Ludwig-Maximilians-Universität München
Faculty of Physics

Bachelor-Thesis

Self Organizing Maps and Bayesian Inference in cosmology

supervised by:
PD Dr. Torsten Enßlin
Max Planck Institute for Astrophysics

submitted by:
Philipp Frank
10753005
14.10.2015

Ludwig-Maximilians-Universität München
Fakultät für Physik

Bachelorarbeit

Selbstorganisierende Karten und bayesianische Schlussfolgerungen in der Kosmologie

betreut von:
PD Dr. Torsten Enßlin
Max-Planck-Institut für Astrophysik

vorgelegt von:
Philipp Frank
10753005
14.10.2015

Table of Contents

1 Introduction

The field of cosmology has witnessed revolutionary scientific progress over the past few decades. Our picture of the Universe at the largest cosmic scales has been a rather theoretical and speculative construct in the past. Recent high precision observations have provided groundbreaking evidence for the modern theory of cosmology.

A major pillar for this theory is the observation of radiation from the origin of our Universe called Cosmic Microwave Background (CMB), still detectable today. This radiation was first predicted by Gamow (1946) based on the theory of the *Big Bang*, and it was first observed in 1965 by A.A. Penzias and R.W. Wilson of the Bell telephone laboratories (see Kolb & Turner (1988)). The Big Bang theory predicts the Universe to originate from a very dense state. Currently it is believed that initially a mysterious energy field called the inflation drove an exponential expansion. During this phase tiny (10^{-5}) anisotropies in the gravitational field were generated from quantum fluctuations of the inflation. Those tiny anisotropies are observed today in the CMB to have almost Gaussian and scale invariant statistics. The shape of the anisotropies was revealed by detailed measurements of the CMB by the Cosmic Background Explorer (COBE) by Bennett et al. (1996), the Planck satellite by the Planck Collaboration et al. (2015) and a number of ground based telescopes. These fluctuations are the seeds for presently observed galaxies and the large-scale-structure (LSS) of the Universe. Detailed calculations show that gravitational evolution of these seed fluctuations in an expanding background explain the presently observed LSS. The expansion of the Universe, called Hubble expansion, has first been discovered in the 1920s by Edwin Hubble who observed that more distant galaxies appear to move away from us more rapidly than closer galaxies.

In addition, more detailed observations of galaxies and the LSS resulted in a disagreement with the theoretical description of the Universe due to the discrepancies between the mass of galaxies determined from their gravitational interaction, and their mass calculated from their luminosity. A possible solution to resolve this disagreement provides the introduction of an additional matter component in the Universe, called *Dark Matter* (DM). This component is called "dark", since it does not emit light and therefore can only be detected according to its gravitational interaction with baryonic matter. The origin and composition of DM is one of the most outstanding questions of our time.

Furthermore, observations of Supernovae reveal the Universe is currently expanding accelerated. A possible theoretical description to account for this effect is the addition of a cosmological constant Λ to the Einstein equation describing the expansion of space. The other common approach is described by an additional energy component in the Universe called *Dark Energy* (DE). A detailed identification of this DE component in terms of a field or particle within a more fundamental theory is still missing.

The combination of the observational facts and Einstein's theory of general rela-

tivity yield the current model of cosmology: the so-called ΛCDM model of a Universe presently dominated by a cosmological constant Λ (or DE) and DM. The model describes a parametrization of the Big Bang model and is based on the Friedman Robertson Walker (FRW) metric for space-time. The FRW metric is the solution to Einsteins's field equations for a spatially homogeneous and isotropic universe. The model consists of three different components.

The first component is the well known and observable *baryonic matter* from which all stars and planets formed. This matter component is only a tiny fraction ($\sim 5\%$) of the total matter in our Universe. The second and larger ($\sim 27\%$) component is DM. The remaining matter content of the Universe ($\sim 68\%$) is assigned to DE. The evolution of the Universe as described by the ΛCDM model follows at least four different epochs of expansion. The expansion measure of the Universe is often expressed in terms of redshift z labeling different epochs of cosmic history.

The initial phase is dominated by an inflation field governing exponential expansion of our Universe. After this phase the Universe is dominated by a "fluid" of radiation and highly-relativistic matter. At this stage the Universe consists mainly of photons, neutrinos, electrons and other massive relativistic particles such as the DM particles. After some expansion and cooling of the Universe, massive particles become non-relativistic, partly annihilate, but their relics finally dominated over the radiation components (photons, neutrinos) in mass from the *equality* epoch at $z_{\mathrm{eq}} \sim 3200$ onwards. Due to the fact that the main matter component produced effectively no pressure, a dust dominated model for the energy component of the Universe can be assumed for $z_{\mathrm{eq}} \sim 3200 < z < \frac{1}{2} \sim z_{\mathrm{DE}}$. During this epoch of the Universe it is dominated by *cold* dark matter (CDM). From $z_{\mathrm{DE}} \sim \frac{1}{2}$ up to the present epoch, the Universe seems to be accelerating again due to DE.

During these epochs all structures of the present Universe formed from quantum fluctuations generated in the early Universe. Those fluctuations were linearly amplified to macroscopic scales in an expanding background. Due to gravitational collapse and non-linear structure formation processes these seed fluctuations formed todays structure of the Universe. The non-linear regime of structure formation involves many different kinds of physical processes such as thermo- and quantum-dynamics. This renders the field of galaxy formation a very complex field, still far from being understood completely. In order to gain further insights into galaxy formation, the connection between the LSS and observable properties of galaxies appear to be of particular interest.

Therefore large redshift surveys such as the Sloan Digital Sky Survey (York et al. (2000)) have been carried out recently. These surveys can be used to improve our understanding of the LSS. A recently proposed method to reconstruct the LSS from SDSS data is the BORG algorithm presented by Jasche & Wandelt (2013). This algorithm results in a reconstruction of the density field of the Universe in a particular region. Building on those results, we present a method for correlation determination between the LSS and

observed galactic quantities. Our analysis method is based on Bayesian inference and is applied to suitably selected sub-samples of galaxy data which exhibit corrections of galaxy and LSS properties in a clear fashion. These sub-samples are generated by an automatic selection algorithm based on a specific kind of artificial neural network called Self Organizing Map.

2 Correlation determination

In this thesis we seek to find a way to distinguish uncorrelated from correlated multivariate data. This can be achieved by using the relative Shannon Entropy or Kullback Leibler Divergence Barnum et al. (2010). The approach described in this section relies on the joint probability distribution function (PDF) of the quantities of interest. In principle it is possible to determine correlations between arbitrary numbers of quantities but for the sake of this work we focused on the two dimensional case. A future generalization towards the N-dimensional case is straightforward.

In order to determine correlation between two quantities x and y we use the fact that their joint PDF $P(x, y)$ carries additional information. To access this information we use the product rule of probability theory:

$$P(x, y) = P(x)P(y|x) \ . \tag{1}$$

Correlation means a dependency of y on x which results in a conditional PDF $P(y|x) \neq P(y)$. For statistical independence the conditional PDF is equal the the marginalized PDF $P(y|x) = P(y)$. Therefore,

$$P(x, y) = P(x)P(y) \tag{2}$$

for statistical independence and

$$P(x, y) \neq P(x)P(y) \tag{3}$$

for correlation. Since we have to deal with a finite set of data in real applications we present discrete versions of the PDFs in the next section. In addition we describe a common approach to compare the joint PDF $P(x, y)$ and the factorized PDF $P(x)P(y)$.

2.a Discretized estimate of joint probabilities

Given a finite set of data points (x_i, y_i) $i \in (1, ..., N)$ it is possible to present an approximate version of the joint PDF $P(x, y)$. This can be achieved by setting up a two dimensional grid with fixed bin size and counting the numbers of data points per pixel.

This yields:

$$P_{ij} = P(x_i \leq x < x_i + \delta x, y_j \leq y < y_j + \delta y) \approx \frac{N_{ij}}{N} \ , \tag{4}$$

where N_{ij} is the number of points in pixel (i, j). The estimate for the marginal distribution $P(x)$ is obtained by summing up all number counts of each pixel for fixed x_i over all y_i, given as:

$$P_i = P(x_i \leq x < x_i + \delta x) \approx \frac{1}{N} \sum_{j=0}^{N-1} N_{ij} \tag{5}$$

and

$$P_j = P(y_j \leq y < y_j + \delta y) \approx \frac{1}{N} \sum_{i=0}^{N-1} N_{ij} \tag{6}$$

respectively.

In order to compare the discretized PDFs we present a discrete version of the relative entropy in the next section.

2.b Relative Shannon entropy (Kullback Leibler Divergence)

The Shannon entropy for the probability distribution of a random variable X is defined as:

$$H[P(X)] := \int P(X) \ \ln(P(X)) \ \mathrm{d}X. \tag{7}$$

Analogously the relative entropy between two distributions is defined as:

$$H[P(X)||Q(X)] := \int P(X) \ \ln\left(\frac{P(X)}{Q(X)}\right) \ \mathrm{d}X, \tag{8}$$

where $Q(X)$ is a reference distribution and $H[P(X)||Q(X)]$ describes the relative gain of information by updating from $Q(X)$ to $P(X)$.

Plugging the joint PDF $P(x, y)$ as $P(X)$ with $X = (x, y)$, and the product of the marginalized PDFs $P(x)P(y)$ as $Q(X)$ into Eq. (8) leads to

$$H[P(x,y)||P(x)P(y)] = \int\int P(x,y) \ \ln\left(\frac{P(x,y)}{P(x)P(y)}\right) \ \mathrm{d}x \ \mathrm{d}y \ . \tag{9}$$

$H[P(x, y)||P(x)P(y)]$ is larger than zero for correlated quantities. The gain of information by updating from $P(x)P(y)$ to $P(x, y)$ leads to a positive divergence. For statistical independence there is no gain of information since $P(x, y) = P(x)P(y)$ (Eq. (2)) leads to

$$H[P(x,y)||P(x)P(y)] = H[P(x)P(y)||P(x)P(y)] = \int\int P(x,y) \ \ln(1) \ \mathrm{d}x \ \mathrm{d}y = 0. \tag{10}$$

As described above for a finite set of parameters x_i and y_i the joint PDF can be estimated on a discrete lattice. In this case we obtain a discretized version of the relative

entropy,

$$H[P(x,y)||P(x)P(y)] \approx \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{ij} \ln\left(\frac{P_{ij}}{P_i P_j}\right) \geq 0. \tag{11}$$

A value $H[P(x,y)||P(x)P(y)] \approx 0$ consequently signals statistical independence.

Theoretically, the approach described above is a simple estimator for correlation, since it is easy to implement. Fig. 1 shows how the method performs on artificial data. The results show that the relative entropy is higher for correlated structures. The entropy for uncorrelated data is zero within numerical limits. The strength of this method relies on its unparameterized estimate for correlation between x and y. Therefore it is applicable in more generic cases than a parametrized model. Unfortunately this entropy-based method is more sensitive to noise than parametrized models. As the noise increases, the entropy decreases until it is approximately zero although there might still be correlations. The only way to increase the empirical relative entropy in the presence of noise is to obtain more data. However, this may be either expensive or not possible at all. Therefore we propose a different way of correlation determination in the following, on the basis of a parametrized model.

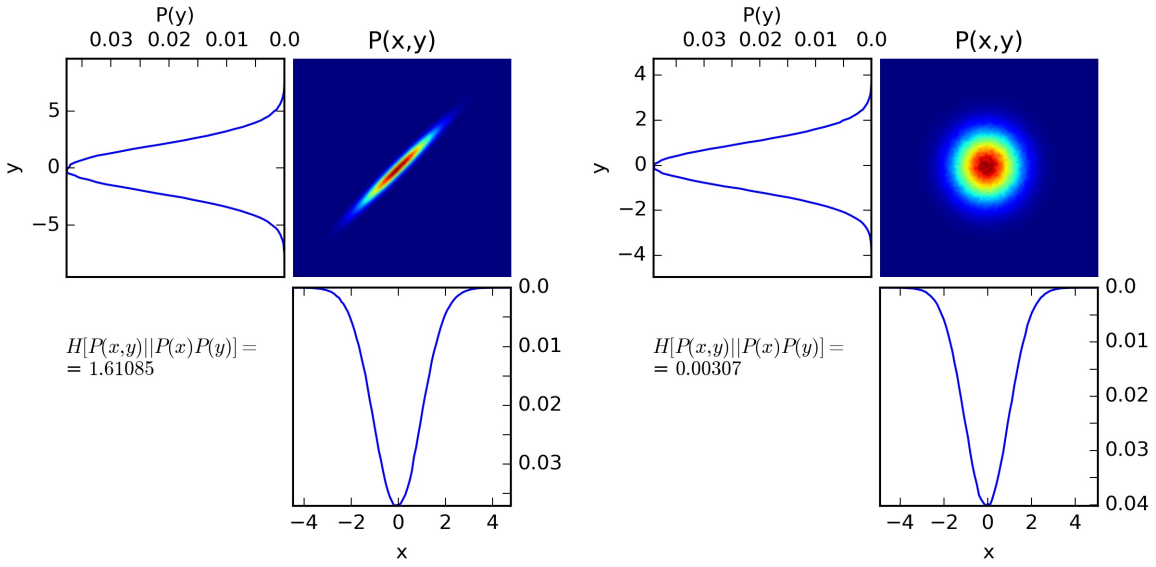

Figure 1: The discretized estimate for the relative entropy. Both figures are generated for a fictional data set. In the right picture x and y are set up to be independent and normal distributed. In the left picture correlation between x and y is set up as: $y = 2x + n$, with x being normal distributed. n is random normal distributed noise with a noise covariance of $\sigma_n = \frac{1}{\sqrt{2}}\sigma_x$ where σ_x is the covariance of x. In both cases we generated 1000000 data points. The relative entropy was calculated according to Eq. (11).

3 A simple parametric model

3.a Theoretical framework

In the previous chapter we discussed the possibility to detect correlations in datasets via parameter-free methods. However, these methods suffer from noisy estimates of the joint distribution which is approximated on a equidistant lattice of N grid nodes. The large number of grid nodes is a particular problem if there is only a very limited amount of measurements available. To counter this problem we propose a simple parametric model to test for correlations. More specifically we assume the relation between x and y can be written as:

$$y = f(x) + n \tag{12}$$

where f is some arbitrary unknown function and n is assumed to be uncorrelated, normally distributed noise. The underlying assumption of this relation is usually that x has a causal impact on y. If it were the other way around, x and y should change roles. In this work, we will often assign y to be the more noisy quantity irrespective of the causal structure. The mild restriction on the generic form of the correlation already helps to regularize the impact of noise significantly, as will be demonstrated.

If f is continuously differentiable then it can be expanded in a Taylor series up to Mth order and equation (12) yields:

$$y \approx \sum_{i=0}^{M} f_i x^i + n \ . \tag{13}$$

Determination of correlations therefore requires to determine optimal coefficients f_i for a given set of data points $(x_i, y_i), i \in [1, ..., U]$. Eq. (13) should hold for every point of the given data and therefore gives a relation for each data point. Combining these N relations into a vector relation by defining vectors $\mathbf{y} := (y_1, y_2, ..., y_U)^T$ and $\mathbf{f} := (f_0, f_1, ..., f_M)^T$ yields:

$$\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_U \end{pmatrix} = \mathbf{y} = \mathbf{Rf} + \mathbf{n} = \begin{pmatrix} x_1^0 & x_1^1 & x_1^2 & ... & x_1^M \\ ... & ... & ... & ... & ... \\ x_U^0 & x_U^1 & x_U^2 & ... & x_U^M \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ ... \\ f_M \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \\ ... \\ n_U \end{pmatrix} \ . \tag{14}$$

Without further knowledge about the noise we assume n to obey Gaussian statistics with zero mean and diagonal covariance. This indicates the noise of individual data points to be uncorrelated. We further add the restriction that each n_i has the same variance p. This is reasonable if there are no locally varying uncertainties in the data space. Therefore the

probability distribution for n is set up as:

$$P(\mathbf{n}|\mathbf{N}) = \mathcal{G}(\mathbf{n}, \mathbf{N}) := \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{n}^T\mathbf{N}^{-1}\mathbf{n}} \tag{15}$$

where $N_{ij} = p\ \delta_{ij}$ and $|\mathbf{N}|$ denotes the determinant of $\mathbf{N}$ which is simply $|\mathbf{N}| = p^U$. Since it is required that $p \geq 0$, it can be parametrized as $p := e^{\eta}$, where the unknown constant $\eta \in \mathbb{R}$ needs to be inferred from the data.

