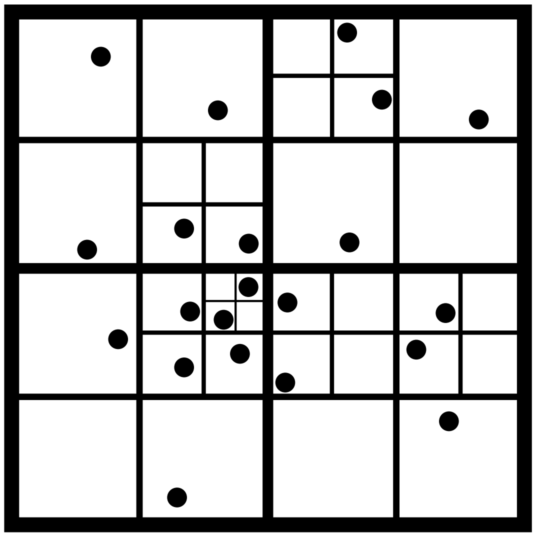

Figure 3.1: A two-dimensional example of particles (black dots) being split into cells using the Barnes-Hut algorithm (Barnes and Hut, 1986). The outermost square is the simulation box, i.e. the root of the tree, and the samaller squares with decreasing line thicknesses are its descendants. Note how each cell contains either one particle, no particles or four smaller subcells.

created, the total mass of the enclosed particles and the location of the mass centre of the particles within it is stored as a pseudoparticle to allow easy calculation of the approximate gravitational effect the particles within a cell have on a distant particle.

To aid understanding the process, a two-dimensional simplification of the simulation box after the tree has been constructed is shown in figure 3.1, together with the corresponding tree in figure 3.2. The thick outer line of the former figure is the simulation box, corresponding to the topmost node of the tree in the latter. This root cell is first divided into four (in contrast to eight in the three-dimensional case) subcells, which is shown with the second-thickest line dividing the simulation box into quarters in figure 3.1. These four cells are the four children of the root node in the tree. Each of the subcells contains more than one particle, so each of those is again split into four quarters, each quarter a child of the original node in the tree.

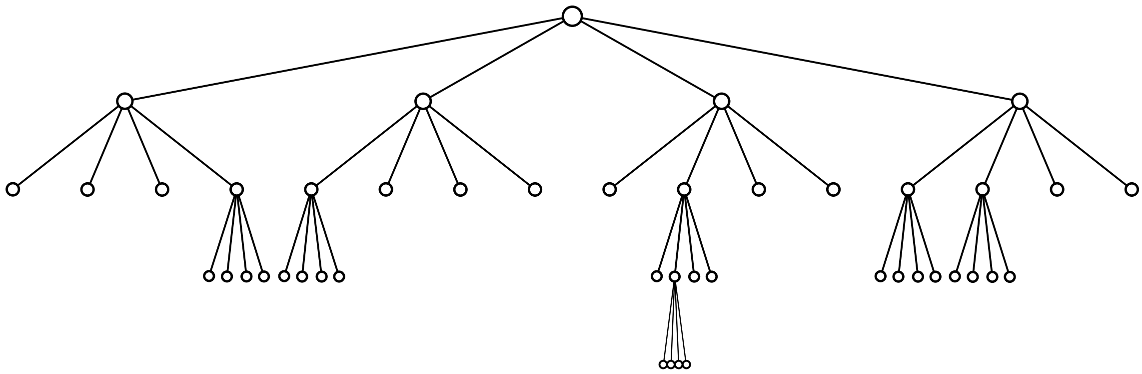

Figure 3.2: Cells of figure 3.1 shown as a tree, each level of refined cells in 3.1 traversed from left to right and up to down corresponding to each level of nodes from left to right. Cells with no particles are not used in force calculations and thus they do not need to be stored in computer memory, but they are shown here to emphasize the structure of the tree.

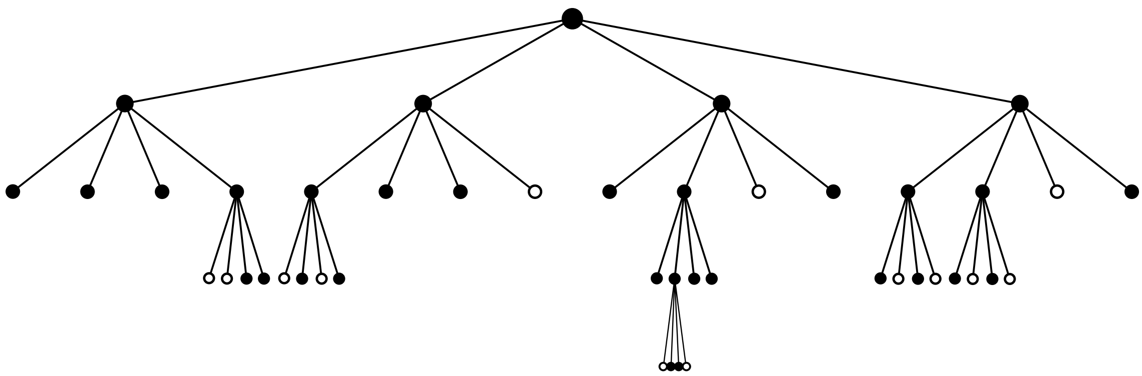

Figure 3.3: Alternative tree where cells with particles are solid, doesn't work as well as it could

These quarters of quarters form the third level of the tree in figure 3.2.

Some of these newly-created cells are empty or only have a single particle, so for them the recursion halts and the nodes of the tree are leaves. The rest are again split and a new layer of the tree is created, but as some cells were complete already the fourth layer of the tree is not full. In this case only one cell requires division beyond the fourth level, producing the last four leaves of the tree on the fifth level.

Often, as is the case in the example, some of the leaf nodes are empty. In the case of a real simulation, these of course do not need to be saved as there is no information to store, but in this example case all are drawn to emphasize the regular structure of the tree. In the three-dimensional case where each internal cell has eight subcells the formed tree is known as an octree, whose two-dimensional analog, the quadtree, is constructed in the previous example.

The tree can then be utilized to speed up the force calculations. When calculating the gravitational acceleration felt by a single particle, the tree is traversed starting from the root. The ratio l/D of the length of a side of the cell and the distance from the particle to the pseudoparticle representing the mass within the cell is then calculated. If the ratio is smaller than a predefined accuracy parameter θ representing the opening angle of the cell, the cell is treated as if everything within it was replaced with the pseudoparticle having the combined mass of the cell. Otherwise the subcells of the cell are examined recursively in the same way until a small enough subcell is found or a leaf of the tree is reached, at which point the pseudoparticle and the real particle in the cell are equivalent and the force can be calculated with the maximum accuracy possible for the simulation.

Using the Barnes-Hut algorithm, the tree construction and the force calculations both have time complexity of $\mathcal{O}(n \log n)$. This is a significant improvement over the $\mathcal{O}(n^2)$ of the direct summation considering that the accuracy cost is fairly small: Barnes and Hut (1986) report accuracy of about 1 % when $\theta = 1$ and the accuracy can be improved by either setting a smaller θ or including multipole moments in the pseudoparticles (Barnes and Hut, 1989). Similar algorithms with even better time complexity of $\mathcal{O}(n)$ have also been introduced such as ones by Dehnen (2002) and Xue (1998).

The algorithm is also straightforward to parallelize as different branches of the tree can be assigned to their own threads, though memory management has to be done carefully when forces outside the current branch of a thread are calculated

(Binney and Tremaine, 2008). One way to circumvent the memory management issue is to use a particle mesh based integrator for long range forces as GADGET-2 does (Springel, 2005). The algorithm can also handle problems where the range of densities in the simulation box is wide which is important for applications such as galaxy mergers and cosmological simulations. This, together with their competitive time complexity, makes tree based codes an appealing tool for many astrophysical simulations (Binney and Tremaine, 2008).

3.1.2 Leapfrog Integrator

A number of different integrators suitable for astronomical and astrophysical applications have been developed for solving different problems (Binney and Tremaine, 2008). No integrator is optimal for every task and thus factors such as the integration time, amount of memory available per particle, smoothness of the potential and the cost of a single gravitational field evaluation should be considered (Binney and Tremaine, 2008). One of the integrators that are well suited for cosmological simulations is the leapfrog integrator, which is also used by GADGET-2 simulation code among others (Springel, 2005).

When a fixed time-step is used, the leapfrog integrator conserves the energy of the system and is time reversible (Binney and Tremaine, 2008). While a variable time-step is possible and often used, it requires some modifications to the algorithm, presented in e.g. Springel (2005). Other benefits of the integrator are its second order accuracy and the fact that it doesn't require excessive amounts of memory per particle as only the current state of the system is needed in calculating the next step(Binney and Tremaine, 2008). With its second order accuracy paired with symplecticity, it rivals fourth order integrators such as the fourth order Runge-Kutta when used for long simulation runs (Binney and Tremaine, 2008).

Timestepping with the leapfrog integrator consists of two phases, the drift and

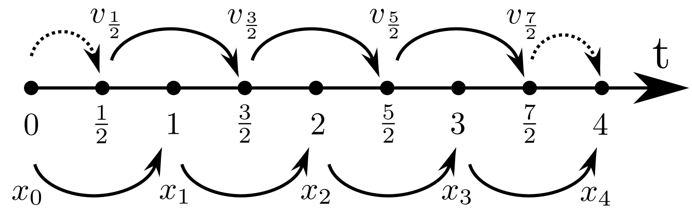

Figure 3.4: Timesteps taken by the leapfrog algorithm, positions (x) updated as indicated with lower arrows and velocities (v) as the upper arrows. It is not important for the algorithm which of the two is chosen to step through the integer times, choice of it being x in this figure is arbitrary. The dashed short arrows depict the half-steps that are needed when a synchronized output is desired.

the kick steps, which are alternated with a half-step offset as shown in figure 3.4 (Binney and Tremaine, 2008). During the kick step, the momenta of the particles are updated and during the drift step the positions of the particles are changed according to the previously calculated momenta (Binney and Tremaine, 2008). When synchronized output of both positions and velocities is desired, a half-timestep advance or backtrack to either of the variables will result in both variables being outputted at the same time. This kind of synchronization steps at the ends of the integration are indicated in figure 3.4 with dashed arrows.

3.1.3 Halo Finding

The data output by a cosmological simulation consists of individual particles tracing the underlying density field. To make comparisons to the real universe, a way of matching structures in the simulation to observable objects is needed. In a dark matter only simulation no structure would obviously be directly observable but luckily many properties of dark matter haloes from simulations can be compared with the estimated dark matter haloes of observed galaxies. Making such comparisons requires naturally such structures to be identified first, which makes structure finding a key step in the data analysis for many astrophysical and cosmological applications

(Knebe et al., 2013).

A number of different halo finders designed to read N-body data and extract locally overdense gravitationally bound systems have been developed to suit different needs, most based either on locating density peaks or collecting and linking together particles located close to each other based on some metric (Knebe et al., 2013). A pair of methods belonging to the two groups respectively, the friends-of-friends (FOF) and SUBFIND algorithms, are discussed here. Both algorithms were developed separately, FOF by Davis et al. (1985) and SUBFIND by Springel et al. (2001), but they work well when used together by inputting FOF results to SUBFIND as a starting point for the subhalo finding (Springel, 2005).

The FOF algorithm is simple and based purely on the spatial separation of the particles: pairs of particles residing closer to each other than a chosen threshold distance called linking length are marked to reside within same group by linking them together (Davis et al., 1985). When all particles have been processed, each distinct subset of particles linked to each other is defined to be a group (Davis et al., 1985). Figure 3.5 presents an example of a set of particles grouped using the FOF algorithm. Depending on the specifics of the data and its intended usage, one might want to discard the smallest groups with only a few particles both because they are more likely than bigger groups to be just realizations of the random noise instead of actual physical structures but also due to their small size as they might not be massive enough to represent the structures that are being studied.

The algorithm is made appealing by its simplicity, small number of free parameters and ability to find structures of arbitrary shape (Davis et al., 1985). However, it is prone to link unrelated structures via thin bridges that might consist of only a single chain of particles. This behaviour can be seen in figure 3.5: the leftmost group has two dense areas that could be cores of two separate groups connected by a single-particle bridge. Removal or even suitable movement of this particle

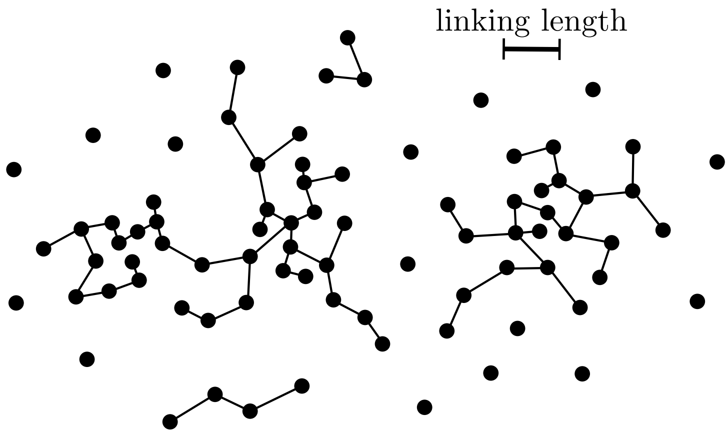

Figure 3.5: Friends-of-friends groups found from a mock data set using the length of the indicator in the upper right area of the illustration as the linking length. Particles are depicted as black dots and the links connecting particles within groups are shown with black lines.

would result in the group separating into two distinct groups. The two big groups also stretch quite close to each other in their lower parts: moving or adding one particle between the chains of particles protruding towards each other would result in the two groups merging. Also the algorithm as is cannot be used to detect substructure within larger objects, though modifications of it such as hierarchical friends-of-friends algorithm by Gottlöber et al. (1999) have been developed Springel et al. (2001).

