Master's thesis
Astronomy

Your Title Here

Anni Järvenpää

September 21, 2017

Tutor: Professor Peter Johansson
Dr. Till Sawala

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	September 21, 2017	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

Think whether LG or LCDM first

2.1 Local Group

Definition of galaxy group, our local group is one of these.

Mass estimate (Li, Yang masses for the LG and MW)

Maybe something about scale of things in our universe, what are galaxy groups made of, what do you get if you go one distance scale up, what's different in galaxy clusters

2.1.1 Structure

Galaxies that are part of LG, distribution of smaller ones around bigger ones

Current mass estimates (at least timing argument, hubble flow and maybe satellites)

2.1.2 Evolution

How have we ended up in a situation described earlier? What will happen in future?

2.2 Expanding universe

2.2.1 Discovery

Make maths, add cosmological constant, make observations, remove cosmological constant

Enough cosmology here or in other sections to make other parts of thesis to make sense and to suffice as master's thesis = basic textbook cosmology and galaxy formation theory

2.2.2 ΛCDM Cosmology

2.2.3 Hubble flow

What is, where seen, what means, how to measure, hotness/coldness

Plot: observations with fitted hubble flow

3. general simulation thingies

Data used here from EAGLE which uses modified GADGET-2 which is a tree-code that uses leapfrog

3.1 N-body simulations

3.1.1 Hierarchical Tree Algorithm

3.1.2 Numerical Integrators

3.1.3 Halo Finding with Subfind

3.2 Description of actual simulations used

Volume, number of particles, compare to other simulations, where better and where maybe worse

Resimulation of interesting regions

Simulation has same parameters as EAGLE 800 Mpc volume used schaye 2015 paper DM-only parts: Volker-Springer Gadget and Gadget 2 papers 1999 and 2005 or something, gravity part is more interesting than SPH Zooms can use multiple meshes, only one is used here gravitational softening

4. Mathematical and statistical methods

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

4.1 Statistical Background

vähän parempi tässä kuin aiemman otsikon alla

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.

kerro mikä α ja N käytössä myöhemmin kunhan tiedät

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 it's 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

ei hyvä, harkitse esim http://puppulause-generaattori.fi/?ava-insana=jakauma-funktio

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)

PDF määritelmä vaikea ymmärtää

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)

esittelet nyt nolosti EDF:n nimeltä kahdesti, mieti ratkaisu

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

lisää johonkin selitys normaalijakauman parametreille

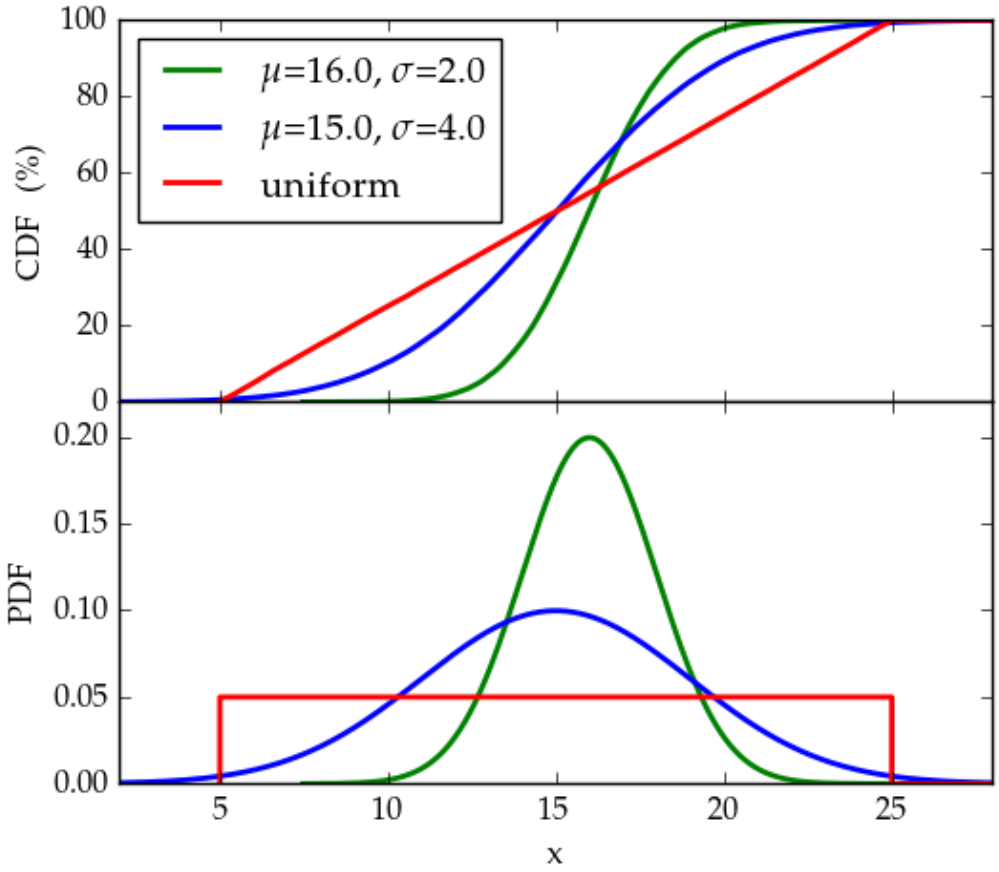

Figure 4.1: Cumulative distribution function (top panel) for three random samples (PDFs shown in the bottom panel) drawn from different distributions, two of which are normal and one is uniform. Parameters μ and σ of the normal distribution describe the mean and the spread of the distribution respectively,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

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

harkitse laittavasi jämpti arvo N:lle kun muut osat valmiita

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

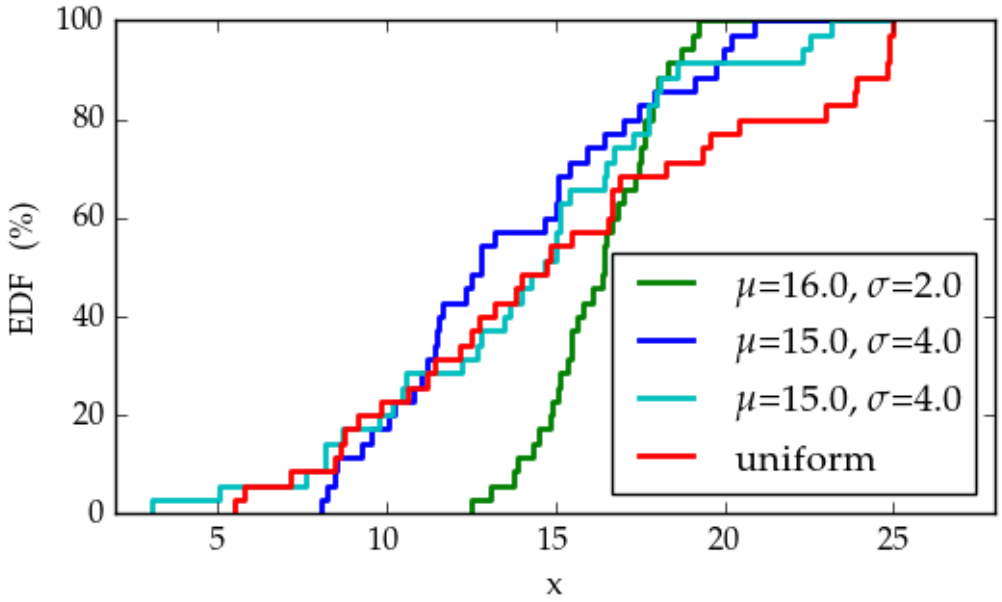

Figure 4.2: Empirical distribution function for four random samples (N=35) drawn from the same distributions as in figure 4.1. Note that both the blue and the cyan data are drawn from the same distribution.

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 $x_1 \dots x_i$ 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

onko otsikko järkevä kun

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

toisiko jotain lisää jos olisi β_1 ja β_2 lausekkeet?

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.

HF: OLS/TLS? Sido PCA:han

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))

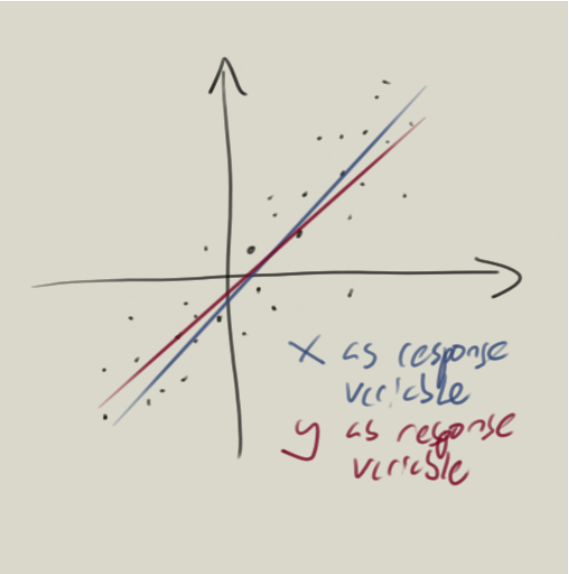

Figure 4.4

regression can be used instead of OLS (**?**). 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).

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

orthogonal vs uncorrelated

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) (Jolliffe, 2002). This approach also solves the problem of sometimes arbitrary choice

Figure 4.5

of division of the data to dependent and independent variables introduced in section 4.2.something Pearson (1901).

not sure if kannattaa kuvaille proseduuria erityisen tarkkaan, selityksestä tulee pitkällinen jos aikoo aloittaa keskimääräisen tähtitieteilijän lähtötasolta, ehkä vain ohjaa lähteeseen

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

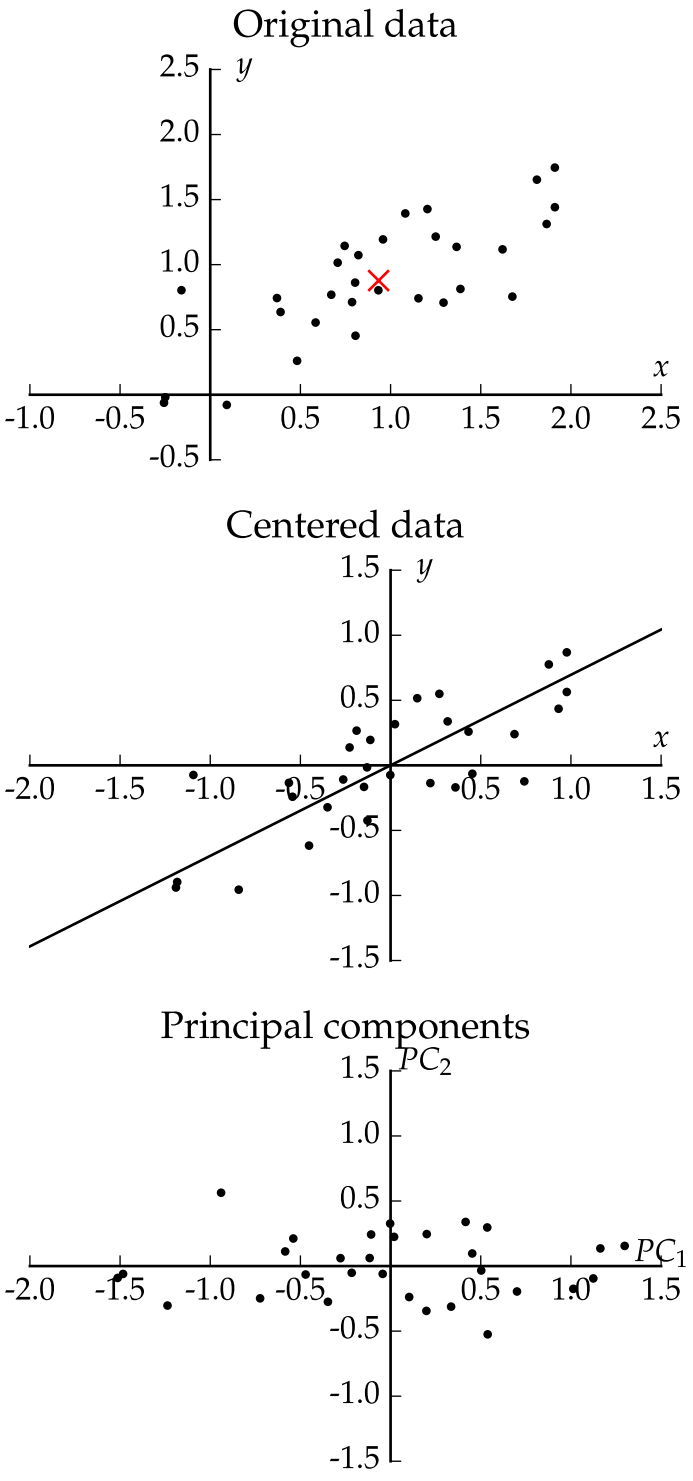
Original data
y
x
2.5
2.0
1.5
1.0
0.5
-0.5
-1.0
-0.5
0.5
1.0
1.5
2.0
2.5
Centered data
y
x
1.5
1.0
0.5
-0.5
-1.0
-1.5
-2.0
-1.5
-1.0
-0.5
0.5
1.0
1.5
Principal components
PC_2
PC_1
1.5
1.0
0.5
-0.5
-1.0
-1.5
-2.0
-1.5
-1.0
-0.5
0.5
1.0
1.5

Figure 4.6: .

jos päädyt puhumaan eigenvektoreista tai kovarianssimatriiseista, selitä ne täällä

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.

mieti mitä monospeissillä ja ole konsistentti

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

maininta bootstrappingista (loppuun?) ja siitä, että maistuu koneoppiminen?

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). This might occur for example when PCA is used to 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 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

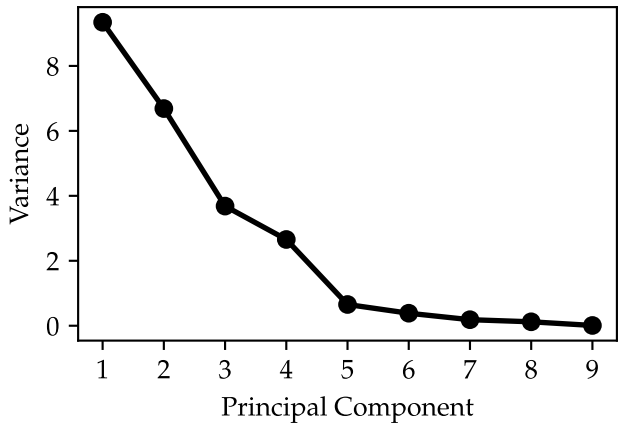

Figure 4.7: Example of a scree plot of randomly generated normally distributed data. In this case the plot has a clear elbow at fifth PC with the PCs 5-9 appearing roughly on a line. Thus the last five PCs could be a good number of PCs to be omitted if dimensionality reduction is desired.

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.

joku kiva lopetus tämän jälkeen?

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

TODO: oispa parempi otsikko. mieti, onko tämä muutenkaan hyvä nyt kun on siirretty yksi otsikkotaso ylöspäin

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

keksi paremmat esimerkit koko kappaleeseen, jotain relevanttia myöhempää tutkimusta ajatellen. katso kommentit paperista sen jälkeen, kaikkia ei täällä vielä

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

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.

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

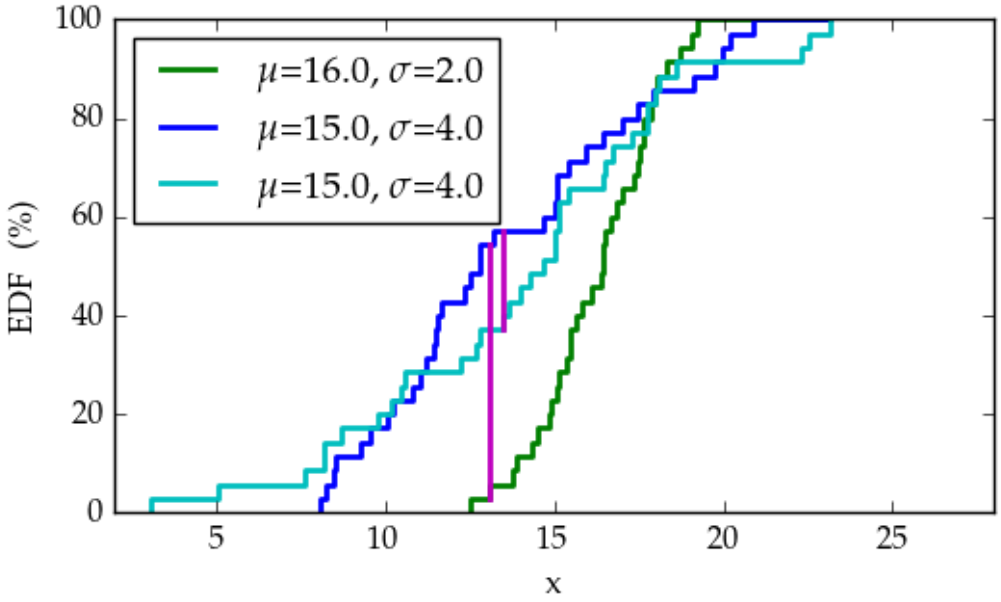

Figure 4.8: KS test parameter values (magenta vertical lines) shown graphically for three samples from figure 4.2.

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

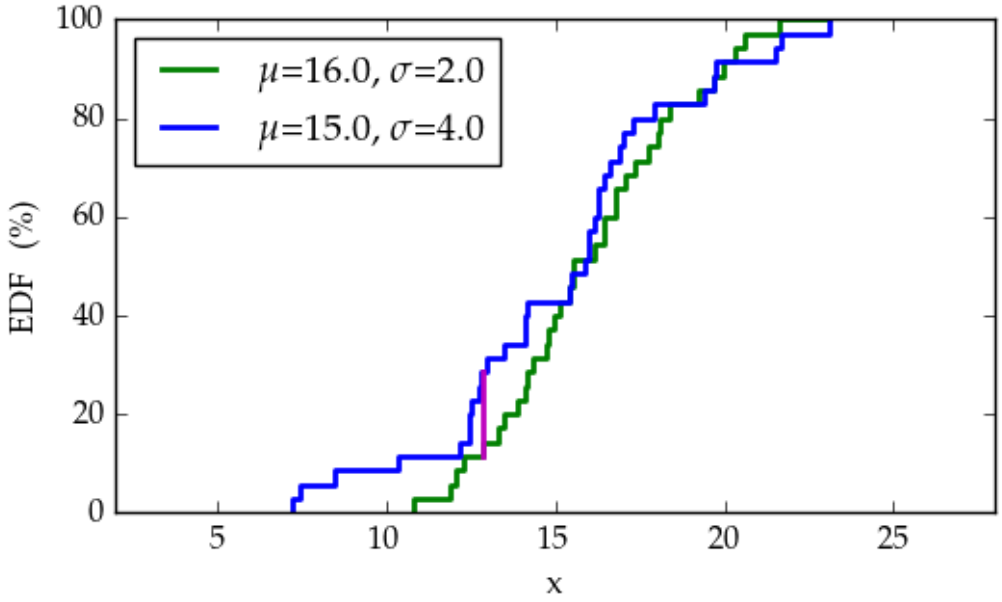

Figure 4.9: KS test ran on another pair of samples drawn from blue and green distributions in figure 4.1.

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 $\alpha = 0.01$.

mistä p-value saadaan, kerro taas aiemmin (tai täällä)

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

"explain better"

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

ehkä vähän lyhyenpuoleisia kappaleita

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)

hmnngh

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.

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

Mieti, pitäisikö kolme viimeistä esittää esim scatterplottina combined mass vs mass in more massive

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,

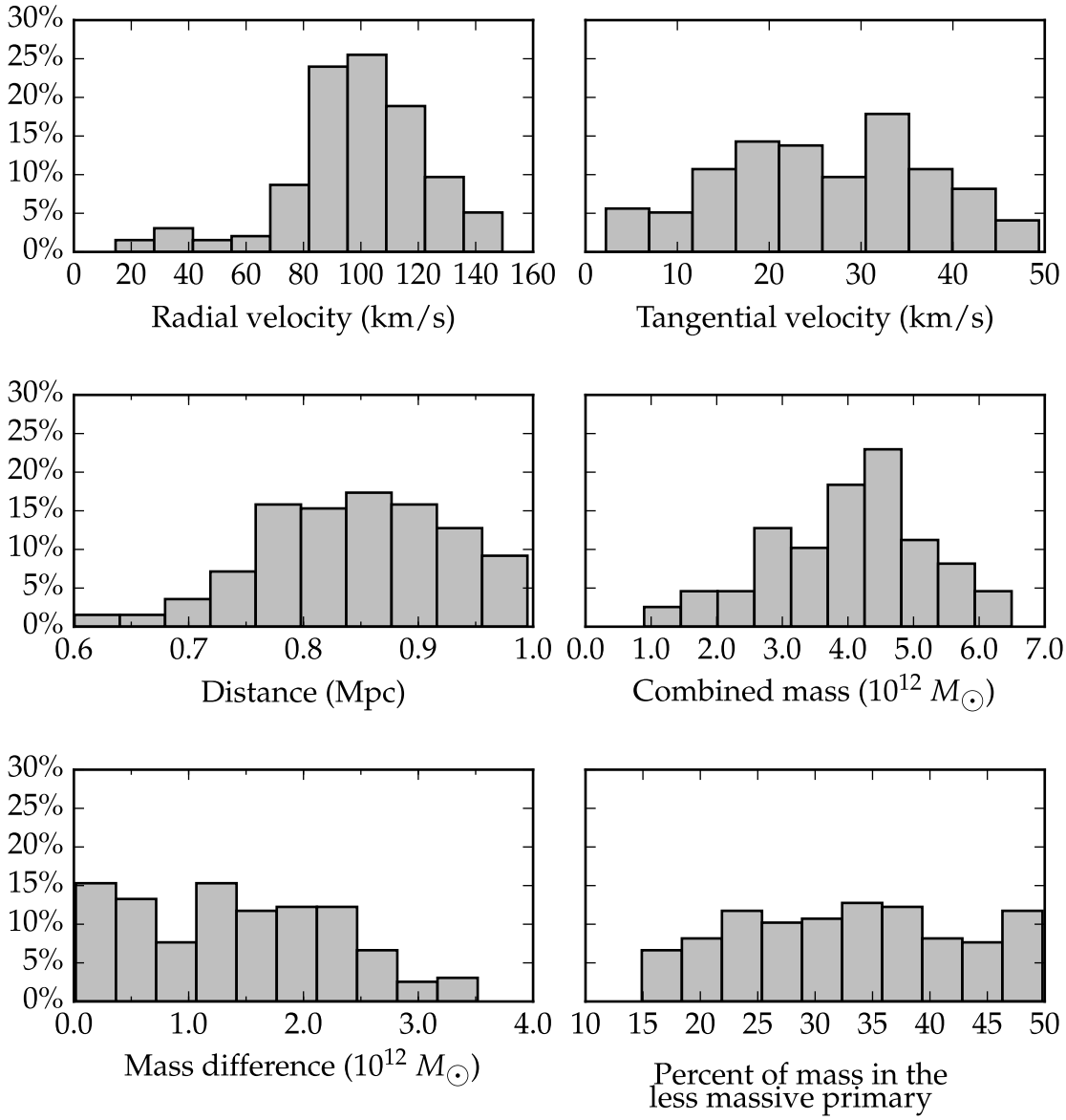

Figure 5.1: Distributions of LG analogue properties. TODO: selitä, mieti miten y-akselin label, binien rajat pitäisi pakottaa suurimpaan ja pienimpään sallittuun arvoon

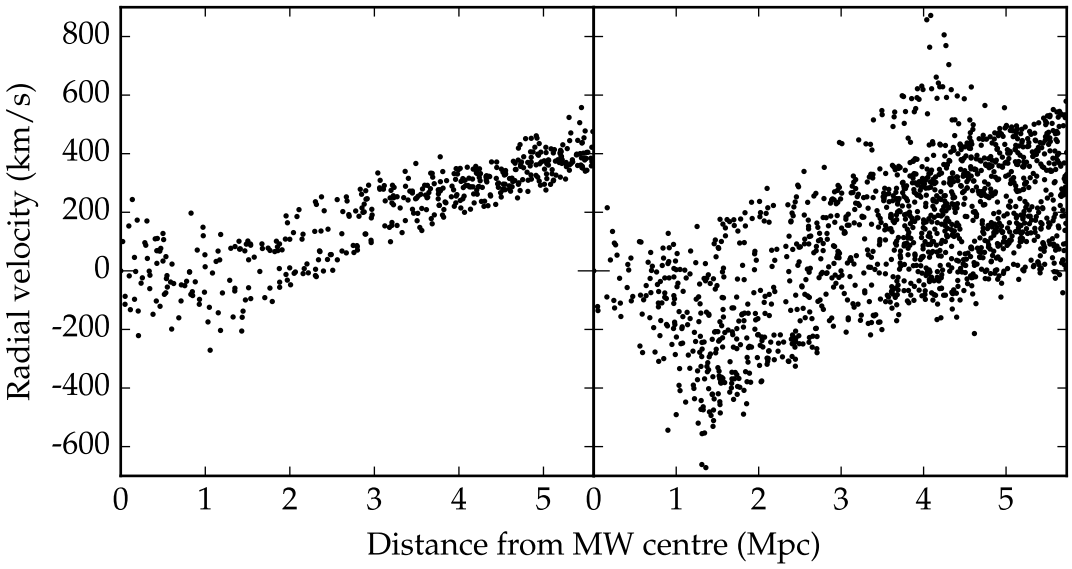

Figure 5.2: Hubble Flows around Milky Way in two simulations.

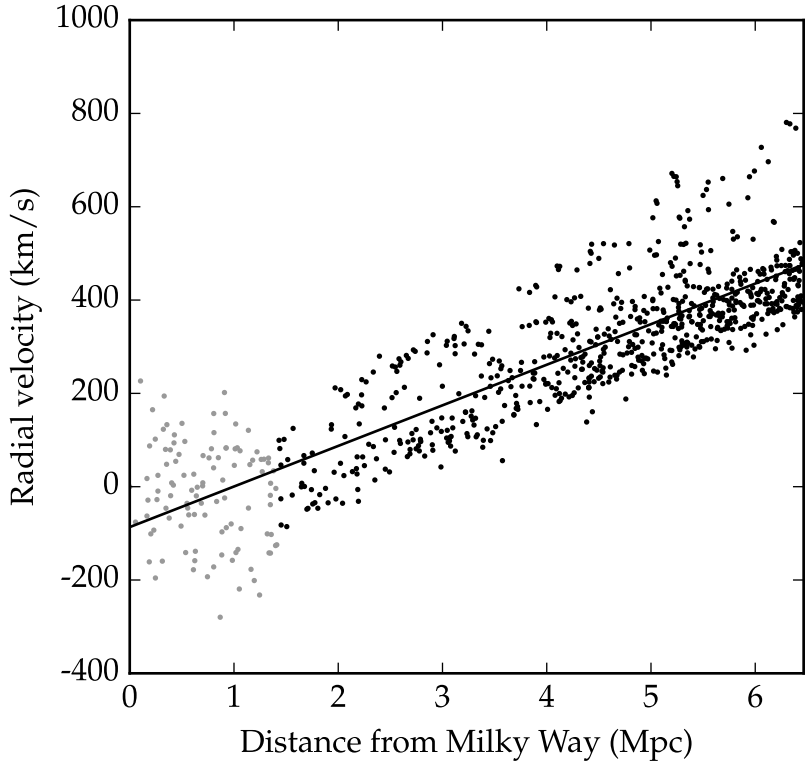

Figure 5.3: HF slope: 86.9929348817

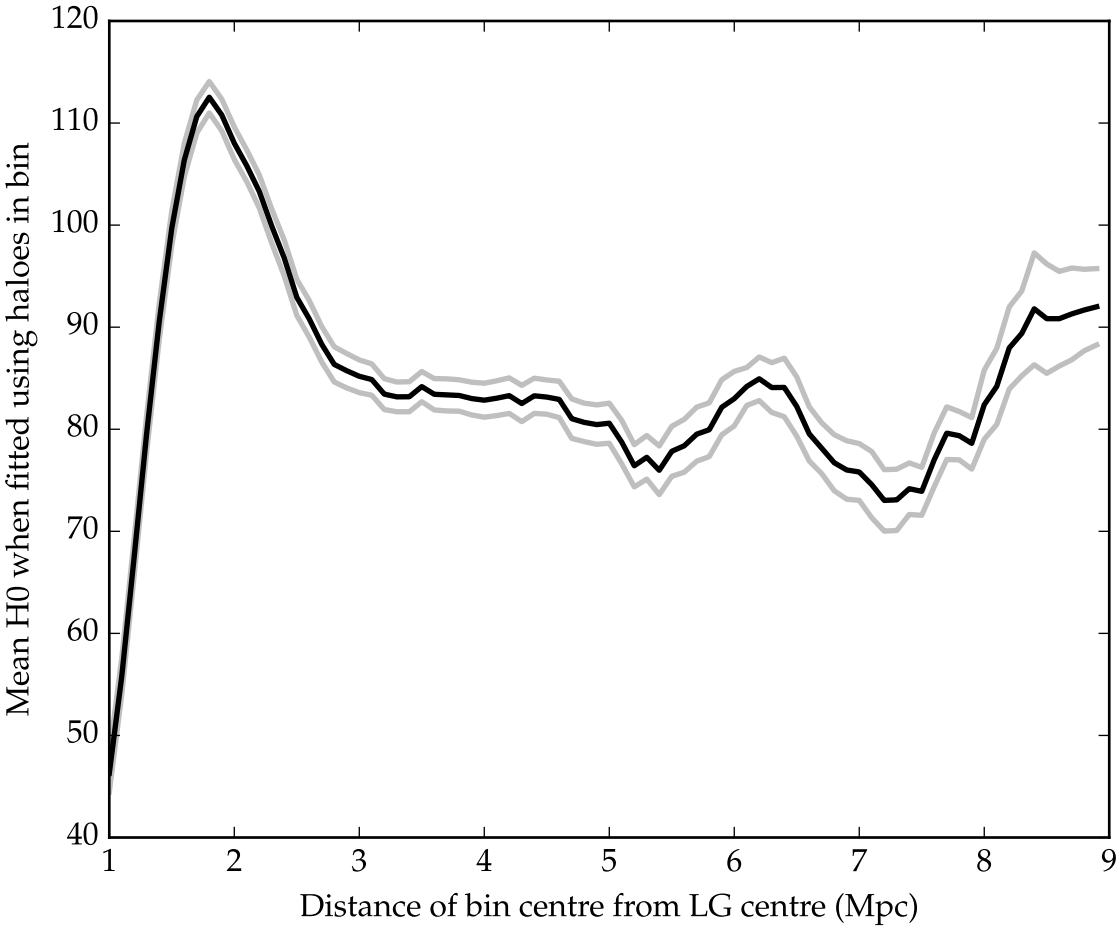

Figure 5.4: Mean H0 in different 2 Mpc bins, grey curves show standard error.

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

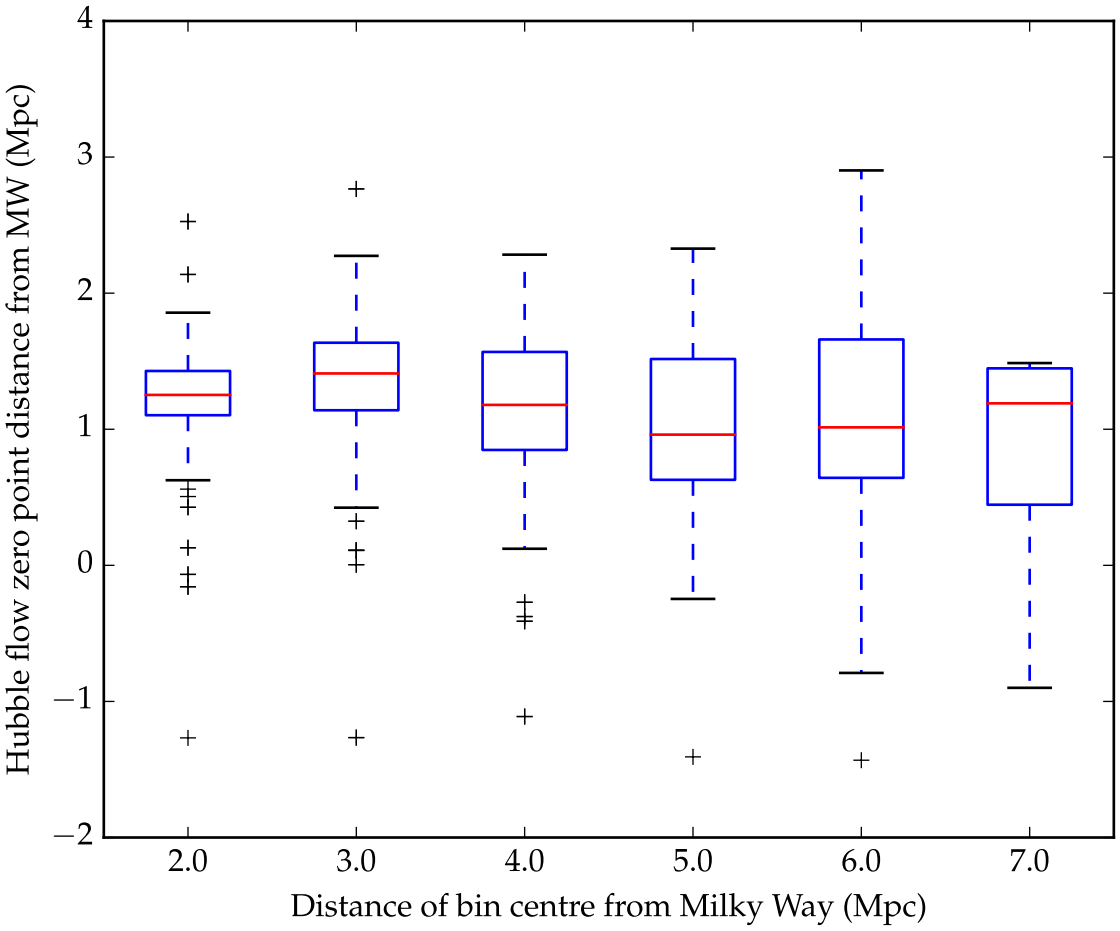

Figure 5.5: HF zero point in different 4 Mpc bins. specify one outlier outside the plot

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

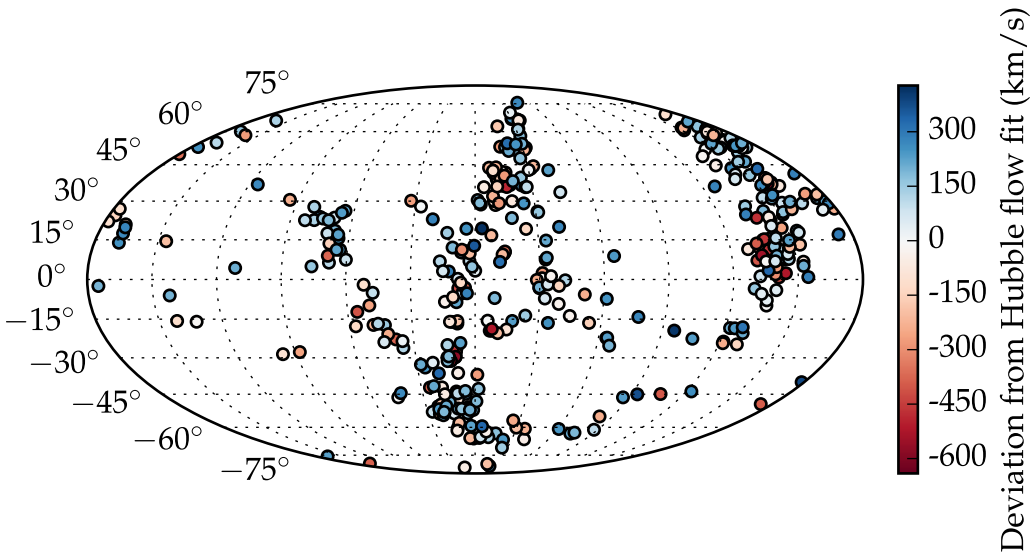

Figure 5.6: Projections of haloes around the less massive LG primary with distances ranging from 1.5 Mpc to 5.0 Mpc.

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 vasemmanpuolim- vähemmän tilaa mainen plottien välissä

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

Mailasdisanako? vaakasuunnassa?

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.

Kaksi eri Keltaiset vähän kynnystä? Liian turhan samanlaisia. kapea ja epätasapainoinen, laita päällekkäin?

5.4 Statistical Estimate of the Local Group Mass

Analysis similar to Fattahi et al 2016 paper

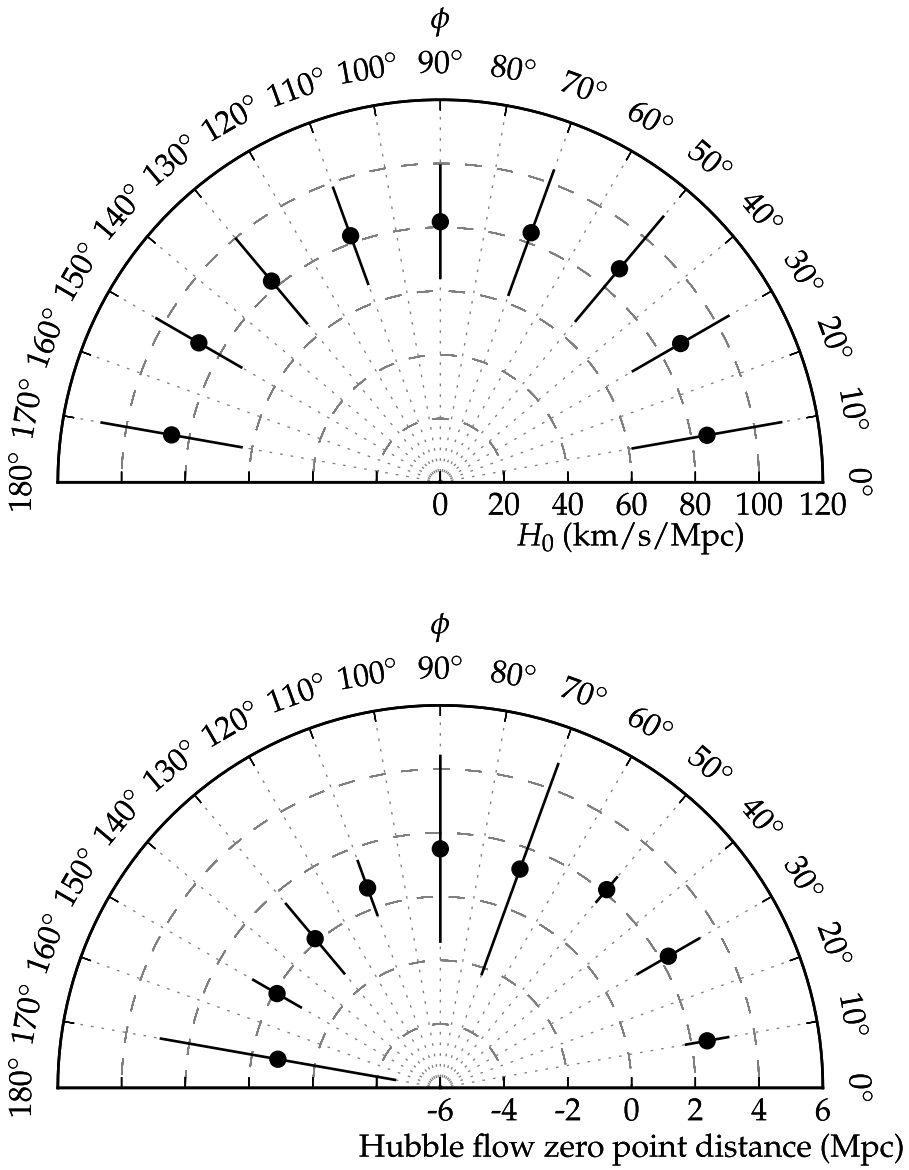

Figure 5.7: Mean Hubble flow slope and zero point as seen from Milky Way analogue in different 20° bins as measured from line connecting Milky Way and Andromeda analogues, direction 0° being towards Andromeda.

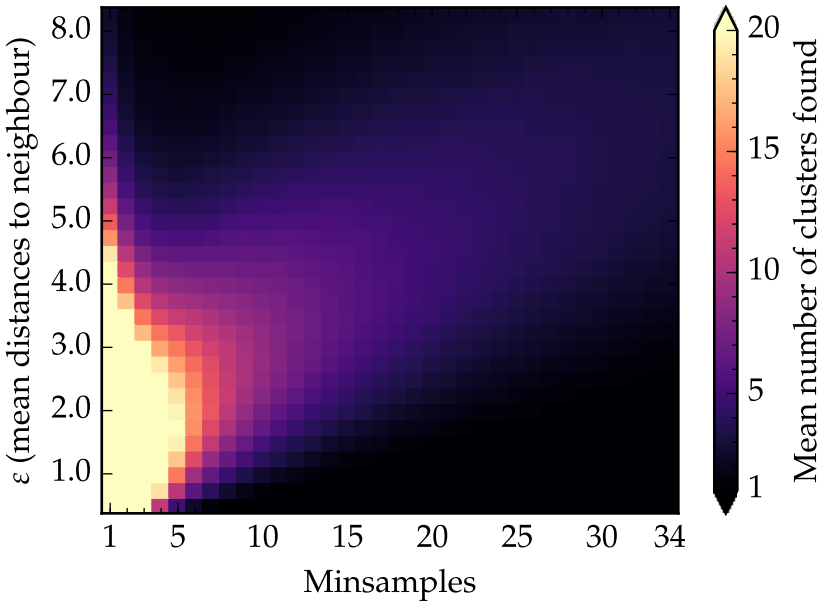

Figure 5.8: Mean number of clusters found for all simulations in dataset with different DBSCAN parameters. In all simulations ε is scaled using the mean distance between closest neighbours.

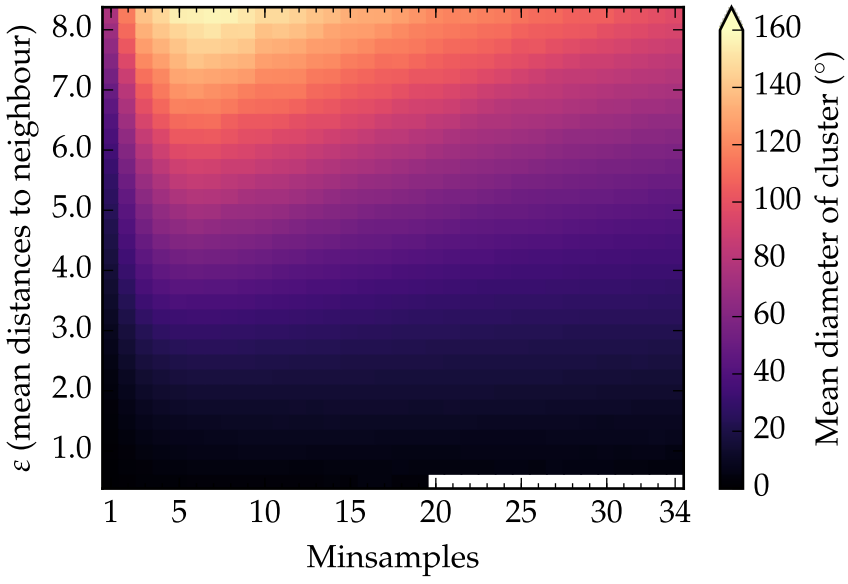

Figure 5.9: Mean diameter of clusters found for all simulations in dataset with different DBSCAN parameters. In all simulations ε is scaled using the mean distance between closest neighbours.

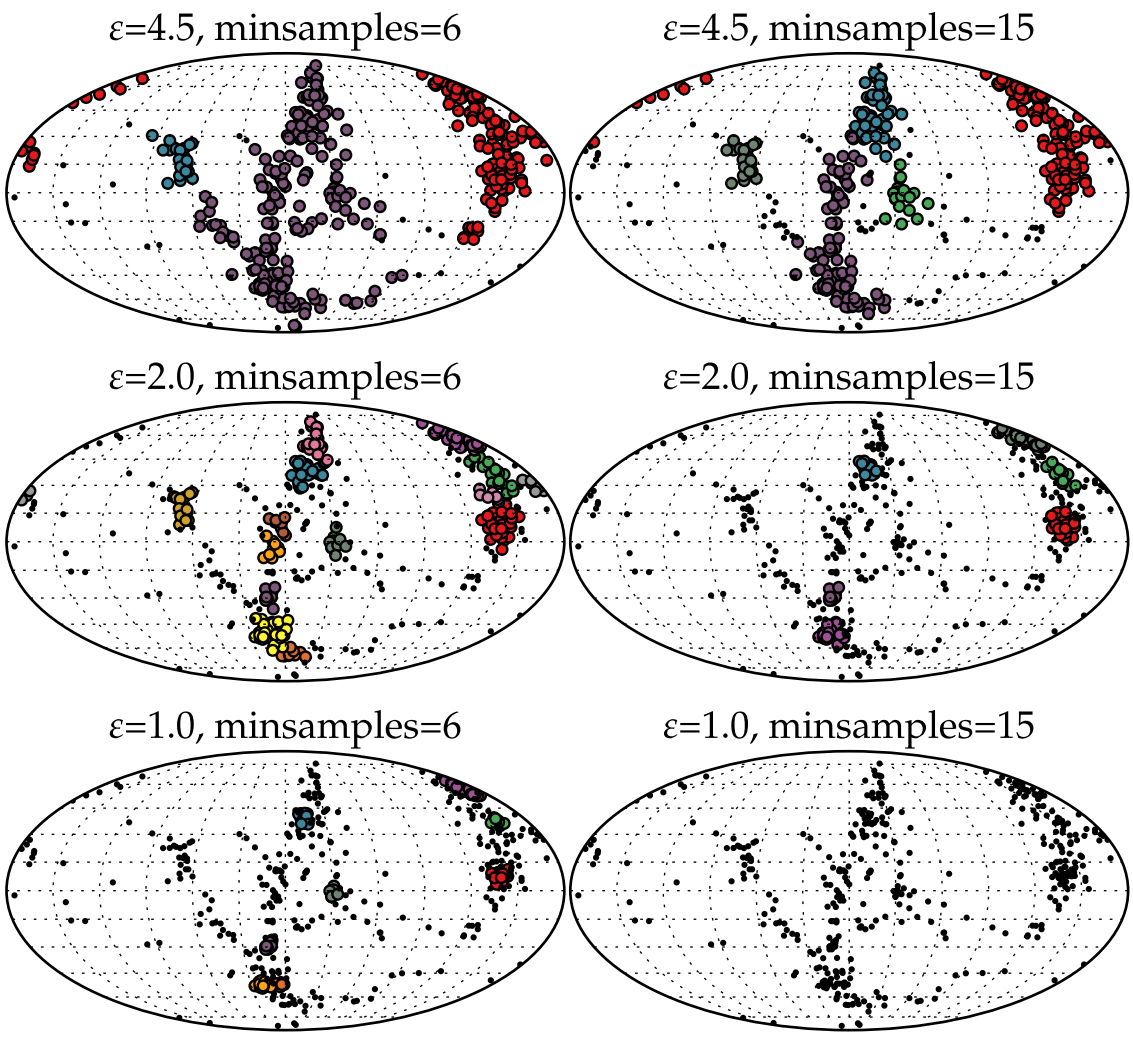

Figure 5.10: Results of DBSCAN clustering on same simulation output with different clustering parameters. TODO: mieti, kuuluuko tämäntyyppinen DBSCANin yleisiä ominaisuuksia esittelevä kuva enemmänkin teoriaosaan. Toisaalta selvästi dataspesifejä juttuja.

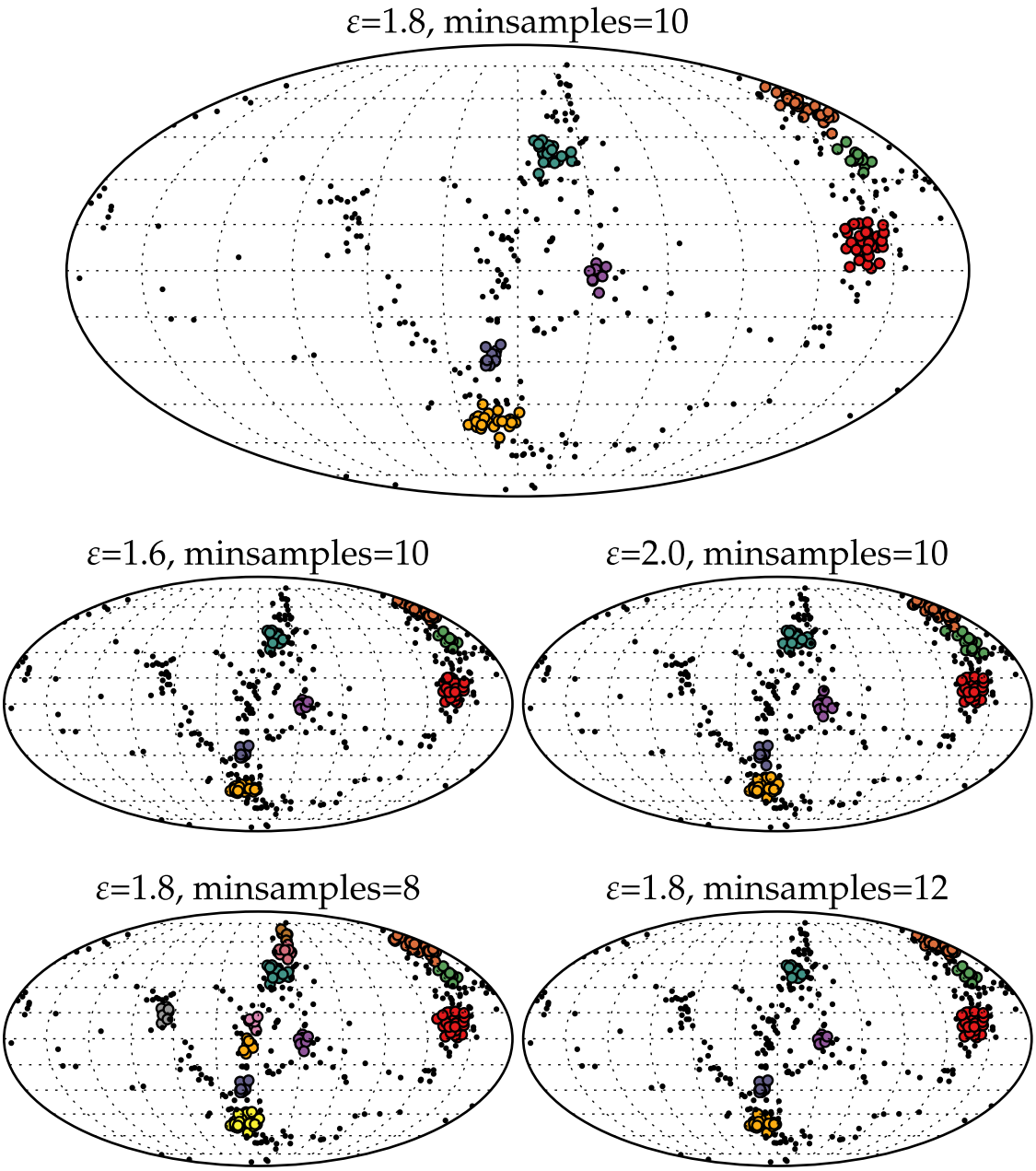

Figure 5.11: The effect of slightly variying the clustering parameters around the values ε=1.8 and minsamples=10 used when analyzing clustered data.

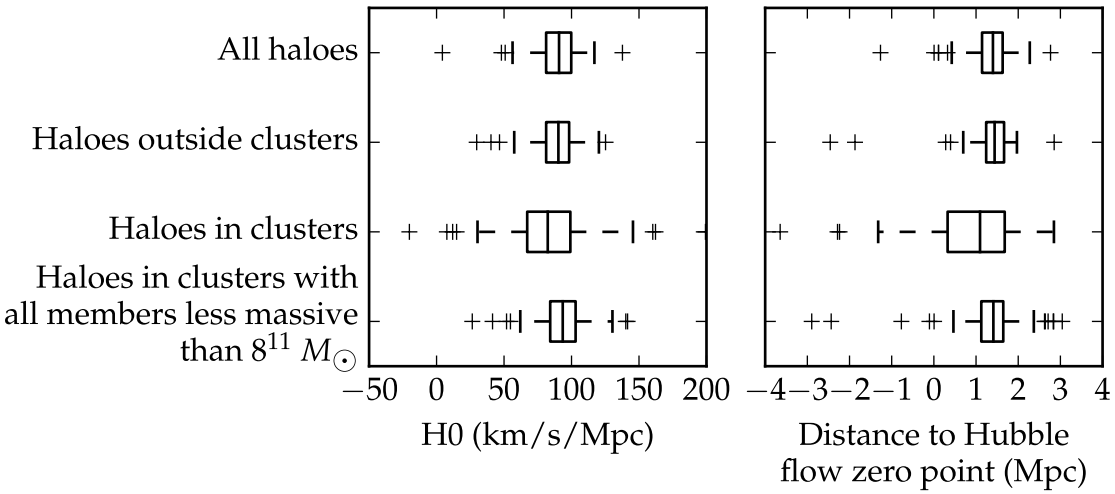

Figure 5.12: Hubble constant and distance to the point at which velocity due to the fitted Hubble flow is zero calculated from different samples. HUOM OBS NB erittele plotin ulkopuolelle jääneet kaukaiset outlierit

6. Conclusions

Bibliography

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.

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.

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

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.

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

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

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.

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.

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