The resulting probability distribution for $\mathbf{f}$ given the data $\mathbf{d}$ and the noise parameter η can be expressed in terms of the joint probability of all these quantities:

$$P(\mathbf{f}|\mathbf{d}, \eta) = \frac{P(\mathbf{f}, \mathbf{d}, \eta)}{P(\mathbf{d}, \eta)} = \frac{P(\mathbf{f}, \mathbf{d}, \eta)}{P(\mathbf{d}|\eta)P(\eta)} \tag{16}$$

The Prior distribution for η is assumed to be flat because there is no information which would make a choice of certain values for this quantity more plausible. Therefore (16) leads to:

$$P(\mathbf{f}|\mathbf{d}, \eta) \propto P(\mathbf{f}, \mathbf{d}|\eta) \tag{17}$$

up to a constant factor which depends on η and $\mathbf{d}$, but not on $\mathbf{f}$. The constant of proportionality will be accounted for once we normalize the distribution.

The joint probability of $\mathbf{f}$,$\mathbf{d}$ and η can be obtained by marginalisation over $\mathbf{n}$ and use of the data model given in Eq.(14):

$$\begin{aligned} P(\mathbf{f}, \mathbf{d}, \eta) = \int P(\mathbf{f}, \mathbf{d}, \eta, \mathbf{n})\ \mathrm{d}\mathbf{n} \quad = \\ \int P(\mathbf{d}|\mathbf{f}, \eta, \mathbf{n})\ P(\mathbf{f})\ P(\mathbf{n}|\eta)\ P(\eta)\ \mathrm{d}\mathbf{n} \propto \int \delta^D(\mathbf{y} - (\mathbf{R}\mathbf{f} + \mathbf{n}))\ \mathcal{G}(\mathbf{n}, \mathbf{N})\ \mathrm{d}\mathbf{n} \quad = \\ \mathcal{G}(\mathbf{y} - \mathbf{R}\mathbf{f}, \mathbf{N}) = \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{R}\mathbf{f})^T\mathbf{N}^{-1}(\mathbf{y}-\mathbf{R}\mathbf{f})}\ . \end{aligned} \tag{18}$$

We further assume the prior on $\mathbf{f}$ to be flat to permit $\mathbf{f}$ to model an arbitrary polynomial of order M. Using completion of the square in the exponent, Eq. (18) can be written as:

$$P(\mathbf{f}, \mathbf{d}, \eta) \propto \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{N}^{-1}\mathbf{y} - \mathbf{j}^T\mathbf{D}\mathbf{j})} e^{-\frac{1}{2}(\mathbf{f}-\mathbf{D}\mathbf{j})^T\mathbf{D}^{-1}(\mathbf{f}-\mathbf{D}\mathbf{j})} \tag{19}$$

with $\mathbf{D} = (\mathbf{R}^T\mathbf{N}^{-1}\mathbf{R})^{-1}$ and $\mathbf{j} = \mathbf{R}^T\mathbf{N}^{-1}\mathbf{y}$. Note that the second exponential function is a Gaussian distribution in $\mathbf{f}$ with mean $\mathbf{D}\mathbf{j}$ and covariance $\mathbf{D}$. If η is known then the proper probability distribution of $\mathbf{f}$ given $\mathbf{d}$ is obtained from (19) by normalization.

An estimate for η can be obtained via a maximum a posteriori (MAP) approach of the marginal probability distribution for η given $\mathbf{d}$. This distribution is obtained by

marginalizing Eq. (19) with respect to $\mathbf{f}$:

$$\begin{aligned} P(\eta|\mathbf{d}) \propto P(\mathbf{d},\eta) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &= \\ \int P(\mathbf{f},\mathbf{d},\eta)\ \mathrm{d}\mathbf{f} = \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}}\ e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}-\mathbf{j}^T\mathbf{D}\mathbf{j})} \int e^{-\frac{1}{2}(\mathbf{f}-\mathbf{D}\mathbf{j})^T\mathbf{D}^{-1}(\mathbf{f}-\mathbf{D}\mathbf{j})}\ \mathrm{d}\mathbf{f} \qquad &= \\ = \frac{|2\pi\mathbf{D}|^{\frac{1}{2}}}{|2\pi\mathbf{N}|^{\frac{1}{2}}}\ e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{N}^{-1}\mathbf{y}-\mathbf{j}^T\mathbf{D}\mathbf{j})} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &(20) \end{aligned}$$

and the negative logarithm of this distribution is:

$$\begin{aligned} \mathcal{H}(\eta|\mathbf{d}) = -\ln(P(\eta|\mathbf{d}) = \frac{1}{2}(\ln(|2\pi\mathbf{N}|) - \ln(|2\pi\mathbf{D}|) + \mathbf{y}^T\mathbf{N}^{-1}\mathbf{y} - \mathbf{j}^T\mathbf{D}\mathbf{j} + \tilde{H}_0 \qquad &= \\ = \frac{1}{2}((U-(M+1))\eta + e^{-\eta}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{y})) + H_0 \qquad\qquad &(21) \end{aligned}$$

where in the following we call $\mathcal{H}(\eta|\mathbf{d})$ a Hamiltonian. Here we used the definitions of D and j and the fact that N is diagonal. Note that $M+1$ is the dimensionality of the signal space and U the dimensionality of the data space. H_0 and $\tilde{H}_0$ are terms independent of η. The MAP solution for η is then given by setting the first derivative of $\mathcal{H}(\eta|\mathbf{d})$ to zero:

$$\frac{\partial\mathcal{H}(\eta|\mathbf{d})}{\partial\eta} = \frac{1}{2}((U-(M+1)) - e^{-\eta}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{y})) \stackrel{!}{=} 0 \qquad (22)$$

and therefore

$$p_{\mathrm{MAP}} = e^{\eta}_{\mathrm{MAP}} = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{y}}{U-(M+1)} \qquad (23)$$

Given these results we obtain the conditional PDF for $\mathbf{f}$ as:

$$P(\mathbf{f}|\mathbf{d},\eta) = \mathcal{G}(\mathbf{f}-\mathbf{D}\mathbf{j},\mathbf{D}) = \frac{1}{|2\pi\mathbf{D}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{f}-\mathbf{D}\mathbf{j})^T\mathbf{D}^{-1}(\mathbf{f}-\mathbf{D}\mathbf{j})}, \qquad (24)$$

where we ignored factors independent on $\mathbf{f}$ since the distribution is now properly normalized. Mean and covariance of this distribution are then given by:

$$\mathbf{f}_{\mathrm{WF}} = \langle\mathbf{f}\rangle_{(\mathbf{f}|\mathbf{d},\eta)} = \mathbf{D}\mathbf{j} = (\mathbf{R}^T\mathbf{N}^{-1}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{N}^{-1}\mathbf{y} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{y} \qquad (25)$$

and

$$\mathbf{D} = \left\langle(\mathbf{f}-\langle\mathbf{f}\rangle)(\mathbf{f}-\langle\mathbf{f}\rangle)^T\right\rangle_{(\mathbf{f}|\mathbf{d},\eta)} = (\mathbf{R}^T\mathbf{N}^{-1}\mathbf{R})^{-1} = \frac{1}{p_{\mathrm{MAP}}}(\mathbf{R}^T\mathbf{R})^{-1}. \qquad (26)$$

These equations resemble the solution of the famous Wiener filtering equation. Note that the mean of $\mathbf{f}$ does not depend on p_{MAP} since the noise is assumed to be zero centered and the prior for $\mathbf{f}$ is flat. This method determines the full a posteriori probability distribution for the coefficients $\mathbf{f}$. For visualization the posterior of $\mathbf{f}$ (Eq. (24)) can be transformed into data space resulting in a PDF for the realizations of the correlation function $f(x)$.

The mean $\langle f(x) \rangle$ is derived as:

$$\bar{f}(x) = \langle f(x) \rangle = \tilde{\mathbf{R}}(x) \langle \mathbf{f} \rangle = \sum_{i=0}^{M} x^i \langle \mathbf{f}_i \rangle = \begin{pmatrix} 1 & x & x^2 & ... & x^M \end{pmatrix} \begin{pmatrix} \langle \mathbf{f}_0 \rangle \\ \langle \mathbf{f}_1 \rangle \\ ... \\ \langle \mathbf{f}_M \rangle \end{pmatrix} \tag{27}$$

with $x \in \mathbb{R}$ and $\tilde{\mathbf{R}}(x) : \mathbb{R}^{M+1} \to \mathbb{R}$. $\tilde{\mathbf{R}}$ has the same structure as $\mathbf{R}$ but the finite dimensional part of the operator, corresponding to the data points x_i, has been replaced by an infinite dimensional part for $x \in \mathbb{R}$.

Analogously we obtain the covariance $\mathbf{Y}$ as:

$$\begin{aligned} \mathbf{Y}_{xy} &= \left\langle (f(x) - \bar{f}(x))(f(y) - \bar{f}(y))^T \right\rangle \\ &= \left\langle (\tilde{\mathbf{R}}(x)\mathbf{f} - \tilde{\mathbf{R}}(x)\mathbf{f}_{\mathrm{WF}})(\tilde{\mathbf{R}}(y)\mathbf{f} - \tilde{\mathbf{R}}(y)\mathbf{f}_{\mathrm{WF}})^T \right\rangle \\ &= \tilde{\mathbf{R}}(x) \left\langle (\mathbf{f} - \mathbf{f}_{\mathrm{WF}})(\mathbf{f} - \mathbf{f}_{\mathrm{WF}})^T \right\rangle \tilde{\mathbf{R}}(y)^T \\ &= \tilde{\mathbf{R}}(x) \mathbf{D} \tilde{\mathbf{R}}(y)^T \\ &= \frac{1}{p_{\mathrm{MAP}}} \tilde{\mathbf{R}}(x) (\mathbf{R}^T \mathbf{R})^{-1} \tilde{\mathbf{R}}(y)^T \end{aligned} \tag{28}$$

Combining these results yields a PDF for the possible realizations of the fitted curve

$$P(f(x)|\mathbf{d}) = \mathcal{G}(f(x) - \tilde{\mathbf{R}}(x)\mathbf{f}_{WF}, \mathbf{Y}) \ , \tag{29}$$

which describes how likely a realization is, given the data. This permits to visualize the fitted curve including corresponding uncertainties in specific areas of the data space. For an illustration see Fig. 2. Here we applied the method on fictional data and compared the reconstruction to the original signal.

The Bayesian implementation of this method has the advantage that it is able to find correlation in more noisy data, compared to the cross-entropy-based model free approach described in the previous chapter. In addition the Wiener filter models a posterior PDF for correlation structures and therefore infers uncertainties more precisely. On the other hand, a successful application of this method needs additional information about the data generation process. For optimal reconstructions the order M of the polynomial describing the signal correlation needs to be known. In contrast, for real data application the underlying model is often not known. Especially in fields where the physical processes causing correlation are not yet understood completely, it is important to have a method which does not need to know the data model in the beginning. Therefore possible ways to infer generation processes from data are described in the next sections.

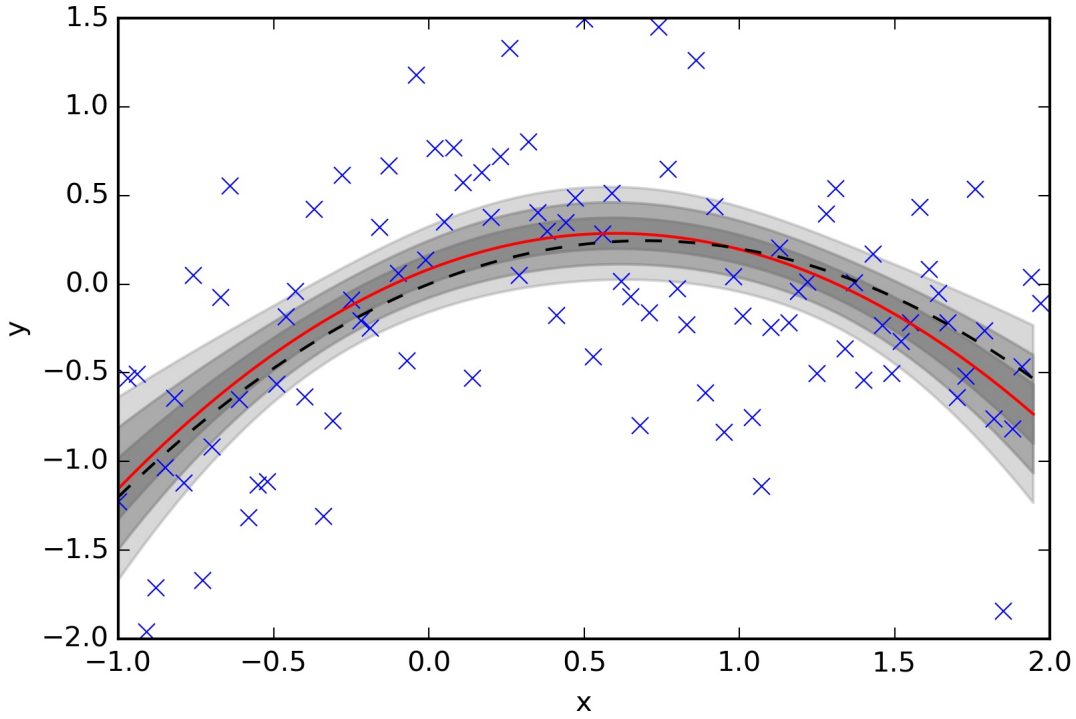

Figure 2: The application of our method to an artificial data set. Blue points correspond to the dataset on which the fit was performed. The data was defined by adding Gaussian noise to a subsample of the original curve which is the black-dashed one in this picture. The red line is the best fit using a data model with $M = 2$ for the given dataset. The gray areas indicate one, two and three times the variance $\sigma = \sqrt{Y}$.

3.b Model selection

Given a specific data model with known deterministic structure between two quantities x, y the parametric method described above is able to determine coefficients of the model constrained by data. However, in practice such data models are often not available, rendering estimations of correlations a challenging task. Especially for noisy data it is a problem to decide which model to prefer.

As an example one could always choose the data model to be a high order polynomial. In principle, fitting this model to data, exhibiting only linear correlation, should yield vanishing coefficients for non-linear terms and appropriate inference of the correlation would still be possible. However using this setup might encounter the following problems: some part of the data has to be used to set the higher order coefficients to zero and therefore less data remains to estimate the relevant coefficients. Therefore the estimated parameters may be poorly conditioned. Another problem occurs for noisy data with a correlation structure of low order in x. If the order of the reconstructed polynomial can be arbitrary large the algorithm regards higher order polynomials to be more likely as these can adopt to any feature in the data, even if this is only caused by noise. This effect

is frequently referred to as "overfitting" and is described in further detail in Section 3.b.1.

To overcome these problems we require that the employed data model to resemble the actual data generation process as close as possible. The optimal approach would be to derive the data model. This requires complete knowledge of the physical theory which explains the correlation function as well as knowledge of the data generation process. But this approach is in contrast to the goal of this thesis to build generic uniform methods. Since the methods we develop should aim at finding trends in data which was generated from unknown processes, we will put forward an approach to estimate the data model from the data directly in the next section.

3.b.1 A posterior model determination

In this section we describe an approach to determine the preferred model from data. In a general framework there might be no prior information about the data model at hand. Therefore we perform a posterior model selection. More precisely, we apply the method described in section 3.a to data for polynomials of different order and compare the likelihood of all models after the fit. To do so, we need to exclude the complication of the data being generated from a superposition of two or more data generation processes with different correlation structure. Therefore data that consists of several components has to be filtered and sub-divided into different groups. Details on this approach will be further discussed in Chapters 5.b and 6.