In contrast to groups found by FOF, the SUBFIND algorithm has been developed to extract physically well-defined subhaloes that are self-bound and locally overdense from a given parent group. SUBFIND can work with an arbitrary parent group, but FOF groups are well-suited for parent groups as the algorithm is simple and with appropriately long linking lengths the FOF groups, while possible to consist of multiple independent physical structures, are unlikely to split a physical structure between FOF groups (Springel et al., 2001).

Unlike FOF, SUBFIND uses a local density estimate instead of individual particle pairs. It labels all locally overdense regions enclosed by an isodensity contour traversing a saddle point as substructure candidates (Springel et al., 2001). This is done by lowering an imaginary density threshold through the range of the density field: particles surrounding a common local density maximum are assigned to a common substructure candidate until two separate substructure regions join in a saddle point of the potential (Springel et al., 2001). When a saddle point is reached the two substructure candidates it connects are both stored individually to be processed further and the saddle point particle is added to a new substructure candidate containing the particles from both of the smaller candidates (Springel et al., 2001). Thus the algorithm is able to identify a hierarchy of substructures within each other (Springel et al., 2001).

A two-dimensional example of this substructure candidate identification is shown in figure 3.6. The algorithm starts from the particle in the most dense area of the simulation. At that point, the particle has no neighbours that would be in higher density than the particle itself, and thus it becomes the first particle of a substructure candidate. All of the particles belonging to this substructure candidate are marked with dashing in the figure. The algorithm iterates through the particles in order of decreasing density, always finding that the next particle has one neighbouring particle in higher density area than the particle itself and thus adding it to the same dashed group, until the second local density maximum is reached. At that point, a new substructure candidate marked with checkerboard pattern is created. The following particles are assigned to their respective substructure candidates based on their single higher potential neighbour until a saddle point between the two is reached, at which point the state of the substructure candidates is shown in figure 3.6. At that point the current substructure candidates are complete and will be saved. Next the particles belonging to the two structures can be joined by the saddle point particle to join a new bigger substructure candidate.

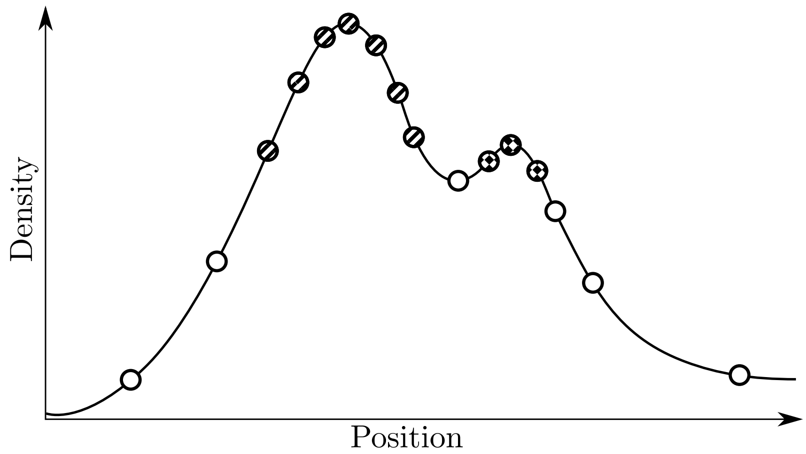

Figure 3.6: Intermediate stage of the SUBFIND algorithm, shown just before it reaches the first saddle point. Circles depict simulation particles and the line the underlying density field. Striped and checkered circles both belong to their own subhalo candidates whereas the white ones are not yet labeled.

Unfortunately, now some particles are assigned to multiple substructure candidates and it is not clear that all particles within one substructure candidate are actually part of an actual physical structure (Springel et al., 2001). Is is very much possible that some particles are just passing by and if the same particles were reexamined at a later time, they would no longer be anywhere near the structure they were supposed to belong to. Hence the next step in the analysis is to eliminate unbound particles by iteratively removing particles with positive energy until all of the remaining particles are bound to each other by their mutual gravitational attraction (Springel et al., 2001). At this stage, each particle is labelled only based on the smallest structure it resides in, which solves the problem of a single particle belonging to multiple structures (Springel et al., 2001).

After the iterative pruning stage some substructure candidates can vanish completely or be left with very few members. These substructure candidates with less than some minimum number of bound particles can be discarded (Springel et al., 2001). The structures candidates surviving the pruning can then be considered to represent physical structures and are labelled as subhaloes (Springel et al., 2001). In this work, all of the analysis is based on catalogues containing such subhaloes.

3.2 Simulation Runs

[insert some form of attribution for the simulations, e.g. some kind of ”in preparation” citation?]. The simulations are cosmological zoom-in simulations, meaning that interesting regions at $z = 0$ were identified from an output from a low-resolution simulation, after which resolution of the interesting volumes was increased and the simulation was run again. In this case, the interesting regions were defined as surroundings of a pair of dark matter haloes resembling ones of the Milky Way and Andromeda galaxies, the two main galaxies of the Local Group.

The simulations contained only dark matter, and for my analysis only the dark matter halo information was needed, so instead of using the full simulation output I conducted my analysis on subhalo catalogues. The subhalo information was extracted as described in section 3.1.3. The resulting data set consists of subhalo catalogues from 448 zooms, each centered on one region resembling the Local Group in the low-resolution simulation. The following sections shortly introduce the code used to run the simulation and the parameters of the simulation for both stages of the zoom-in simulations.

3.2.1 Modified GADGET-3

The simulation code that was used to run the simulations on which all of the analysis in this master’s thesis is based on was a modified version of GADGET-3, an earlier version of which is described in Springel (2005). The code is same as used in the EAGLE project, so detailed descriptions of the changes can be found in Schaye et al. (2015). However, the changes mostly affect the handling of baryonic matter so understanding the basic GADGET-2 gives a good basis for understanding the simulations (Schaye et al., 2015).

GADGET uses a TreePM algorithm to compute forces, meaning that short-range forces are calculated using a tree method as described in section 3.1.1 and a particle mesh is employed for long-range forces (Springel, 2005). As the code is parallel, normal tree-construction algorithm is problematic in regards to splitting the nodes of the octree between processors (Springel, 2005). To ensure a balanced workload amongst all processors, the particles are split between processors by constructing a space-filling fractal curve known as Peano-Hilbert curve, and splitting it to segments with approximately equal number of particles on each segment (Springel, 2005). The properties of the curve ensure that particles close to each other along the curve are also near each other in the 3D space, which means that close-range forces can frequently be calculated without need to access memory belonging to other processors (Springel, 2005). In regions where an octree constructed from particles belonging to a processor should contain particles assigned to other processors, a pseudoparticle resembling in principle the pseudoparticles used when calculating forces for cells where $l/D < \theta$ is inserted instead of the full particle information (Springel, 2005).

Updating the positions of the particles is done using an integrator resembling the leapfrog integrator described in section 3.1.2, but the integrator is modified to allow using variable time-step lengths (Springel, 2005). These modifications are important, as in cosmological simulations it is not sensible to use the same time step in all parts of the simulation as there are both high-density regions and sparse void areas, first of which requiring a time-step so small that a lot of computational time is wasted while integrating the latter with more detail than is needed.

TODO: wrap up, maybe by providing the reasoning behind choosing gadget

Box size	800 Mpc/h
Number of particles	0
Particle mass	0
Softening length	0

Table 3.1: Properties of the parent simulation.

H	67.77 km/s/Mpc
Ω_m	0.307
Ω_Λ	0.693

Table 3.2: Cosmological parameters used in the simulation. TODO: planck + lensing 2013

3.2.2 Parent simulation

First step of a zoom-in simulation running process is to construct initial conditions and run a low-resolution box. A large box with 800 Mpc/h was chosen to ensure that the box is big enough to contain a reasonable sample of Local Group analogues and that the large scale structure of the universe is represented in the simulation.

TODO: simulation properties, both cosmology and resolution related, starting resolution

When the low-resolution box reached redshift $z = 0$, haloes were identified using the procedure described in section 3.1.3. From the resulting halo catalogue, halo pairs resembling the Local Group were identified. TODO: the criteria used

3.2.3 Zoom simulations

	min	mean	median	max
High resolution particles	min	mean	median	max
High resolution particle mass	min	mean	median	max
High resolution softening length	min	mean	median	max
Total particles	min	mean	median	max

Table 3.3: Properties of the zoom simulations.

4. Mathematical and statistical methods

täällä tarvittavat esitiedot ja önnönnöö, listaa mm. mitä aiot kertoa kunhan tiedät itsekään

4.1 Statistical Background

Precision of the used equipment limits accuracy of all data gathered from physical experiments, simulations or observations. Therefore the results are affected by the measurement process and the results have to be presented as estimates with some error, magnitude of which is affected by both number of data points and accuracy of the measurement equipment (Bohm and Zech, 2010).

Estimating errors for measured quantities offers a way to test hypotheses and compare different experiments (Bohm and Zech, 2010). This is done using different statistical methods, of which the main methods relevant for this thesis are covered here. The methods are shortly introduced in the following sections together with basic statistical concepts that are necessary to understand the methods.

4.1.1 Hypothesis testing and p-values

A common situation in scientific research is that one has to compare a sample of data points to either a model or another sample in order to derive a conclusion from the dataset. In statistics, this is known as hypothesis testing. For example, this can mean testing hypotheses such as ”these two variables are not correlated” or ”this sample is from a population with a mean of 1.0” (J. V. Wall, 2003). Next paragraphs shortly introduce the basic concept of hypothesis testing and methods that can used to test the hypothesis ”these two samples are drawn from the same distribution” following the approach of (Bohm and Zech, 2010) and (J. V. Wall, 2003).

The process of hypothesis testing as described by begins with forming of a null hypothesis H_0 that is formatted such that the aim for the next steps is to either reject it or deduce that it cannot be rejected with a chosen significance level. Negation of the null hypothesis is often called research hypothesis or alternative hypothesis and denoted as H_1. For example, this can lead to H_0 ”this dataset is sampled from a normal distribution” and H_1 ”this dataset is not sampled from a normal distribution”. Choosing the hypothesis in this manner is done because often the research hypothesis is difficult to define otherwise.

After setting the hypothesis one must choose an appropriate test statistic. Ideally this is chosen such that the difference between cases H_0 and H_1 is as large as possible. Then one must choose the significance level α which corresponds to the probability of rejecting H_0 in the case where H_0 actually is true. This fixes the critical region i.e. the values of test statistic that lead to the rejection of the H_0.

This kind of probability based decision making is always prone to error. It is easy to see that α corresponds to the chance of H_0 being rejected when it is true. This is known as error of the first kind. However, this is not the only kind of error possible. It might also occur that H_0 is false but it does not get rejected, which is known as error of the second kind.

There is no one optimal way of choosing α, but instead one should try to find a balance between false rejections of null hypothesis and not being able to reject null hypothesis based on the dataset even if in reality it might not be true. When sample size (often denoted N) is large, smaller values of α can often be used as decisions get more accurate when N grows. For example tässä työssä α oli jokin ja N jotain muuta.

It is crucial not to look at the test results before choosing α in order to avoid intentional or unintentional fiddling with the data or changing the criterion of acceptance or rejectance to give desired results. Only after these steps should the test statistic be calculated. If the test statistic falls within the critical region, H_0 should be rejected and otherwise stated that H_0 cannot be rejected at this significance level. The critical values for different test statistics are widely found in statistical textbooks and collections of statistical tables or they can be calculated using statistical or scientific libraries available for many programming languages.

Despite statistical tests having a binary outcome "H_0 rejected" or "H_0 not rejected", a continuous output is often desired. This is what p-values are used for. The name p-value hints towards probability, but despite its name p-value is not equal to the probability that the null hypothesis is true. These p-values are functions of a test statistic and the p-value for a certain value t_{obs} of a test statistic gives the probability that under the condition that H_0 is true, the value of a test statistics for a randomly drawn sample is at least as extreme as t_{obs}. Therefore if p-value is smaller than α, H_0 is to be rejected.

4.1.2 Distribution functions

Some statistical tests such as the Kolmogorov-Smirnov test and the Anderson-Darling test make use of distribution functions such as cumulative density function (CDF) and empirical distribution function (EDF) in determining the distribution from which a sample is drawn.