Even within such a single group we need to identify the preferred model. Instead of comparing the full PDF of the likelihood we restrict the discussion to a comparison of the maximum of the likelihood as a proxy. Specifically, we compare negative logarithm of the maximum for reasons that become clear later. This denotes the minimum of the Hamiltonian $\mathcal{H}(d|\mathbf{f},\eta)$ leading to

$$\mathcal{H}_{\min} := \mathcal{H}(d|\mathbf{f}_{\mathrm{WF}},\eta_{\mathrm{MAP}}) = -\ln(P(\mathbf{d}|\mathbf{f}_{\mathrm{WF}},\eta_{\mathrm{MAP}}))\ . \tag{30}$$

For the parametric model with unknown noise covariance (Chapter 3.a) $\mathcal{H}_{min}$ becomes:

$$\begin{aligned}\mathcal{H}_{\min} &= -\ln(P(\mathbf{d}|\mathbf{f}_{\mathrm{WF}},\eta_{\mathrm{MAP}})) \propto -\ln(P(\mathbf{d},\mathbf{f}_{\mathrm{WF}},\eta_{\mathrm{MAP}}))\\ &= -\ln(\mathcal{G}(\mathbf{y}-\mathbf{R}\mathbf{f}_{\mathrm{WF}},\mathbf{N}(\eta_{\mathrm{MAP}})))\end{aligned} \tag{31}$$

using the fact that the prior for $\mathbf{f}$ and η is assumed to be flat. Plugging in the definition of the joint probability Eq. (18) leads to:

$$\mathcal{H}_{\min} = \frac{1}{2}\frac{1}{p_{\mathrm{MAP}}}(\mathbf{y}-\mathbf{R}\mathbf{f}_{WF})^T(\mathbf{y}-\mathbf{R}\mathbf{f}_{WF}) + \frac{1}{2}U\ln(p_{\mathrm{MAP}}) \tag{32}$$

where U denotes the dimension of the data space.

It is straightforward to see that the maximum likelihood shows the adequacy of chosen models in light of the data: If the model explains the data well, the mean realization of the model $f(x) = \mathbf{R}\mathbf{f}_{WF}$ is close to the $\mathbf{y}$ values of the data. Therefore the likelihood becomes a narrow Gaussian and its maximum value is higher compared to a broader distribution of a less adequate model. Therefore $\mathcal{H}(\mathbf{d}|\mathbf{f}, \eta)$ minimizes for the best fitting model. Figure 3 underlines how important the selection of the appropriate model is. A lower order polynomial might not have the quality to reconstruct all structures of the signal. On the other hand a high order polynomial adds additional structure to the reconstruction although the fitted features might be noise. Fig. 5 shows the effect of "over-fitting" more precisely.

To be robust against over-fitting, the selection process should be able to decide how much structure can be extracted from the dataset. We propose a possible solution in the next section.

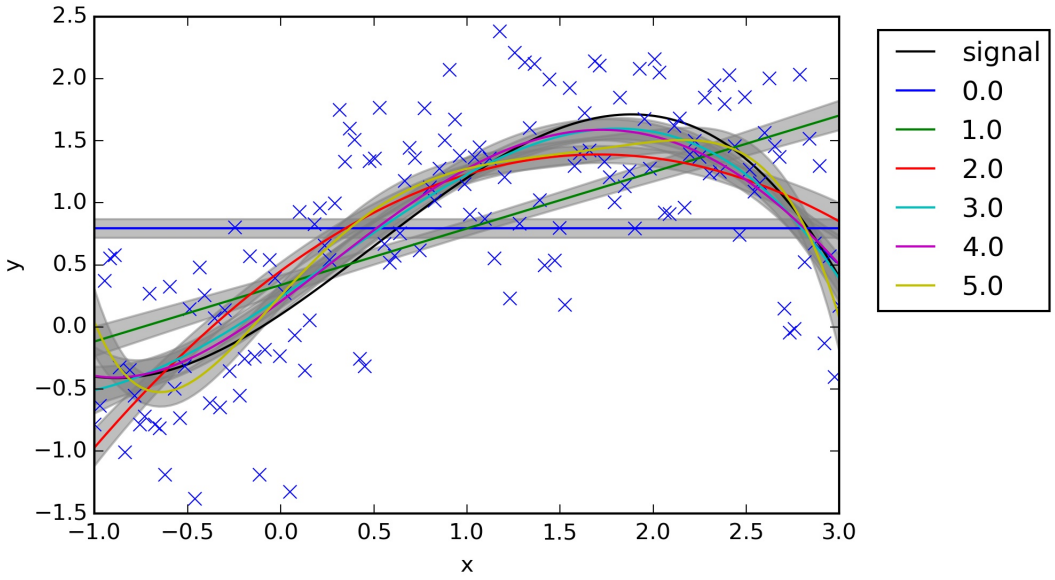

Figure 3: Fits for different polynomials to a mock data set. The data was generated according to $y = f(x) + n$ (12) with n being Gaussian noise with a variance of $\sigma_n = 0.6$. The function $f(x)$ denotes the signal (black line) and was set up as $f(x) := -0.2x^3 + 0.3x^2 + x + 0.1$. The gray areas indicate the uncertainties for each fit. The numbers in the legend correspond to the order of the reconstructed polynomial.

3.b.2 Bayesian information criterion BIC

The Bayesian information criterion (BIC) is a possible solution of the over-fitting problem (see Liddle (2007) for further information). The BIC value is associated with the likelihood entropy by introducing a penalty for higher order polynomials. Specifically we assume:

$$BIC = 2\mathcal{H}_{min} + k\ \ln(\dim(\mathbf{d})) =$$
$$= \frac{1}{p_{\mathrm{MAP}}}(\mathbf{y} - \mathbf{R}\mathbf{f}_{\mathrm{WF}})^T(\mathbf{y} - \mathbf{R}\mathbf{f}_{\mathrm{WF}}) + U\ln(p_{\mathrm{MAP}}) + (M+2)\ln(U) \tag{33}$$

with k being the number of fitted parameters. Note that if the order of the polynomial is M then $k = M + 2$ since there are $M + 1$ parameters for the polynomial and 1 for the noise covariance. Fig. 5 shows how the correction to the entropy takes care of the "overfitting" and Fig. 4 shows the BIC value compared to $\mathcal{H}_{\mathrm{min}}$ of different models for the mock data used in Fig. 3.

3.c Summary and possible fields of application

In this chapter we discussed the properties of parametric models including model selection processes. Combining steps from model determination to perform the best fit in this model gives a tool which is able to find the best estimated polynomial for a given, noisy dataset. If the data does not support higher orders the method will always prefer lower order polynomials even though the actual correlation might be of higher order. This is a perfectly valid information theoretical result. More detailed information on the correlation requires to take more data. To demonstrate this effect we show in Fig 6 how the selected order decreases with increasing noise. Other tests to see whether the method behaves as expected in theory, are performed in the next section.

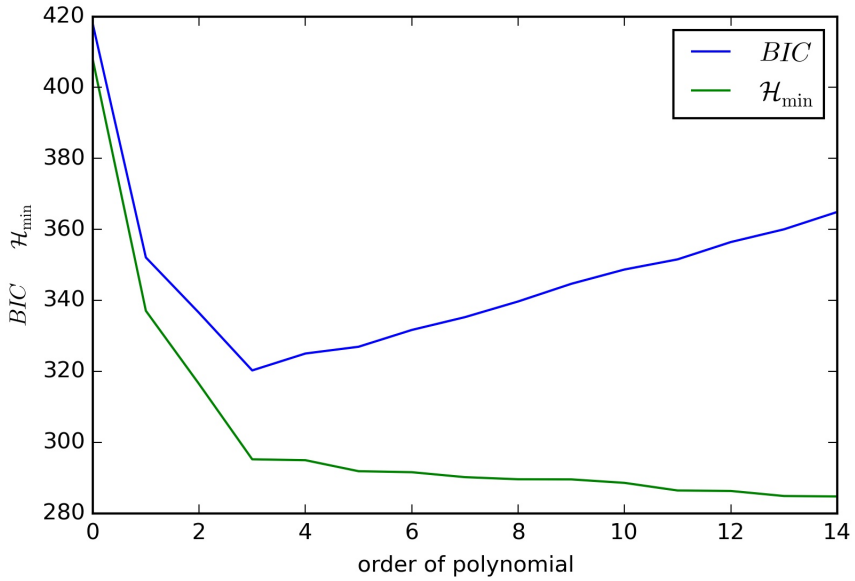

Figure 4: $\mathcal{H}_{\min}$ and the BIC value as a function of polynomial order. The same data as in Fig. 3 is used. We see that the BIC has its minimum for a third order polynomial which is exactly the order of the signal as seen in Fig. 3. In contrast, $\mathcal{H}_{\min}$ decreases with increasing order.

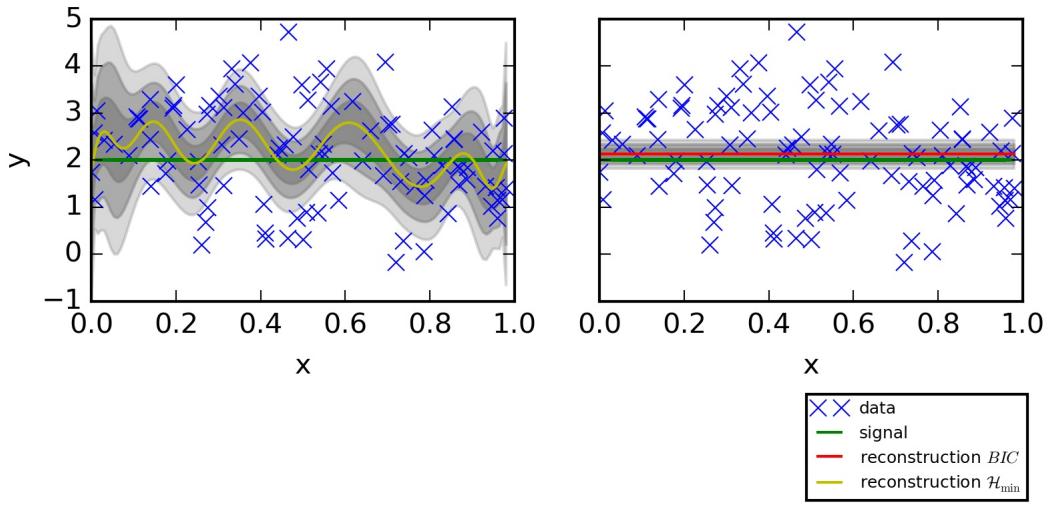

Figure 5: Fits for selected data models from $\mathcal{H}_{\min}$ (left) and the minimal BIC value (right). We restricted the order of the polynomial to be 15 at most, which is exactly the order selected according to $\mathcal{H}_{\min}$. Without restrictions, the model which connects every data point has a minimal Hamiltonian. In contrast, the BIC is minimal for an order of zero.

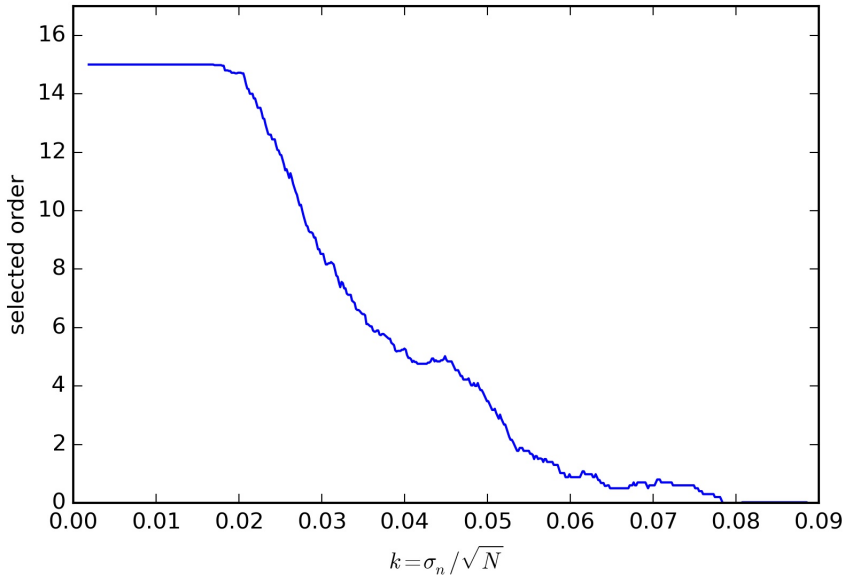

Figure 6: Histogram of recovered polynomial order for different inverse signal to noise ratios k. $N = 1000$ and denotes the sample size. The noise variance σ_n ranges from ≈ 0 to ≤ 3. The signal was generated according to Eq. 13 as a fifteenth order polynomial. We see that the most adequate order selected by the BIC decreases with increasing k. Note that the selected model depends on the specific data realization, therefore we averaged over reconstructed orders with similar k

3.d Method testing

In this section we demonstrate the performance of our method in different test cases. The test data was generated in two different ways:

- **Self consistent tests** are tests with mock data that was generated in a way that it respects the assumptions made in order to develop this algorithm. These tests should show how the method behaves within its limits.
- **Inconsistent tests** are tests with datasets generated from a setup where the assumptions of our method do not hold any more. These tests check whether the method works also reliably on datasets with unknown properties.

3.d.1 Tests with consistent mock data sets

The first test was generated to visualize the general behaviour of our method. Therefore we applied it to mock datasets of different quality. More precisely we generated data sets of various sizes and different noise covariances from the same signal. The data was generated consistent to the data model described in Eq. (13). The correlation coefficients are set up according to Table 1.

Fig. 8 shows the reconstructions for all generated samples. The plots indicate how the precision of the reconstructed signal decreases with increasing noise covariance and

Table 1: correlation coefficients for the generated data in Fig. 8

f_0	f_1	f_2	f_3
0.0	-0.5	-1.0	0.3

increases with increasing sample size N. More precisely the overall variance of a parameter fit behaves like: $\sigma_{\mathbf{f}} \approx \frac{\sigma_n}{\sqrt{N}}$.

In addition the model selection process is involved. For low quality samples a lower order polynomial was selected. This behaviour is in perfect consensus with our information theoretical results described in Section 3.b.

In Figure 7 we show a fit for a mock data set with Gaussian distribution among the x axis. Since we did not restrict the distribution of data among the x axis in our method development, the reconstructions are still valid. In addition the uncertainties increase in regions with less data points.

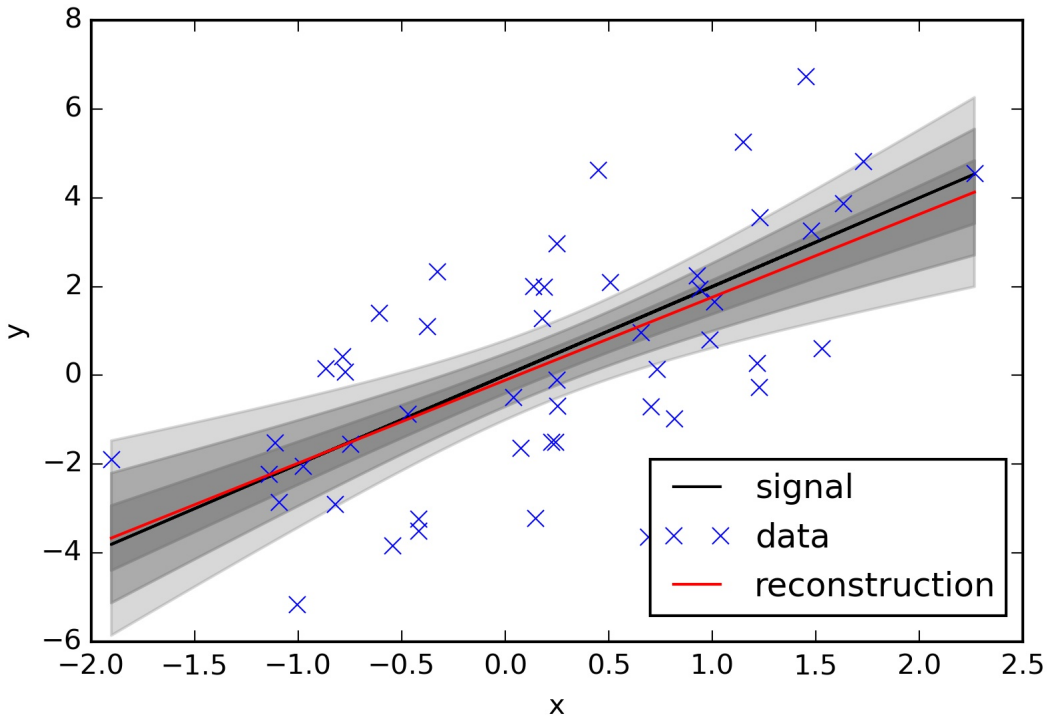

Figure 7: Reconstruction from mock data. The data was set up to be Gaussian distributed among the x axis with a variance of $\sigma_x = 1$. The y values were generated consistent to Eq. (12). More precisely, $y = 2x + n$ where n denotes Gaussian noise. The noise variance was set up to be $\sigma_n = 2\sigma_x$.

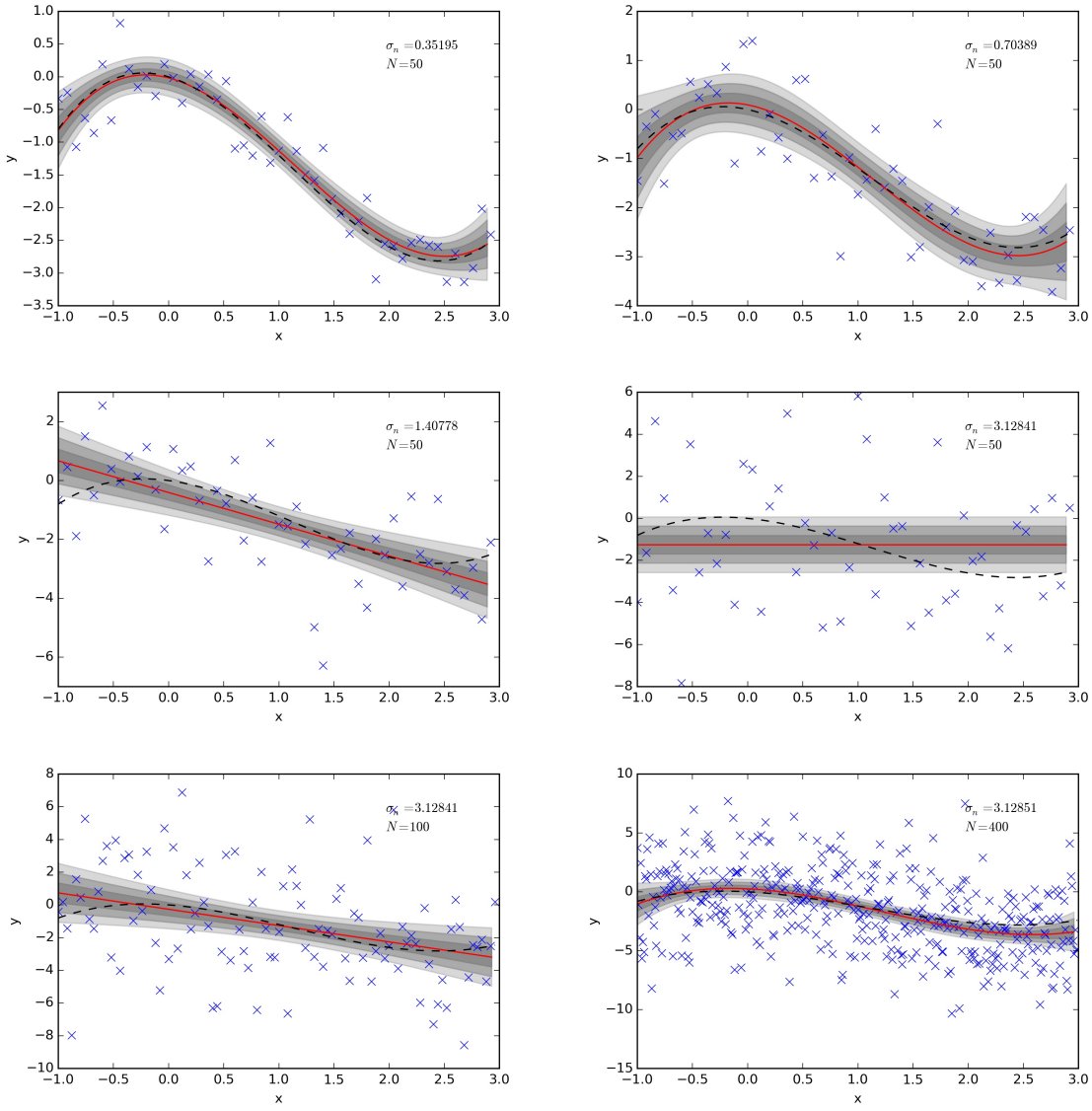

Figure 8: The panels show the reconstructions for different mock datasets. The data was generated to be consistent with the data model in Eq. (13). The correlation coefficients we used are shown in Table 1.

3.d.2 Tests with inconsistent mock data sets

As discussed above, the algorithm relies on the assumption that there exists a unique deterministic correlation between two quantities. Generally two ore more different processes may be combined in a dataset, yielding a superposition of structures. To test the method behaviour in cases where data has a superposition structure, we set up two different data samples.

In the first setup we generated data that clusters in two different regions on the x axis of a two dimensional data space. More precisely, we generate data which is distributed according to two Gaussians among x. Each region has its own correlation structure with the second dimension y. Each subsample of data is consistent with the data model described in Eq. 12, but the combined dataset is not consistent any more. Fig. 9 shows how the method performs on this kind of data. The reconstruction acts on the combined data and therefore is not able to reconstruct both correlation functions. It rather reconstructs a curve similar to the correlation structure dominating the data in this region of the data space. In the intermediate region between the two regions, the reconstructed function does not follow any existing correlation, but rather interpolates between them. This is the major problem that occurs with such a heterogeneous dataset. The reconstruction might indicate a correlation trend which is not existent in any of the signals.

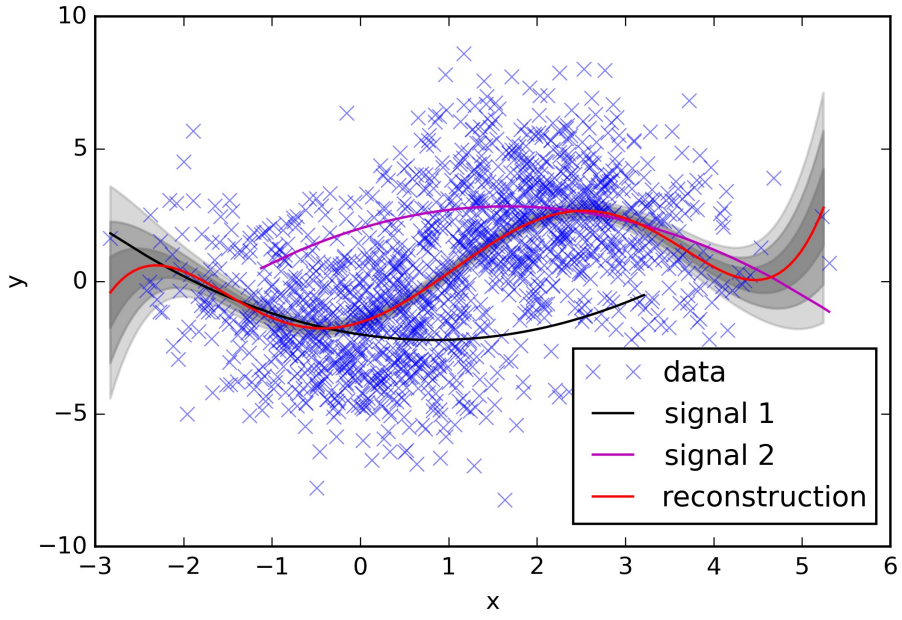

Figure 9: Fit for data generated from an inconsistent model. Signal 1 and signal 2 are both second order polynomials. The explicit coefficients of the correlation functions are shown in Table 2. Subsets of data were drawn from each signal according to Eq. (13) and the data presented to the algorithm is the combination of those subsets.

Table 2: Coefficients for the correlation functions used in Fig 9

	f_0	f_1	f_2
Signal 1	-2	-0.5	0.3
Signal 2	2	1	-0.3

Another inconsistent setup with respect to our data model Eq. (12) is a combined set of data clustering in a higher dimensional set of data. More precisely, we set up a three dimensional data space with data points holding three different properties x,y,z. In addition, we separate two different sub-samples of data along the z axis. One sample is Gaussian distributed among $z = 0$, the other sample among $z = 4$. Each subsample has its own correlation function between x and y, both consistent with Eq. (12). Combining these sub-samples into one data sample results in inconsistent data for correlation determination between x and y. Fig. 10 shows the performance of our reconstruction method on this dataset. Since the information about the separation of the data is only available in z direction, correlation determination fails for x and y for the combined data. As indicated in the picture, distinguishing the sub-samples in the projection of the data to the x-y-plane is not possible. Therefore the reconstruction does not support the properties of the signals. In our explicit example of Fig. 10 the signals were set up in a way so that their dependencies annihilate each other. Therefore the reconstruction indicates that there is no correlation between x and y although there is correlation for each sub-sample.

Table 3: Coefficients for the correlation functions used in Fig 10

	f_0	f_1	f_2
Signal 1	0	0	2
Signal 2	0	0	-2

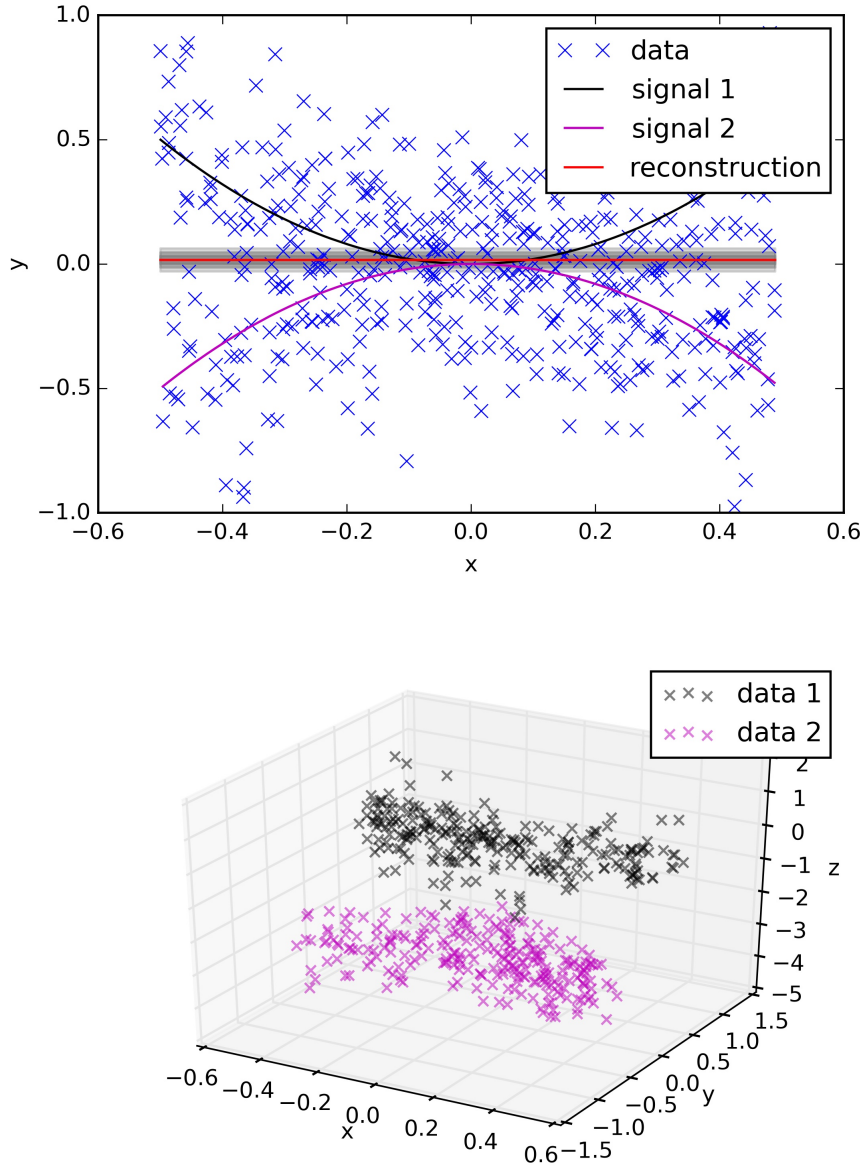

Figure 10: Fit for a dataset from an inconsistent model with a clustered separation in z. The right picture shows the full data space with mock data from two different models for correlation between x and y. Data drawn from different models are indicated by different colors. In the right picture we show the projection of the data in the x-y-plane. The correlation functions have the same color as the corresponding data on the right. The correlation functions were generated according Eq. 13. Table 3 denotes the correlation coefficients. The reconstruction method applied on the combined data indicates no correlation between x and y.

4 Connecting large-scale-structure properties with the Galaxies of the SDSS DR7 main galaxy sample

The main goal of this thesis is to find correlations between properties of the large-scale-structure (LSS) and properties of galaxies. The data for galactic properties is based on the results of Sloan Digital Sky Survey (SDSS) and is described in further detail in the following.

4.a Galactic data

The dataset used in this work is constructed from the sample DR7.2 of the New York University Value Added Catalog (NYU-VAGC) Blanton et al. (2005). This catalog is based on DR7 Abazajian et al. (2009), the seventh data release of the SDSS York et al. (2000). The sample consists of 527 372 galaxies in total, in a redshift range of $0.001 < z < 0.4$. Each data point holds angular positions in ecliptic coordinates. Together with the redshifts this gives the 3d positions for galaxies. Table 4 shows the ranges of the catalog in the r-band Petrosian apparent magnitude r, the logarithm of the stellar mass $\log(M_*)$ in units of the solar mass $M_* = \frac{M}{M_S/h^2}$ and the absolute magnitude $M_{0.1r}$. $M_{0.1r}$ is corrected to its $z = 0.1$ value according to the K-correction code Blanton & Roweis (2007) and the luminosity evolution model Blanton et al. (2003).

The full catalog contains a wide range of galaxies with different properties. Therefore a sub-division of the data into several sub-samples can be helpful to distinguish different correlation structures of the data. In fact the correct sub-sampling of the data set plays an important role in the data analysis sections of this thesis. In sections 5.b and 6.d we compare some of the galactic properties to the LSS and describe different methods of sub-sampling. The reconstruction samples of the LSS used for correlation determination are described in the next Section.

Table 4: Property ranges of the galactic data

	z	r	$M_{0.1r}$	$log(M_*)$
min	0.001	10.1	−18.8	6.6
max	0.4	18.8	−23.0	11.6

4.b BORG reconstruction maps

The reconstructions for the density field are based on a non-linear, non-Gaussian full Bayesian LSS analysis and is based on the same dataset as used for the correlation study (see Jasche & Wandelt (2013)). The method used by Jasche & Wandelt (2013) is based on a Markov Chain Monte Carlo sampling called BORG and results in a set of density

contrast field samples δ_i. The density contrast δ is the normalized difference of the density of the Universe ρ to its mean $\bar{\rho}$. Specifically,

$$\rho = \bar{\rho}(1 + \delta) \ . \tag{34}$$

The density contrast samples can be recombined to an approximate estimate for the PDF of the density contrast. Specifically,

$$P(\delta|\mathbf{d}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta^D(\delta - \delta_i) \ . \tag{35}$$

A correct treatment of uncertainties in the density field when applying the reconstruction methods described in Section 3 can be found in the next Section.

4.b.1 Probability distribution for correlation functions with the LSS

As described in the previous section the BORG algorithm provides an ensemble of density contrast field realizations that capture observational uncertainties. In order to treat the uncertainties in the density contrast correctly, the reconstruction algorithm has to be applied to each realization independently. This yields a PDF $P(\mathbf{f}|\delta_i\mathbf{d})$ for each δ_i. The dependency of the realizations has to be marginalized out of the PDF's in order to obtain the final PDF for the correlation function $P(\mathbf{f}|\mathbf{d})$. This results in a Gaussian mixture for the posterior PDF. Specifically,

$$\begin{aligned} P(\mathbf{f}|\mathbf{d}) &= \int P(\mathbf{f}, \delta|\mathbf{d}) \mathrm{d}\delta = \int P(\mathbf{f}|\delta, \mathbf{d}) P(\delta|\mathbf{d}) \mathrm{d}\delta \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \delta^D(\delta - \delta_i) P(\mathbf{f}|\delta_i, \mathbf{d}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{G}(\mathbf{f} - \mathbf{m}_i, \mathbf{D}_i) \end{aligned} \tag{36}$$

where δ_i denotes one of the N realizations of the density contrast and $\mathbf{m}_i$ and $\mathbf{D}_i$ denote the corresponding mean and covariance for each fit.