To understand CDF and EDF, one must first be familiar with probability density function (PDF). As the name suggests, PDF is a function the value of which at some point x represents the likelihood that the value of the random variable would equal x. This is often denoted $f(x)$. Naturally for continuous functions the probability of drawing any single value from the distribution is zero, so these values should be interpreted as depicting relative likelihoods of different values. For example if $f(a) = 0.3$ and $f(b) = 0.6$ we can say that drawing value b is twice as likely as drawing value a. (Heino et al., 2012)

Another way to use the PDF is to integrate it over semi-closed interval from negative infinity to some value a to obtain the CDF, often denoted with $F(x)$:

$$F(x) = \int_{-\infty}^{x} f(x')\,dx'. \tag{4.1}$$

This gives the probability of a random value drawn from the distribution having value that is smaller than x. Relation between the PDF and the CDF is illustrated in figure 4.1, where PDFs and CDFs are shown for three different distributions. It is easy to see the integral relation between PDF and CDF and how wider distributions have wider CDFs. (Heino et al., 2012)

Both the PDF and the CDF apply to whole population or the set of all possible outcomes of a measurement. In reality the sample is almost always smaller than this. Therefore one cannot measure the actual CDF. Nevertheless, it is possible to calculate a similar measure of how big a fraction of measurements falls under a given value. This empirical counterpart of the CDF is known as empirical distribution function (EDF), often denoted $\hat{F}(x)$, and for a dataset $X_1, X_2, ..., X_n$ containing n samples it is defined to be

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x] \tag{4.2}$$

where I is the indicator function, value of which is 1 if the condition in brackets is

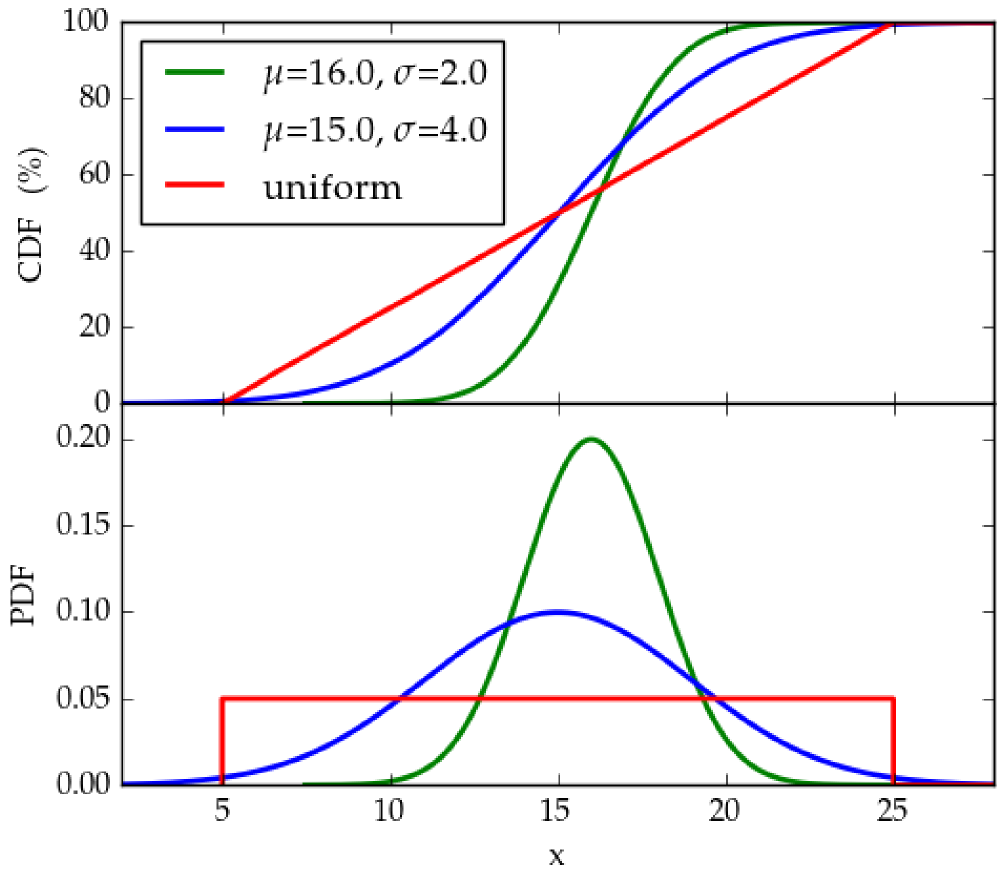

Figure 4.1: Cumulative distribution function (top panel) for three random samples (PDFs shown in the bottom panel) drawn from different distributions, two of which are normal and one is uniform. Parameters μ and σ of the normal distribution describe the mean and the spread of the distribution respectively, large values of σ resulting in wide distribution.

true, otherwise 0. (Feigelson and Babu, 2012)

Due to the EDF being a result of random sampling, it may deviate from the underlying CDF considerably as can be seen by comparing CDFs in figure 4.1 and corresponding EDFs in figure 4.2. This example is somewhat exaggerated with its N=35 as the actual dataset used in this thesis has N>100, but reducing the sample size makes seeing the effects of random sampling easier. The latter figure also has EDFs corresponding to two random samples drawn from the distribution of the

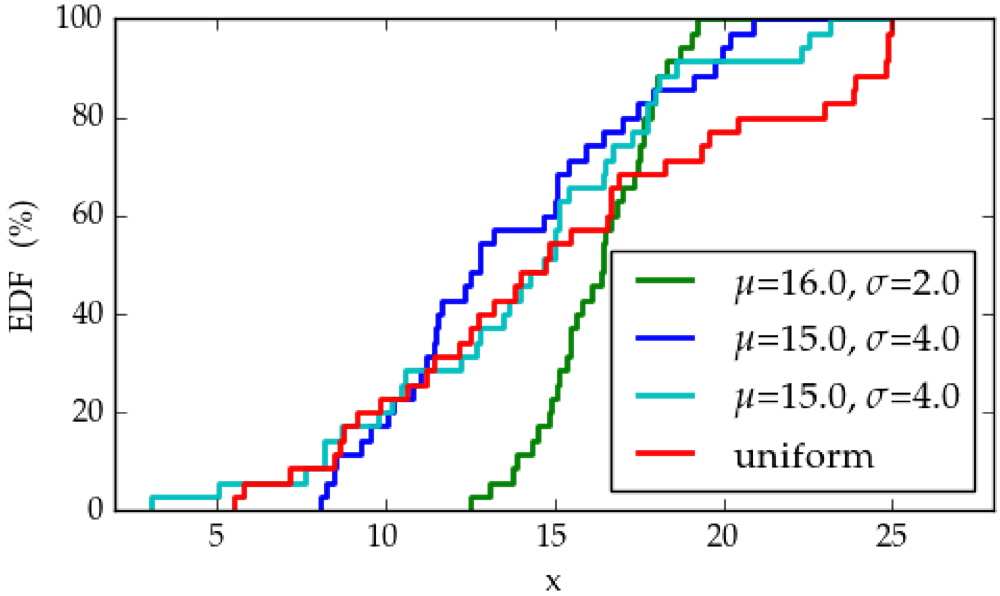

Figure 4.2: Empirical distribution function for four random samples (N=35) drawn from the same distributions as in figure 4.1. Note that both the blue and the cyan data are drawn from the same distribution.

green curve in the first figure to further illustrate the differences that can arise from random sampling. This randomness also makes determining whether two samples are drawn from the same distribution difficult.

4.2 Linear Regression

Regression analysis is a set of statistical analysis processes that are used to estimate functional relationships between a response variable (denoted with y) and one or more predictor variables (denoted with x in case of single predictor or $x_1 \ldots x_i$ if there are multiple predictor variables) (Feigelson and Babu, 2012). In this section, we will cover both simple regression where there is only one response variable and multiple linear regression where there are more than one response variables. The models also contain ε term that represents the scatter of measured points around the fit. One of the models used is linear regression model, which can be used to fit any relationship where the response variable is a linear function of the model parameters (Montgomery, 2012). In addition to the widely known and used models where the relationship is a straight line, such as

$$y = \beta_0 x + \varepsilon \tag{4.3}$$

all models where relationship is linear in unknown parameters β_i are linear (Montgomery, 2012). Thus for example the following are linear models

$$y = \beta_0 x^2 + \varepsilon \tag{4.4}$$

$$y = \beta_0 e^x + \beta_1 \tan x + \varepsilon \tag{4.5}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4.6}$$

On the other hand, all models where the relationship is not linear and therefore

$$y = x_0^\beta + \varepsilon \tag{4.7}$$

$$y = \beta_0 x + \cos(\beta_1 x) + \varepsilon \tag{4.8}$$

are nonlinear.

4.2.1 Simple linear regression

Simple linear regression is a model with a single predictor variable and a single response variable with a straight line relationship, i.e.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4.9}$$

where parameter β_0 represents the y axis intercept of the line and β_1 is the slope of the line (Montgomery, 2012). The parameters can be estimated using method of least squares, where such values are found for the parameters that the sum of squared differences between the data points and the fitted line is minimized (Montgomery, 2012).

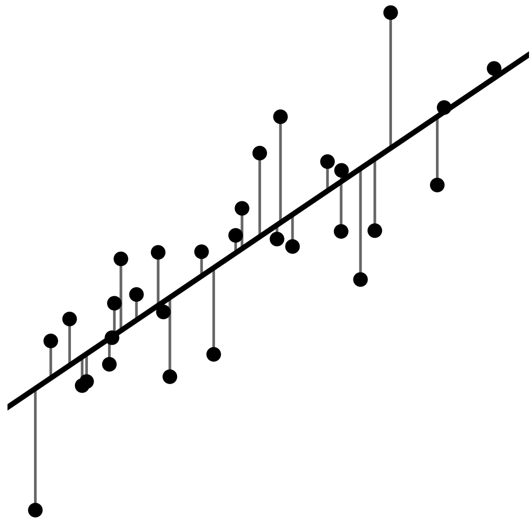

Figure 4.3: TODO: caption

The best-known method of minimizing the sum of squared error is the ordinary least-squares (OLS) estimator. The OLS method uses distances measured vertically as shown in figure 4.3 and thus the minimized sum is

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \tag{4.10}$$

where x_1 and y_i are single values of the measured quantities (Feigelson and Babu, 2012). This approach requires that the values of the predictor variable are known exactly without error and all uncertainty is in the values of the response variable (Feigelson and Babu, 2012). In those situations where this assumption is not valid, results acquired using OLS may be counterintuitive. This can be seen for example in figure 4.4 where OLS is used to calculate two linear fits: one where x is used and predictor variable and y as response variable and another where y is the predictor and x the response.

When dividing the variables to the independent variable with no error and a response variable with possible measurement error is not a justifiable choice OLS should not be used. One alternative for OLS is total least squares (TLS, also known as orthogonal least squares in some sources such as (Feigelson and Babu, 2012))

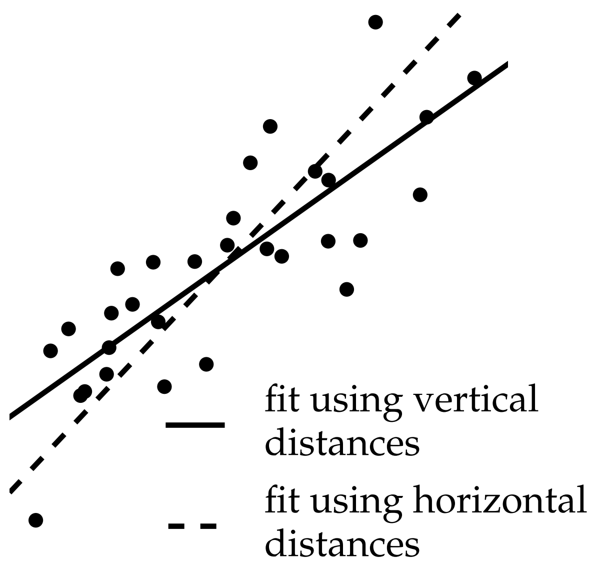

Figure 4.4: TODO: caption

regression can be used instead of OLS (Markovsky and Huffel, 2007). The major difference between OLS and TLS is that instead of vertical distance, the minimized squared distance is measured between a point and its projection to the fitted line, thus providing minimum of the sum of the squared orthogonal distances from the line (Feigelson and Babu, 2012). These minimized distances are shown in figure 4.5.

4.2.2 Multiple linear regression

gradun sovellus: onko PCR relevantti?

ongelman kuvailu, esim OLS:lle yleistys, jälleen liittyy PCA

4.3 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure first introdueced by Pearson (1901) to aid physical, statistical and biological investigations where fitting a line or a plane to n-dimensional dataset is desired. When performing PCA, one transforms a data set to new set of uncorrelated variables i.e. ones represented by orthogonal basis vectors. These variables are called principal components (PCs)

Figure 4.5

(Jolliffe, 2002). This approach also solves the problem of sometimes arbitrary choice of division of the data to dependent and independent variables introduced in section 4.2.something Pearson (1901).

PCA can be used to both reduce and interpret data (Johnson, 2007). Often PCA alone does not produce the desired result, but instead PCs are used as a starting point for other analysis methods such as factor analysis or multiple regression (Johnson, 2007). These applications are introduced in the following subsections together with a short description of performing PCA and interpreting its results. In addition to these applications, PCA is also used in image compression, face recognition and other fields (Smith, 2002).

4.3.1 Extracting Principal Components

In order to understand the process of obtaining principal components of a data set let us follow the procedure on a two-dimensional data set shown in the top panel of figure 4.6 with black dots. First step of finding the PCs is to locate the centroid of the dataset i.e. the mean of the data along every axis (Smith, 2002). This is marked

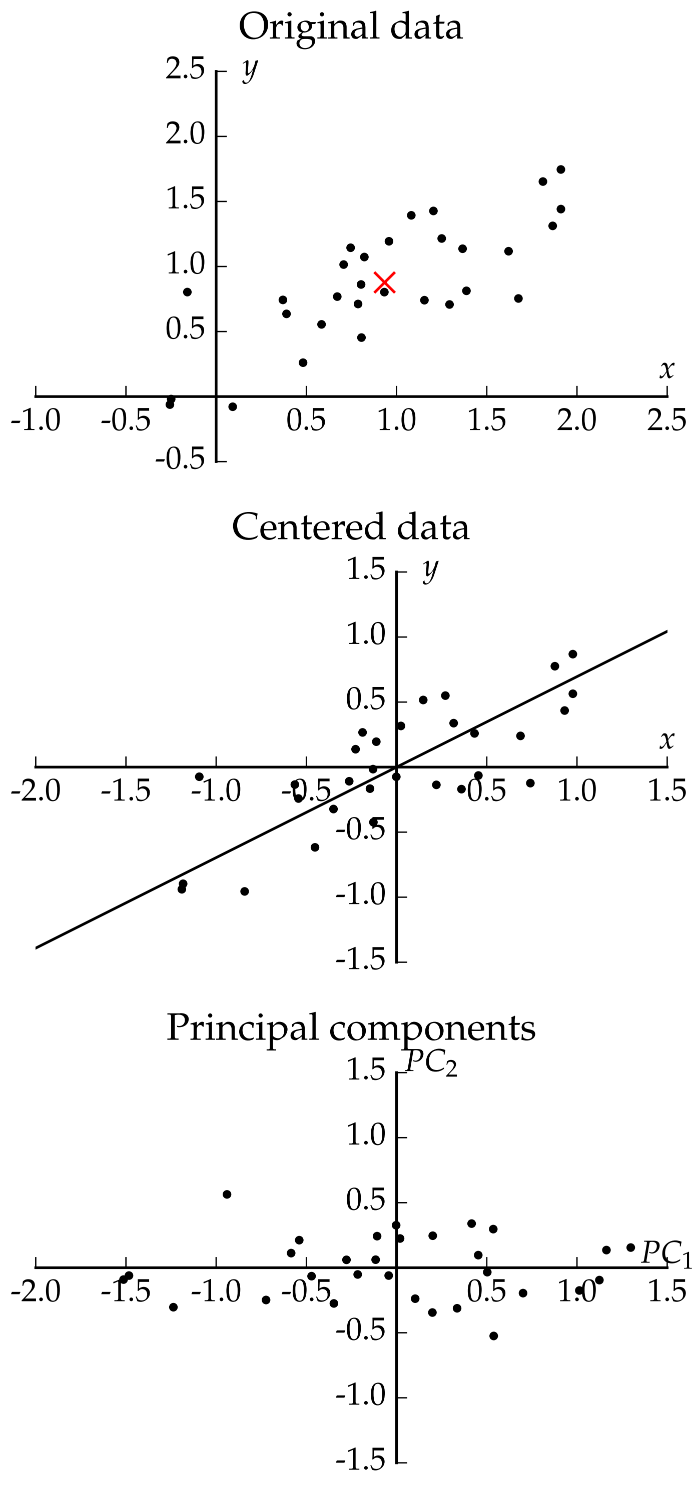
Original data
y
x
2.5
2.0
1.5
1.0
0.5
-0.5
-1.0
-0.5
0.5
1.0
1.5
2.0
2.5
Centered data
y
x
1.5
1.0
0.5
-0.5
-1.0
-1.5
-2.0
-1.5
-1.0
-0.5
0.5
1.0
1.5
Principal components
PC_2
PC_1
1.5
1.0
0.5
-0.5
-1.0
-1.5
-2.0
-1.5
-1.0
-0.5
0.5
1.0
1.5

Figure 4.6: .

with a red x in the top panel of figure 4.6.

The best-fit line and therefore the PCs always pass through the centroid of the system (Pearson, 1901), so subtracting the location of the centroid from the data is a natural next step, as this ensures that in the next step only the slope has to be optimized. This is done in the middle panel of the figure 4.6. If the variables have different units, each variable should be scaled to have equal standard deviations (James et al., 2013) unless the linear algebra based approach with correlation matrices, as explained in e.g. (Jolliffe, 2002), is used.

If this scaling is not performed, the choice of units can arbitrarily skew the principal components. This is easy to see when considering for example a case where one has distances to galaxies in megaparsecs and their masses in units of $10^{12}\ M_{\odot}$, both of which might result in standard deviations being of the order of unity and PCA might thus yield principal components that are not dominated by neither variable alone. Now, say another astronomer has a similar data set, but distances are given in meters. In this case, most of the variation is in the distances, so distances will also dominate the PCs. If all variables are measured in the same units, scaling can be omitted in some cases (James et al., 2013).

Now the first PC can be located by finding the line that passes through the origin and has the maximum variance of the projected data points (Jolliffe, 2002), shown with a black line in the middle panel of figure 4.6 for our data set. PCs are always orthogonal and intersect at the origin, so in the two-dimensional example case the second and final PC is fully determined. The data set can now be represented using the PCs as is shown in the bottom panel of the figure 4.6.

Had the data set had more than two dimensions, the second PC would have been chosen such that it and the first PC are orthogonal and that variance along the new PC is again maximised (Jolliffe, 2002). This can be repeated for each dimension of the data set or, if dimensionality reduction is desired, only for a smaller number of dimensions.

This level of understanding is often enough to successfully apply PCA to a problem, because PCA has ready-made implementations for many programming languages such as `prcomp` in R (James et al., 2013) and `sklearn.decomposition.PCA` in scikit-learn library (Pedregosa et al., 2011) for Python. If a more mathematical approach is desired, Smith (2002) explains PCA together with covariance matrices, eigenvectors and eigenvalues required to understand the process very clearly. Jolliffe (2002) also includes a very thorough description of PCA.

4.3.2 Excluding Less Interesting Principal Components

Even though a data set has as many principal components as there are measured variables, one is often not interested in all of them as the last principal components might explain only a tiny fraction of the total variation in the data (James et al., 2013). Reducing the dimensionality of the problem also greatly eases visualizing and interpreting the data. Thus one might want to retain only first few of the PCs when PCA is used to for example compress, visualize or just interpret the data set at hand (James et al., 2013; Johnson, 2007). Unfortunately, many of the rules and methods used to determine the number of PCs to retain are largely without a formal basis or require assuming a certain distribution which is often not justifiable with the data (Jolliffe, 2002). With careful consideration these methods can nevertheless aid a researcher in making informed decisions and reasoned conclusions, so some rules are introduced in this section.

If the PCA is performed to aid visualizing the data set, retaining only the two first PCs can be a justified choice as two is the maximum number of dimensions that are easy to visualize on two-dimensional media such as paper and the two first PCs determine the best-fit plane for the data (Jolliffe, 2002). Of course the question whether the two PCs are sufficient to describe the data reasonably well still remains

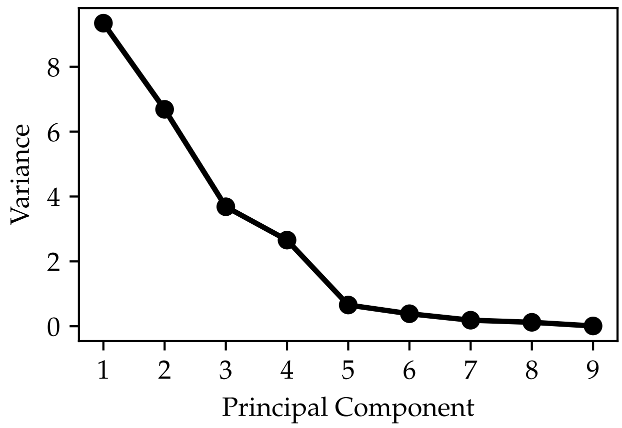

Figure 4.7: Example of a scree plot of randomly generated normally distributed data. In this case the plot has a clear elbow at fifth PC with the PCs 5-9 appearing roughly on a line. Thus the last five PCs could be a good number of PCs to be omitted if dimensionality reduction is desired.

unanswered in this case. Fortunately it can be addressed using some of the following methods used in general case of determining how many PCs to retain.

One widely used technique was introduced by Cattell (1966) to be used in factor analysis, but is also very much applicable to PCA (Jolliffe, 2002). This so called Cattell scree test involves plotting the variance of the data points along each PC versus the index of the PC. These plots tend to look similar to what is shown in figure 4.7, resembling a steep cliff with eroded material accumulated at the base, which is why these plots are known as scree plots and the nearly linear section of the plot is called the scree.

When the scree plot has two clearly different areas, the steep slope corresponding to the first PCs and a more gently sloping scree for the latter PCs, locating this elbow in the plot connecting the two areas will give the number of PCs that should be included (Jolliffe, 2002), which in case of figure 4.7 would yield five PCs. Some sources such as (Cattell, 1966) suggest that in some cases the PC corresponding to the elbow should be discarded, which will result in one less PC.

Unfortunately, as Cattell also acknowledges in his paper, all cases are not as easy to analyze as the one in figure 4.7 and may prove difficult to discern for an inexperienced researcher. This problem might arise from for example noise in the linear part of the plot or scree line consisting of two separate linear segments with different slopes. The first case has no easy solution, but in the latter case Cattell suggests using the smaller number of PCs.

Another straightforward method for choosing how many PCs to retain is to examine how much of the total variation in data is explained by first PCs and including components only up to a point where pre-defined percentage of the total variance is explained (Jolliffe, 2002). Whereas the previous method posed a challenge in determining which PC best matches the exclusion criteria, when using this approach the problem arises from choosing the threshold for including PCs. Jolliffe (2002) suggests that a value between 70 % and 90 % of the total variation is often a reasonable choice, but admits that the properties of the data set may justify values outside this range. Unfortunately, the suggested range is quite wide, so it may contain pultiple PCs and therefore it is up to the researcher to determine the best number of PCs, while the criterion again acts as only an aid in the process.

4.3.3 Principal Component Regression

4.4 Error analysis

4.5 Comparing two samples drawn from unknown distributions

A common question in multiple fields of science is whether two or more samples are drawn from the same distribution. The most relevant methods that can be used to address this problem are introduced here following (Bohm and Zech, 2010) and (Feigelson and Babu, 2012) apart from introducing the χ^2 test which is mostly based on the approach of (Corder, 2014).

Questions related to comparing samples can emerge for example when comparing effectiveness of two procedures, determining if the instrument has changed over time or whether observed data is compatible with simulations. There are multiple two-sample tests that can address this kind of questions, e.g. χ^2, Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests.

In addition to comparing two samples, these tests can be used as one-sample tests to determine whether it is expected that the sample is from a particular distribution. However, some restrictions apply when using the one-sample variants. Some of these tests use categorical data, i.e. data where variables fall in pre-defined categories, and compares numbers of samples in different categories, whereas the others are applied to numerical data and compare empirical distribution functions (EDF) of the datasets.. Examples of such categories might be for example ”galaxies that are active” or ”data points between values 1.5 and 1.6”.

4.5.1 χ^2 test

Astronomical data often involves classifying objects into categories such as ”stars with exoplanets” and ”stars without exoplanets” or the spectral classes of stars (Feigelson and Babu, 2012). One tool for analyzing such categorical data is χ^2 test, which can be used both to determine whether a sample can be drawn from a certain distribution and to test whether two samples can originate from a single distribution.

The method described here is sometimes referred to as Pearson's χ^2 test due to existence of other tests where χ^2 distribution is used. In some cases, such as with small 2×2 contingency tables and when expected cell counts are small, other variants of χ^2 test should be used. For example the Yates's χ^2 test or the Fisher

Stellar class	Number of observed planetary systems
A	6
F	38
G	39
K	134

Table 4.1: Example of categorical data.

Stellar class	Observations (f_o)	Theory (f_e)
A	6	6
F	38	28
G	39	71
K	134	112
total	217	217

Table 4.2: Data of table 4.1 together with expected values if null hypothesis was true.

exact test work better in these cases than the χ^2 test.

For one-sample test, the χ^2 test uses the number of measurements in each bin together with a theoretical estimate calculated from the null hypothesis. For example one might have observed exoplanets and tabulated the number of planet-hosting stars of different spectral class as is shown in table 4.1 and now wants to test the observations against null hypothesis "Distribution of stellar classes for observed exoplanet-hosting stars is equal to that of main sequence stars in solar neighbourhood as given by Ledrew (2001)" using significance level $\alpha = 0.01$. The data is categorical, so now χ^2 test is a justified choice.

In this case the first step would be to calculate the expected observation counts for each bin according to the null hypothesis. Table 4.2 contains these expected counts (f_e) together with the observations (f_o). These observed and expected values are then used to calculate the χ^2 test statistic, defined as

$$\chi^2 = \sum_i \frac{(f_o - f_e)^2}{f_e}. \tag{4.11}$$

With the data given above this results in $\chi^2 \approx 23.6$. The data has four bins, so the degree of freedom is $4 - 1 = 3$. Next one can compare the calculated χ^2 value to a tabulated critical value for our significance level $\alpha = 0.01$. These tabulated values can be widely found in statistics textbooks and books specifically dedicated to statistical tables.

In this case according to Corder (2014) the critical value is 11.34, which means that as $23.6 > 11.34$ one can reject the null hypothesis and conclude that at 1% significance level the distribution of stellar classes for observed exoplanet-hosting stars is not equal to that of main sequence stars in solar neighbourhood. This of course can either be due to exoplanets being more numerous around some stellar classes than others or arise from some observational effect such as the observer observing more of the later type stars and thus arbitrarily skewing the distribution of the exoplanet finds.

The χ^2 test can also be used to test for independence of two or more samples. The data is again tabulated and now the χ^2 test statistic is calculated as

$$\chi^2 = \sum_i \sum_j \frac{(f_{oij} - f_{eij})^2}{f_{eij}} \tag{4.12}$$

where f_{oij} denotes the observed frequency in cell (i, j) and f_{eij} is the expected frequency for that cell. The expected frequency can be calculated using the following formula

$$f_{eij} = \frac{R_i C_i}{N} \tag{4.13}$$

where R_i is the number of samples in row i, C_j is the number of samples in column j and N is the total sample size.

According to Corder (2014), the degrees of freedom is $(R-1)(C-1)$ where R is the number of rows and C is the number of columns in tabulated data. This is true in many if not most cases, but the way of collecting data can affect the degrees of freedom in both one-sample and multi-sample cases, as Press et al. (2007) explains. For example, if the one-sample model is not renormalized to fit the total number of observed events or, in two-sample case, the sample sizes differ, the degrees of freedom equal to number of bins N_b instead of $N_b - 1$.

Before performing the χ^2 test on a dataset, it is important to confirm that the data meets the assumptions for χ^2 test, given for example in (Bock et al., 2014) and (Heino et al., 2012). First of all, the data has to consist of counts i.e. not for example percentages or fractions. These counts should be independent of each other and there has to be enough of them, generally > 50 is sufficient. Bins should also be chosen such that all bins have at least five counts according to the null hypothesis. If the last condition is not met, one can can consider combining bins.

4.5.2 Kolmogorov-Smirnov test

For astronomers one of the most well-known statistical test is the Kolmogorov-Smirnov test, also known as the KS test. It is computationally inexpensive to calculate, easy to understand and does not require binning of data. It is also a nonparametric test i.e. the data does not have to be drawn from a particular distribution.