4.b.2 Mapping the SDSS data onto reconstructed density fields

The density field inference was applied to the northern region of the sky as covered by the SDSS survey. More precisely, the inference is performed on a cube with 750 $\frac{\mathrm{Mpc}}{\mathrm{h}}$ side length with a grid resolution of ≈ 3 $\frac{\mathrm{Mpc}}{\mathrm{h}}$ in the co-moving frame. This results in a cubic grid with 265^3 voxels. Table 5 denotes the boundaries of this box.

In order to compare the properties of the LSS, we map the galaxies onto the cubic grid and extract the information about the LSS for each galaxy. More precisely, we look for the position of each galaxy and store the properties of the LSS in the voxel hosting the galaxy. All galaxies within one voxel are assigned the same LSS information. This results

Table 5: Boundaries of the cubic grid in the co-moving frame

Axis	Boundaries ($\frac{\text{Mpc}}{\text{h}}$)	
x	-700	50
y	-375	375
z	-50	700

in an extended data catalog, containing the intrinsic properties of the galaxies as well as the properties of the LSS in the surrounding area of each galaxy. Note that this procedure is perfectly applicable for all kinds of cosmological data as long as there is information about the 3D position of the objects in the data.

In order to map the data catalog of the SDSS onto the grid, we need to transform co-ordinates of galaxies from redshift space to co-moving frame. Redshifts z_i are transformed to co-moving distances d_{com} according to:

$$d_{\text{com}} = \int_0^{z_i} \frac{1}{cH(z)} \mathrm{d}z, \tag{37}$$

where c is the speed of light and $H(z)$ denotes the Hubble parameter. $H(z)$ is given as:

$$H(z) = H_0 \sqrt{\Omega_m (1+z)^3 + \Omega_c (1+z)^2 + \Omega_\Lambda}, \tag{38}$$

under the assumption of a concordance ΛCDM model with the cosmological parameters $\Omega_m = 0.24$, $\Omega_c = 0.00$, $\Omega_\Lambda = 0.76$, $h = 0.73$ and $H_0 = h\ 100\ \frac{\text{km}}{\text{s MPc}}$ (see Spergel et al. (2007)). We used this set of parameters instead of more recent ones in order to match the cosmology used for the LSS reconstructions.

As a final step we calculate the Cartesian coordinates for each galaxy:

$$x = d_{\text{com}} \cos(\delta) \cos(\alpha) \tag{39}$$

$$y = d_{\text{com}} \cos(\delta) \sin(\alpha) \tag{40}$$

$$z = d_{\text{com}} \sin(\delta), \tag{41}$$

where α and δ are the right ascension and declination of the ecliptic frame, respectively. In Figure 12 we show slices through samples of the density field as well as slices through the mean density field. The mean density contrast $\langle\delta\rangle$ was calculated according to:

$$\langle\delta\rangle \approx \frac{1}{N} \sum_{i=1}^{N} \delta_i \ . \tag{42}$$

By looking at the plot it becomes clear that the uncertainties in the reconstructed maps increase with increasing distance to us. Therefore the mean of the reconstructed samples average out to an over-density of $\langle\delta\rangle = 0$ in areas where there is no structural information.

In order to exclude areas of too high uncertainties in the further analysis of correlation determination, we excluded all galaxies of the SDSS sample above a certain distance $d_{\rm lim} = 450$ Mpc. This results in a sub-sample including only galaxies with redshifts between $0.001 < z < 0.38$. Figure 11 shows galaxies mapped on different slices of the density field for the full sample and the limited sub-sample. The plot indicates that most of the galaxies are located within this limit. This becomes clear by taking into account that the LSS reconstruction is based on the same SDSS sample. Therefore uncertainties in the density field decrease in areas where more galaxies are located.

The density field allows a derivation of many important quantities of the LSS. Some important examples are: the gravitational potential, the tidal-shear tensor and the web type classification. The rescaled gravitational potential Φ is given as

$$\Delta\ \Phi = \delta \tag{43}$$

and the tidal-shear tensor T_{ij} is given by the Hessian of Φ:

$$T_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j}\ . \tag{44}$$

4.b.3 Web type classification of the LSS

The web type is a classification of different structure types of the LSS. Various classification methods have been presented in literature. In this thesis we split the LSS into four different types: voids, sheets, filaments and clusters. The web type was classified according to the eigenvalues of the tidal-shear tensor (for further information see Hahn et al. (2007)). Table 6 shows the explicit classification rules and Fig. 13 shows the classification of a reconstructed sample according to these rules.

Table 6: Web type classification according to the ordered eigenvalues of the tidal shear tensor. In this thesis we used $\lambda_{\rm th} = 0$

Classification	
Void	$\lambda_{\rm th} > \lambda_1, \lambda_2, \lambda_3$
Sheet	$\lambda_1 > \lambda_{\rm th} > \lambda_2, \lambda_3$
Filament	$\lambda_1, \lambda_2 > \lambda_{\rm th} > \lambda_3$
Cluster	$\lambda_1, \lambda_2, \lambda_3 > \lambda_{\rm th}$

The structural classification as well as the density field reconstruction itself contain information about the LSS of an area a galaxy is located at. These quantities are used in order to compare galactic properties with the LSS. To do so, we apply the methods developed in Section 3 to our extended data set in the following.

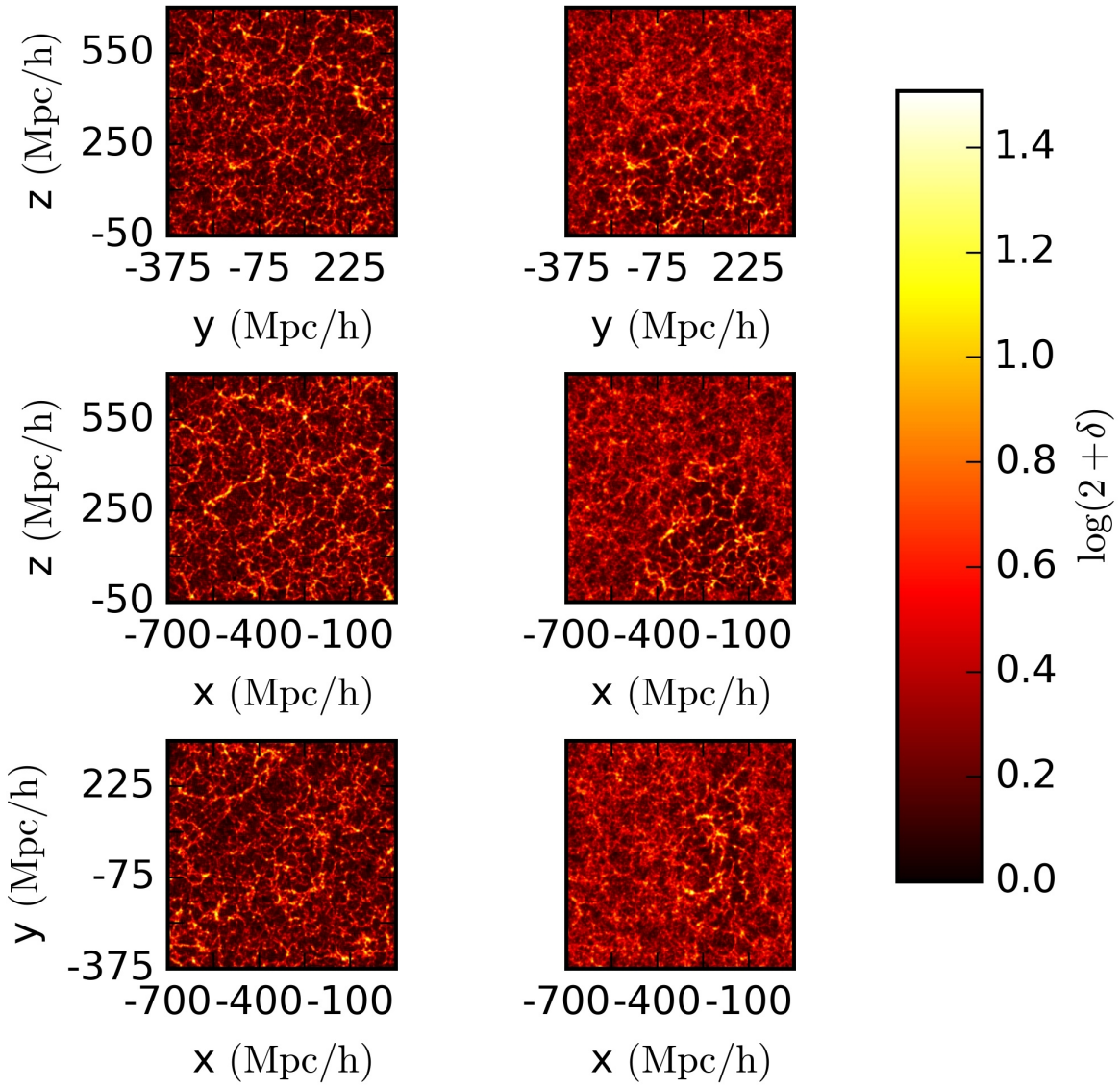

Figure 11: Slices through the 3D reconstructed density field volume. Upper panels show the results for one realization of the density contrast δ_i while lower panels depict the ensemble mean density contrast $\langle\delta\rangle$. In order to ensure the displayed quantity to be positive we used $\log(2+\delta)$ for the respective non-linear density contrasts since $\delta \in [-1, \infty[$.

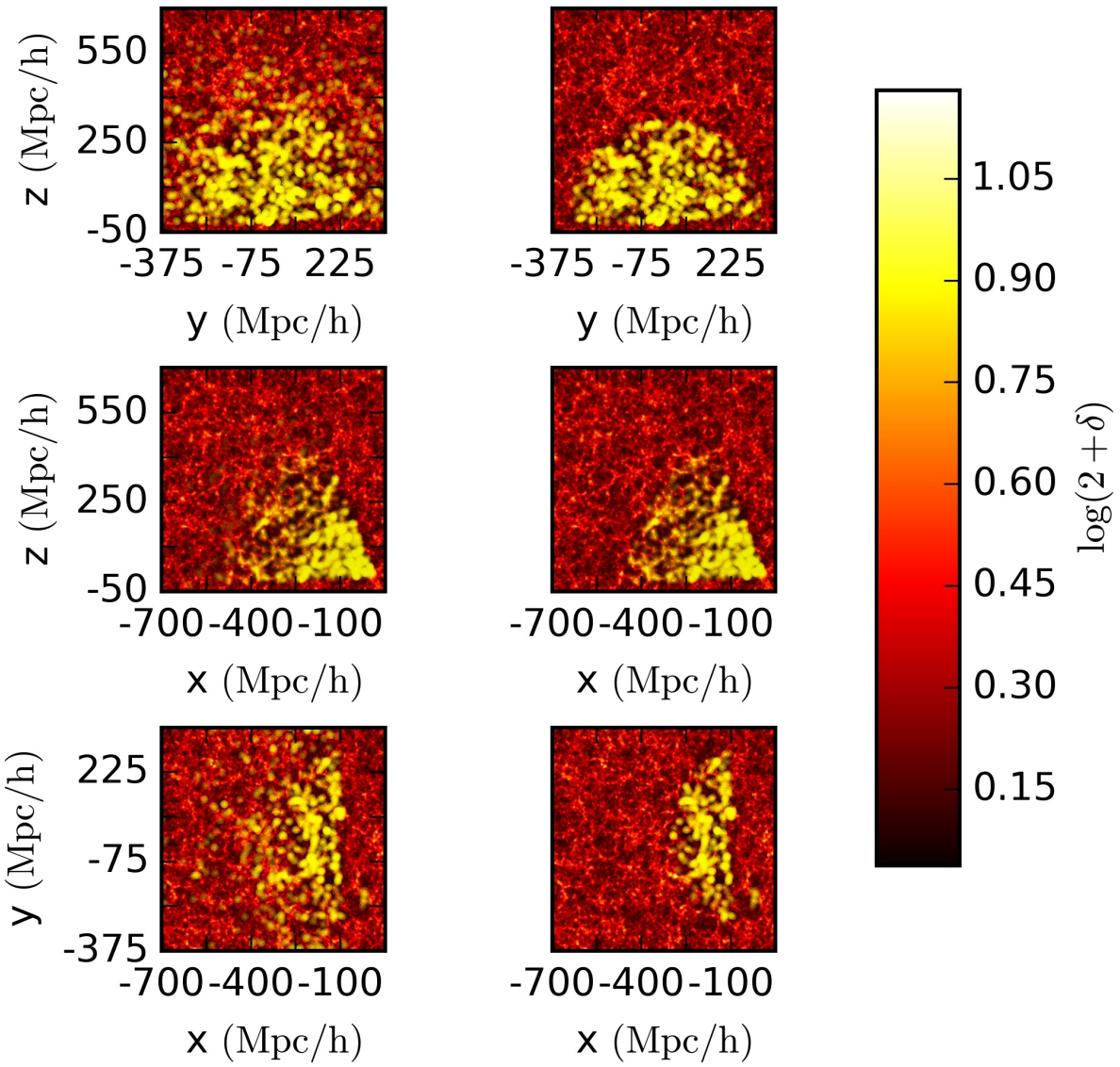

Figure 12: Each yellow dot correspond to a galaxy which is mapped to slices of the 3D reconstructed density field volume. Upper panels show the full SDSS sample mapped to the density field while lower depict a sub-sample where we exclude all galaxies above $d_{\rm lim}$ as described in Section 4.b.2.

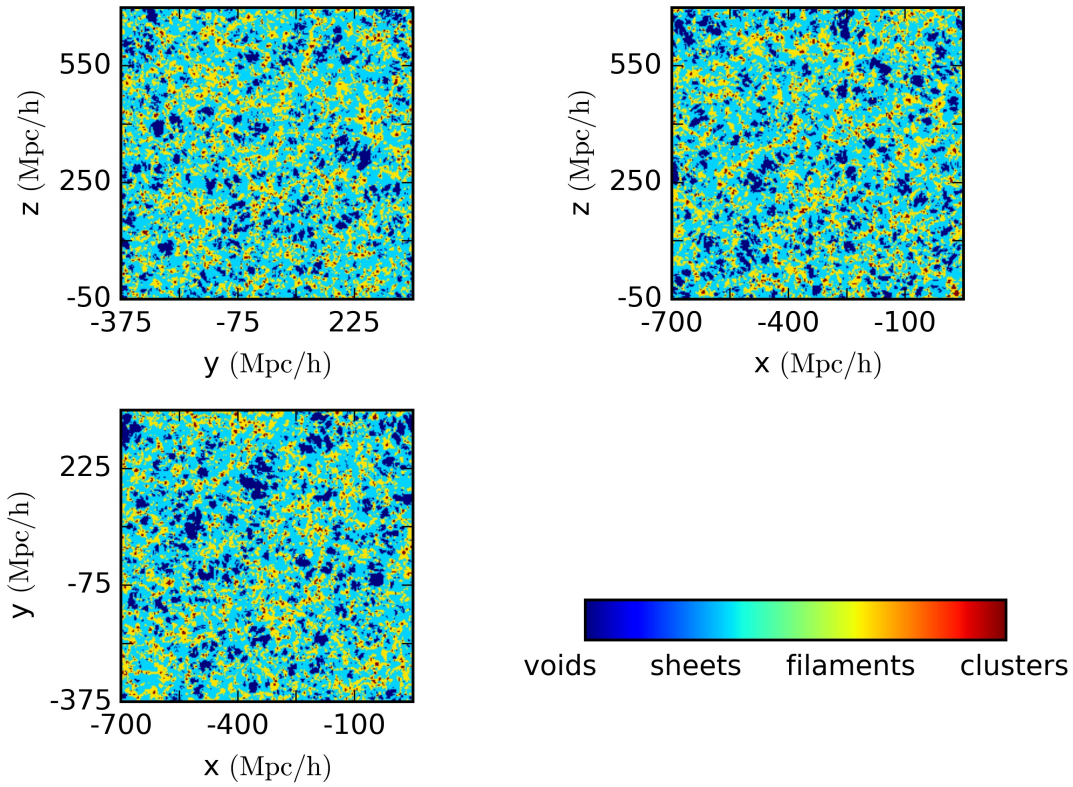

Figure 13: Slices of the 3D LSS reconstruction. We distinguish the LSS according to the web-type classification described by Table 6

5 Data analysis

In this section we describe the application of our method to the datasets provided by the SDSS sky survey and the BORG reconstruction maps. Since the section 3.d only contains tests with mock data sets it is important to make some more consistency checks with real data in order to verify the functionality of the method provided.