In the astrophysical context this is often important because astrophysical models usually do not fix a specific statistical distribution for observables and it is common to carry out calculations with logarithms of observables, after which the originally possibly normally distributed residuals will no longer follow a normal distribution. When using the KS test, the values on the x-axis can be freely reparametrized: for example using $2x$ or $\log x$ on x-axis will result in same value of the test statistic

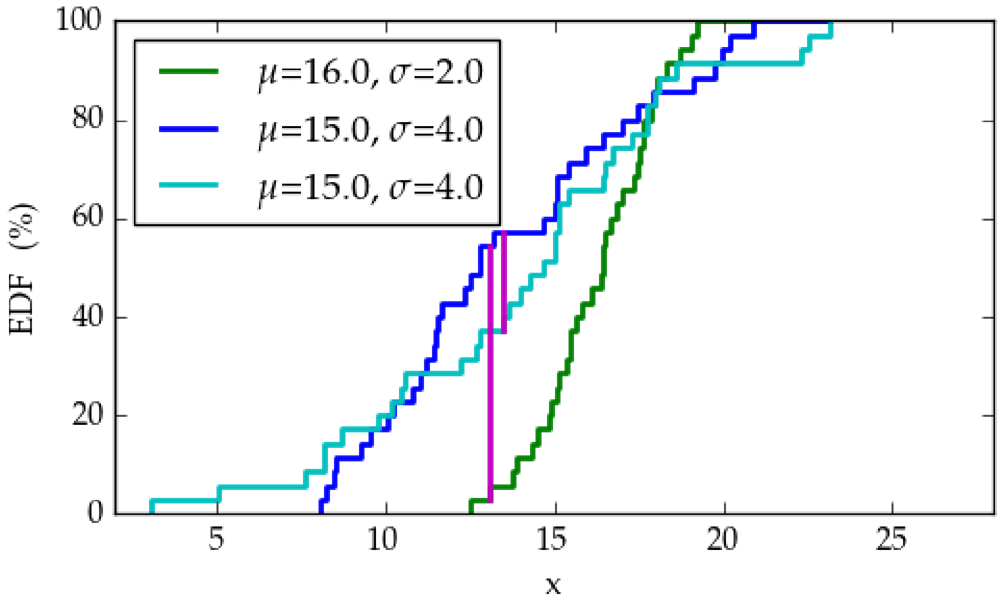

Figure 4.8: KS test parameter values (magenta vertical lines) shown graphically for three samples from figure 4.2.

as using just x (Press et al., 2007).

The test can be used as either one-sample or two-sample test, both of which are very similar. For two-sample variate the test statistic for the KS test is calculated based on empirical distribution functions $\hat{F}_1$ and $\hat{F}_2$ derived from two samples and the test statistic

$$D = \sup_x |\hat{F}_1(x) - \hat{F}_2(x)| \tag{4.14}$$

uses the maximum vertical distance of the EDFs. This test statistic is then used to determine the p-value and thus decide whether the null hypothesis can be rejected. For one-sample variate the procedure is similar, but EDF $\hat{F}_2$ is substituted with the CDF that corresponds to the null hypothesis.

As an example, let us consider two pairs of samples from figure 4.2: green and blue (two samples drawn from different normal distributions) and blue and cyan (two samples drawn from same normal distribution). We can formulate the test and null hypotheses for both pairs as H_0="the two samples are drawn from the same

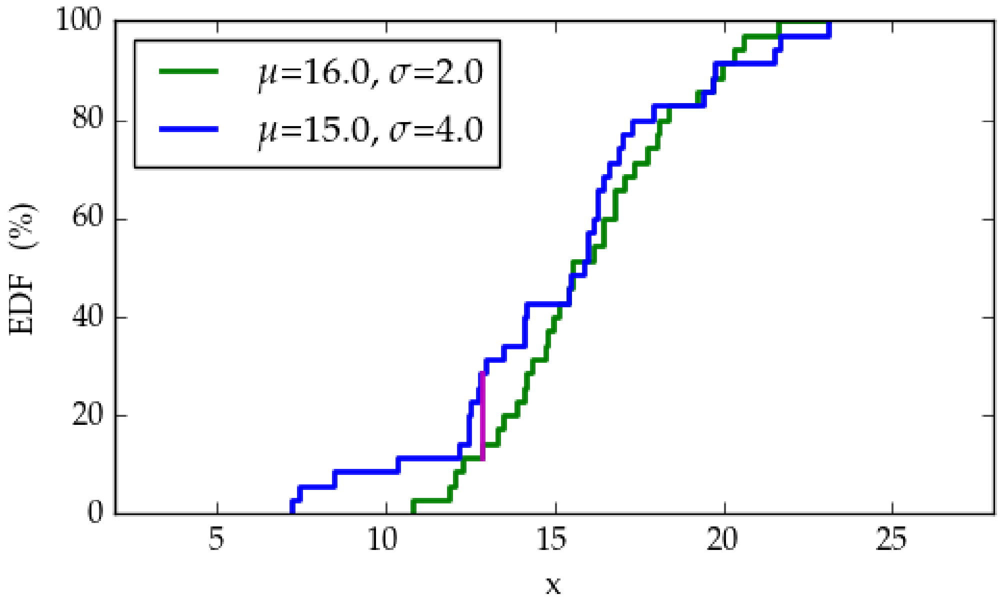

Figure 4.9: KS test ran on another pair of samples drawn from blue and green distributions in figure 4.1.

distribution" and H_1="the two samples are not drawn from the same distribution" and choose a significance level of for example $\alpha = 0.05$ or $\alpha = 0.01$.

The test statistic is then calculated and for these samples we get $D = 0.51$ for the green-blue pair and $D = 0.20$ for the blue-cyan pair. Test statistics are illustrated in figure 4.8 where the test statistics D are shown as vertical magenta lines. According to Python function `scipy.stats.ks_2samp`, these values of D correspond to p-values 9.9×10^{-5} and 0.44 respectively, which means that the null hypothesis "green and blue samples are drawn from the same distribution" is rejected at both 0.05 and 0.01 significance levels but the null hypothesis "blue and cyan samples are drawn from the same distribution" cannot be rejected.

In this case the KS test produced result that matches the actual distributions from which the samples were drawn. Using a different random realization might have resulted in a different conclusion, for example one shown in figure 4.9 results in $D = 0.17$ that corresponds to a p-value of 0.64 i.e. null hypothesis could not have been rejected using the α specified earlier. In a similar manner there can be cases where two samples from one distribution are erroneously determined not to come from the same distribution if the samples differ from each other enough due to random effects.

The latter example case also illustrates one major shortcoming of the KS test: it is not very sensitive to small-scale differences near the tails of the distribution. For example in figure 4.9 the blue sample goes much further left, but because EDF is always zero at the lowest allowed value and one at the highest one the vertical distances near the tails are small and the test is most sensitive to differences near the median value of the distribution. On the other hand, the test performs quite well when the samples differ globally or have different means. (Feigelson and Babu, 2012)

The KS test is also subject to some limitations and it is important to be aware of them in order to avoid misusing it. First of all, the KS test is not distribution free if the model parameters, e.g. mean and standard deviation for normal distribution, are estimated from the dataset that is tested. Thus the tabulated critical values can be used only if model parameters are determined from some other source such as a simulation, theoretical model or another dataset.

Another severe limitation of KS test is that it is only applicable to one-dimensional data. If the dataset has two or more dimensions, there is no unique way of ordering the points to plot EDF and therefore if KS test is used, it is no longer distribution free. Some variants that can handle two or more dimensions have been invented, such as ones by Peacock (1983) and Fasano and Franceschini (1987), but the authors do not provide formal proof of validity of these tests. Despite this, the authors claim that Monte Carlo simulations suggest that the methods work adequately well for most applications.

4.5.3 Other tests based on EDFs

Unsatisfactory sensitivity of the KS test motivates the use of other more complex tests. Such tests are for example the Cramér-von Mises test (CvM) and Anderson-Darling (AD) test, both of which have their strengths. Similar to KS test, both of these can be used as one-sample or two-sample variants.

First of these tests integrates over the squared difference between the EDF of the sample and CDF from the model or two EDFs in case of two-sample test. The test statistic W^2 for one-sample case can be expressed formally as

$$W^2 = \int_{-\infty}^{\infty} [\hat{F}_1(x) - F_0(x)]^2 \, dF_0(x) \tag{4.15}$$

For two-sample version, the theoretical CDF F_0 has to be replaced with another empirical distribution function $\hat{F}_2$.

Due to integration, the CvM test is able to differentiate distributions based on both local and global differences, which causes it to often perform better than the KS test. Similar to the KS test, the CvM test also suffers from EDFs or an EDF and a CDF being equal at the ends of the data range, which again makes the test less sensitive to differences near the tails of the distribution.

In order to achieve constant sensitivity over the entire range of values, the statistic has to be weighted according to the proximity of the ends of the distribution. The AD test does this with its test statistic defined as

$$A^2 = N \int_{-\infty}^{\infty} \frac{[\hat{F}_1(x) - F_0(x)]^2}{F_0(x)[1 - F_0(x)]} \, dF_0(x) \tag{4.16}$$

where N is the number of data points in sample. This weighing makes the test more powerful than the KS and CvM tests in many cases. (Bohm and Zech, 2010; Feigelson and Babu, 2012)

Also other more specific tests exist, such as the Kuiper test which is well suited for cyclic measurements. The test should always be chosen to match the dataset such that it best differentiates between the null and research hypotheses.

4.6 Cluster Analysis

DBSCAN

5. Findings from DMO Halo Catalogue Analysis

5.1 Selection of Local Group analogues

criteria, how many found, what are like (some plots maybe? distributions of masses, separations, velocity components, number of subhaloes within some radius or correlations between two of those?). Some of this might be part of previous chapter too (relevant to resimulation)?

TODO: selitykset sille, miten osa on keskittynyt tiettyihin arvoihin sallitulla välillä ja osa jakautunut tasaisemmin. Mieti, pitäisikö kolme viimeistä esittää esim scatterplottina combined mass vs mass in more massive

Figure 5.1 shows how different features of the found LG analogues are distributed. TODO: selitykset sille, miten osa on keskittynyt tiettyihin arvoihin sallitulla välillä ja osa jakautunut tasaisemmin.

5.2 Hubble Flow Measurements

HF, local H_0, H_0 within shells, zero-point, are previous consistent with what went into the simulation

Figure 5.2: two different simulations, MW-centered, huom obs nb how different they are: scatter, number of haloes, changes in scatter (bound structures)

Figure 5.3 shows haloes included and excluded in fitting, how is the process done

Figure 5.4 shows H0 at different radii. First bump is clear, latter ones not,

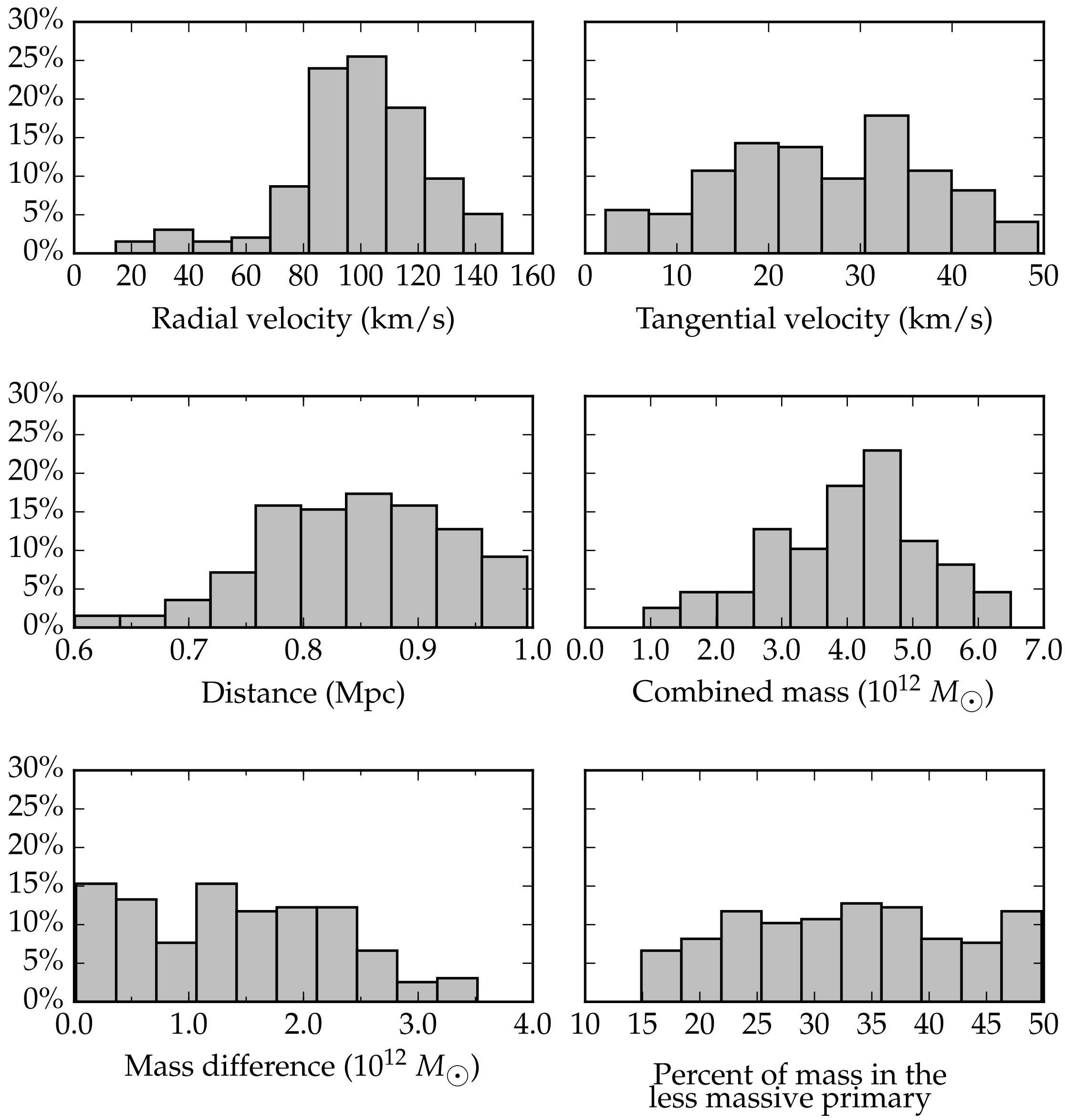

Figure 5.1: Distributions of LG analogue properties. TODO: selitä, mieti miten y-akselin label, binien rajat pitäisi pakottaa suurimpaan ja pienimpään sallittuun arvoon

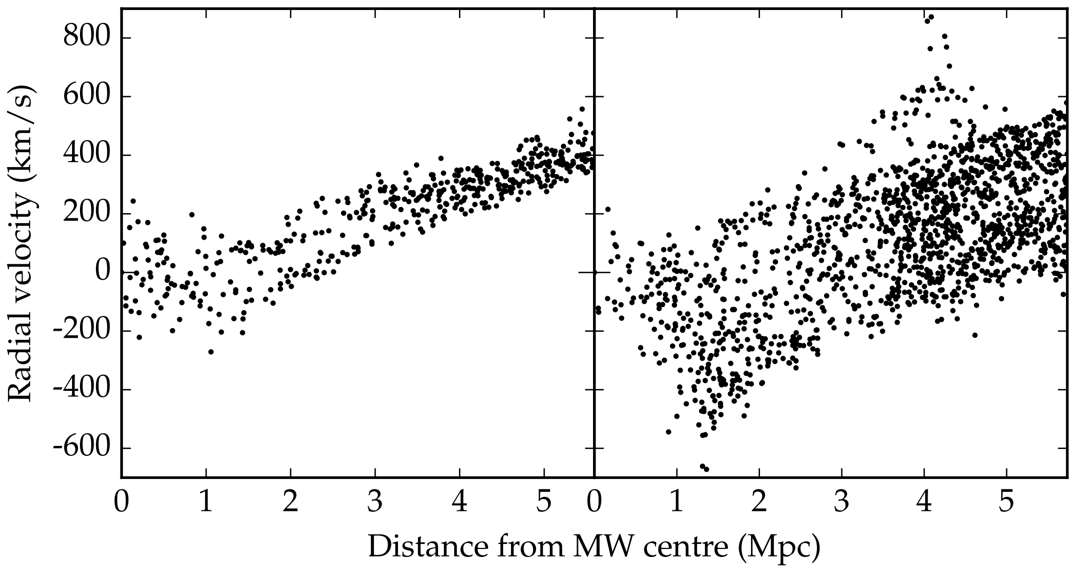

Figure 5.2: Hubble Flows around Milky Way in two simulations.

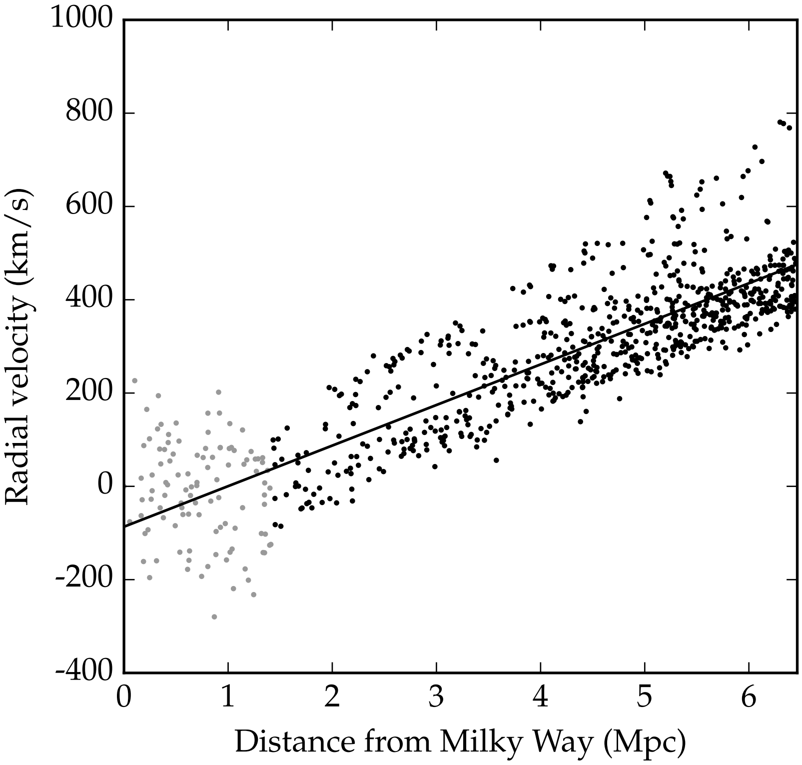

Figure 5.3: HF slope: 86.9929348817

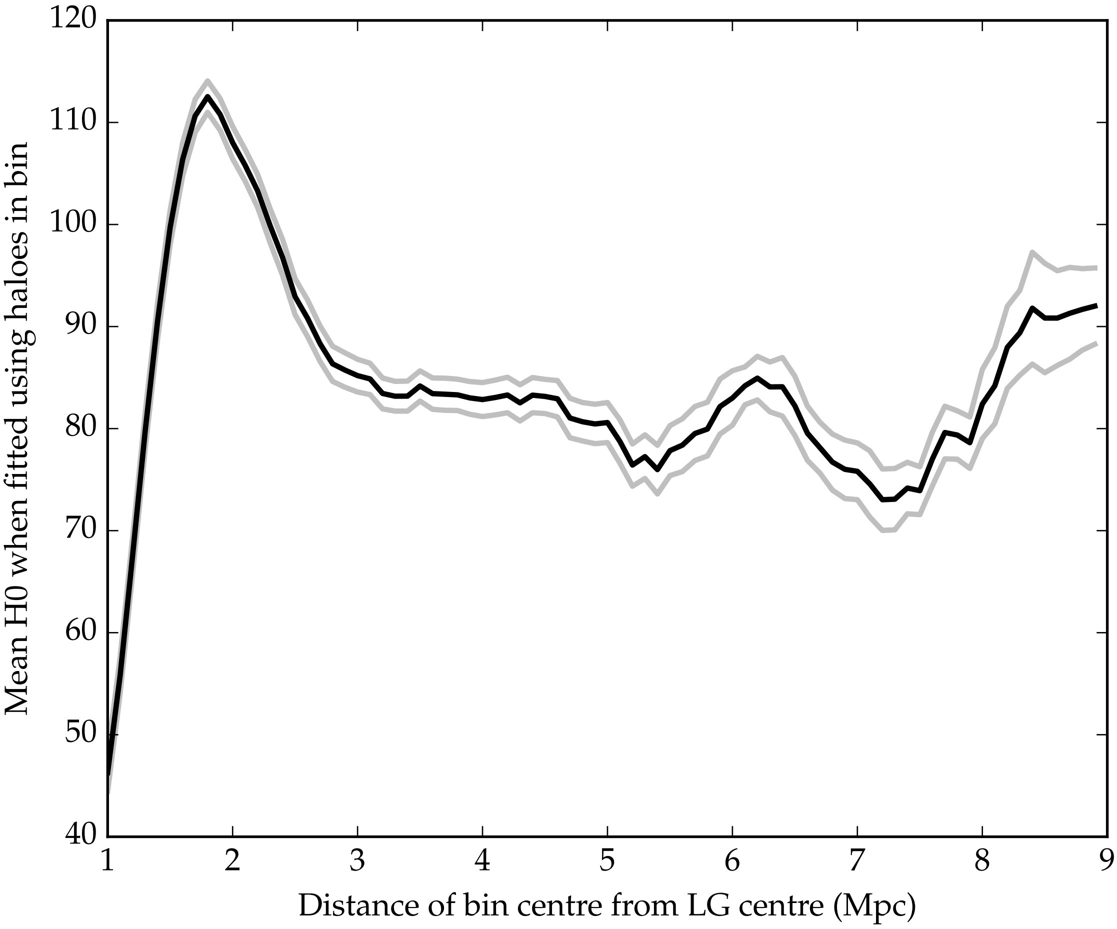

Figure 5.4: Mean H0 in different 2 Mpc bins, grey curves show standard error.

H0>67.7 km/s, why. First ones have 350 samples, last ones only seven, remember to explain standard error. Figure 5.5 has bigger bins and shows zero points. Think whether both should use samy plot type and which is better (line vs boxplot). If boxplot stays, change colours to all-black? At least explain what is what in plot.

X

5.3 Anisotropy of Hubble flow

isotropy + randomness or anisotropy? esittele konsepti. plots: see notebook last pages

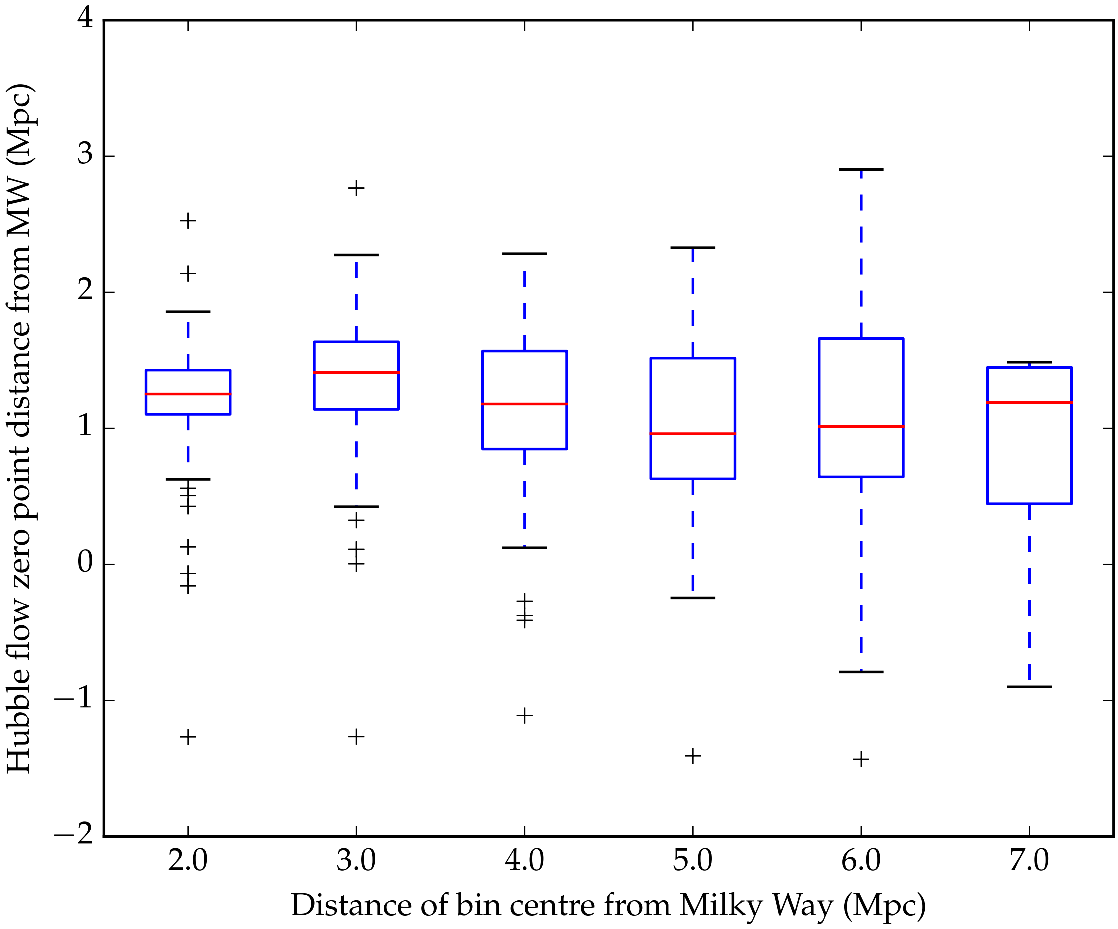

Figure 5.5: HF zero point in different 4 Mpc bins. specify one outlier outside the plot

Simulaatio 97, esittele jo tässä näkyvät klöntit joissa paljon samaa väriä, näytä myös klusterointi ja vertaa löytöjä siihen

Figure 5.6 shows distribution of haloes around Milky Way analogue from one simulation with haloes closer than 1.5 Mpc away from center excluded to avoid cluttering the view with Andromeda counterpart and its satellites.

5.3.1 Clustering

Used DBSCAN introduced in [earlier chapter], angular distances of projections on sky as seen from MW.

Figure 5.8 shows the effect of varying minsamples and ε on number of clusters found in each simulation (1.5 Mpc < r < 5.0 Mpc again). Regions where there are

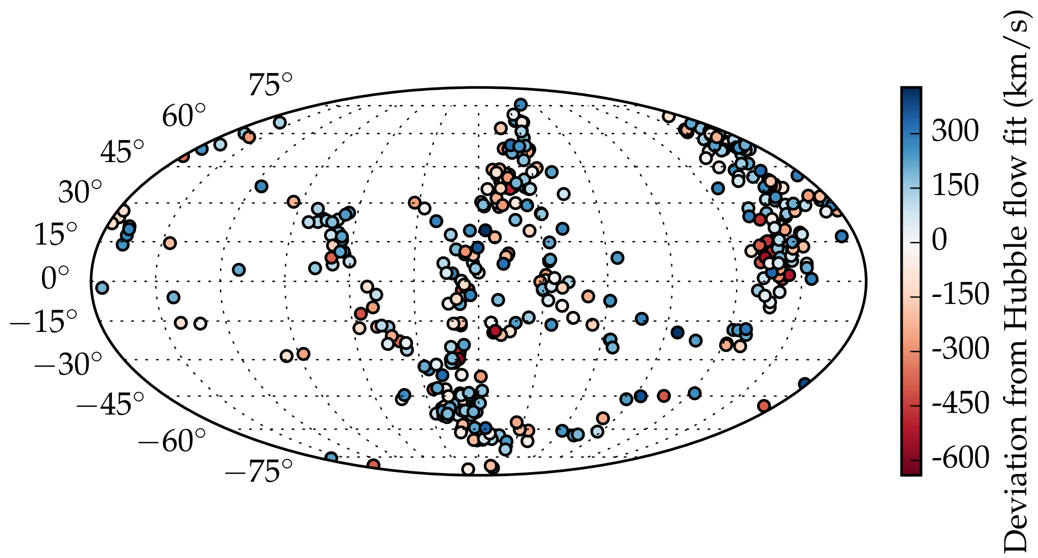

Figure 5.6: Projections of haloes around the less massive LG primary with distances ranging from 1.5 Mpc to 5.0 Mpc.

ridiculously many clusters and ones where there are one or zero, relevant region in between, some areas have similar number of clusters but do the clusters look the same, see plots that don't exist yet

TODO: mieti laitatko samaan figureen, vertaile kuitenkin, selitä

Figure 5.9 shows the change in mean diameter (supremum of angular distance between haloes) in cluster when ε and minsamples are varied. White areas where no clusters are found in any simulation.

Ehkä vähän vähemmän tilaa plottien välissä vaakasuunnassa? Keltaiset vähän turhan samanlaisia.

vasemmanpuolimmainen Maksimisarake? laitakynnys? Kaksi eri kynnystä? Liian kapea ja epätasapainoinen, laita päällekkäin?

Figures 5.10 and 5.11 show how the clustering results vary when clustering parameters are varied.

Figure 5.12 shows how derived values of slope and zero-point for the Hubble flow change when the Hubble flow fitting is carried out on partial data chosen based on the cluster membership of the haloes.

5.4 Statistical Estimate of the Local Group Mass

Analysis similar to Fattahi et al 2016 paper

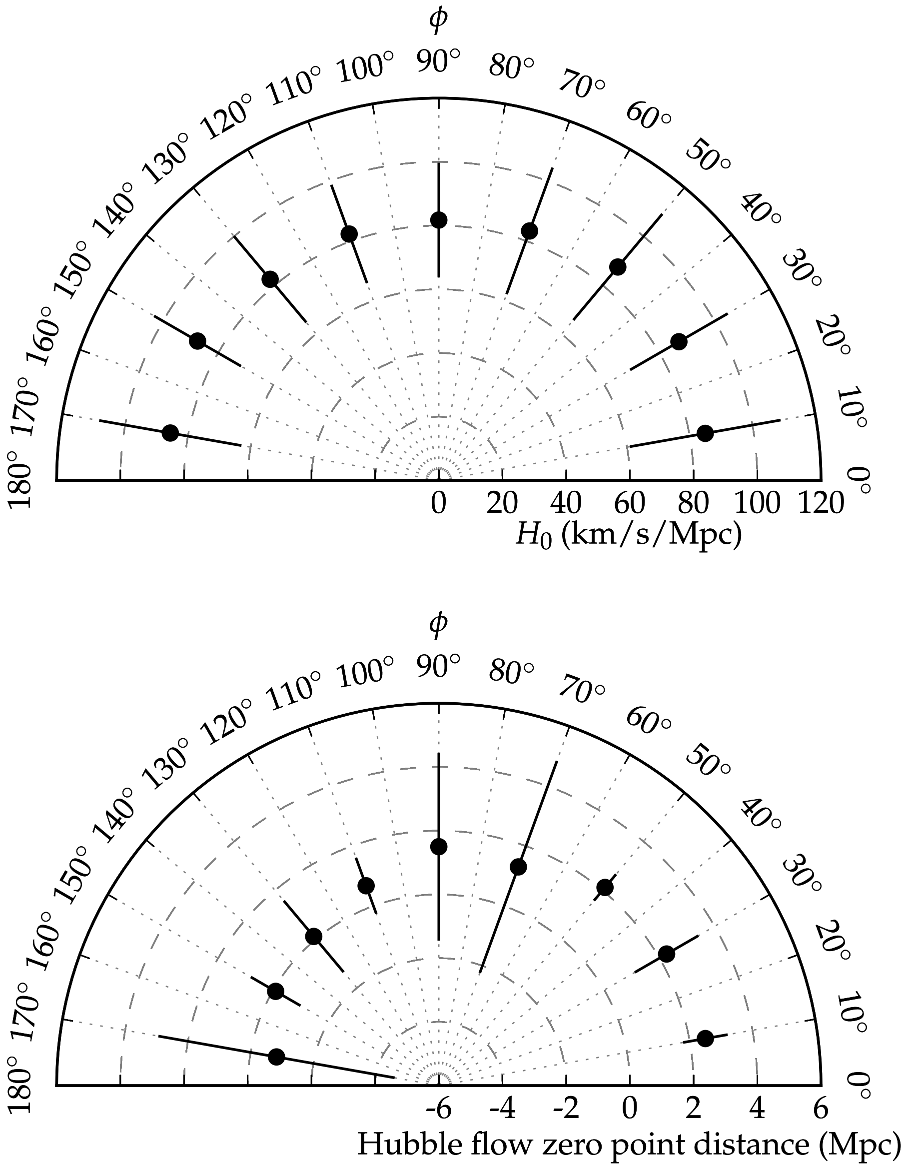

Figure 5.7: Mean Hubble flow slope and zero point as seen from Milky Way analogue in different 20° bins as measured from line connecting Milky Way and Andromeda analogues, direction 0° being towards Andromeda.

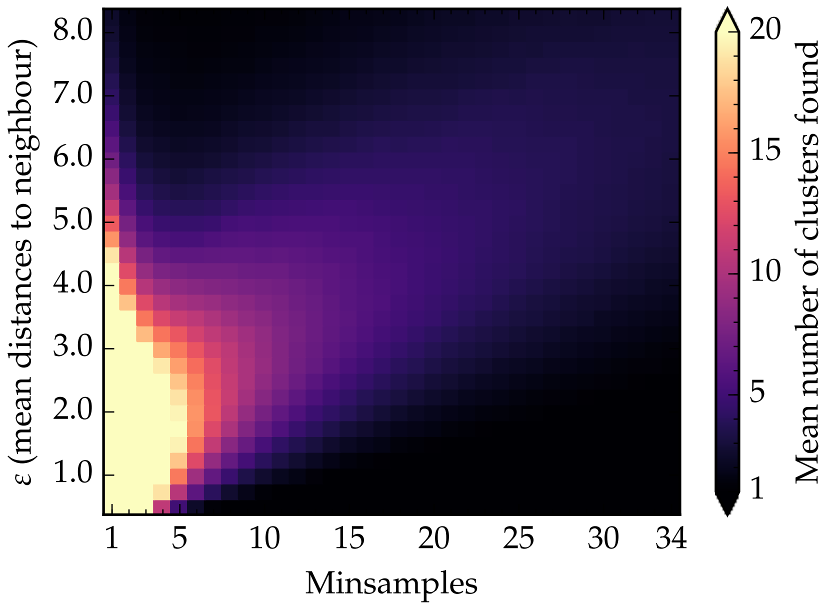

Figure 5.8: Mean number of clusters found for all simulations in dataset with different DBSCAN parameters. In all simulations ε is scaled using the mean distance between closest neighbours.

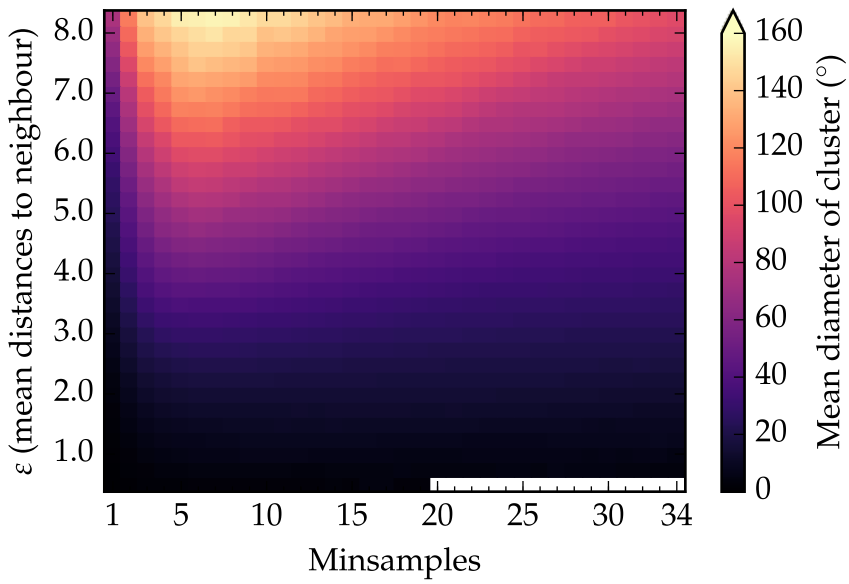

Figure 5.9: Mean diameter of clusters found for all simulations in dataset with different DBSCAN parameters. In all simulations ε is scaled using the mean distance between closest neighbours.

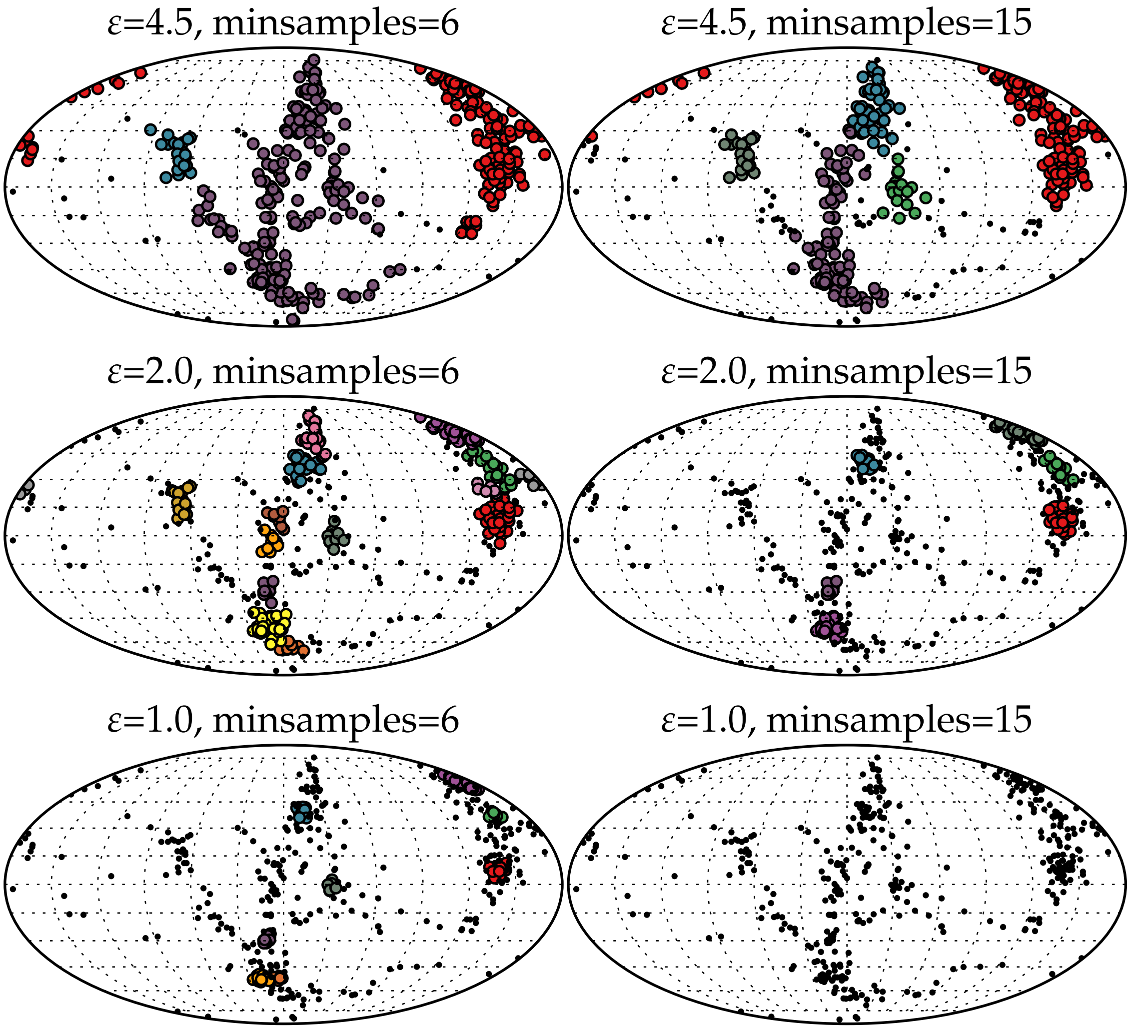

Figure 5.10: Results of DBSCAN clustering on same simulation output with different clustering parameters. TODO: mieti, kuuluuko tämäntyyppinen DBSCANin yleisiä ominaisuuksia esittelevä kuva enemmänkin teoriaosaan. Toisaalta selvästi dataspesifejä juttuja.

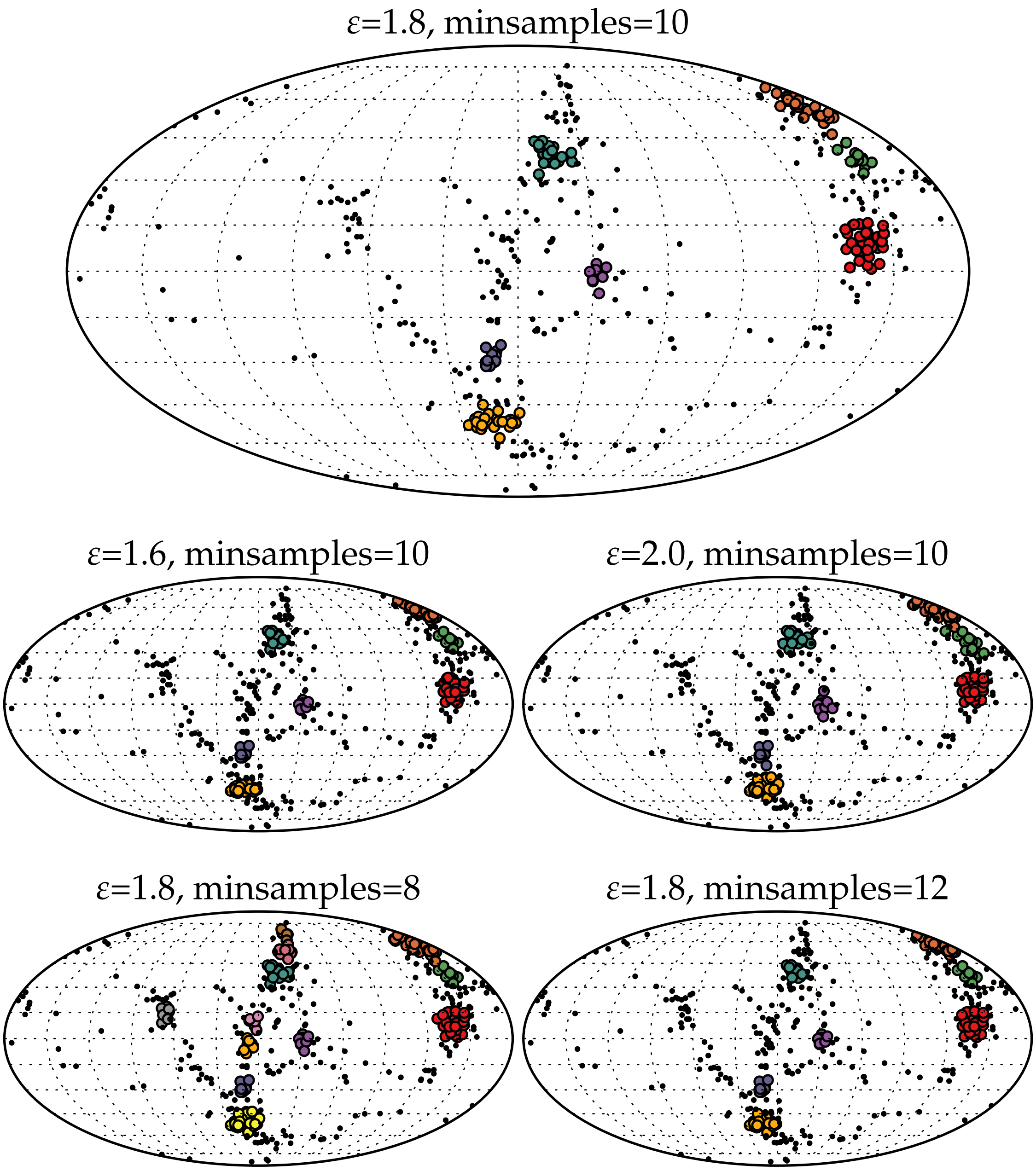

Figure 5.11: The effect of slightly variying the clustering parameters around the values ε=1.8 and minsamples=10 used when analyzing clustered data.

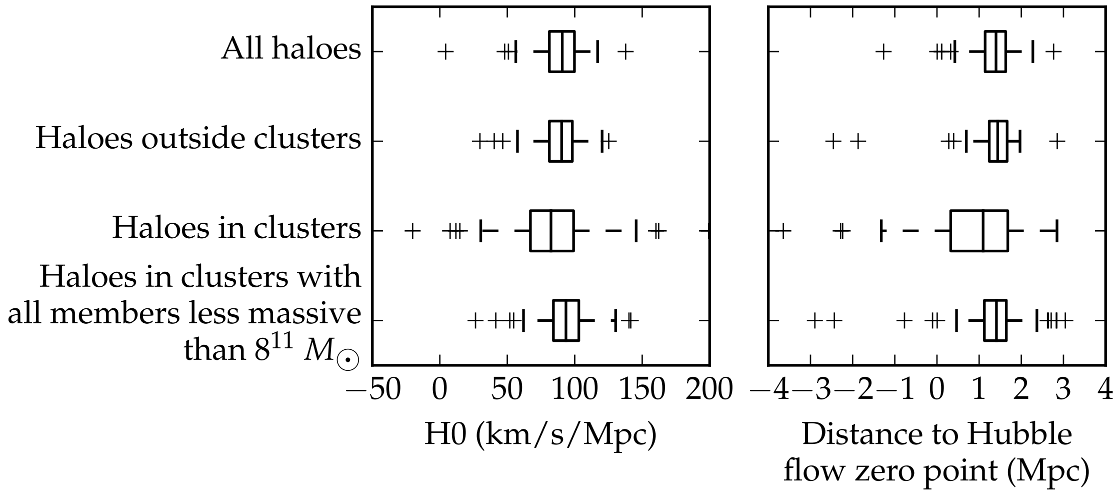

Figure 5.12: Hubble constant and distance to the point at which velocity due to the fitted Hubble flow is zero calculated from different samples. HUOM OBS NB erittele plotin ulkopuolelle jääneet kaukaiset outlierit

6. Conclusions

Bibliography

A. W. Appel. An Efficient Program for Many-Body Simulation. *SIAM Journal on Scientific and Statistical Computing, vol. 6, no. 1, January 1985, p. 85-103.*, 6: 85–103, January 1985.

J. Barnes and P. Hut. A hierarchical O(N log N) force-calculation algorithm. *Nature*, 324:446–449, December 1986. doi: 10.1038/324446a0.

J. E. Barnes and P. Hut. Error analysis of a tree code. *Astrophysical Journal Supplement*, 70:389–417, June 1989. doi: 10.1086/191343.

James Binney and Scott Tremaine. *Galactic dynamics.* Princeton series in astrophysics. Princeton University Press, Princeton, 2nd edition edition, 2008. URL `http://login.libproxy.helsinki.fi/login?url=http://site.ebrary.com/lib/helsinki/Doc?id=11217466`.

D. Bock, P. Velleman, and R De Veaux. *Stats: Modeling the World.* Pearson, third edition edition, 2014.

G. Bohm and G. Zech. *Introduction to statistics and data analysis for physicists.* DESY, 2010. ISBN 9783935702416. URL `http://www-library.desy.de/preparch/books/vstatmp_engl.pdf`.

Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

Gregory W. Corder. *Nonparametric statistics : a step-by-step approach.* Wiley, Hoboken, New Jersey, second edition edition, 2014.

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

Walter Dehnen. A hierarchical o(n) force calculation algorithm. *Journal of Computational Physics*, 179(1):27 – 42, 2002. ISSN 0021-9991. doi: https://doi.org/10.1006/jcph.2002.7026. URL `http://www.sciencedirect.com/science/article/pii/S0021999102970269`.

G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225:155–170, March 1987. doi: 10.1093/mnras/225.1.155.

Eric D. Feigelson and G. Jogesh Babu. *Modern Statistical Methods for Astronomy: With R Applications.* Cambridge University Press, 2012. doi: 10.1017/CBO9781139015653.

S. Gottlöber, A. A. Klypin, and A. V. Kravtsov. Halo evolution in a cosmological environment. In G. Giuricin, M. Mezzetti, and P. Salucci, editors, *Observational Cosmology: The Development of Galaxy Systems*, volume 176 of *Astronomical Society of the Pacific Conference Series*, page 418, June 1999.

Michael Griebel, Stephan Knapek, and Gerhard Zumbusch. *Numerical Simulation in Molecular Dynamics.* Springer-Verlag, Berlin, Heidelberg, 2007.

R. Heino, K. Ruosteenoja, and J. Räisänen. *Havaintojen tilastollinen käsittely.* Department of Physics (University of Helsinki), 2012.

Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929.

C. R. Jenkins J. V. Wall. *Practical Statistics for Astronomers*. Cambridge Observing Handbooks for Research Astronomers. Cambridge University Press, illustrated edition edition, 2003. ISBN 9780521454162,0521454166.

G. James, D. Witten, T. Hastie, and R Tibshirani. *An introduction to statistical learning : with applications in R*. Springer texts in statistics. Springer, New York, 2013.

Richard Arnold Johnson. *Applied multivariate statistical analysis*. Pearson Prentice Hall, Upper Saddle River, 6th ed edition, 2007.

I. T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer, New York, 2nd edition edition, 2002.

Alexander Knebe, Frazer R. Pearce, Hanni Lux, Yago Ascasibar, Peter Behroozi, Javier Casado, Christine Corbett Moran, Juerg Diemand, Klaus Dolag, Rosa Dominguez-Tenreiro, Pascal Elahi, Bridget Falck, Stefan Gottlöber, Jiaxin Han, Anatoly Klypin, Zarija Lukić, Michal Maciejewski, Cameron K. McBride, Manuel E. Merchán, Stuart I. Muldrew, Mark Neyrinck, Julian Onions, Susana Planelles, Doug Potter, Vicent Quilis, Yann Rasera, Paul M. Ricker, Fabrice Roy, Andrés N. Ruiz, Mario A. Sgró, Volker Springel, Joachim Stadel, P. M. Sutter, Dylan Tweed, and Marcel Zemp. Structure finding in cosmological simulations: the state of affairs. *Monthly Notices of the Royal Astronomical Society*, 435(2):1618–1658, 2013. doi: 10.1093/mnras/stt1403. URL `http://dx.doi.org/10.1093/mnras/stt1403`.

G. Ledrew. The Real Starry Sky. *Journal of the Royal Astronomical Society of Canada*, 95:32, February 2001.

Malcolm S. Longair. *Galaxy formation*. Astronomy and astrophysics library. Springer, Berlin; New York, 2nd ed edition, 2008.

Ivan Markovsky and Sabine Huffel. Overview of total least-squares methods. 87: 2283–2302, 10 2007.

Houjun Mo, Frank Van den Bosch, and Simon White. *Galaxy formation and evolution*. Cambridge University Press, Cambridge; New York, 2010.

Douglas C. Montgomery. *Introduction to linear regression analysis*. Wiley series in probability and statistics. John Wiley & Sons Ltd, Hoboken, New Jersey, fifth edition edition, 2012.

M. L. Norman, G. L. Bryan, R. Harkness, J. Bordner, D. Reynolds, B. O'Shea, and R. Wagner. Simulating Cosmological Evolution with Enzo. *ArXiv e-prints*, May 2007.

J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202:615–627, February 1983. doi: 10.1093/mnras/202.3.615.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Planck Collaboration. Planck 2013 results. xvi. cosmological parameters. *Astronomy and Astrophysics*, 571:A16, 2014. doi: 10.1051/0004-6361/201321591. URL `https://doi.org/10.1051/0004-6361/201321591`.

Planck Collaboration. Planck 2015 results. I. Overview of products and scientific results. *Astronomy and Astrophysics*, 594:A1, September 2016. doi: 10.1051/0004-6361/201527101.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, editors. *Numerical recipes: The art of scientific computing*. Cambridge University Press, New York, third edition edition, 2007.

Paul Renteln. *Manifolds, Tensors, and Forms: An Introduction for Mathematicians and Physicists*. Cambridge University Press, Cambridge; New York, 2013.

Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, 2015. doi: 10.1093/mnras/stu2058. URL `+http://dx.doi.org/10.1093/mnras/stu2058`.

Lindsay I Smith. A tutorial on principal components analysis. 2002.

Volker Springel. The cosmological simulation code GADGET-2. *Monthly Notices of the Royal Astronomical Society*, 364:1105–1134, December 2005. doi: 10.1111/j.1365-2966.2005.09655.x.

Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies – i. results at z = 0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750, 2001. doi: 10.1046/j.1365-8711.2001.04912.x. URL `http://dx.doi.org/10.1046/j.1365-8711.2001.04912.x`.

Simon D. M. White, Carlos S. Frenk, and Marc Davis. *Is the Universe Made of*

Massive Neutrinos? Springer Netherlands, Dordrecht, 1984. ISBN 978-94-009-7245-2. doi: 10.1007/978-94-009-7245-2_8.

Guoliang Xue. An o (n) time hierarchical tree algorithm for computing force field in n-body simulations. *Theoretical Computer Science*, 197(1):157–169, 1998.

A. Principal Components

PC	H_0	HF zero	HF zero (clustered)	HF zero (not clustered)	σ_{radvel}	σ_{radvel} (clustered)	σ_{radvel} (not clustered)	$v_{r,LG}$	$v_{t,LG}$	r_LG
1	-0.386	-0.449	-0.324	-0.379	-0.393	-0.211	-0.384	-0.211	0.097	0.013
2	0.147	0.221	0.248	0.150	-0.064	-0.599	-0.056	-0.599	-0.103	0.332
3	0.287	0.145	0.186	0.223	-0.531	0.258	-0.535	0.258	0.115	0.313
4	0.009	0.020	-0.152	0.102	0.151	-0.047	0.136	-0.047	0.934	0.221
5	-0.024	-0.171	-0.273	-0.105	0.140	0.134	0.227	0.134	-0.270	0.840
6	0.015	-0.092	0.706	-0.663	0.049	0.065	0.072	0.065	0.147	0.127
7	-0.848	0.156	0.323	0.343	-0.074	0.071	0.060	0.071	-0.001	0.128
8	-0.162	0.582	-0.206	-0.315	0.456	0.024	-0.529	0.024	-0.018	0.061
9	0.041	-0.572	0.239	0.326	0.549	-0.008	-0.452	-0.008	-0.016	0.031
10	0.000	-0.000	0.000	0.000	-0.000	0.707	-0.000	-0.707	0.000	-0.000

Table A.1: component H0s zeropoints inClusterZeros outClusterZeros allDispersions clusterDispersions unclusteredDispersions radialVelocities tangentialVelocities LGdistances