5.a Comparison of our analysis to state of the art results

As a first consistency check we follow a similar philosophy as described by Lee & Li (2008). Lee & Li looked for links between the LSS as characterized by the LSS map of Erdoğdu et al. (2004) and the galaxies of a previous SDSS survey (DR4 by Adelman-McCarthy et al. (2006)). We first bin this data sample into several sub-samples for example using the absolute magnitude as a discrimination and compare the mean of the density field for each sample with the respective magnitude.

Since in this thesis we used the BORG reconstruction and the data release 7 of the SDSS for cross correlation, the first consistency check is to apply the methods used by Lee & Li (2008) to this new data set and to compare the results. To obtain an ensemble mean density field, we average over the LSS reconstruction samples provided by the BORG algorithm. Since the reconstructed field of Lee and Li is only available for nearby galaxies,

a redshift threshold is taken at $z < 0.04$. In Fig. 14 the mean density field was averaged over four different bins in the absolute magnitude and the stellar mass (Table 7 shows the explicit bin range). In addition we applied our reconstruction method (Chapter 3) to the data. The results show correlations between galaxy properties and the density field that appear to follow the same trend as described by Lee & Li (2008). However, our recovered amplitudes of the correlation trends are stronger. This is because the resolution of the reconstructed density field of the BORG algorithm is higher compared to the density field reconstruction maps used by Lee & Li (2008). Therefore the amplitudes of the density contrast δ are increased there.

Despite this difference, the results of our method seem to be consistent with the results published by Lee & Li (2008). Therefore we extend the data catalog up to the distance limit of the BORG reconstructions in the following.

Table 7: chosen bins in absolute magnitude $M_{0.1r}$ and in the logarithm of the stellar mass $\log(M_*)$

Sample	$M_{0.1r}$	$\log(M_*)$
1	> -17.5	< 8.9
2	$[-18.2, -17.5]$	$[8.9, 9.4]$
3	$[-19.1, -18.2]$	$[9.4, 19.0]$
4	< -19.1	> 10.0

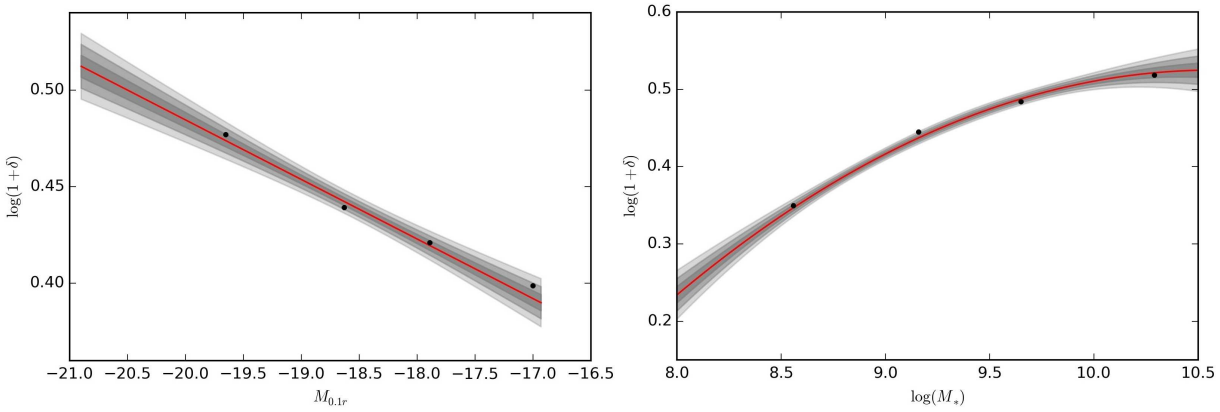

Figure 14: The dots correspond to the mean values of the logarithm of the density contrast in each bin according to Table 7. The red line is the best linear fit to the given sample and the gray areas indicate uncertainties.

5.b Sub-dividing the galaxy sample

In Section 3 we pointed out that for our method for correlation determination to work properly it is important to use data generated from a single process of the form given by Eq. (12) only. Therefore we propose to sub-divide the galaxy sample into sub-samples ac-

cording to different separation rules. These rules should be chosen in a way to distinguish different data generation processes as far as possible.

Important effects for large surveys are flux limitations of telescopes. This means that at larger distances only the brightest galaxies are detected (see e.g. Mo et al., 2010). For this reason flux limitations introduce a distance dependent bias on observed galaxies. To remove this effect a volume limitation of the sample was applied, yielding a uniform selection of galaxies within the chosen sub-volume. Volume limitation means to limit the data set of a flux limited survey in a way to ensure that in this sub-sample all existing galaxies are included in the sample. A possible way to accomplish this division is to include only galaxies to the sample brighter than a certain absolute magnitude limit M_{lim} and below a certain redshift limit z_{lim}. Here z_{lim} is the distance at which a galaxy with absolute magnitude M_{lim} has an apparent magnitude equal to the limit of the survey m_{lim}. More precisely:

$$M_{\mathrm{lim}} = m_{\mathrm{lim}} - 5\log\left(\frac{r_{\mathrm{lim}}}{r_0}\right) \tag{45}$$

with r_{lim} being the luminosity distance corresponding to z_{lim} and $r_0 = 10$ pc conventionally (see Mo et al. (2010)).

Fig. 15 shows correlations for different volume limitations of the sample. The correlation functions indicate that separation into different volume limited samples also separates different correlation structures. It is interesting to notice that the correlation function for the full sample is always some high order polynomial which (as described in the method testing section 3.d.2) might be due to the fact that the full data set is a superposition of subsets generated by different processes with different correlation structures. Indeed this fact is revealed by the correlation functions for the different volume limitations. As shown in the picture, for sub-samples with lower magnitude the correlation function between the absolute magnitude and the density field shows linear behaviour. The samples for higher magnitudes still show complex correlation structures which indicates that in these regions we may still see a superposition of effects. To observe separated correlation signals in regions of higher magnitudes, further subdivisions of the sample seems to be required.

This could be achieved by sorting the data with respect to its web type classification in the large-scale environment. This means to split each volume limited catalog into sub-samples containing only the galaxies in voids, sheets, filaments and clusters respectively. Since properties of galaxies in these regions differ, the processes causing correlations for observed quantities may also differ in those regions. Fig. 16 shows the correlation functions for the division according to these restrictions for a specific volume limitation at $M_{0.1r} = -19.0$ and Fig. 17 for a volume limitation at $M_{0.1r} = -20.5$. Again this subdivision tends to distinguish some of the different processes responsible for correlation, but not all of them.

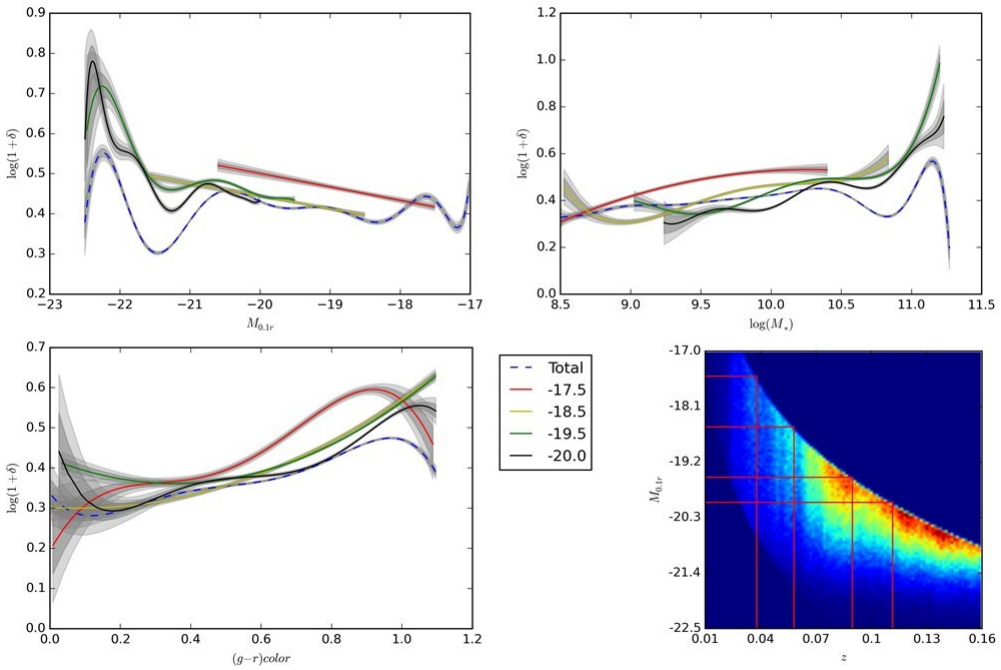

Figure 15: Correlation functions for different galactic properties with the density contrast. Note that each line correspond to a fit for a specific subset of data points corresponding to the volume limitations shown in the last picture. The specific absolute magnitude limit for each correlation function is shown in the legend. The last picture also shows the density distribution of the galaxies in the magnitude redshift space. Redder areas correspond to higher number counts of galaxies in this region and blue areas to lower counts, respectively. Instead of the density contrast we used the logarithm of the density contrast, since all compared galactic properties are on a logarithmic scale.

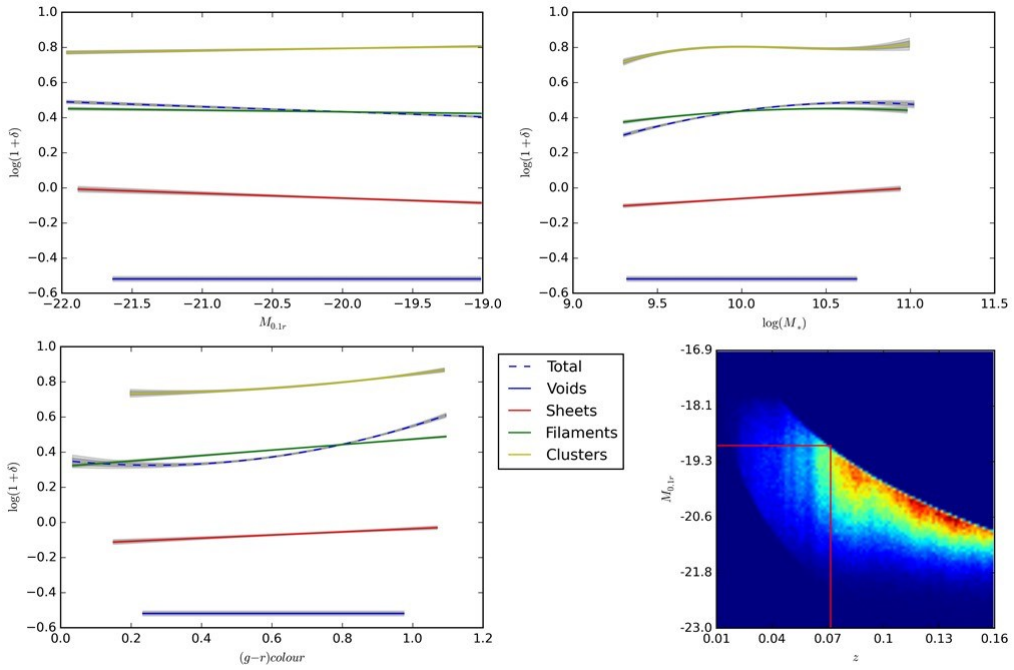

Figure 16: The plots show the correlation functions for different subsamples of the galaxy data now being separated after their web-type classification according to the LSS for one specific volume limited sample. The limitation was applied at an absolute magnitude threshold of $M_{0.1r} = -19.0$.

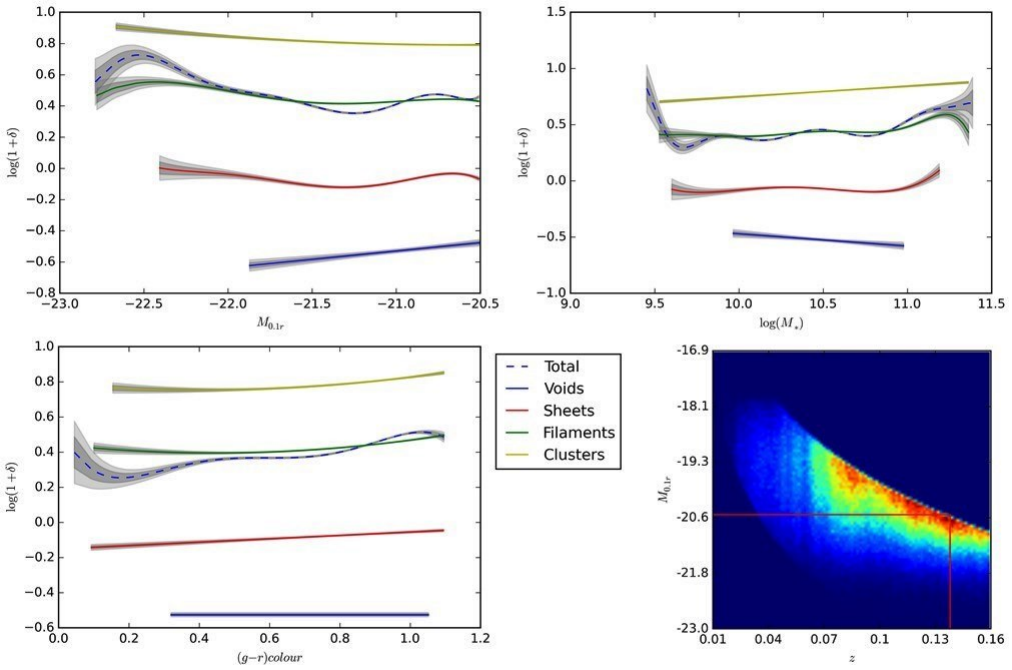

Figure 17: Same plots as in Fig. 16 but for a different volume limitation at $M_{0.1r} = -20.5$.

The range of possible subdivisions of a data catalog may be arbitrary large rendering manual treatment of the data impractical. Consequently a main goal of this thesis is to develop a flexible and general method to compare different kinds of cosmological data with the LSS properties. Therefore we seek to find an automatic approach to subdivide a data set and perform correlation analyses.

Since manual selection always results in data which appear spatially separate in data space, it is reasonable to set the selection process according to such different regions. Specifically here we propose to use a specific kind of artificial neural network called "Self Organizing Map" (SOM). SOMs are used to group sub-samples of data by similarity. One realization of a SOM is described in the next section.

6 Self Organizing Maps

A Self Organizing Map (SOM) is an artificial neural network specifically designed to identify clusters in high dimensional data sets. More precisely it compares data space positions and combines data with similar positions into a sub-sample of data. To accomplish this division, the SOM has to adopt the properties of the spatial distribution of the data. This chapter provides an overview on the SOMs as will be employed in the remainder of this thesis. [1]

6.a Method

A SOM is an artificial neural network which is able to determine the structure of a dataset in an high dimensional space. The network has a specific topological structure. For example each neuron of the network is linked to its neighbours in a square lattice pattern with a neighbourhood function representing the strength of those links. The network is trained by data with a training algorithm which gets repeated for every data point multiple times resulting in a learning process.

Before the process can start the network has to be linked to data space. Therefore each neuron holds a vector $\mathbf{W} = (W_1, W_2, ..., W_N)^T$ in the N dimensional data space, called weights. It is important to point out that the neurons live in two different spaces: the data space with the position represented by its weight and the network pattern where each neuron is linked to each other by a neighbourhood function. The SOM used in this work is initialized with a square lattice pattern.

Since in the beginning no information about the data space has been provided to the network, weights are initialized randomly in data space. After initialization the actual learning process starts. Each iteration of the learning process follows the same structure,

[1] The SOM analysis method used in this work is based on a SOM implementation provided by the python package pymvpa (www.pymvpa.org/generated/mvpa2.mappers.som.SimpleSOMMapper.html).

as described in the following.

First the "Best matching unit" (BMU) is calculated for a randomly chosen data vector $\mathbf{V} = (V_1, V_2, ..., V_N)^T$. The BMU is defined to be the closest neuron to $\mathbf{V}$ in terms of similarity, as expressed by a data-space distance measure. The measurement in this thesis is based on the Euclidean distance D with modified scales for each dimension. Specifically

$$D = \sqrt{\sum_{i=1}^{N} \left(\frac{V_i - W_i}{\sigma_i} \right)^2} \ , \tag{46}$$

where σ_i being the scale factor for each component i. If the components of $\mathbf{V}$ are of different shape (meaning that they have different measurement units) it is important to normalize them. This ensures that all components have the same shape. σ_i is defined as:

$$\sigma_i := V_{i\ \max} - V_{i\ \min} \tag{47}$$

where $V_{i\ \max}$ and $V_{i\ \min}$ are the maximum and minimum values of the ith component of all data vectors. D is a measurement for the distance between two vectors in an Euclidean space.

The weight of the neuron for which D gets minimal is modified according to the value of $\mathbf{V}$. Therefore the new weight for the BMU at iteration step $t + 1$ is:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + L_t(\mathbf{V} - \mathbf{W}_t), \tag{48}$$

where $\mathbf{W}_t$ is the previous weight and L_t is the "learning rate". The learning rate is a decreasing function of t and hence quantifies how strong an input vector should influence the weights at a specific iteration step. It has to be a decreasing function since the tth vector presented to the network should not change the weight of a neuron as much as the previous ones to ensure converging information updates. There are two convenient shapes for learning rates: a linear and an exponential decay. In this work we chose to use the exponential decay with L_t given as:

$$L_t = L_0 e^{-\frac{t}{\lambda}} \ . \tag{49}$$

L_0 is the initial learning rate and λ is a tunable parameter to adopt the change of the learning rate for each iteration.

Since neurons are linked to each other, adaptation of individual neurons will also affect the weights of all other neurons. The strength of the modification of those weights should decrease with distance to the BMU in the specified topology of the network. Therefore

the size of the neighbourhood of a single neuron for a specific iteration step t is

$$\sigma_t = \sigma_0 e^{-\frac{t}{\lambda}} \ , \tag{50}$$

where σ_0 is the initial neighbourhood size. Note that the size decreases with t in order to ensure that the modification of the vicinity of the BMU gets less important with increasing t. The neighbourhood size σ defines the influence rate Θ of one iteration:

$$\Theta_t = e^{-\frac{d_{\mathrm{BMU}}^2}{2\sigma_t^2}} \ , \tag{51}$$

where d_{BMU} is the distance between the position of the updated neuron and the BMU of the tth iteration step in the square lattice pattern. It is important to distinguish d_{BMU} from D, since d_{BMU} is the distance between two neurons in the network pattern and D is the euclidean distance in data space. Note that Θ assumes a value of one for the BMU itself therefore modification functions can be combined, yielding

$$\mathbf{W}_{t+1} = \mathbf{W}_t + L_t \Theta_t (\mathbf{V} - \mathbf{W}_t) \ . \tag{52}$$

These steps are repeated for every single vector in the dataset. Note that for a single process the order of the vectors presented to the network matters. The first vectors influence positions of the neurons in data space stronger than vectors which are presented to the network at later steps. To avoid biasing weights to the first subset of data, the whole learning process has to be repeated multiple times. The final result is then given by averaging the weights for each learning process.
This results in an iterative process including the following steps:

- Repeat multiple times:
  - Initialization of the network patter
  - Initialization of the weights for all neurons randomly
  - Repeat for all N data vectors $\mathbf{V}_t$, $t \in (1, .., N)$:
    - Calculate the BMU for $\mathbf{V}_t$
    - Update the weights of the neurons according to $\mathbf{V}_t$
- Average the weights for each learning process

For large sets of data this learning process is numerically expensive. But once completed, the trained SOM is a numerically fast and powerful tool to represent the structure of datasets. For a new vector $\mathbf{V}'$ it is now easy to classify this vector by simply comparing it to the neurons. More precisely the neuron which holds the weight closest to $\mathbf{V}'$ (in terms of the Euclidean distance) represents the region of the data space $\mathbf{V}'$ lies in.

Therefore a SOM can be used to find clusters in a high dimensional data space. After the SOM has been trained, each training vector is presented to the trained SOM and all vectors closest to the weight of a specific neuron are stored in a sub sample of the data. Each sub sample now holds a set of data vectors which are all similar to the weight of their neuron. The average properties of this region are then represented by the data space position given by the weight of the linked neuron.

6.b Test

Real data often come with the artifact that they appear to be a combination of clustered regions in data space. Each region might have different correlation structures. In low dimensional data sets this clustering is often easily detect visually, but this becomes increasingly more challenging in high dimensions. The method described in section 3 requires data that was generated from a unique linear process. Therefore it is reasonable to apply the SOM to data and treat identified sub-samples independently for correlation analysis. Fig 18 shows the results after applying a SOM to the problem described in Section 3.d.2 using the same data setup. The Figure indicates that the SOM is able to determine different regions of the data space. In addition we applied the reconstruction method described in Section 3 to each sub-sample of the data separated by the SOM. The reconstructions show that for each sub-sample a successful inference of the original signal is possible.

6.c Application

In Section 5.b sub-dividing of data samples (*i.e.* Galaxy properties) was performed manually by using knowledge about the cosmological data which should be compared to the large scale environment. This requires knowledge on the data generating process (for example how the telescope works and which systematics it includes into the data) as well as some knowledge about the physical properties of the analyzed objects (for example knowing which regions of the data are expected to have similar correlation structures). The goal of this thesis is to develop a method which is able to find trends for the correlation between the LSS of our Universe and the objects visible in the sky today, a field where not all physical processes are understood yet. In addition a correct statistical formalization of the data generation process in this context is a very complicated task and might not be worth the effort when the goal is to find trends only. This is why the SOMs are a powerful tool in this context since they are able to find simple and strong systematics of the data generation as well as physical properties which cause different correlation structures and are able to split the data in order to ensure that only one correlation structure remains for each generated subsample. Of course the results generated by this method should not be seen as being exactly the real physical correlation, but as a means to find new

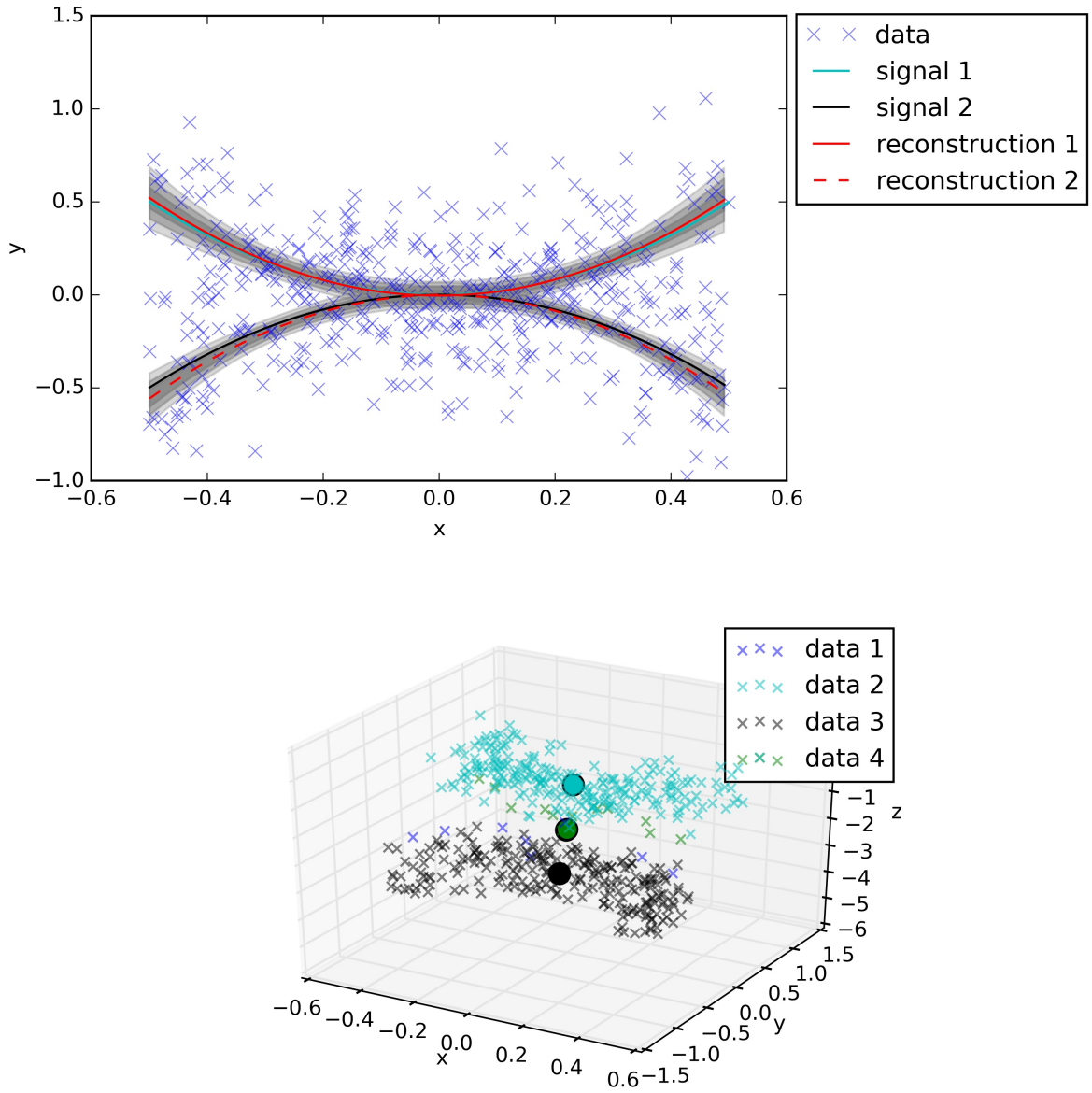

Figure 18: The SOM applied to the mock data described in Section 3.d.2. The right figure shows the assignment of the data to different neurons after the learning process. All points associated with a specific neuron are drawn in the color of this neuron. Note that in this case a 2×2 grid for the neurons was set up but only 2 different clustered regions exist. The SOM links two neurons to the clusters and the remaining neurons to a few intermediate data points. The left plot shows the generated signals and the reconstructions. The reconstruction method described in Section 3 was applied to each sub-sample of the data according to the selection of the SOM.

and unknown trends. These might help to understand the formation and the evolution of galaxies better. This method might also help determining regions in the sky to find objects with specific properties of the cosmological large-scale structure. As our SOM will not only use the intrinsic properties of the galaxies but also properties of the cosmological environment for classification.

6.d Data analysis

An important issue for the actual training of a SOM with a specific dataset is the best possible selection of the data space to train it with. In order to find as many separated regions in data as possible, it seems to be reasonable to include all available properties of the galaxies and the LSS to the training space to include as much information as possible. But using dimensions which are redundant to each other can lead to problems during the training process. Since all elements of the data vector get treated the same, this redundant information puts a higher weight onto their separation. Therefore the optimal setup should include enough complementary data properties in order to gain all the information necessary for data cluster determination, but should avoid redundant information. Therefore for the SDSS catalog a reasonable setup is to include redshifts z, r-band absolute magnitudes $M_{0.1r}$ and colors of galaxies. To include properties of the LSS we extended the training space with the logarithm of the density field $\log(1 + \delta)$ and the three eigenvalues of the tidal shear tensor at the location of each galaxy. This setup seems to be reasonable, since many properties of the LSS (for example the web type classification and the ellipticity) depend on these quantities. The logarithm of the stellar mass $\log(M_*)$, another common property of galaxies, was excluded from the training process since it appears to be proportional to the absolute magnitude up to a constant only. The usage of the logarithm of the density field instead of the density field arises from the fact that the included galactic properties are on a logarithmic scale and therefore the denpendencies should be estimated on this scale as well.

After the division of the full data sample of the SDSS digital sky survey into sub samples according to the classification by the SOM we applied the method presented in Section 3 to determine correlations in respective data samples. Fig. 19 and 20 show some of the results for specific neurons. Note that the model selection process described in Section 3.b is still involved in the analysis and therefore the shown correlation structure was selected by the algorithm itself.

We see that for many sub-samples generated by the SOM, brighter galaxies seem to be located in denser regions, while galaxies of lower brightness are located in less dense regions. This results in a linear correlation between the density field and absolute magnitudes. The trend remains in different sub-samples classified by the SOM, ranging from blue galaxies in low density regions associated with filaments (Fig. 19) to red galaxies

in high density regions associated with clusters (Fig. 20). The correlation strength seem to decrease for high density regions. In extreme cases we found sub-samples indicating an inverted trend in dense clusters (see Fig. 21).

Due to our interpretation of the results, we conclude that those inverted trends do not constitute physical processes. We rather propose that this effect is caused by redshift distortions in the data sample. These distortions arise from peculiar velocities δv of galaxies, which introduce a Doppler shift to the redshift measurement (see Kaiser (1987)). This effect causes galaxy clusters to appear stretched along the line of sight, an effect frequently referred to as "Fingers of God". Assuming virialized structures, the velocity dispersion increases in high density regions according to

$$\Phi \propto \frac{\delta v^2}{c^2} \ , \tag{53}$$

where Φ denotes the gravitational potential. The velocity dispersion rises up to $\delta v \sim 1000 \frac{\mathrm{km}}{\mathrm{s}}$. Introducing a redshift uncertainty $\delta z \approx \frac{\delta v}{c}$ leads to uncertainties in the co-moving frame up to $\delta d_{\mathrm{com}} \approx 14$ Mpc. Since the resolution of the BORG reconstruction maps is ~ 3 Mpc, a galaxy can be mapped 4 voxels away from its actual position in an extreme case. The density field reconstructions are corrected to the "Fingers of God" effect due to the fact that the reconstruction process involves a non-linear reconstruction of the density field from the initial conditions. Therefore a heavy galaxy near the center of a cluster can be mapped to its actual vicinity. Since the properties of such galaxies differ from others, the SOM groups those galaxies into a sub-sample according to their similarity. This sample shows the inverted trend shown in Fig. 21 since heavier galaxies get mapped further away from the cluster into low density regions.

Solutions to the redshift distortions problem are beyond the scope of this thesis, but problematic regions can be identified according to the classification of the neuron which holds the sub-sample of the data. In lower density regions the distortion is smaller than the grid resolution and therefore the results remain reliable. A further discussion about possible ways of including this effect into the analysis and future fields of application of this method are provided in the next Section.

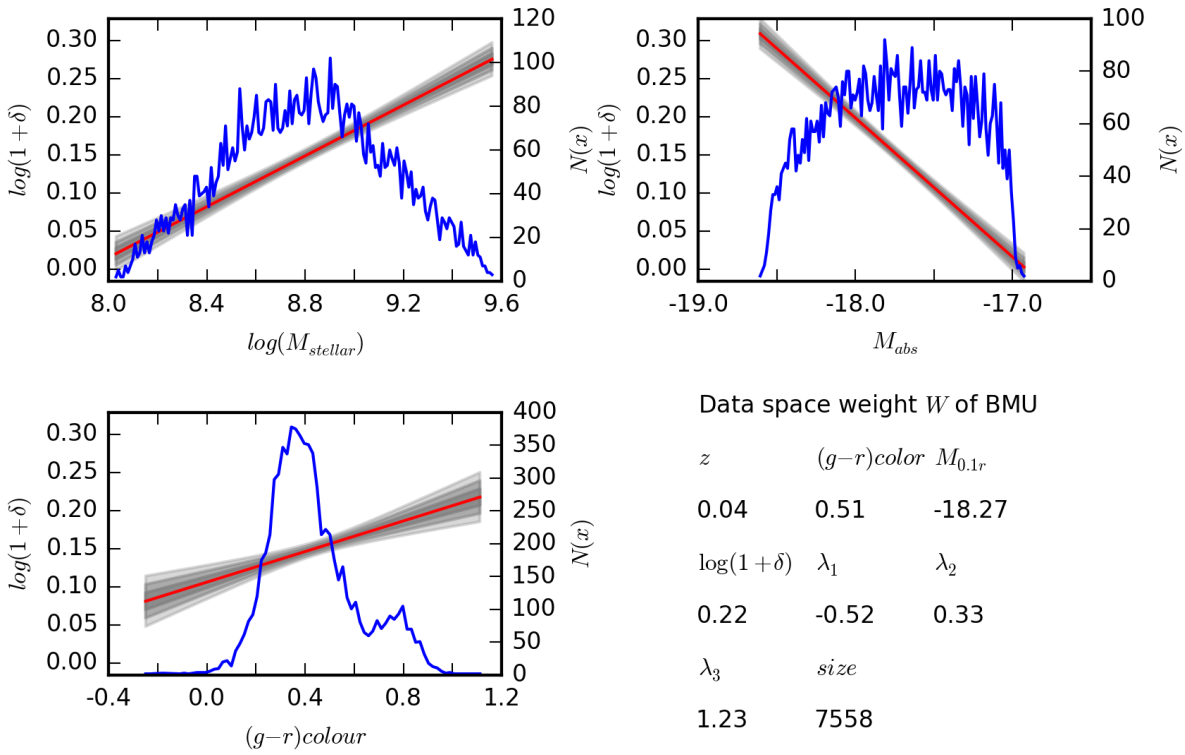

Figure 19: Each plot shows the fit for the correlation function between one galactic property (here the logarithm of the stellar mass $\log(M_*)$, the absolute magnitude $M_{0.1r}$ and the (g-r) color) and the logarithm of the density field $\log(1+\delta)$. The reconstruction method described in Section 3 is applied to a sub-sample of data which clusters according to the SOM. The blue line is a histogram of the galaxy distribution according to the x axis. All data points used for this fit are associated with one neuron after the training process. The bottom right part shows the weight W of the BMU which is the mapped data-space position of the neuron associated to the data cluster used for correlation.

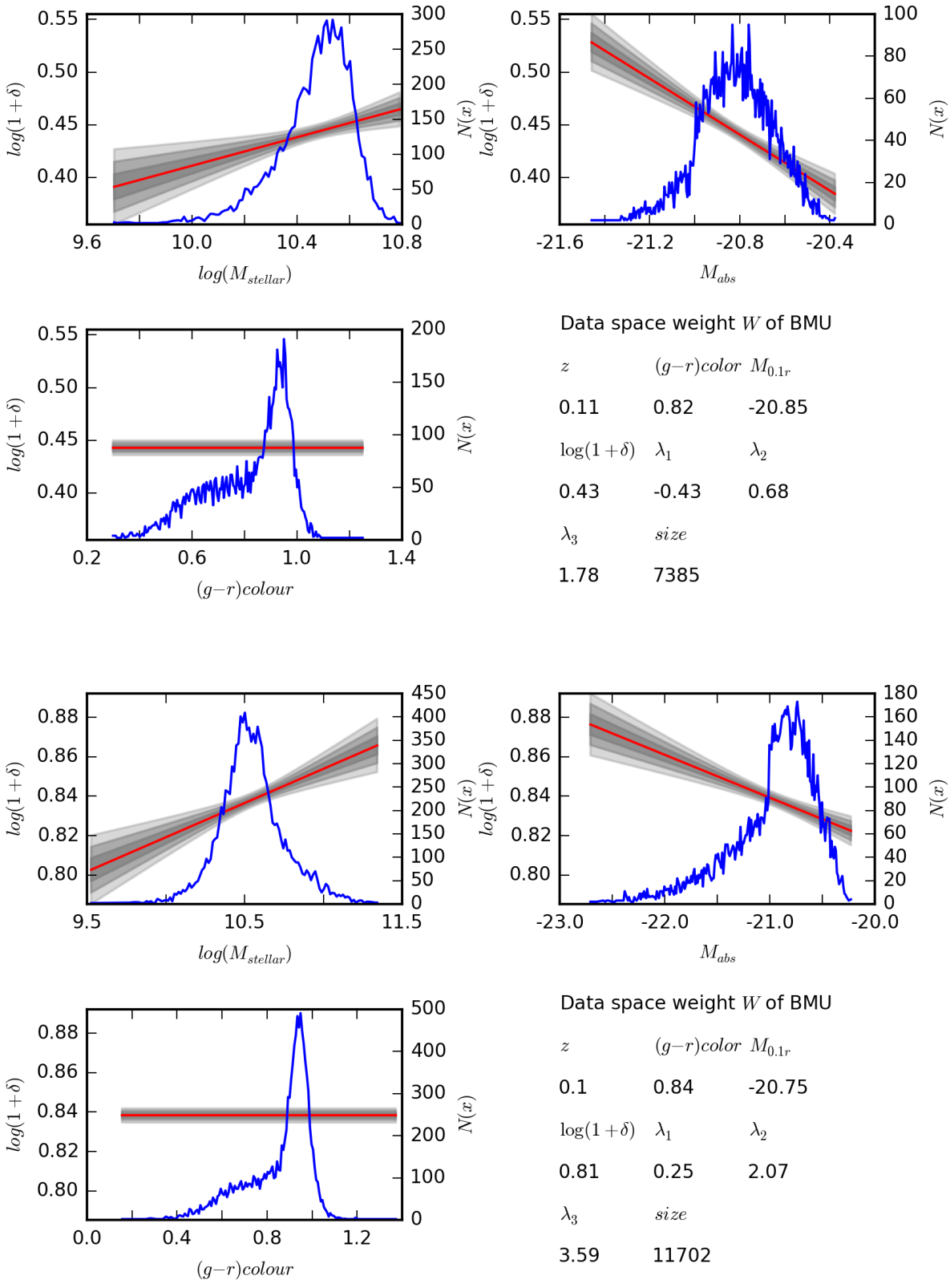

Figure 20: The plots have the same structure as in Fig. 19 but the fitting was applied to sub-samples of data for different neurons. Note that for the lower pictures the neuron is in a region associated with galaxy clusters (or DM halos) according to the eigenvalues of the tidal shear tensor λ_i.

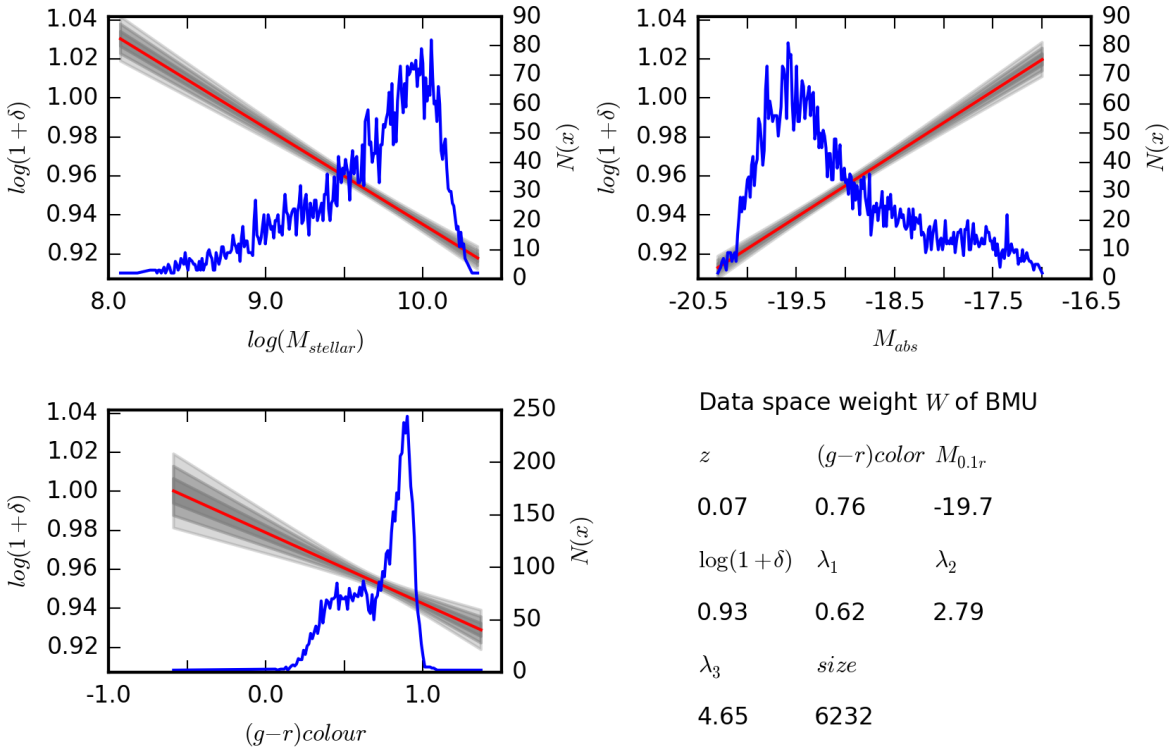

Figure 21: The plots have the same structure as in Fig. 19 but the fitting was applied to subsamples of data for different neurons. Note that this data cluster shows an inverted correlation trend compared to the correlation structure of the data clusters in Fig. 19 and 20.

7 Summary and outlook

7.a Summary

This thesis presents the development and application of a Bayesian inference approach to search for correlations between different observed quantities in cosmological data sets. Specifically we focus on identifying relations between the various properties of galaxies and the cosmic large-scale-structure (LSS). This is of particular scientific interest, since the properties of galaxy formation and evolution are assumed to be directly linked to the LSS of our Universe. Studying the correlation between galaxies and their LSS environment will hence give further insight into the process governing galaxy formation.

In order to determine correlation structures we implemented a parameter free method based on a discretized version of the Shannon entropy in Chapter 2. In principal, this method is able to determine correlations without specific knowledge about the correlation structures and the data generation process, but suffers from noisy data. To counter this problem we introduce a parametrized approach to test for correlation in data in Section 3.

This results in a method based on Bayesian inference including the assumption on correlation structures to be parametrized as a polynomial with unknown order. The method infers a posterior PDF of the coefficients describing the correlation polynomial via a Wiener Filter approach. To automatically choose the order of polynomials as supported by the data, we employ a model selection method based on the "Bayesian information criterion" (BIC). The BIC compares the likelihood of different models matching the data. Apart from our initial restrictions, this allows us to compare galaxy properties to properties of the LSS without prior information about correlation structures.

In this work we rely on galaxy data provided by the Sloan Digital Sky Survey (SDSS DR7 Abazajian et al. (2009)). The data set includes various galactic properties such as r-band absolute magnitude, stellar mass, color, redshifts and angular positions. Our developed method analyses the connection between these quantities and selected properties of the LSS. LSS information, as used in this work, is provided by the BORG algorithm, a fully Bayesian inference framework to analyze the 3D density fields in observations. The method provides detailed 3D density fields including a detailed treatment of uncertainties inherent to observations. This is achieved by providing a set of data constrained realizations of the LSS via an efficient implementation of a Monte Carlo Markov Chain algorithm (see Jasche & Wandelt (2013)).

The density fields obtained by the BORG algorithm in combination with our developed methods provide a generic framework to relate cosmological objects to the properties of the LSS. Inferred density fields permit to derive a variety of different properties including eigenvalues and eigenvectors of the tidal shear tensor of the gravitational potential, web type classifications and other properties. In order to study correlations, we match loca-

tions of galaxies in the SDSS to the density field and assign LSS properties to respective galaxies resulting in a combined galaxy-LSS dataset described in Section 4. In this work the analysis is limited to the nearby northern sky, up to a redshift of $z = 0.38$.

The application of our method to the combined data in Section 5 verified that there exists correlation. These correlations appear to be superpositions of different correlation structures. We assume different correlation structures to belong to different data generation processes. Therefore we seek to find a way to distinguish sub-samples of data belonging to different processes. Dividing the combined data into sub-samples according to different properties of data attempts to disentangle these processes manually. Manual organization of such datasets appears to be a challenging task. Therefore we propose to use automatic approaches to adequately and accurately sub-divide data.

Specifically, in this work we relied on a specific kind of artificial neural network called "Self Organizing Map" (SOM) in order to tackle the sub-division problem. A SOM seeks to classify and distinguish data clusters. The strength of this method is its ability to separate sub-samples of data in noisy and highly structured observations. We applied the SOM to our combined galaxy-LSS data in Section 6. The application results in sub-samples of galaxies which appear to hold unique properties of the data space indicating a clustered structure of galaxy data in the chosen properties. These results provide a new approach to a more natural classification of galaxies within their LSS environments in comparison to manual sub-division.

In order to reveal correlation structures belonging to specific sub-samples, we applied our correlation analysis method to individual data clusters in Section 6.c. The reconstructed trends indicate varying correlation structures of different sub-samples. As an example we found that the correlation between the density field and absolute magnitudes appears to be stronger in filaments compared to clusters. In contrast, other sub-samples, although separate in data space, appear to show similar correlations. The combined results ranging from the classification of galaxies according to galaxy and LSS properties to the revealed correlation structures provide insight into galaxy formation in specific cosmic environments.

The generic framework of our method allows a simple analysis for many different kinds of datasets, including highly structured and noisy data. In addition, our methods for classification and correlation determination are applicable in different but related fields. In the following we give a brief overview over possible future extensions and applications.

7.b Outlook and possible fields of application

This work discusses a proof of concept for the proposed methods, enabling to us address many different data analysis tasks. Our analysis indicates that the combined method was successfully applied to specific galaxy and LSS properties. Therefore as a first step

we propose to enlarge the data space by additional galaxy properties such as the star formation rates, the 4000 Å break strength, and other common properties. Other LSS properties available from the reconstructions, for example the ellipticity of the gravitational potential, can be included as well. Obtained results may give further insights into galaxy formation in relation to the LSS. The proposed method provides a general data analysis pipeline that can readily be applied to coming and more detailed density fields obtained by the BORG algorithm.

Another interesting project consists in applying the classification methods of the SOM to a dataset containing the full spectra of galaxies in addition to their LSS properties. Inferred galaxy properties, used in this work, are derived from observed spectra involving assumptions on galaxy formation and evolution. Therefore classifying galaxies in their LSS environment according to their spectrum with SOMs seems to be a natural next step towards less model dependent analysis. Such a classification will result in spectral groups, each associated to a sample of galaxies and could therefore be of major interest to the field of studying galaxies in their LSS environment.

A particularly interesting aspect of the BORG reconstruction is the fact that it also provides information on the cosmological initial conditions from which the present LSS formed. Initial LSS properties could be included in the analysis to connect properties of present day galaxies to the conditions in the early Universe and thereby connect primordial physics with late time cosmology.

References

Abazajian K. N., Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Allende Prieto C., An D., Anderson K. S. J., Anderson S. F., Annis J., Bahcall N. A., et al. 2009, , 182, 543

Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Anderson K. S. J., Anderson S. F., Annis J., Bahcall N. A., Baldry 2006, , 162, 38

Barnum H., Barrett J., Orloff Clark L., Leifer M., Spekkens R., Stepanik N., Wilce A., Wilke R., 2010, New Journal of Physics, 12, 033024

Bennett C. L., Banday A. J., Gorski K. M., Hinshaw G., Jackson P., Keegstra P., Kogut A., Smoot G. F., Wilkinson D. T., Wright E. L., 1996, , 464, L1

Blanton M. R., Hogg D. W., Bahcall N. A., Brinkmann J., Britton M., Connolly A. J., Csabai I., Fukugita 2003, , 592, 819

Blanton M. R., Roweis S., 2007, , 133, 734

Blanton M. R., Schlegel D. J., Strauss M. A., Brinkmann J., Finkbeiner D., Fukugita M., Gunn J. E., Hogg D. W., Ivezić Ž., Knapp G. R., Lupton R. H., Munn J. A., Schneider D. P., Tegmark M., Zehavi I., 2005, , 129, 2562

Erdoğdu P., Lahav O., Zaroubi S., Efstathiou G., Moody S., Peacock J. A., Colless M., Baldry I. K., Baugh C. M., Bland-Hawthorn J., Bridges T., Cannon 2004, , 352, 939

Gamow G., 1946, Physical Review, 70, 572

Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, , 375, 489

Jasche J., Wandelt B. D., 2013, , 432, 894

Kaiser N., 1987, , 227, 1

Kolb E. W., Turner M. S., 1988, The early universe. Collection of reprints

Lee J., Li C., 2008, ArXiv e-prints

Liddle A. R., 2007, , 377

Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution

Planck Collaboration Ade P. A. R., Aghanim N., Arnaud M., Ashdown M., Aumont J., Baccigalupi C., Banday A. J., Barreiro R. B., Bartlett J. G., et al. 2015, ArXiv e-prints

Spergel D. N., Bean R., Doré O., Nolta M. R., Bennett C. L., Dunkley J., Hinshaw G., Jarosik N., Komatsu E., Page L., Peiris H. V., Verde L., Halpern M., Hill R. S., Kogut 2007, , 170, 377

York D. G., Adelman J., Anderson Jr. J. E., Anderson S. F., Annis J., Bahcall N. A., Bakken J. A., Barkhouser R., Bastian S., Berman 2000, , 120, 1579

Hiermit erkläre ich, die vorliegende Arbeit selbständig verfasst zu haben und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel benutzt zu haben.

Datum: München, Unterschrift: