Orlin Koop

Higher order effects in cross-correlation based estimates of the galaxy luminosity function

Masters thesis

June 30, 2020

Thesis supervisor: Dr. M. P. van Daalen

Leiden University
Sterrewacht

Abstract

Large galaxy surveys bring in enormous amounts of photometric data of which spectroscopic follow-up would be desirable, but is practically infeasible. Still we would want to know the redshift of these sources to infer information on the evolution of our universe. Many methods to assign redshifts to photometric data are, however, biased, but [van Daalen and White, 2018] conceptually proved the effectiveness of their extension of existing methods for using cross-correlations to derive redshift distributions for photometric galaxies. Their method simultaneously yields a redshift-dependent galaxy luminosity function, based on a parametric model using a Schechter function to fit to data. Still, higher order effects like K-corrections and other theoretical defects play a role. Results of additions to the model to counteract these defects are presented and explained. These include the addition of a second Schechter function, the addition of a new parameter to fit the luminosity dependent galaxy bias factor and the incorporation of the K-corrections. The extended model with the second Schechter is shown to work sufficiently on a simulated mock galaxy catalogue. The addition of the luminosity bias parameter seems to have a negative effect on results, but can be helpful when applying the method to actual data. Methodology for incorporating K-corrections is presented and preliminary results are shown, but these show that the model is not flexible enough yet. We present possible causes hereof and discuss future paths research should thus take.

Contents

Introduction

Over the last few decades, cosmological studies have made significant progress due to technological advancement enabling researchers to perform larger and more detailed cosmological simulations to predict observations progressively more precise using constraints from others, and to constrain the importance of physical processes in the history of our universe. Current state-of-the-art cosmological simulations are quite able to recover observational constraints from large imaging surveys, which is advantageous for furthering our knowledge of the evolution of the Universe. Simulating the universe enables us to take a snapshot at any given cosmological timestamp, or *redshift*, and compare it with current observations from more distant regions of the observable universe. Still the underlying physics could be simulated and constrained to greater detail, to further our understanding of the origin and evolution of our universe.

One of the key constraints in this process is recovering the *galaxy luminosity function* (GLF), $\phi(L)dL$, of the galaxy population, offering powerful constraints on models of galaxy evolution since all galaxy formation theories and simulations should return a GLF equal to the observed one. Additionally, the GLF is linked to a selection function on how many galaxies are visible to us given their redshift and luminosity, and thus linked to the redshift distribution of galaxies. Observing and constraining this redshift distribution helps with studying cosmological parameters, galaxy scaling relations and gravitational lensing effects.
The GLF describes the number density of galaxies with luminosities in the range $L \pm dL/2$. An example is seen in Figure 1.

The GLF commonly seems to follow a power law that is truncated by an exponential cutoff at the bright end (lower absolute magnitudes). This behaviour is observed in all wavebands, so that the universal GLF is commonly fitted by a *Schechter function* of the form:

$$\phi(L)dL = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} \exp\left(-\frac{L}{L^*}\right) \frac{dL}{L^*}, \tag{1}$$

or, transformed to absolute magnitudes by $M - M^* = -2.5\log(L/L^*)$:

$$\phi(M)dM = 0.4\ln(10)\phi^* 10^{0.4(\alpha+1)(M^*-M)} \exp\left[-10^{0.4(M^*-M)}\right]. \tag{2}$$

Here M^* (L^*) is a characteristic magnitude (luminosity), α is the faint-end slope, and ϕ^* is an overall normalization. An example is seen in Figure 2

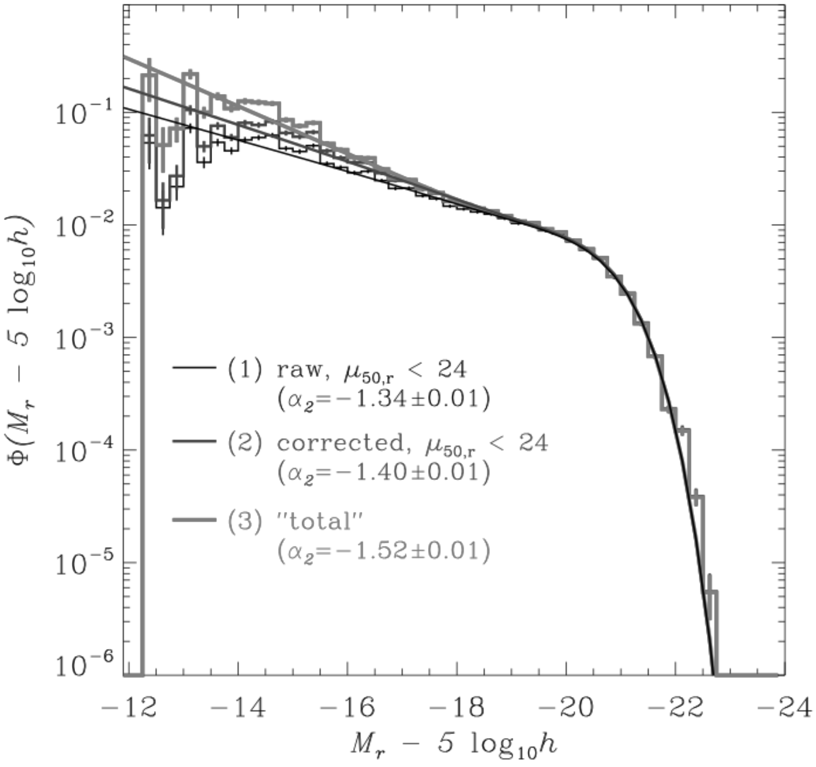

Figure 1: *(From [Blanton et al., 2005]) Example GLF for r-band magnitudes taken from the SDSS galaxy survey. $\mu_{50,r}$ here signifies the half-light brightness and α_2 is the low-luminosity slope.*

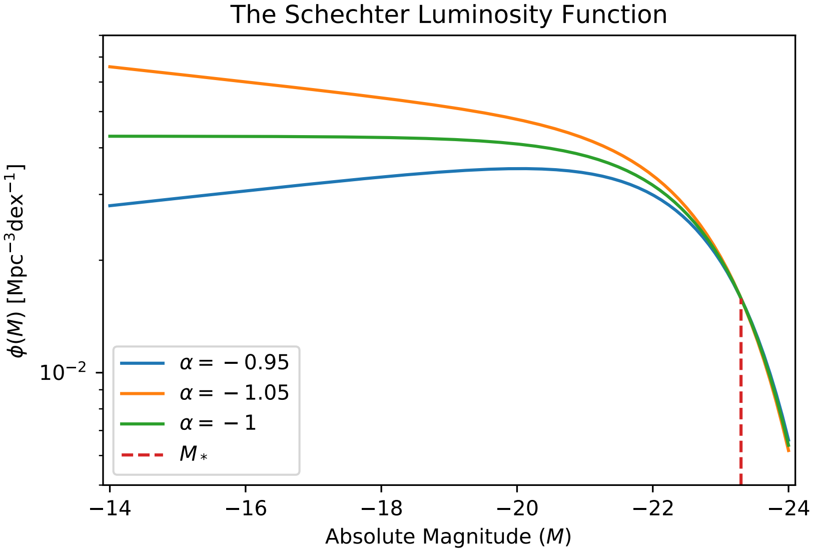

Figure 2: *Schechter functions for various values of the low luminosity slope parameter α.*

Technically, the parameters of a GLF depend not only on the waveband, but also on the morphological type, color, redshift and environment of a galaxy. Therefore, one of the most challenging problems in galaxy formation and cosmology is to explain the dependence of these parameters on galaxy properties. It should be noted that the universal shape, which we will study here, can be explained by feedback processes on top of a halo mass function with the same shape (power law and exponential cutoff) due to the effect of galaxy evolution through cosmic time. This we can simulate given certain physical models and compare to observations. Therefore, to further our understanding of this evolution, a well-observed universal GLF through cosmic time is very valuable. Of course, if techniques for constraining this universal GLF work, they can also be applied to datasets selected on galaxy type, stellar mass, star formation rate etc.

To measure this universal GLF accurately, we need cosmological volumes that are as extensive as possible (choosing observed scales to allow the GLF to be as universal as possible), especially at the brighter ends of the luminosity function, where galaxies are rare. These large datasets are offered by general large imaging surveys, but the photometric redshifts derived from these lead to non-negligible uncertainties in the absolute magnitude of the galaxies, see e.g. [Bolzonella et al., 2000]. Therefore the aforementioned advantage offered by simulations cannot be exploited as well as we would like. Spectroscopic surveys, on the other hand, provide accurate redshift measurements, but can only do this for far fewer galaxies, because spectroscopy can only be performed on bright enough targets. Additionally, targets need to be far enough apart on the sky to avoid fiber collision on spectroscopes, and taking spectra is expensive, meaning that even if we have an ideal sample of bright, far apart galaxies, we can only take spectra for a fraction of them.

Many methods for deriving this redshift information indirectly from the vast photometric surveys already available have been developed, usually using a library of Spectral Energy Distributions (SEDs) and/or spectroscopic sources to train algorithms assigning a redshift to each galaxy. In general, however, these methods do not yield unbiased redshift distributions, because the data used to train algorithms consists of a biased population, namely those galaxies of specific (spectral) types, or those bright and/or close enough, see e.g. [Cunha et al., 2009, Bezanson et al., 2016].

Naturally, methods countering these problems in application have been researched by e.g. [Lima et al., 2008], but avoiding photometric redshifts is the preferable option. One way to do this is by using information about how strongly photometric galaxies cluster with sources with a known redshift. Even if there is a difference in galaxy bias (with respect to the underlying matter density field) between the two samples, both should still trace the same underlying large-scale structure as the overall galaxy population, making it statistically likely that two galaxies being close on the sky means that they are close along the line of sight as well. In figure 3 an example of this can be seen. Therefore, we could infer a statistical redshift distribution for photometric galaxies from clustering measurements.

Techniques exploiting clustering information to obtain redshift information have been used to characterize the errors of photometric redshift catalogues, reconstruct the density field or derive redshift distributions from clustering directly, e.g. [Padmanabhan et al., 2007, Choi et al., 2016, Cucciati et al., 2016] and [Matthews and Newman, 2010, Schulz, 2010, McQuinn and White, 2013], and [Ménard et al., 2013, Morrison et al., 2017]. Using clustering information one could again compare the observationally derived redshift-distribution with data from simulations (like Semi-Analytic Methods as in [Bates et al., 2019]), again prone to selection bias as before. Furthermore, galaxy bias factors with respect to the matter density field are impactful in this context, since there is no natural way of correcting for them. Therefore, in this thesis we fit the observations to data derived from modelling the GLF by the aforementioned Schechter function, therefore enabling us to model the bias evolution with redshift as well.

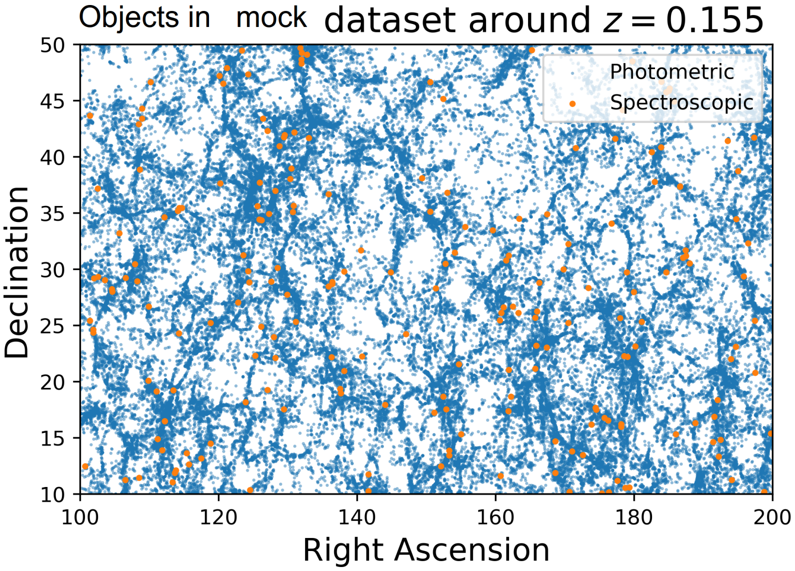

Figure 3: *Positions of objects around a redshift of $z = 0.155$ from a Mock Galaxy survey extracted from the Planck Millennium all-sky lightcones from [Henriques et al., 2015]. Blue points are randomly selected objects satisfying conditions for observation of apparent magnitude, and Orange points are randomly selected viable candidates for spectroscopic measurements. For details on selection, see Section 2.3. Notice the higher (lower) density of orange dots where blue dots are more (less) dense, signifying both samples tracing a similar underlying density field.*

We will improve the method introduced in [van Daalen and White, 2018] (which we will abbreviate to vDW) by implementing further flexibility in their model (which we refer to as the 'vDW (fiducial) model), beginning with the addition of a second Schechter function to be fitted. Furthermore, we implement K-corrections [1] to the derived absolute magnitudes by binning our data over colour and discuss how independent the bias evolution can be from the model, or if it should be given a free parameter as well. These higher order effects are of importance when applying the model to observational data, while it has been proven conceptually to work on simulated data altered to fit a single Schechter Function in vDW. Since losing information on properties of galaxies in each bin seemed to not be a problem, efforts have already been made to include colour binning in the analysis by [Rahman et al., 2016, Bates et al., 2019].

To this end, in the first chapter we present the underlying theory exploited by this model. This involves the theories surrounding the formation of the Large Scale Structure (LSS), the methodology behind clustering and the definition and details about the GLF.

In the second chapter we will explain the inner workings of the fiducial method introduced by vDW. We will give an overview of the theoretical background, process and the ways in which this model was tested. Hereby we thus proved that by cross-correlating two types of surveys, photometric and spectroscopic, luminosity functions for large volumes can be derived with smaller redshift uncertainties than otherwise possible.

Then, having explained the fiducial model, in the third chapter we explain the alterations that we make to the fiducial model, and show how these affect the outcome, discussing outcomes and future

[1] Corrections (due to the galaxy spectrum not being flat) to the distance modulus depending on the redshift of the spectrum of the galaxy between wavebands. For a general overview, see [Hogg et al., 2002].

prospects.

Chapter 1

Galaxy Clustering and the Luminosity Function

As explained in the introduction, to discuss our fiducial model we first need to explain the main theoretical context we derive this model in. This includes a longer introduction about the luminosity function, the formation of LSS and the phenomenology behind clustering. We will do this in the context of our assumed cosmology, the ΛCDM model from [Planck Collaboration, 2016].

1.1 The Galaxy Luminosity Function

We have already introduced the GLF $\phi(M)dM$ in the introduction (equation (2)), but have spared further details and remarks for this section. A general overview of previous research concerning the GLF is found in [Johnston, 2011].

Suppose we have a magnitude limited sample including all galaxies in a patch of sky with $m < m_{\mathrm{lim}}$, where m_{lim} denotes the limiting apparent magnitude of our sample. The absolute magnitude corresponding to a given m is then given by:

$$M = m - 5\log(d_{\mathrm{L}}(z)/\mathrm{Mpc}) - 5 - K(z) - E(z) + L, \tag{1.1}$$

where $E(z)$ is a correction factor for dust extinction, L is a term for lensing magnification due to gravitational lensing and $K(z)$ is the *K-correction*, correcting the observed flux into a fixed rest-frame band, so that absolute magnitudes are the same for identical galaxies at different redshifts. Note that $K(z) = 0$ if galaxy spectra are flat. K-corrections depend on the type of galaxy and its spectra to correct for the magnitude difference due to redshifting of this spectrum. Dust extinction and lensing magnification effects are outside the scope of this thesis and will be ignored from now on.

Because the survey is magnitude limited, a galaxy with a luminosity L will only be part of the survey if it is located within a maximum luminosity distance d_{max}, or z_{max}, such that:

$$5\log(d_{\mathrm{max}}(L)/\mathrm{Mpc}) = m_{\mathrm{lim}} - M^* - 5 + 2.5\log(L/L^*) - K(z_{\mathrm{max}}), \tag{1.2}$$

where $2.5\log(L/L^*) = M^* - M$ for a reference object with absolute magnitude (luminosity) M^* (L^*). Thus we can introduce a maximal volume V_{max} out to z_{max} in which an object with certain luminosity L can be detected:

$$V_{\mathrm{max}}(L) = \frac{\Omega}{3}\left[\frac{d_{\mathrm{max}}(L)}{1+z}\right]^3, \tag{1.3}$$

where Ω is the area of the sky covered by the survey. Then $\phi(L)dL = \sum_i \frac{1}{V_{\max}(L_i)}$, where i ranges over all galaxies in $L \pm dL/2$. To understand this, recall the short introduction and definition of the GLF as the number density of galaxies per unit luminosity (or magnitude) in equation (2). A natural question to ask now is what the dependencies of the GLF are, and what it generally looks like, which we discuss below. The discussion is based on [Mo et al., 2010, pp. 654-658].

Firstly, the observed GLF is generally bimodal ([Binney and Tremaine, 2008, pp. 757 & 777]), which is explained due to its dependence on galaxy color and thus observed waveband. Redder galaxies are better fitted with a Schechter function with brighter characteristic magnitude M^*, and a shallower faint-end slope than the bluer galaxies. In [Blanton et al., 2005] it is shown that the red population does seem to show an upturn at the faint end, but these galaxies usually have low surface brightness, and the significance of this is difficult to quantify due to cosmic variance effects due to the small cosmic volume in which these galaxies can actually be detected. Still, both populations seem to be fitted well by a Schechter function separately (possibly ignoring the upturn).

Additionally, morphology of galaxies is strongly correlated with their color, and the bright end of the GLF is dominated by ellipticals, while the intermediate range is by spirals, and the faint by irregulars and dwarf ellipticals.

Secondly, galaxy environment could be of impact on the shape of the GLF. The GLF has been estimated for a number of nearby galaxy clusters, and compared to those in the general field. The observational results show the GLF of cluster galaxies has a steeper slope, with a marked faint-end upturn, as well as a more extended tail at the bright end. Within the same morphology class, however, the GLFs do agree between cluster and field galaxies. This means the relative amplitudes of the GLFs of different morphological subclasses change with environment.

Lastly, the dependence of the GLF on redshift has been studied, and current telescopes are able to give complete samples of galaxies out to very faint magnitude limits, in various bands, up to higher redshifts. The red and blue populations show different evolution through time, and as a consequence the GLF has indeed changed significantly. For instance, the total stellar mass density in the red population has doubled since $z \sim 1$ while that in the blue stays roughly constant, so galaxies form in the blue regime, but when star formation is quenched they join the red population.

There is quite a noticeable difference between observed galaxy counts and the predicted galaxy counts in a non-evolving GLF model in non-expanding space. Two different kinds of evolution can explain this. The first of these is luminosity evolution of galaxies, where the comoving number of galaxies does not change, but at higher redshift the intrinsic brightness of the galaxies increases, meaning at higher redshifts galaxies are brighter. The second explanation would be number evolution, due to mergers and creation of galaxies. In the case of only luminosity evolution, galaxy counts can be obtained from the local GLF together with an assumption about how luminosities of galaxies evolve with redshift. To include this effect, one often adds an E-correction, $E(z)$, in the distance modulus, signifying the change in magnitude in the observational pass-band for a galaxy at redshift z when it is evolved to present time.

As can be seen, there are many effects to take into account when trying to determine the observational GLF, and the question is which of these is most important in the context of the fiducial model as mentioned in the Introduction.

Since we are searching for a universal GLF, only limited by magnitude and spatial scales chosen (as will be discussed later on), the effects due to color, morphology or environment should not impact our results too much, although it is good to keep them in mind and know where they come from.

Furthermore, due to random selection of targets, which are sampled in the same manner, only differing in conditions for objects we have photometric and/or spectroscopic information on, our model will be independent of bias that could enter through selection effects. This we will explore in more detail in chapter 2 and 3.

1.2 Formation of Large Scale Structure

The discussion in this section is based on [Mo et al., 2010, pp. 202-204,206,263,283-285,679-681] and [Coles and Lucchin, 2002, pp. 339-342].

Since we will be discussing the distribution of galaxies in our Universe, we need to understand up to some level how they even formed and what processes lie behind the current observable state of this Universe. Therefore we will shortly discuss the current theory behind galaxy formation. In the ΛCDM model, we assume the existence of a Big Bang, and quickly afterwards a period of rapid expansion termed *inflation.* From observations of the cosmic microwave background (CMB), electromagnetic radiation dating from the epoch of recombination, we know the universe began in a hot, dense and nearly uniform state, but this has not persisted, since we observe all kind of non-uniform structures on all scales, from stars and planets up to galaxy clusters and voids, walls and filaments. This structure forms from gravitational instabilities of small, early density fluctuations that slightly altered the isotropy of the universe at that time. These are, in the context of cosmic inflation, formed as amplifications of minute, random quantum fluctuations in the pre-inflation plasma.

The next step should be to understand the time evolution of these perturbations in the expanding universe. For structures with sizes much smaller than the horizon scale (measuring the distance from which one could retrieve information, i.e., the scale at which structures are causally connected), so that causality is instantaneous, having density contrasts relative to the background much smaller than unity, we can use Newtonian perturbation theory. The relativistic extension of this should be used for superhorizon fluctuations or when the matter in the perturbation is not a non-relativistic fluid. Decomposing the perturbations in Fourier modes shows that some are amplified while others are damped, so the evolution acts as a filter on the primordial perturbations and can be described by a transfer function. Once the overdensity has grown to a certain scale, evolution becomes non-linear due to gravitational instability.

Usually, we describe the *density perturbation field* as:

$$\delta(\mathbf{x}, t) = \frac{\rho(\mathbf{x}, t)}{\overline{\rho}(t)} - 1, \tag{1.4}$$

where $\rho(\mathbf{x}, t)$ is the density at position $\mathbf{x}$ and time t, and $\overline{\rho}(t)$ is the density averaged over space. In the linear regime, density perturbations evolve independently of each other, and the amplitudes depend on the aforementioned transfer function. When we want to observe the current structure, we need to find the statistical properties of this cosmic density field. Specifying $\delta(\mathbf{x})$ (or $\delta_{\mathbf{k}}$ for all $\mathbf{k}$, where $\mathbf{k}$ is a wave-vector when having transformed into the Fourier domain) is impractical and unnecessary, for we consider the mass density field in the universe as one realization of a random process (the aforementioned quantum fluctuations produce a random field well approximated by a homogeneous and isotropic Gaussian random field). In similarity with statistical physics, we thus seek to describe properties of a 'gas' for which we do not need to know all positions and velocities, but only their distribution functions and statistical properties. In other words, the statistical properties of the random perturbation field $\delta(\mathbf{x})$ are specified if the probability for any realization $\delta(\mathbf{x})$ is known. When the universe is divided into n infinitesimal cells centered at $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$, we can characterize $\delta(\mathbf{x})$ by the probability distribution:

$$\mathcal{P}_x(\delta_1, \delta_2, ..., \delta_n) d\delta_1 d\delta_2 ... d\delta_n, \tag{1.5}$$

giving the probability that δ has values in $[\delta_i, \delta_i + d\delta_i]$ at positions $\mathbf{x}_i$. If we know all the moments $\langle \delta_1^{l_1} \delta_2^{l_2} ... \delta_n^{l_n} \rangle$, where l_i are non-negative integers, we completely determine this distribution function. Since the cosmological principle requires equivalence of all positions and directions, the cosmological density field has to be statistically homogeneous and isotropic as well, which we also observe to a high degree in the CMB. In other words, the moments need to be invariant under spatial translation and rotation. The most straightforward moments are of course $\langle \delta(\mathbf{x}) \rangle = 0$ (which follows from the definition of $\delta(\mathbf{x})$ and $\sigma^2 = \langle \delta^2(\mathbf{x}) \rangle$, which should be independent of $\mathbf{x}$ due to assumed ergodicity[1]). For the discussion in this thesis, the most important moment is the *two-point correlation function*:

$$\xi(x) = \langle \delta_1 \delta_2 \rangle; \qquad x \equiv |\mathbf{x}_1 - \mathbf{x}_2|, \tag{1.6}$$

of which we see that $\xi(0) = \sigma^2$. When finding the second moment of the Fourier transform of the density perturbation field, also known as the *power spectrum*, we can see that it is the Fourier transform of the two-point correlation function. It can be shown that a homogeneous and isotropic Gaussian random field is completely determined in a statistical sense by its power spectrum. We dive a little bit deeper into this power spectrum when introducing the galaxy bias factors below.

1.2.1 Clustering

Usually when wanting to 'measure' clustering, one starts with a sample for which sky positions and redshifts are listed for all members, providing a non-uniform sampling of the true galaxy distribution throughout a finite volume. It should be noted here already that, for naive sample selection, only intrinsically bright galaxies are included at high redshift, while more different types of galaxies are included at lower redshift, so observational criteria induce strong selection effects, which could lead to a bias. We, however, will be exploiting the clustering between samples with accurate redshift measurements and samples without redshift measurements. A *selection function* $S(\mathbf{x})$ could be defined, describing which galaxies are or are not included in the sample. It is the probability that a 'random' galaxy located near $\mathbf{x}$ is included in the sample. It could, for example, vary strongly in $|\mathbf{x}|$ or z because of apparent magnitude limits. Once we characterize this selection function, it can be used to correct for the missed galaxies due to selection effects.

Example 1.2.1. (From [Mo et al., 2010]) Consider a survey which is magnitude limited in the sense that it only selects all galaxies with $m < m_{lim}$, so brighter than m_{lim}. Galaxies with apparent magnitude m at redshift z have absolute magnitude $M = m - 5\log(d_{\mathrm{L}}(z)/\mathrm{Mpc}) - 5$, where $d_{\mathrm{L}}(z)$ is the luminosity distance corresponding to the measured redshift, and we ignore dust extinction and K-correction effects. Therefore, m_{lim} corresponds to a luminosity limit L_{lim}, and the selection function can be written as:

$$S(z) = \frac{\overline{n}(z)}{\overline{n}_0} = \frac{\int\limits_{L_{\mathrm{lim}}}^{\infty} \phi(L) dL}{\int\limits_{0}^{\infty} \phi(L) dL},$$

[1] Ergodicity means that the dynamical system has the same behaviour averaged over time as averaged over probability space, meaning the state after a long time is independent of its initial state

where $\phi(L)$ is the galaxy luminosity function. ◀

We have briefly seen the two-point correlation function $\xi(x)$ above, but that is derived from theory. Now we will consider how to recover the two-point correlation function from observations. One can write the joint probability dP_{12} of finding one galaxy in a small volume dV_1 and another in dV_2, separated by $\mathbf{r}_{12}$ as:

$$dP_{12} = n_{\mathrm{g}}^2[1 + \xi_{\mathrm{gg}}(\mathbf{r}_{12})]dV_1 dV_2, \tag{1.7}$$

where n_{g} is the mean number density of galaxies, and $\xi_{\mathrm{gg}}(r)$ is the two-point galaxy-galaxy spatial correlation function. Note that, because of assumed statistical homogeneity and isotropy, it should only depend on the modulus of $\mathbf{r}_{12}$.

Example 1.2.2. Equation (1.7) implies that the mean number of galaxies $\langle N\rangle_r$ within a distance r of a given galaxy is equal to:

$$\langle N\rangle_r = \frac{4}{3}\pi n_{\mathrm{g}} r^3 + 4\pi n_{\mathrm{g}} \int\limits_0^\infty \xi(r'_{12}) r'_{12} dr'_{12}, \tag{1.8}$$

where the second term represents the excess number compared to a uniform random distribution. ◀

If one only has a projected, two dimensional catalogue, as we will have in later chapters, one can define the *two-point galaxy-galaxy angular correlation function* $w(\theta)$:

$$d^2P_2 = n_\Omega^2[1 + w(\theta_{12})]d\Omega_1 d\Omega_2, \tag{1.9}$$

where n_Ω is the mean number of galaxies per unit solid angle and d^2P_2 is now the joint probability of finding two galaxies in small elements of solid angle $d\Omega_1$ and $d\Omega_2$ separated by an angle θ_{12} on the celestial sphere. These two correlation functions can be related through the Limber equation, providing an integral relation between the two for small angles.

Estimators

There are several ways to estimate the two-point correlation function, and at small distances these provide similar performance. However, at larger distances, some of them could be biased, mostly due to magnitude and survey volume limitation in the galaxy sample. Therefore, a good estimator needs an edge-correction, mostly for the larger scales, at which only a small fraction of galaxies enters the estimation ([Kerscher et al., 2000, Landy and Szalay, 1993]). Given a complete galaxy sample in a given volume with N objects and a Poisson catalog (a binomial process with N_{rd} points, generated within the same boundaries), possible estimators are:

$$\begin{aligned} \hat{\xi}_{\mathrm{DP}}(r) &= \frac{N_{\mathrm{rd}}}{N}\frac{\mathrm{DD}(r)}{\mathrm{DR}(r)} - 1; \quad \hat{\xi}_{\mathrm{HAM}}(r) = \frac{\mathrm{DD}(r)\mathrm{RR}(r)}{|\mathrm{DR}(r)|^2} - 1; \\ \hat{\xi}_{\mathrm{LS}}(r) &= 1 + \frac{N_{\mathrm{rd}}(N_{\mathrm{rd}}-1)}{N(N-1)}\frac{\mathrm{DD}(r)}{\mathrm{RR}(r)} - 2\frac{N_{\mathrm{rd}}-1}{N}\frac{\mathrm{DR}(r)}{\mathrm{RR}(r)}, \end{aligned} \tag{1.10}$$

known as the *Davis and Peebles*, *Hamilton* and *Landy-Szalay* estimators. Here, $\mathrm{DD}(r)$ is the number of pairs of galaxies with separation within $r \pm dr/2$, $\mathrm{DR}(r)$ is the number of pairs between a galaxy and a point of the Poisson catalog and $\mathrm{RR}(r)$ is the number of pairs in the Poisson catalog with the same distance interval. Often, the N are absorbed into the definition of DD, DR and RR, so then the Landy-Szalay estimator becomes:

$$\hat{\xi}_{\mathrm{LS}}(r) = \frac{\mathrm{DD}(r) - 2\mathrm{DR}(r) + \mathrm{RR}(r)}{\mathrm{RR}(r)}. \tag{1.11}$$

In the case of the angular correlation function the same estimators can be used, but with pairs separated by an angle θ. This sets up some of the terminology and ideas used in the discussion of the galaxy bias in the next subsection, where we will show that the galaxy bias consist of two main parts, a halo bias and a luminosity bias.

1.2.2 Galaxy bias

The origin of cosmological perturbations has not been unraveled enough to have a refined theory of it, so we cannot constrain the overall amplitude of the linear power spectrum from theory, and thus need observations, for instance from the CMB, which is still linear. We know how to evolve the power spectrum from the CMB using the theoretical growth factor. One could compare with N-body simulations or linear theory on sufficiently large scales, but one cannot be sure about the accuracy of this methodology for finite amplitude fluctuations, especially because linear theory only holds on large scales, where sampling noise will make measuring the small fluctuation difficult. Non-linear evolution of dark matter fluctuations is very precise nowadays, so the only real remaining problems pertain baryonic matter, and thus include galaxies and feedback processes.

One needs to make sure the sample one uses to measure clustering is large enough to be representative of the universe as a whole. If a finite sample is used, the value of the statistic could differ when one would take a sample of the same size at a different position. This effect is known as *cosmic variance.*

These problems are, however, overshadowed by the bias that could exist between galaxy fluctuations and mass fluctuations. It is not far-fetched to assume galaxies should not form randomly purely according to the local matter density, but at specific locations where collapse, cooling and star formation can occur, for instance in peaks of the density field. Suppose the matter density field δ_M, smoothed on some scale M to define a mass scale of a galaxy, is Gaussian with variance equal to σ_M^2, then $\xi_{\mathrm{mm},M}(r)$, the two-point cross-correlation function of the matter at these mass scales, is equal to $\langle \delta_M(\mathbf{x})\delta_M(\mathbf{x}')\rangle$, where $r = |\mathbf{x} - \mathbf{x}'|$. If galaxies trace the mass, $\xi_{\mathrm{gg}}(r) = \xi_{\mathrm{mm},M}(r)$. Now imagine galaxies only form in high-density regions above some threshold $\delta_{\mathrm{c}} = \nu\sigma_M$, where we take the result from linear theory that $\delta_{\mathrm{c}} = 1.68$ is the threshold for non-linear collapse, and would thus be needed for structure formation. The correlation function for points exceeding $\nu_{\mathrm{c}} = \delta_{\mathrm{c}}/\sigma$ is then, for large ν_{c}, equal to $\xi_{\nu_{\mathrm{c}}} \simeq \exp[\nu_{\mathrm{c}}^2 w(r)] - 1$, where $w(r) = \xi(r)/\sigma^2$.

Using this as a qualitative model, one could define $\xi(r)_{\mathrm{gg}} = b^2\xi(r)_{\mathrm{mm}}$, where b is known as the *bias factor*. One could also define $\delta_{\mathrm{g}} = b\delta_{\mathrm{m}}$, the *linear bias model*, which does not follow from the first definition, but does imply it.

Linear Bias and the Power Spectrum

In this linear bias model it makes sense to look back at the power spectrum and see how bias enters in the spatial correlation function through there. Remember that the power spectrum is the second moment of the Fourier transform of the density perturbation field, thus (in the Fourier domain):

$$\xi_{\mathrm{mm}}(x) = \frac{1}{V}\sum_{\mathbf{k}} P(k)e^{i\mathbf{k}\cdot\mathbf{x}} = \frac{1}{(2\pi)^3}\int P(k)e^{i\mathbf{k}\cdot\mathbf{x}}d^3\mathbf{k}. \tag{1.12}$$

Where then $P(k) = V\langle|\delta_{\mathbf{k}}|^2\rangle$ is the power spectrum, and the second expression above follows from taking V to infinity in the usual Fourier transform conventions. Integrating out the angle between $\mathbf{k}$ and $\mathbf{x}$ we find:

$$\xi_{\mathrm{mm}}(x) = \frac{1}{2\pi^2}\int_0^\infty k^3P(k)\frac{\sin(kx)}{kx}\frac{dk}{k} = \int_0^\infty \Delta^2 j_0(kx)\frac{dk}{k}, \tag{1.13}$$

where we redefine a dimensionless power spectrum $\Delta^2(k) = k^3P(k)/2\pi^2$ and $j_0(kx) = \sin(kx)/kx$ is the spherical Bessel function.

In the linear bias model, we can now easily implement the bias factor into this expression to relate the linear power spectrum Δ with the two-point (auto)correlation function of the spectroscopic sample (having bias factor b_{s}) and the photometric sample (having bias factor b_{p}) as follows:

$$\xi_{\mathrm{ss}}(r) \approx \int_0^\infty \frac{dk}{k} b_{\mathrm{s}}^2\Delta^2(k)j_0(kr) \tag{1.14}$$

$$\xi_{\mathrm{ps}}(r) \approx \int_0^\infty \frac{dk}{k} b_{\mathrm{p}}b_{\mathrm{s}}\Delta^2(k)j_0(kr) \tag{1.15}$$

For linear scales, we can thus give the relationship between the two correlation functions as $\xi_{\mathrm{ps}}(r) = b_{\mathrm{p}}\xi_{\mathrm{ss}}(r)/b_{\mathrm{s}}$. This will be used and explored in further detail in the next Chapter. A different, somewhat more natural way to see this linear bias model is viable can be found in the Appendix.

All in all, however accurately one could predict mass fluctuations analytically and however robustly one can measure galaxy fluctuations, one cannot compare the two without assuming a relationship between them in the form of some bias model. The linear bias model is the simplest one, and not entirely correct. Instead, one could do a second-order bias correction, or make the bias scale-dependent. Knowing b would then eliminate the problem, but the linear bias model only holds on large scales (and perhaps not even then), and there could be regions where a model like this fails completely, for instance where the bulk of the matter is non-baryonic, non-luminous material, thus resulting in very little correlation between galaxies and concentrations of mass. There are ways to circumvent this. For instance, looking at peculiar velocities of galaxies (due to not only the luminous material) or by analyzing the amplitude of CMB fluctuations.

1.2.3 Luminosity Bias

Aside from the galaxy bias discussed in the previous section, there can also be a bias due to which types of galaxies form in a given overdensity (or halo) given that this overdensity is high enough that galaxies can form. This bias is named the *luminosity bias*, and we will introduce it here, to explain in the next chapter how we are going to use this.

One could write the GLF at a redshift z as:

$$\phi(L,z)dL = dL\int \Phi(L|M,z)n(M,z)dM, \tag{1.16}$$

where $n(M,z)$ is the mass function of dark matter halos at redshift z, and $\Phi(L|M,z)$ is the conditional luminosity function at z, specifying the average number of galaxies with luminosity in the range $L \pm dL/2$ residing in a halo of mass M at redshift z. $n(M,z)$ can be obtained from from i.e. simulations or accurate non-spherical collapse models.

Since the large-scale clustering amplitude of a galaxy population is determined by the mass distribution of their host halos, the observed galaxy correlation amplitude can be used to constrain the mass distribution of halos. Given a population around redshift z with luminosity function $\phi(L,z)$, combined with equation (1.16), we find:

$$b_{\mathrm{g}}(L,z) = \frac{1}{\phi(L,z)}\int \Phi(L|M,z)b_{\mathrm{h}}(M,z)n(M,z)dM, \tag{1.17}$$

for

$$b_{\mathrm{g}}(z) = \frac{1}{n_{\mathrm{g}}(z)}\int b_{\mathrm{g}}(L,z)\phi(L,z)dL; \quad n_{\mathrm{g}}(z) = \int \phi(L,z)dL. \tag{1.18}$$

When we can determine L completely and deterministically by M (and possibly z), $\Phi(L|M,z)$ becomes a Dirac delta function and $b_{\mathrm{g}}(L,z) = b_{\mathrm{h}}(M,z)$, and $L = L(M)$. Then, in the limit of large r, we can determine $b_{\mathrm{g}}(L,z)$ from:

$$b_{\mathrm{g}}(L,z) = \sqrt{\frac{\xi_{\mathrm{gg}}(r|L,z)}{\xi_{\mathrm{mm}}(r|z)}}, \tag{1.19}$$

where ξ_{gg} is the two-point correlation function of galaxies of luminosity L at redshift z and $\xi_{\mathrm{mm}}(r|z)$ that of matter at redshift z in the assumed model of structure formation.
As can be seen here, it is not unreasonable to assume that the galaxy bias can be split up into two components, one depending on the mass of the dark matter halo, which ultimately depends on z, and one only dependent on the luminosity of the observed object, without further dependence on luminosity or z.

In summary, the current large scale structure and galaxy distribution are determined by several ingredients interacting in a complicated way. There is a background cosmological model, a breakdown of matter into baryonic and non-baryonic matter, the latter of which could be hot, cold or a mixture. These underlying conditions supply a transfer function for the early random field of density fluctuations. These result in the initial power spectrum, which characterizes the linear regime of the fluctuation spectrum. The observed galaxy distribution is a biased tracer of the mass distribution caused by the growth of the fluctuations through cosmic time. This growth of the initial linear fluctuations becomes non-linear through gravitational collapse, forming structures such as filaments and haloes, which house the galaxies we observe. We analyzed how then to measure galaxy clustering, and how the galaxy bias factors into this, and how the bias depends on physical quantities. This is of importance to the model introduced in chapter 2, which uses cross correlations between galaxy samples to characterize the cosmic structure, so we will need to correct for the different biases between these galaxy samples.

Chapter 2

The Fiducial Model and Methodology

In this chapter we explain the basic idea of the model we use to find the universal redshift distribution and luminosity function of the universe, as mentioned, from clustering data by applying tomography to the luminosity function. We will start by explaining how we handle the clustering and the cross-correlation signal, and go on by explaining why we need to specify a model for $N_{\mathrm{p}}(m, z)$, the number of photometric galaxies in a given bin in redshift and apparent magnitude. Furthermore, we will finally specify this model itself. We explain basic results of the vDW fiducial model and explain choices made in the model as well as the reasoning behind them before talking about improvements on the model in chapter 3. We present here a combination of results from vDW, [Schulz, 2010] and [Ménard et al., 2013].

2.1 Cross-correlations

We will consider the observed number density distribution of a photometric sample of galaxies over apparent magnitude $n(m)$, as well as the distributions of galaxies over redshift in bins of apparent magnitude $n_{\mathrm{m}}(z)$, as projections of the underlying luminosity function $\phi(M, z)$, where M now is absolute magnitude and not mass, as a function of redshift, and can thus reconstruct this luminosity function from these projections. An added advantage of fitting a luminosity function this way is that we do not use the information about magnitude and redshift separately, but simultaneously, meaning we can account for effects we can only probe by using both sources of information. For instance: galaxies that appear bright are unlikely to be at high redshift, so we are less likely to overestimate the exponential cutoff using this methodology.

We denote photometric samples and their corresponding quantities with subscript p, and spectroscopic with subscript s. Note that we do not know the distribution in magnitude and redshift for the photometric sample, but we want to derive it. We do assume that we know the redshift distribution of the spectroscopic sample, albeit for a limited number of galaxies, significantly smaller than the number of photometric galaxies.

It is quite straightforward to state that the number of photometric-sample galaxies in apparent magnitude bin m_λ and redshift bin z_i is given by:

$$N_p(m_\lambda, z_i) = \int\limits_{z_{i,\min}}^{z_{i,\max}} \int\limits_{m_{\lambda,\min}}^{m_{\lambda,\max}} \frac{dN_{\mathrm{p}}}{dmdz}(m, z)dmdz, \tag{2.1}$$

where the $\cdot_{,\mathrm{min}}$ and $\cdot_{,\mathrm{max}}$ denote the edges of a given bin. To be able to derive the luminosity function we need to find the fraction of galaxies in the photometric sample $f_\mathrm{N}(m_\lambda, z_i)$ in a bin of apparent magnitude m_λ that also reside in a bin of redshift z_i. This fraction will serve as a weighting factor for the model value of the clustering $\widetilde{w}_{ps}(m_\lambda, z_i)$, as well specify the fraction of galaxies in our photometric sample with given absolute magnitude, i.e. the GLF. We can simply write this fraction $f_\mathrm{N}(m_\lambda, z_i)$ as:

$$f_\mathrm{N}(m_\lambda, z_i) = \frac{N_\mathrm{p}(m_\lambda, z_i)}{N_\mathrm{p}(m_\lambda)}; \quad N_\mathrm{p}(m_\lambda) = \sum_i N_\mathrm{p}(m_\lambda, z_i). \tag{2.2}$$

We already know the number counts $N_\mathrm{p}(m_\lambda)$ from the observations. In the model, however, the numbers $N_\mathrm{p}(m_\lambda)$ are not regarded as constraints that need to be enforced, but as draws from a Poisson distribution with mean $\widetilde{N}_\mathrm{p}(m_\lambda)$, which is the model value for $N_\mathrm{p}(m_\lambda)$ we will discuss in the next section. Therefore, when writing f_N from now on we mean it is the ratio of the model values for N_p.

We will now first limit the discussion to only deriving a redshift distribution of the photometric sample to explain the technique of using correlations to recover a redshift distribution, $\psi_\mathrm{p}(z)$, and later again extend to also finding a luminosity function.

Note that we can also write the redshift distribution of the photometric sample as:

$$\psi_\mathrm{p}(z) = \frac{dN_\mathrm{p}}{dzd\Omega}\left[\int_0^\infty \frac{dN_\mathrm{p}}{dzd\Omega}dz\right]^{-1}, \tag{2.3}$$

where $dN_\mathrm{p}/d_z d\Omega$ is the number of photometric galaxies per unit redshift and steradian, and the term in brackets is the total number to ensure $\psi_\mathrm{p}(z)$ is normalized. When we divide the survey in redshift bins, however, we denote the fraction of the total in the ith z-bin as $\psi_\mathrm{p}(z_i)$.

The other constraint, or 'signal', we have from the data is the aforementioned angular cross-correlation function of all photometric galaxies in apparent magnitude bin m_λ with spectroscopic galaxies in redshift bin z_i, $w_\mathrm{ps}(m_\lambda, z_i, \theta)$, because we can only observe this projected spatial distribution, not knowing the redshifts of the photometric sample.

Note that we can write:

$$w_\mathrm{ps}(z_i, \theta) = \int_0^\infty \xi_\mathrm{ps}(r(z, z_i, \theta))\psi_\mathrm{p}(z)dz \tag{2.4}$$

This now is where the analysis from the previous chapter becomes important. The $\xi_\mathrm{ps}(r(z, z_i, \theta))$ is the 3D cross-correlation function between the entire photometric set, and the spectroscopic galaxies in redshift bin i. This is not observable, because we do not know the redshifts of the photometric sample accurately enough. The key assumption now factors into the simple statement that $\xi_\mathrm{ps}(r) \propto \xi_\mathrm{ss}(r)$, where $\xi_\mathrm{ss}(r)$ is the 3D autocorrelation of the spectroscopic sample, which is observable. Of course, on large (linear) scales, we find, using the linear bias model from the previous chapter:

$$\xi_\mathrm{ps}(r) = \frac{b_\mathrm{p}}{b_\mathrm{s}}\xi_\mathrm{ss}(r) \qquad \Rightarrow \qquad w_\mathrm{ps}(\theta, z_i) = \int_0^\infty \frac{b_\mathrm{p}(z)}{b_\mathrm{s}(z)}\xi_\mathrm{ss}(r(z, z_i, \theta))\psi_\mathrm{p}(z)dz. \tag{2.5}$$

Since $b_\mathrm{s}(z)$ can be fitted with the spectroscopic data, we can, without further assumptions or techniques, only invert and solve this relation for the product $b_\mathrm{p}(z)\psi_\mathrm{p}(z)$ in terms of the observables.

To thus reliably find the redshift-dependent GLF, we need to know how b_p depends on z. Problem here is that for a magnitude limited sampling of the galaxies in the universe, $b_\mathrm{p}(z)$ is certainly not constant. If it was, it would scale out when normalizing $\psi_\mathrm{p}(z)$ to 1. The population of galaxies being examined at high redshift will certainly consist of the more brighter, rarer and more biased objects than those at low redshifts. This presents a degeneracy with the recovered redshift distribution. We will explain (here and in the following subsection) how we try to get rid of this degeneracy.

Even if the photometric survey were also volume limited there would be effects of evolution in intrinsic luminosities and densities of the tracers leading to the redshift dependent bias factor. One way to resolve the degeneracy could be by also measuring the angular autocorrelation function of the photometric sample w_pp, which can be written in terms of ξ_ss as:

$$w_\mathrm{pp}(\theta) = \int dz_1 \int dz_2 \psi_\mathrm{p}(z_1)\psi_\mathrm{p}(z_2) \left(\frac{b_\mathrm{p}(z_1)b_\mathrm{p}(z_2)}{b_\mathrm{s}(z_1)b_\mathrm{s}(z_2)}\right) \xi_\mathrm{ss}(\theta, z_1, z_2) \propto \int dz \psi_\mathrm{p}(z)^2 \frac{b_\mathrm{p}^2(z)}{b_\mathrm{s}^2(z)} \xi_\mathrm{ss}(\theta, z), \quad (2.6)$$

where we change variables into a central redshift and take note that ξ_ss vanishes for too large $\Delta z = z_1 - z_2$. Thus in this case, the new observable $w_\mathrm{pp}(\theta)$ cannot break the degeneracy. Also, for w_pp, satellites and environment become way more important, since the samples we correlate are less independent. Therefore, we need to find another observable, or appeal to some model of the redshift dependence of the bias.

The reason this is significant can be seen by considering estimators of, i.e., the mean redshift of a sample, which would be affected by assuming the wrong functional form of the bias as function of redshift. If we know that we recover $b_\mathrm{p}(z)\psi_\mathrm{p}(z)$, our estimator of the mean redshift would be:

$$\overline{z}_{est} = \int_0^\infty z \frac{b_\mathrm{true}(z)}{b_\mathrm{est}(z)} \psi_\mathrm{p}(z) dz, \quad (2.7)$$

while the true $\overline{z}_\mathrm{true}$ does not take into account these bias factors.
As noted by [Schulz, 2010], the bias may not be particularly smooth in its redshift evolution if different samples are used for different redshifts. We do assume a smooth bias, which does not have to mean the binned bias varies slowly. The only way a variation between bins could then impact our analysis is if the effective selection of the spectroscopic sample suddenly changes with redshift, not being captured in a comparable change in the photometric sample (and thus in the parameter K introduced in section 2.1.2), the general variation $f(z)$ or the luminosity bias (both also introduced in section 2.1.2).

2.1.1 Spatial Scales

Because we consider a spatial correlation function, we need to give constraints to which spatial (or angular) scales in distance we pay attention in this model. Essentially, too large scales are not sampled well enough at smaller redshifts, and too small scales are not of interest for a global GLF. To explore the details of how to best assign a redshift distribution to the photometric sample, let us first assume an ideal case where all photometric galaxies are at the same redshift z_0, so $\psi_\mathrm{p}(z) = N_\mathrm{p}\delta_D(z - z_0)$, where δ_D is the Dirac delta function, and N_p the amount of photometric galaxies. Later, we will see how to extend this. If we now split the spectroscopic sample in redshift bins δz_i, and for each i measure $w_\mathrm{ps}(\theta, z_i)$, which can be given by an estimator like:

$$w_\mathrm{ps}(\theta, z_i) = \frac{\langle n_\mathrm{p}(\theta, z_i)\rangle}{n_\mathrm{p}} - 1, \quad (2.8)$$

where the $\langle n_{\mathrm{p}}(\theta, z_i)\rangle$ denotes the mean density estimate of the photometric sample around the spectroscopic objects at redshift z_i. We thus are searching for some signal within a redshift bin, for all photometric galaxies are at the same redshift, so $\psi_{\mathrm{p}}(z) \propto w_{\mathrm{ps}}(z_i)$, and through the normalization condition $\int \psi_{\mathrm{p}}(z)dz = N_{\mathrm{p}}$ we find the amplitude.

The optimization of the sensitivity of this estimator depends on the spatial scales involved in the analysis, and is mostly limited by shot noise induced by the finite size of the samples, or cosmic variance. Therefore, including as many scales available to measurements is advantageous for increasing the sensitivity. Therefore, we integrate the angular cross-correlation function over multiple angular scales to recover:

$$\overline{w}_{\mathrm{ps}}(z) = \int\limits_{\theta_{min}}^{\theta_{max}} w_{\mathrm{ps}}(\theta, z)W(\theta)d\theta, \tag{2.9}$$

where $W(\theta)$ is a weight function for each scale, whose integral is normalized to unity, and it is aimed at optimizing the overall S/N. If $W(\theta) = \theta^{-1}$, there is equal amount of clustering information per logarithmic scale.

Now we need to note that we need to set θ_{min} and θ_{max} to match a range of projected radii we want to consider, $r_{\mathrm{p,min}}$ and $r_{\mathrm{p,max}}$, combined with the redshift range of the sample. As this angular scale becomes comparable to the mean separation of spectroscopic (reference) objects, the amount of useful clustering information decreases since number counts become correlated. Additionally, information at these scales is often subject to systematic effects due to dust extinction or fluctuations in the zero point of the photometry. Therefore, we limit to scales smaller than several Mpc, and to scales larger than the maximum between the typical size of sources and the point spread function of the survey.

Spread in Redshift

More generally, the photometric sample is of course spread out over a range Δz, so the spatial cross-correlations with the spectroscopic objects will depend on more characteristics like the type of objects in both samples, their relative clustering amplitude with respect to the dark matter field, the redshift dependence of these quantities and the observed scales of correlation. This was all introduced in the discussions about bias throughout the previous text. Therefore, we get that:

$$\overline{w}_{\mathrm{ps}}(z_i) \propto \psi_{\mathrm{p}}(z_i)\overline{b}_{\mathrm{p}}(z_i)\overline{b}_{\mathrm{s}}(z_i)\overline{w}_{\mathrm{DM}}(z_i), \tag{2.10}$$

where all quantities are now integrated over a range of scales. It should be noticed that if the relative variation of $\psi_{\mathrm{p}}(z)$ dominates over that of $\overline{b}_{\mathrm{p}}(z)$ in the redshift range, we approach the regime where $\psi_{\mathrm{p}}(z) \propto \delta_{\mathrm{D}}(z - z_0)$, but only up to finite accuracy can $\psi_{\mathrm{p}}(z)$ then be estimated. Also, only knowledge of the derivative of $\overline{b}_{\mathrm{p}}$ and $\overline{b}_{\mathrm{s}}$ is needed to characterize $\psi_{\mathrm{p}}(z)$. Constraints on $\overline{b}_{\mathrm{p}}(z)$ can be derived from its autocorrelation function through $\overline{w}_{\mathrm{ss}} = \overline{b}_{\mathrm{s}}^2\overline{w}_{\mathrm{DM}}(z)$ as explained above.

Note that this relation only holds on scales where galaxies are linearly biased w.r.t. the dark matter field. [Schmidt et al., 2013] prove with numerical simulations that inclusion of smaller scales only provides a small departure from this fact. Still this would depend on how small you actually go and which galaxies you include, so we should be careful in selection of the scales. Note that we can characterize $\overline{w}_{\mathrm{DM}}(z)$ from theory, but not observe it directly.

The main limitation is thus, as already explained before, the lack of knowledge of $\overline{b}_{\mathrm{p}}$. Proposed methods of constraining this quantity up until now have been using the autocorrelation of the photometric sample (which, as explained above, is not quite effective), using a redshift averaged value, deproject its redshift dependence through iterative techniques, or minimizing its contribution altogether while only estimating the error induced by approximating $\psi_p\mathrm{p}(z)$ without it. In the next subsection, we explain how we, in our extended model, can use another method of constraining this bias.

Going back to the model with bins not only in redshift, but also in apparent magnitude, we can extend equation(2.9) to:

$$\overline{w}_{\mathrm{ps}}(m_\lambda, z_i) = \int_{\theta_{\min}}^{\theta_{\max}} w_{\mathrm{ps}}(m_\lambda, z_i, \theta) W(\theta) d\theta, \tag{2.11}$$

where we will again refer to a model value of $\widetilde{\overline{w}}_{\mathrm{ps}}(m_\lambda, z_i)$ which we find by extending equation (2.10) (while discretizing it) to:

$$\widetilde{\overline{w}}_{\mathrm{ps}}(m_\lambda, z_i) = \sum_j f_{\mathrm{N}}(m_\lambda, z_j) \frac{\overline{b}_{\mathrm{p}}(m_\lambda, z_j)}{\overline{b}_{\mathrm{s}}(z_j)} \overline{w}_{\mathrm{ss}}(z_i, z_j), \tag{2.12}$$

where we use that both samples trace the same underlying density field, and $f_{\mathrm{N}}(m_\lambda, z_j)$ is a weight factor of the fraction of galaxies in apparent magnitude bin m_λ that reside in redshift bin z_j. This f_{N} is what we need to extract to find the redshift- and luminosity distributions. Both $\overline{w}_{\mathrm{ps}}$ and $\overline{w}_{\mathrm{ss}}$ can be directly found by using pair counting from the data, using an estimator like in equation (1.10), but the biases are a priori unknown, for we cannot observe directly $\overline{w}_{\mathrm{DM}}$ and thus cannot derive them.

2.1.2 Biases again

Let us first recap the context of the vDW model. We assume to have observed a set of photometric galaxies and a set of spectroscopic galaxies. Of both, we know the positions on the sky, and can thus find angular auto- and cross-correlation functions of the samples. Because we have more accurate redshift values for the spectroscopic sets, we can use them to constrain a most likely redshift to galaxies in the photometric set.

Assuming both samples trace the same underlying dark matter field, their clustering strength is dependent on their biases with respect to this field. Denoting the clustering function of the matter with itself between two redshift bins as $\overline{w}_{\mathrm{mm}}(z_i, z_j)$, we can write:

$$\overline{w}_{\mathrm{ss}}(z_i, z_j) = \overline{b}_{\mathrm{s}}(z_i)\overline{b}_{\mathrm{s}}(z_j)\overline{w}_{\mathrm{mm}}(z_i, z_j), \tag{2.13}$$

and:

$$\overline{w}_{\mathrm{ps}}(m_\lambda, z_i) = \overline{b}_{\mathrm{p}}(m_\lambda, z)\overline{b}_{\mathrm{s}}(z_i)\overline{w}_{\mathrm{mm}}(z_i, z), \tag{2.14}$$

where writing just z means we take the entire sample into account. This happens since we only take certain targets into account, which are biased tracers of the underlying field, mostly due to clustering effects.

We now want to know the clustering $\overline{w}_{\mathrm{p}_j,\mathrm{s}}$, where the subscript p_j denotes the set of photometric galaxies assigned to redshift bin z_j, which can be written as:

$$\overline{w}_{\mathrm{p}_j,\mathrm{s}}([m_\lambda, z_j], z_i) = f_{\mathrm{N}}(m_\lambda, z_j)\overline{b}_{\mathrm{p}}(m_\lambda, z_j)\overline{b}_{\mathrm{s}}(z_i) w_{\mathrm{mm}}(z_i, z_j), \tag{2.15}$$

where $[m_\lambda, z_j]$ denotes that we consider these galaxies only. Therefore, the underlying dark matter field cancels when dividing by $\overline{w}_{\mathrm{ss}}(z_i, z_j)$:

$$\frac{\overline{w}_{\mathrm{p}_j,\mathrm{s}}([m_\lambda, z_j], z_i)}{\overline{w}_{\mathrm{ss}}(z_i, z_j)} = \frac{\overline{b}_{\mathrm{p}}(m_\lambda, z_j)\overline{b}_{\mathrm{s}}(z_i)}{\overline{b}_{\mathrm{s}}(z_i)\overline{b}_{\mathrm{s}}(z_j)} f_{\mathrm{N}}(m_\lambda, z_j) = \frac{\overline{b}_{\mathrm{p}}(m_\lambda, z_j)}{\overline{b}_{\mathrm{s}}(z_j)} f_{\mathrm{N}}(m_\lambda, z_j). \tag{2.16}$$

This is what lies behind the expression in equation (2.12). To be able to reduce the degeneracy between the bias ratio and f_N we assume now that both $\overline{b}_{\mathrm{p}}$ and $\overline{b}_{\mathrm{s}}$ evolve similarly with respect to redshift at a fixed luminosity, so we can write $\overline{b}_{\mathrm{p}}(m,z) = \overline{b}_{\mathrm{p},0} b_{\mathrm{L,p}}(m,z) f(z)$ and $\overline{b}_{\mathrm{s}}(z) = \overline{b}_{\mathrm{s},0} b_{\mathrm{L,s}} f(z)$, where $b_{\mathrm{L,p}}$ is some function of luminosity, and thus apparent magnitude and z, assumed to be known independently or to be fitted from the spectroscopic sample, which is observable.[1]

Note that $b_{\mathrm{L,s}}$ has no dependences except directly on L, which is observable, it is approximately constant in z and the height of the peak for the galaxies formed at that position. This all is not unreasonable to assume, for the bias conceptually is a measure of how likely it is to form a galaxy at a given point in space given the dark matter field. In other words, there is a certain threshold of density above which galaxies can form, and how probable a peak is with the required density encountered at a given point in space and time. This threshold evolves with redshift, thus $\overline{b}$ certainly depends on the critical density and the virial density at a given redshift. Only considering this factor, the *halo bias* (which scales with $f(z)$), is not enough for us, for we observe two physically distinct populations, because the spectroscopic population only consists of those galaxies bright enough to be spectroscopically analyzed. Therefore, we add another factor, a *luminosity bias* ($b_L(m,z)$), signifying that not only the initial overdensity, but also the amplitude of the excess and the surroundings of the peak have influence on the type of galaxy formed. These factors all impact the brightness of the object.

Now the last assumptions made about the bias factors is that these are the only dependencies of $\overline{b}$, and there is no residual dependency on redshift or apparent magnitude except through the relations and functions defined above.

This means we find:

$$\begin{aligned}\frac{\overline{w}_{\mathrm{p}_j,\mathrm{s}}([m_\lambda, z_j], z_i)}{\overline{w}_{\mathrm{ss}}(z_i, z_j)} &= \frac{\overline{b}_{\mathrm{p},0} b_{\mathrm{L,p}}(m_\lambda, z_j) f(z_j)}{\overline{b}_{\mathrm{s},0} b_{\mathrm{L,s}} f(z_j)} f_{\mathrm{N}}(m_\lambda, z_j) = \frac{\overline{b}_{\mathrm{p},0}}{\overline{b}_{\mathrm{s},0} b_{\mathrm{L,s}}} b_{\mathrm{L,p}}(m_\lambda, z_j) f_{\mathrm{N}}(m_\lambda, z_j) \\ &\equiv K b_{\mathrm{L,p}}(m_\lambda, z_j) f_{\mathrm{N}}(m_\lambda, z_j) \equiv f'_{\mathrm{N}}(m_\lambda, z_j).\end{aligned} \tag{2.17}$$

Here we have defined a normalization factor K, dependent on i.e. formation criteria of galaxies and selection of the spectroscopic sample. This allows us to rewrite equation (2.12) as:

$$\widetilde{\overline{w}}_{\mathrm{ps}}(m_\lambda, z_i) = \sum_j f'_{\mathrm{N}}(m_\lambda, z_j) \overline{w}_{\mathrm{ss}}(z_i, z_j), \tag{2.18}$$

where K is assumed to be unknown for the model, and thus one of the parameters.

Now due to the approach we take, which is trying to find a method which can fit a GLF to a given sample of photometric objects given a sample of spectroscopic objects only, we have to provide some model if we do not want to use (often biased sets of) templates from simulations or previous observations. This model is twofold, we have to provide a model for the luminosity bias $b_{\mathrm{L}}(m,z) \equiv b_{\mathrm{L,p}}(m,z)$ as well as for $N_{\mathrm{p}}(m,z)$.

We assume $b_{\mathrm{L}}(m,z)$ to be monotonous in L, since we do not expect more extreme objects to have a lower bias. We use a simple form (motivated by [Benoist et al., 1996, Peacock et al., 2001, Norberg et al., 2001] to have a simple monotonically increasing shape) of:

$$b_{\mathrm{L}}(m,z) = 1 + \frac{L(m,z)}{L'}, \tag{2.19}$$

[1] Alternatively, as mentioned previously, $\overline{b}_{\mathrm{s}}(z)$ can be estimated from data, which has the complication of propagating observational uncertainties. Alongside this, $\overline{b}_{\mathrm{p}}(m,z)$ should then be modelled as e.g. a polynomial in redshift.

where L' is a normalization parameter, to be fitted from i.e. the spectroscopic samples or previous observations. In the vDW model this is set equal to the luminosity of a galaxy with $M' = -23.3$. Assuming the naive distance modulus we only have to calculate b_L once for each bin (m_λ, z_i) since the model is agnostic of redshift and therefore luminosity. Indirectly, the normalization of this function is also controlled by K, which is already a parameter of the model.

Note that we do change from a space where we know clustering in these (m, z)-bins to a space with M-bins if we translate to a GLF as function of M or L. We thus have to carefully consider whether f'_N is still well defined after this transformation, i.e. if we can indeed fit b_L with one parameter L independent of the exact set of galaxies, or if we need to perform more dedicated fits. For purposes of the vDW model one fit from the spectroscopic set is enough. In the next section we will introduce the model for $N_\mathrm{p}(m, z)$, which we need to be able to fit to the redshift distribution and luminosity function simultaneously, which is needed to avoid biased and unphysical results.

2.2 Model for N_p

As can be seen, equation (2.18) is just a linear system of equations, which could be solved for every m_l independently, however, this disregards information inherent in these apparent magnitudes. Due to degenerate solutions most likely existing and clustering measurements having uncertainties, some galaxies with bright apparent magnitude may be placed at high redshift, corresponding to an unphysically high luminosity. Therefore, we need a model for $N_\mathrm{p}(m, z)$ to use to fit both redshift and luminosity distributions together. In this section, we will denote $N_{\mathrm{p},\lambda i} \equiv N_\mathrm{p}(m_\lambda, z_i)$ for brevity. Greek subscripts thus refer to apparent magnitude, and Latin to redshift bins.

The quantity $N_{\mathrm{p},\lambda i}$, the number of survey objects at apparent magnitude m_λ and redshift z_i, is constrained by the luminosity function. When we assume a given cosmology (in our case the flat ΛCDM Universe with parameters from the Planck Millennium survey [Planck Collaboration, 2016]), this luminosity function then also constrains the redshift distribution. Therefore, knowing both the cosmology and the redshift distribution we can infer the shape of the luminosity function through time.

To get to number densities, we need to know the survey volume in a given redshift bin. This is given by the integral over the observed area and the comoving distance to the binedges of the redshift bin. In other words:

$$\begin{aligned} V_i &= \int_A \int\limits_{d_{i,\min}}^{d_{i,\max}} d_\mathrm{c}(z)^2 dd_\mathrm{c}(z) dA \\ &= 4\pi f(A) \int\limits_{z_{i,\min}}^{z_{i,\max}} \frac{d_\mathrm{c}(z)^2 c}{H_0\sqrt{\Omega_{\mathrm{m},0}(1+z)^3 + \Omega_{\Lambda,0}}} dz, \end{aligned} \tag{2.20}$$

because:

$$d_\mathrm{c}(z) = \int\limits_0^z \frac{c}{H_0\sqrt{\Omega_{\mathrm{m},0}(1+z')^3 + \Omega_{\Lambda,0}}} dz' \tag{2.21}$$

is the comoving distance for a flat ΛCDM universe. Here A is the area of the survey on the sky, and $f(A)$ is the fraction of steradian on the sky covered by the survey. The limits of integration $d_{i,\min}, d_{i,\max}$

and $z_{i,\mathrm{min}}, z_{i,\mathrm{max}}$ signify the minimum and maximum distance and redshift values of redshift bin i.

Now the matter is selecting a suitable model for the shape of the luminosity function. In the vDW model this is a single Schechter function. This choice is further discussed in the next chapter. A *Schechter function* is characterized as (as already seen in equation (2) in the Introduction):

$$\phi(M) = 0.4\ln(10)\phi_*(z)10^{0.4(\alpha(z)+1)(M_*(z)-M(m,z))}e^{-10^{0.4(M_*(z)-M(m,z))}}, \tag{2.22}$$

where $\phi(M)$ is the number of galaxies in the volume V per unit luminosity per unit volume and $M(m,z)$ is the absolute magnitude corresponding to a galaxy with apparent magnitude m at redshift z. Note that this conversion is naively calculated with the distance modulus:

$$M(m,z) = m + 5[1 - \log_{10}(d_{\mathrm{L}}(z))], \tag{2.23}$$

where $d_{\mathrm{L}}(z)$ is the luminosity distance (in pc) given the cosmology. This means we implicitly assume flat galaxy spectra, and the effect of this assumption is discussed in Chapter 3.

The free parameters in equation (2.22) are $\alpha(z), M^*(z)$ and $\phi_*(z)$. Note that on small scales, inhomogeneities could impact ϕ_* and it will always (but more strongly on smaller scales) depend on the specific volume selected. Due to the assumption of a homogeneous universe we have a universal luminosity function independent of volume. In surveys, the sample is limited by apparent magnitude, and thus the volume V (if considered to be a ball centered at earth) decreases as M increases, and in comparison, the number density of higher-magnitude (lower luminosity) galaxies at higher redshifts decreases purely due to survey limitations.

We assume the redshift evolution of the parameters α (*low-luminosity power slope*) and M_* (*turn-over (or characteristic) magnitude*) is linear, so we use four free parameters to describe these: $\alpha(z) = \alpha_0 + \alpha_{\mathrm{e}} z$ and $M_*(z) = M_{*0} + M_{*\mathrm{e}} z$.
The parameter $\phi_*(z)$ is called the normalization of the luminosity function, and will be modelled as the exponential of a 5th-order polynomial:

$$\phi_* = \exp\left(\sum_{j=0}^{j=5} \zeta_j \left[\frac{2z}{z_{\mathrm{max}}} - 1\right]^j\right), \tag{2.24}$$

where we have six more free parameters ζ_j, and z_{max} is the maximal redshift considered.

Here we note that, since we have a GLF evolving with redshift, the integral over the sky and GLF do not separate. Furthermore, the choice of six parameters for the normalization seems arbitrary, but has been shown in vDW to allow enough versatility while not adding degeneracies because it is smaller than the amount of redshift bins considered, ensuring it varies smoothly. By choosing the form of the exponential of a polynomial the expression is numerically easy to handle.

Additionally, observationally, it seems that galaxies actually follow a GLF that looks more like a double Schechter function, the sum of two Schechter functions ([Johnston, 2011, Peng et al., 2010]). The impact of this change is discussed in Chapter 3.

To now find, in this context, the model value $\widetilde{N}_{\mathrm{p},\lambda i}$ for $N_{\mathrm{p},\lambda i}$ we will calculate the integrals. Therefore, to avoid divergence of the integrals and because there is a physical lower limit to what classifies as a galaxy, we define a limiting galaxy absolute magnitude $M_{\mathrm{lim}} = -16$. Then the integrated number density of galaxies in apparent magnitude bin m_λ and redshift bin z_i is:

$$\Phi_{\lambda i} = 0.4\ln(10) \int_{z_{i,\min}}^{z_{i,\max}} \phi_*(z) \int_{M_1}^{M_2} \frac{10^{0.4(M_*(z)-M(m,z))(\alpha(z)+1)} e^{-10^{0.4(M_*(z)-M(m,z))}}}{\Gamma\left(\alpha(z)+1, 10^{0.4(M_*(z)-M_{\text{lim}})}\right)} dMdz, \tag{2.25}$$

where we define the limits of integration $M_1 = \min\{M(m_{\lambda,\min}; z), M_{\text{lim}}\}$ and $M_2 = \min\{M(m_{\lambda,\max}; z), M_{\text{lim}}\}$. Also, we define the number density such that:

$$\int_{-\infty}^{M_{\text{lim}}} \frac{d\Phi_{\lambda i}}{dM} dM \equiv \int_{-\infty}^{M_{\text{lim}}} \phi_i(M) dM = \int \phi_*(z) dz_i, \tag{2.26}$$

i.e. we define it such that the integral of $\phi_i(M)$, the GLF in number density of galaxies in bins m_λ and z_i per unit magnitude, equals the integral over the entire bin of ϕ_* for all bins.
Now to derive from this an expected number of galaxies in the given bin, we simultaneously integrate the volume and the luminosity function, to arrive at:

$$\begin{aligned}
\widetilde{N}_{\text{p},\lambda i} &= \int_{z_{i,\min}}^{z_{i,\max}} \int_{m_1}^{m_2} \frac{d\phi_i(M)}{dz_i} \frac{dV}{dz} dmdz \\
&= \frac{2}{5}\ln(10) B \int_{z_{i,\min}}^{z_{i,\max}} \frac{d_{\text{c}}(z)^2 \phi_*(z)}{\sqrt{\Omega_{\text{m},0}(1+z)^3 + \Omega_{\Lambda,0}}} \times \\
&\int_{m_1}^{m_2} \frac{10^{(\frac{2}{5})(M_*(z)-M(m,z))(\alpha(z)+1)} e^{-10^{\frac{2}{5}(M_*(z)-M(m,z))}}}{\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M_{\text{lim}})}\right)} dmdz,
\end{aligned} \tag{2.27}$$

where we define $B = 4\pi f(A)c/H_0$, $m_1 = \min\{m_{\lambda,\min}; m(M_{\text{lim}}, z)\}$ and $m_2 = \min\{m_{\lambda,\max}; m(M_{\text{lim}}, z)\}$. It should be noted here that the modulation of observed galaxy number densities due to lensing magnification, causing a magnification bias, is ignored, since incorporating this in the current model is highly non-trivial and outside our scope.

This integral (2.27) over apparent magnitude has an analytical solution, thus we need only integrate over redshift.[2] Afterwards, summing over all redshift bins, we recover the number of galaxies (according to the model) in apparent magnitude bin m_λ at any redshift, $\widetilde{N}_{\text{p},\lambda}$ (where $z_{\min}$ and $z_{\max}$ are the redshift limits of the entire sample):

$$\begin{aligned}
\widetilde{N}_{\text{p},\lambda} = \sum_i \widetilde{N}_{\text{p},\lambda i} = B \int_{z_{\min}}^{z_{\max}} \frac{d_{\text{c}}(z)^2 \phi_*(z)}{\sqrt{\Omega_{\text{m},0}(1+z)^3 + \Omega_{\Lambda,0}}} \left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M_{\text{lim}})}\right)\right]^{-1} \times \\
\left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_2,z))}\right) - \Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_1,z))}\right)\right] dz.
\end{aligned} \tag{2.28}$$

This quantity can be directly compared to the $N_{\text{p},\lambda}$ of the data. This, along with the clustering signal, constrain our model, making it able for us to fit the model to our observations. How this is done exactly is discussed in the next section.

[2] This is done numerically, since the expression in equation (2.27) is non-analytic. We subdivide each redshift bin in smaller redshift bins where we can assume the GLF constant and sum these contributions.

2.2.1 Fitting

In summary, the vDW model uses 11 free parameters. One of these signifies the bias ratio (K), six of these (the ζ_j from equation (2.24)) are for the normalization of the GLF, and 4 ($\alpha_0, \alpha_\mathrm{e}, M_{*0}$ and $M_{*\mathrm{e}}$) are for the shape parameters of the Schechter function and their redshift-evolution. To find the optimal parameters, the two observable quantities $\overline{w}_{\mathrm{ps},\lambda i}$ and $N_{\mathrm{p},\lambda}$ are fitted to the model values $\widetilde{\overline{w}}_{\mathrm{ps},\lambda i}$ (equation (2.18), as derived from the integrated observed autocorrelation of the spectroscopic sample $\overline{w}_{\mathrm{ss},ij}$) and $\widetilde{N}_{\mathrm{p},\lambda}$ (equation (2.28)). We use an updated likelihood compared to the vDW model (equation (15) from vDW), namely:

$$\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\ln(|C(\boldsymbol{\theta})|) - \frac{1}{2}(\overline{\mathbf{w}}_\mathrm{ps} - \widetilde{\overline{\mathbf{w}}}_\mathrm{ps})^T C^{-1}(\overline{\mathbf{w}}_\mathrm{ps} - \widetilde{\overline{\mathbf{w}}}_\mathrm{ps}) + \sum_{j=0}^{n_\mathrm{m}-1}\left[N_{\mathrm{p},j}\ln\left(\widetilde{N}_{\mathrm{p},j}(\boldsymbol{\theta})\right) - \widetilde{N}_{\mathrm{p},j}(\boldsymbol{\theta})\right], \tag{2.29}$$

where C is the joint covariance matrix, combining three sources of uncertainty, $\boldsymbol{\theta}$ is the vector of parameters, n_m is the number of apparent magnitude bins and $\widetilde{\overline{\mathbf{w}}}_\mathrm{ps}$ is the vector that solves equation (2.18) when we write it as:

$$\widetilde{\overline{\mathbf{w}}}_\mathrm{ps}(m_\lambda) = \mathbf{X} f'_\mathrm{N}(m_\lambda), \tag{2.30}$$

collecting the equations in a vector of length n_z (the amount of redshift bins) and a matrix of size $n_\mathrm{z} \times n_\mathrm{z}$, where $X_{ij} = \overline{w}_\mathrm{ss}(z_i, z_j)$. A more general likelihood form is found in the Appendix, describing how we got to equation (2.29).

The uncertainties entering in C are: uncertainties in the integrated cross-correlation function of the photometric and spectroscopic galaxies $\overline{w}_{\mathrm{ps},\lambda i}$, in the integrated cross-correlation of spectroscopic galaxies in different redshift bins $\overline{w}_{\mathrm{ss},ij}$ and the number of galaxies in magnitude and redshift bins $\widetilde{N}_{\mathrm{p},\lambda i}$. These sources enter through the form of the model from equation (2.18).
For the first two sources, bootstrap resamplings are used to calculate full covariance matrices, and the last source is modelled as a Poisson variable as mentioned before. For a more detailed discussion, see vDW.
This last source of uncertainty depends on the quantities fitted by the last term only, though these quantities do have an effect on the first terms.

2.3 Mock Catalogue, Fiducial Model Results and Prospects

In this section we briefly summarize the results for the vDW model as presented in vDW. We start this discussion with discussing the mock catalogue used in this work and then continue to present briefly the results of the vDW model, briefly mentioning other tests presented in vDW that we assume to hold in our analysis as well, to not have to expand on them in the next chapter for brevity. This thus all sets the context for the next chapter, and gives material to compare further results to, since those are based on what is presented in this section.

The mock galaxy survey is extracted from the publicly available Planck Millennium all-sky lightcones simulation released with [Henriques et al., 2015]. We use these mock surveys to test the model, and make a selection along the same lines as in vDW. It should be noted here that vDW alter the mock survey such that the data exactly fits a Schechter luminosity function to prove the effectiveness of the method conceptually, so it is in fact a simulation. We will only consider the unaltered mock galaxy catalogue in chapter 3, since our purpose is to extend the model to be applicable to actual data.

The selection is such that we consider galaxies with $m \leq 21$ and $z \leq 0.8$, in the sky-area with right ascension in $[100, 200]$ and declination in $[10, 50]$ (about 8% of the sky). We select on i-band absolute magnitude $M_{\rm I} < -16$. Then the spectroscopic sample is constituted by the brightest galaxies in each bin also subject to a constraint to a minimal star formation rate and stellar mass, such that the spectroscopic sample is a small (density of at most $10^{-4}(\text{Mpc}/h)^{-3}$ at each redshift) and highly biased subset. Figure 2 (not reproduced here) in vDW shows the differences in magnitude and redshift distribution between the two selected samples.
There will be $n_{\rm z} = 16$ redshift and $n_{\rm m} = 16$ apparent magnitude bins, with widths $\Delta_z = 0.05$ and $\Delta_m = 0.5$, in the range $z \in [0, 0.8]$ and $m \in [13, 21]$.

In figure 2.1, the results of the vDW model with updated likelihood are shown. A somewhat more detailed discussion of this figure can be found in the Appendix.

The fiducial model most visibly fails in regimes of low sampling, under which we understand the low redshift, low luminosity end, as well as the high redshift high luminosity end. The first is lowly sampled due to some important scales in clustering being undersampled at low redshift due to the limit on the observed cosmological volume, and thus cosmic variance becomes a problem. The second is lowly sampled due to the lack of objects that exists with high luminosities to correctly sample this regime. Still, it fits the distribution accurately, including even the drop-off at low luminosity due to the limit in apparent magnitude.

To summarize, the vDW model extends previous models by deriving a GLF and redshift distribution simultaneously, along with their redshift-evolution. An input GLF for a (modified) mock galaxy survey can be accurately recovered, independent of how the spectroscopic sample has been selected with respect to the photometric sample. The results are not degenerate with galaxy bias, and by further development this model could be applied to real data. This development includes the addition and correction for effects like magnification bias[3], dust extinction and K-corrections[4] in the conversion between apparent and absolute magnitude. Additionally, other assumptions made could have an effect. Firstly, this model assumes that the form of a (candidate) GLF is known, but [Peng et al., 2010] show that, generally, a sum of Schechter functions is a better fit to real data.[5]
Secondly, the simple luminosity bias relation assumed in equation (2.19) has a known, fixed parameter L' (or M'), and we assume the redshift evolution of the remaining terms cancels out. These have not been imposed on the mock sample, but results were still sufficiently valid, so only if the real data has some residual dependence on m or z that the mock data does not would this give a problem. While L' is fixed now, it could also become a free parameter of the model, since it is currently fitted to

[3] Galaxies appearing closer and/or brighter due to gravitational lensing

[4] The simulation from [Henriques et al., 2015] does take dust extinction models and reverse K-corrections into account when calculating apparent magnitudes. Since the results of the current model are promising, these should only serve as minor corrections.

[5] It should be noted that [Peng et al., 2010] does this for a galaxy mass function, and not a GLF. Of course these are linked by the mass-luminosity ratio, but for different objects and eras the mass-luminosity ratio can vary significantly. Still it is a valid assumption that the underlying physical processes dictating the structure of the double Schechter mentioned in [Peng et al., 2010], namely the two components being due to mass quenching and environmental effects, as well as an existing distinction between star-forming and quiescent galaxies, result in the case of a universal GLF better fitted by a sum of two (or multiple) Schechter functions, e.g. [Johnston, 2011, Muzzin et al., 2013, Tomczak et al., 2014, Bonne et al., 2015].

Fiducial model on altered mock sample with Likelihood eq. (2.29)

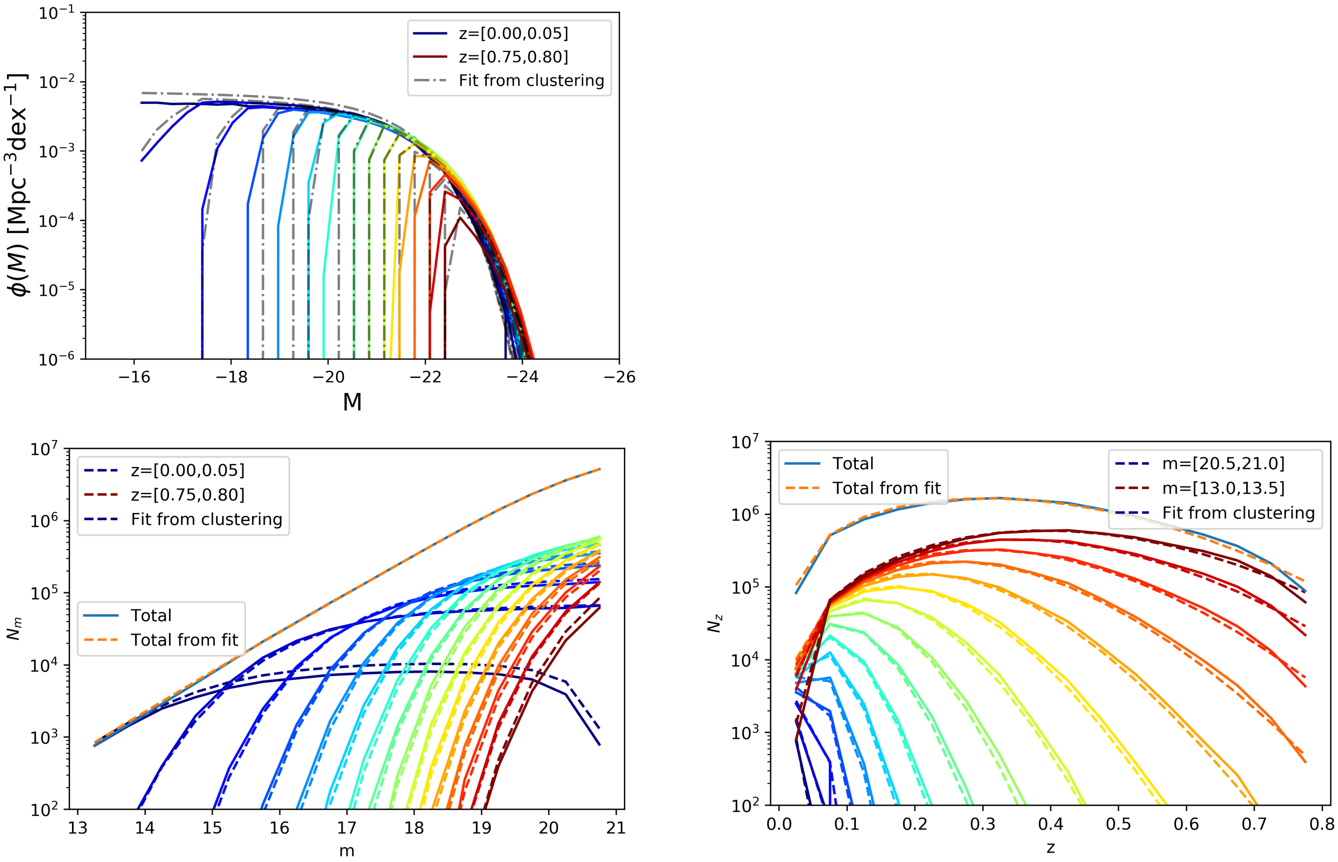

Figure 2.1: *The results for the vDW model, minimized using eq.* (2.29) *as constrained by the cross-correlation signal between photometric and spectroscopic galaxies and the total number of photometric galaxies in each bin of apparent magnitude.* ***Top left:*** *The number of galaxies in different bins of redshift as a function of absolute magnitude. Solid lines show the data, colors show the redshift bins. Dot-dashed grey lines show the outcome of the model.* ***Bottom left:*** *The number of galaxies in different bins of redshift as a function of apparent magnitude. The total over all redshifts, shown by the solid blue line above the rest, is one of the constraints of the model.* ***Bottom right:*** *The number of galaxies in different bins of apparent magnitude as a function of redshift. The solid blue line shows the total over all apparent magnitudes. This is as close as we can come to visualizing the constraint on clustering.*

the spectroscopic data, but could thus be fitted to all data (with assigned redshifts from the model) instead for completeness.

Chapter 3

Changes to the Model and Results

In this chapter we discuss the changes made to the vDW model, aside from the updated likelihood discussed in the last chapter, to ensure an effective application of the model to realistic datasets. These changes have all been introduced during the discussion in previous chapters, but will now be expanded upon in greater detail. We present three higher order effects or alterations that could improve the vDW model when wanting to apply it to actual data (although we will only test it on a mock galaxy catalogue for purposes of this research). These include the addition of a second Schechter function (section 3.1), the addition of a parameter fitting the luminosity dependent galaxy bias b_{L} (section 3.2) and the addition of K-corrections to the naive distance modulus (section 3.3). We mentioned before these three additions all become important when applying the model to actual data, so we will implement them all together. We did check the effect of the additions individually, but the shown results were the best results of the runs we performed. For all the results we will present, we present the starting location of the parameters included in the model and collect the best-fit parameters in a table at the end of the chapter. These starting parameters have been fitted to the data by trial-and-error to find a sufficiently good starting position.

3.1 Double Schechter function

As mentioned at the end of the previous chapter, [Peng et al., 2010] finds that, generally, a sum of Schechter functions is a good fit to real data.[1] The form of the luminosity function then becomes very versatile and is allowed to contain many parameters to be constrained at once. Our first alteration will thus be to see if implementing a second set of parameters belonging to a second Schechter function allows for a sufficient fit on the selected mock galaxy catalogue.

Modelling the GLF with a sum of two Schechters means that a second term, $\widetilde{N}'_{\mathrm{p},\lambda}$, should be added to equation (2.28), equal to:

$$\begin{aligned}\widetilde{N}'_{\mathrm{p},\lambda} = \sum_i \widetilde{N}'_{\mathrm{p},\lambda i} = B \int\limits_{z_{\mathrm{min}}}^{z_{\mathrm{max}}} \frac{d_{\mathrm{c}}(z)^2 \phi'_*(z)}{\sqrt{\Omega_{\mathrm{m},0}(1+z)^3 + \Omega_{\Lambda,0}}} \left[\Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M_{\mathrm{lim}})}\right)\right]^{-1} \times \\ \left[\Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M(m_2,z))}\right) - \Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M(m_1,z))}\right)\right] dz,\end{aligned} \tag{3.1}$$

[1] Note that they prove this for a galaxy mass function, and physically explain their findings by principles of mass quenching. While this could also be a naive cause for comparable structure in the GLF, the M/L-ratio becomes important in conversion. Since this ratio depends strongly on galaxy type and environment, effects of this should be studied in more detail.

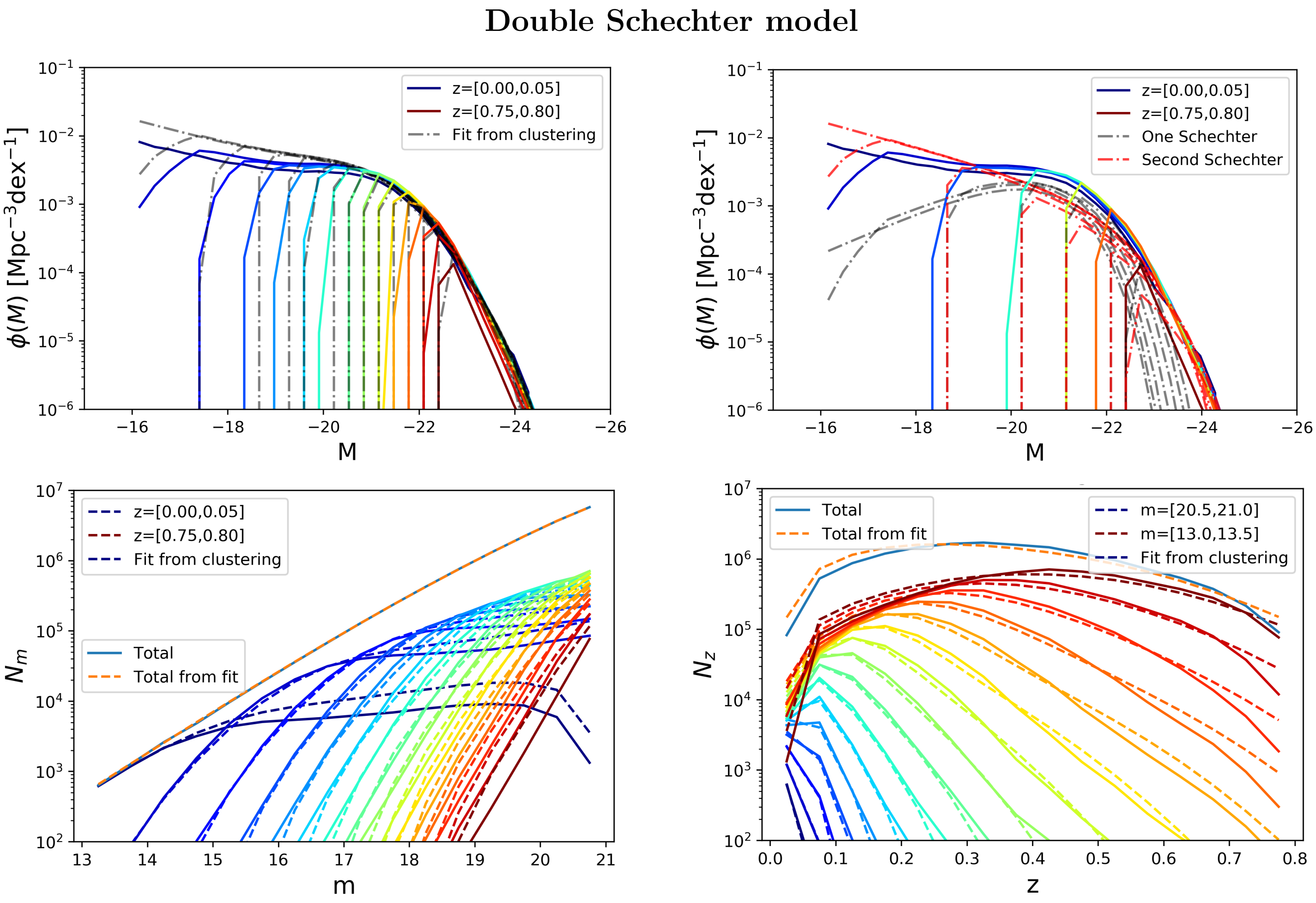
Double Schechter model

Figure 3.1: *As figure 2.1, but results for the double Schechter model (equation* (2.28) *with added equation* (3.1)*) on the non-altered mock galaxy catalogue, minimized using equation* (2.29)*. Note the difference in the data due to using an unaltered mock sample now.* ***Top right:*** *The same as the top left figure, but with the two Schechters plotted separately for an evenly spaced subset of the redshift bins.*

where $\alpha'(z) = \alpha'_0 + \alpha'_\mathrm{e} z$, $M'_*(z) = M'_{*0} + M'_{*\mathrm{e}} z$ and

$$\phi'_*(z) = \exp\left(\sum_{j=0}^{j=5} \zeta'_j \left[\frac{2z}{z_\mathrm{max}} - 1\right]^j\right). \tag{3.2}$$

Therefore, 10 new parameters $\left(\alpha'_0, \alpha'_\mathrm{e}, M'_{*0}, M'_{*\mathrm{e}} \text{ and the } \zeta'_j\right)$ are added. The results are shown in Figure 3.1.

As can be seen here, the GLF of the data differs with respect to figure 2.1, due to using an unaltered mock catalogue. The double Schechter function fits the overall shape of the GLF by having one of the two functions (try to) fit the low-luminosity slope as well as the high-luminosity end. The second Schechter then adds galaxies in the midrange to fill up the 'bump' not fitted well enough by the first. Looking at the top-left panel, it most visibly overestimates the low redshift, low-luminosity end as for the fiducial model, while the constraint in the bottom left panel is fitted exceptionally. It thus does so by compensating for misfitting the low-luminosity slope with less galaxies at intermediate luminosities and redshifts, again overestimating at higher redshifts and low luminosities. Overall, the shape is fitted quite well though, and the misfit can be explained by the mock data having significantly less galaxies than average in the lowest redshift bin and significantly more galaxies than average in the second redshift bin (as stated in vDW). The impact hereof worsens when not altering the data to fit a single Schechter function.

Furthermore, the misfit could be due to degeneracy between different parameters, for example the degeneracy between the high-z normalization and the low-luminosity-slope-parameter α_0 (and α_e) of the Schechter function. This is a degeneracy since the low-luminosity slope is barely probed at higher redshifts and thus effectively in this regime only serves to normalize the profile. This was already apparent in vDW, and we now added only more opportunity for degeneracy by adding a second Schechter with independent normalization and slope.

Still, as we show in the Appendix, fitting only a single Schechter to the unaltered galaxy catalogue results in a worse fit, so definitely implementing a second Schechter is needed for fitting to actual data.

3.1.1 Choice of Parameters

Inspired by [Peng et al., 2010] we implement the second Schechter such that $M'_{*0} = M_{*0}, M'_{*\mathrm{e}} = M_{*\mathrm{e}}$ and $\zeta'_j = \zeta_j$. To still enable a difference in normalization, we do add another parameter f_ϕ such that $\phi'_*(z) = f_\phi \phi'_*(z)$. This new set of parameters we call the 'simple' double Schechter model.[2] The results are seen in figure 3.2.

Although the general shape of the GLF seems to be better reproduced and $N_p(m)$ is fitted as well, the mismatch over the bins is apparent in the bottom panels, most evidently in the bottom right one. This points to the model not having enough freedom with this reduced set of parameters, and we thus lose the constraints where $M'_{*0} = M_{*0}$ and $M'_{*\mathrm{e}} = M_{*\mathrm{e}}$, making M_* a free parameter again, to find the results in figure 3.3.

While this model still has its faults, most importantly in overestimating amounts of galaxies in higher redshift bins while underestimating those in lower redshift bins, the general shape of the GLF is reproduced better than in figures 3.1 and 3.2. Starting the model at a different set of input parameters does have an impact on the result, pointing to left-over degeneracies between parameters, which already were present in vDW. This means that different starting points could possibly lead to a better fit. The most important difference one should consider is if the two Schechters should indeed fit the GLF as they do now, or if we need one Schechter to fit the high-luminosity end while the other only fits the lower luminosity end. As mentioned in chapter 2 and in [Peng et al., 2010], for galaxy mass functions of quiescent galaxies often the two Schechter components are linked to different sources of quenching, namely mass quenching and environment quenching. These effects could be of importance for this universal GLF as well, and with a better understanding dictate a more theoretically substantiated starting point. With these considerations in mind, it is still valid to assume $\zeta_j = \zeta'_j$ since the relative number densities of these galaxies can be assumed to evolve similarly through cosmic time for the redshifts we consider.

3.2 M' (L') as a free parameter

In the previous results, we assumed a simple luminosity bias relation (equation (2.19)), where M' was fitted to the spectroscopic sample. This could be fitted better by including all data in the fit,

[2] During later runs after implementing K-corrections, we noticed f_ϕ becoming negative. Of course this is unphysical, so we reimplemented f_ϕ as $\phi'_*(z) = e^{f_\phi}\phi'_*(z)$. This did not have a significant effect on the results of the original runs.

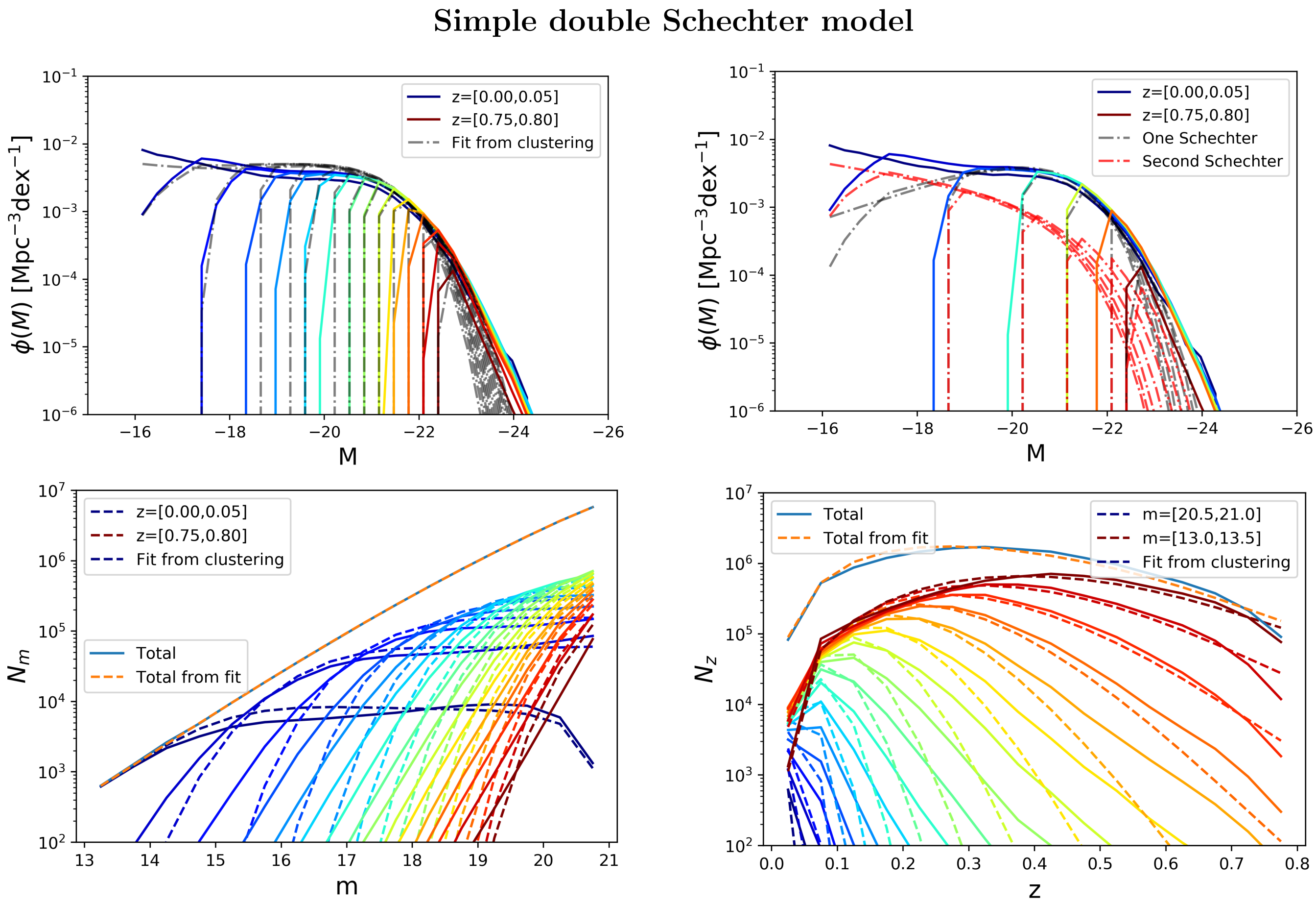

Figure 3.2: *As figure 3.1, but for the simple double Schechter model. Now both the high and low absolute magnitude regimes are underestimated, which is compensated for by overestimating the midrange. Most worryingly the bottom right panel shows a significant difference between model and data for all but the lowest three magnitude bins.*

but should then be done by the model itself. While the parameter M' is certainly going to have degeneracies with K, this could give the model more freedom to fit the clustering constraint better. Of course M' could still always be fitted to the dataset, but by making it a free parameter we account for complications that arise with this. To see what these are, we rewrite equation (2.18) to free $b_{\rm L}(m_\lambda, z_i)$ so we find:

$$b_{\rm L}(m_\lambda, z_i) = \frac{-\sum_{j\neq i} K b_{\rm L}(m_\lambda, z_j)\dfrac{N_{\rm p}(m_\lambda, z_j)}{N_{\rm p}(m_\lambda)}\bar{w}_{\rm ss}(z_i, z_j) + \bar{w}_{\rm ps}(m_\lambda, z_i)}{K\dfrac{N_{\rm p}(m_\lambda, z_i)}{N_{\rm p}(m_\lambda)}\bar{w}_{\rm ss}(z_i)}, \tag{3.3}$$

where we now use the observed $\bar{w}_{\rm ps}(m_\lambda, z_i)$ as $\tilde{\bar{w}}_{\rm ps}(m_\lambda, z_i)$ for fitting M'. This is a high (in our case with 16 m and z bins 256-) dimensional linear system of equations. Therefore, for simplicity, we assume the $\sum_{j\neq i}$-term is negligible, and plot the found $b_{\rm L}$:

$$b_{\rm L}(m_\lambda, z_i) = \frac{N_{\rm p}(m_\lambda)}{K N_{\rm p}(m_\lambda, z_i)}\frac{\bar{w}_{\rm ps}(m_\lambda, z_i)}{\bar{w}_{\rm ss}(z_i)} \tag{3.4}$$

in figure 3.4 for the mock galaxy sample.

It can be seen that results are spread out significantly, and since all $b_{\rm L}$ should follow the same line much closer, this is most likely due to the $\sum_{j\neq i}$-term not being negligible. This means we should

Simple double Schechter model with M_* free

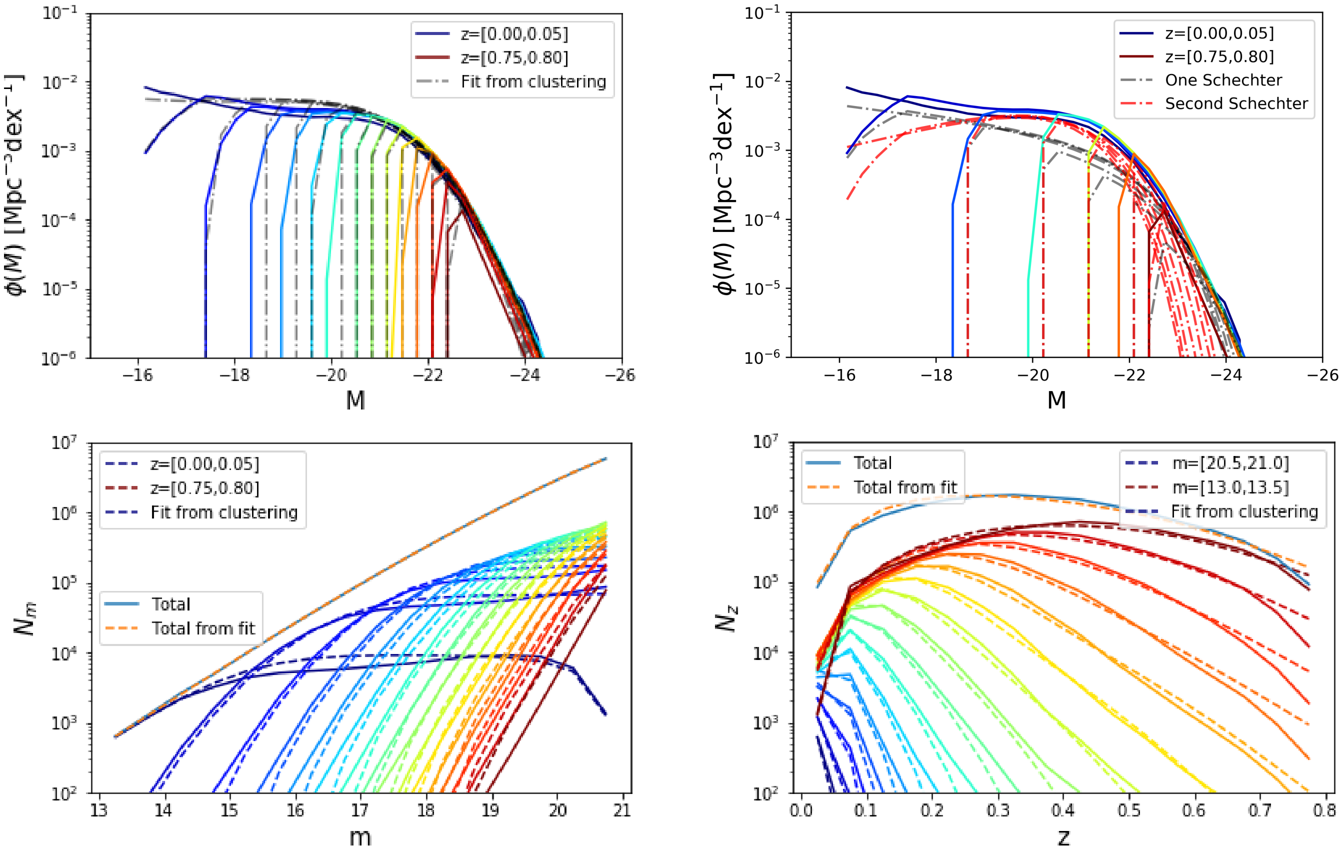

Figure 3.3: *As figure 3.1, but for the simple double Schechter model with M_* a free parameter again. The most apparent problems with both previous figures have been solved. Still the model overestimates the amount of galaxies with midrange absolute magnitude, as well as those in high-redshift bins while underestimating the amount in low redshift bins (see bottom left). Note the 'roles' of the two Schechters have been reversed due to a change in starting position to point out degeneracies.*

actually solve the 256-dimensional linear system of equations. This is the first of the complications. The second is the fact that there exist negative correlations between bins in $\bar{w}_{\rm ps}$, which we cannot take into account here. Furthermore for actual data only being able to use the spectroscopic set means we have statistically insufficient data to accurately fit this parameter, leading to insufficient statistics, which we already had to account for by only plotting points with $\bar{w}_{\rm ps}(m_\lambda, z_i) > 0$, $\bar{w}_{\rm ss}(z_i) > 0$ and $N_{\rm p}(m_\lambda, z_i) > 1000$ in figure 3.4. However, where and how we should set these constraints can not be sufficiently substantiated. Therefore, it is more objective to make M' a free parameter and fit $b_{\rm L}$ each iteration of our model to the entire dataset given the parameters at that point. Results of the simple double Schechter with M_* free and M' free are found in figure 3.5. In conclusion, the addition of M' introduces more degeneracies between parameters. During testing the model with M' reached a local minimum much faster, but the final likelihood was always higher than when running the same model without M' free. Still, letting this parameter be free to fit $b_{\rm L}$ to the entire dataset in each iteration seems more physically substantiated, but of course fitting it in advance to running the model is always a possibility.

These results show once again how degenerate the likelihood-landscape is. The bottom right figure shows a better correspondence for high luminosity, high redshift objects, at the cost of the midrange redshift fit of these quantities. This is understood by looking at figure 3.4, where the fit already showed

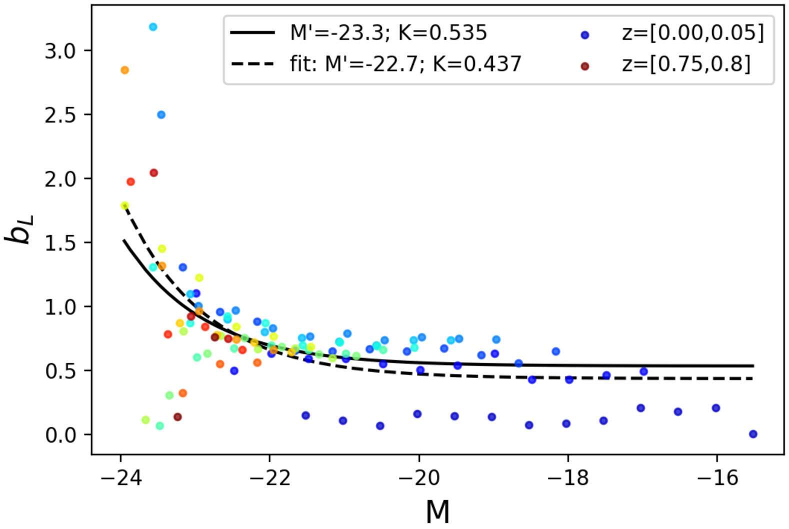

Figure 3.4: *Values for* b_L *calculated from equation* (3.4) *for the mock galaxy sample (of which we know all quantities, while for actual observed data this only holds for the spectroscopic set). The black line is equation* (2.19) *for constant parameter* M' *and the fitted* K *from the fiducial model. The dashed line is the fit to the data with both* K *and* M' *free. Only points with* $\bar{w}_\mathrm{ps}(m_\lambda, z_i) > 0$, $\bar{w}_\mathrm{ss}(z_i) > 0$ *and* $N_\mathrm{p}(m_\lambda, z_i) > 1000$ *have been plotted to reduce the effect of insufficient statistics.*

higher values for b_L at high luminosities than for the case where $M' = -23.3$ was fixed.

3.3 K-corrections

As introduced in chapter 2, K-corrections are of importance as a higher order correction in our model, since they have an impact on the distance modulus as seen in equation (1.1), while up until now we assumed a naive distance modulus from equation (2.23). K-corrections allow us to transform from the observed wavelength when observed through a particular filter at a redshift z, into the emitted wavelength in the rest frame at $z = 0$. Of course if a galaxy spectrum is flat, the K-correction is zero, but this is never the case exactly. K-corrections exactly correct for the redshift of the spectrum, which means a different waveband of the spectrum is redshifted to the waveband of observation, while without K-corrections we implicitly assume that the spectrum looks the same in both wavebands. The derivation in [Hogg et al., 2002] serves as a great overview of the theoretical background. The question now is how to most effectively introduce K-corrections into the model. The past decade, most methodology concerning K-corrections in galactic astrophysics is ultimately based on the `kcorrect` code introduced in [Blanton and Roweis, 2007] (other methods have been presented in e.g. [Chilingarian et al., 2010, Loveday et al., 2012, O'Mill et al., 2011, Beare et al., 2014]., where the last one gives a good general overview of past methods). In summary, they reduce (using Principle Component Analysis and observations) model spectra to a base set of general spectra that generate the low-dimensional subspace galaxy spectra are observed to reside in.[3] Then they find for a given

[3] It should be noted that this is of course based on spectroscopic measurements, so the bias we tried to avoid by not letting our model depend on SEDs now enters through the calculation of K-corrections, but this is unavoidable, since

Simple double Schechter model with M_* and M' free

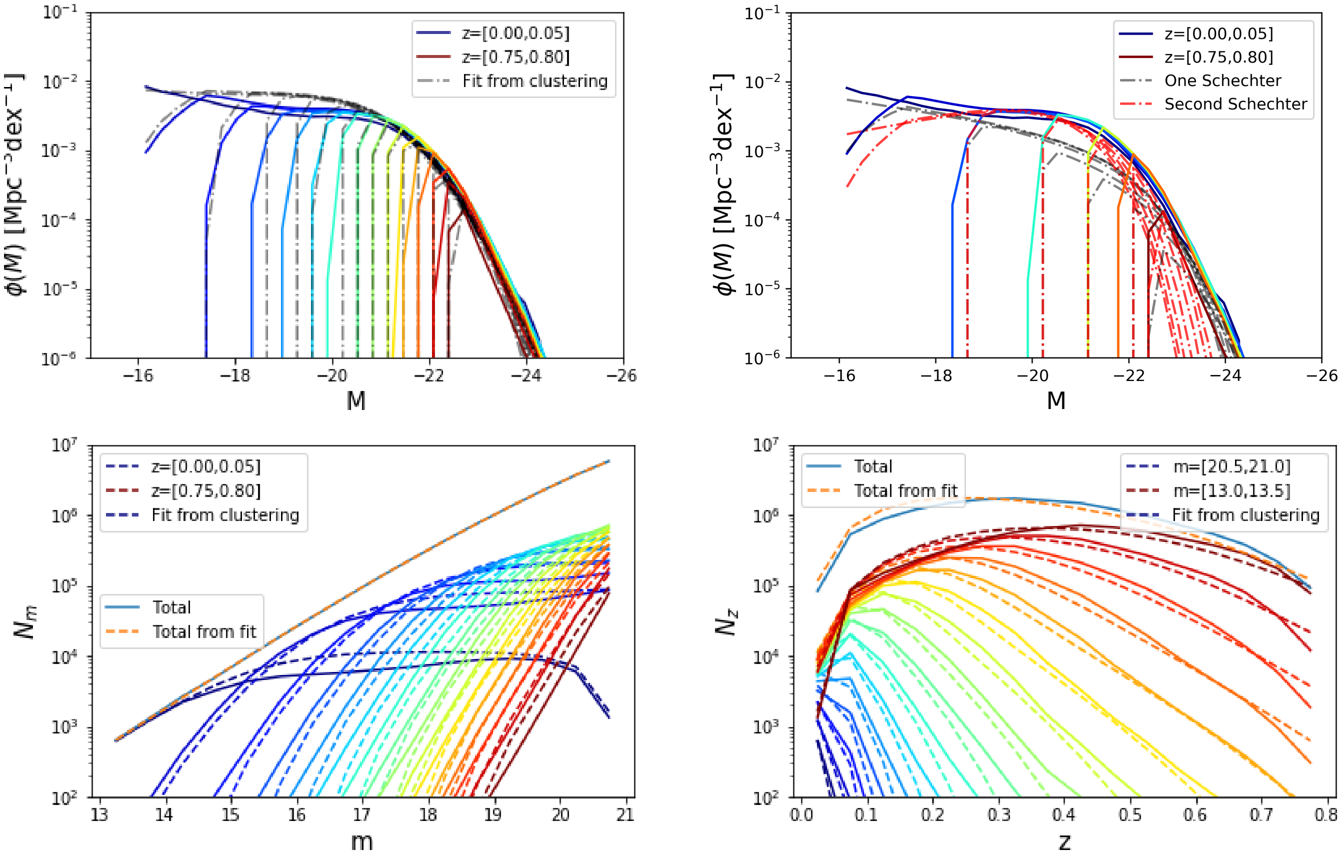

Figure 3.5: *As figure 3.1, but for the simple double Schechter model with M_* and M' free. Results are comparable to figure 3.5, except the bottom right figure shows a better fit of brighter, high redshift galaxies, although the low luminosity slope is overestimated significantly now. Note the 'roles' of the two Schechters have been reversed due to a change in starting position to point out degeneracies.*

spectrum the linear combination of the 5 base spectra that reproduces this, and using the known forms of the base spectra, K-corrections can be estimated by redshifting those base spectra before linear combination and then finding integrated fluxes over the linear combination. They show that only knowing the apparent magnitudes in three wavebands is enough to robustly fit galaxy spectra, with the exception of some specific absorption or emission lines, and thus calculate K-corrections.

Since the process of calculating K-corrections in this way is quite involved, we do not want to calculate K-corrections in every iteration of our model as this would be computationally too heavy. Therefore, we introduce more bins, in two observed colours, to correct the distance modulus correctly for each bin in apparent magnitude, redshift and two observed colors, where we only have to calculate the K-corrections once for the entire code given the ranges of the bins. To see if this is viable, we do need to confirm that K-corrections change negligibly over the range of one bin in apparent magnitude, redshift and the two observed colors.

We take $r-i$ and $r-z$ as our observed colors to be able to cover the largest range in possible redshifting of the spectrum with the range of wavebands we use. Since we used apparent magnitudes in the i-band, adding the r and z magnitudes achieves this purpose without going outside the limit of

they exactly depend on the spectrum of the object. We will show below that this, however, is not a problem for our purposes.

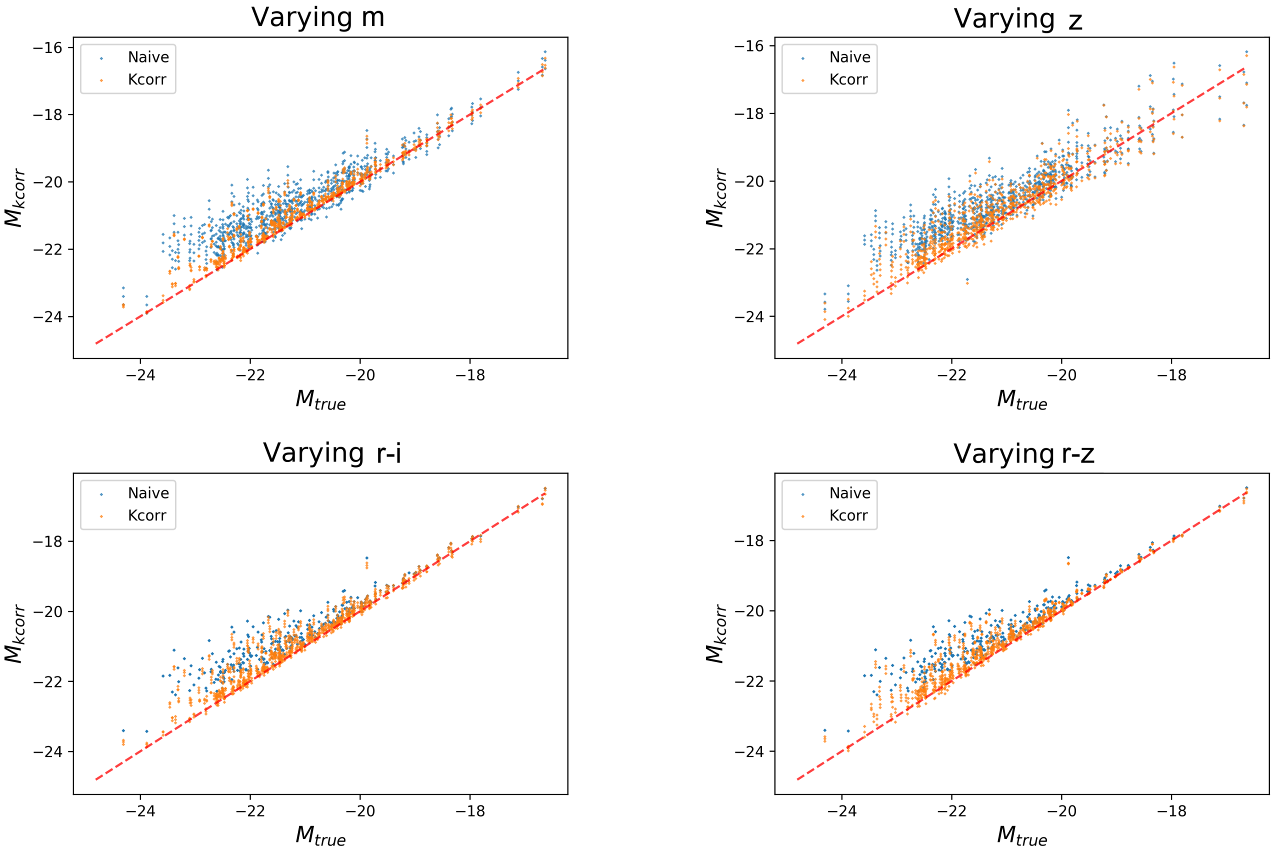

Figure 3.6: *Variation of* `kcorrect` *in the* r*-band over the bins of the model. The mock galaxy sample galaxies are assigned to their bin and then their naively estimated absolute magnitude* M_{Kcorr} *(blue) and K-corrected absolute magnitude (orange) are plotted for the lower and upper edge and the midpoint of the corresponding bin, given the actual absolute magnitude* M_{true} *extracted from the simulation. The red dashed line shows where all points should be were the distance modulus estimated perfectly.* ***Top left:*** *Variation over the bin in apparent magnitude* m*, showing of course multiple blue points as well, and little spread in K-corrected points over the width of the bin.* ***Top right:*** *Variation over the bin in redshift* z*, which has the largest spread in naive as well as K-corrected values, due to K-corrections only serving as a small correction on top of the naive distance modulus, which is governed by redshift variation. It can thus also be seen the variation over the blue and orange points is comparable.* ***Bottom left:*** *Variation over the bin in* $r - i$ *color, where we see only one blue point since color does not affect apparent magnitude or redshift and we keep the* r*-magnitude constant.* ***Bottom right:*** *Same as the bottom left, but for the* $r - z$ *color. Note the small variation over a bin in both figures.*

SDSS filters.[4] The range of the observed colours in our mock galaxy survey is $r - i = [-0.3, 1.55]$ and $r - z = [-0.5, 2.3]$. The amount of bins we now take for the two colors depends on the variation of K-correction values over the range of such a bin, but also we do not want to take too many bins to avoid insufficient statistics. We take $n_{r-i} = n_{r-z} = 8$ bins, $\Delta_{r-i} = 0.23125$ and $\Delta_{r-z} = 0.35$ wide. In figure 3.6 we show the variation of `kcorrect` over the bins of our model.

Here we see that the most significant variation is seen over a redshift bin, while we can safely assume that for 8 bins in each observed color, the K-correction is constant over the entirety of the bin in apparent magnitude and two observed colors. Of course the big variation over z comes mostly

[4] It should be noted that this thus only works up to some redshift $z \sim 1$ because galaxies further away are redshifted so much that the wavelength range of some SDSS filters is shifted into the IR or even Radio regimes, thus complicating observations. Since we limit ourselves to $z \leq 0.8$ this can be a cause of error in implementing K-corrections, but only significantly for the highest redshift bin.

from the naive distance modulus, since K-corrections only serve as a small correction, also apparent from the variation over the bin being approximately as large in the naive calculation as it is in the K-corrected calculation. To thus implement K-corrections in the model we need to subdivide over the redshift bins up to the limit where we can assume the GLF to be constant over z as well, since the expression in equation (2.27) is non-analytic and has to be numerically integrated. This was already done as mentioned in chapter 2, so implementation is straightforward. Note that, with this methodology, we only need to calculate K-corrections for each bin in apparent magnitude m_λ, redshift z_{i,sub_j} and observed colours $r-i$ and $r-z$ once, where sub_j denotes the j-th subbin of redshift bin z_i.

3.3.1 Updated model for N_p

When we do not ignore K-corrections and set $c_1 \equiv r-i$ and $c_2 \equiv r-z$, equation (2.27) becomes:

$$\begin{aligned}\widetilde{N}_{\mathrm{p},\lambda i c_1 c_2} = \frac{2}{5}\ln(10)B \int\limits_{z_{i,\min}}^{z_{i,\max}} \frac{d_\mathrm{c}(z)^2\phi_*(z)}{\sqrt{\Omega_{\mathrm{m},0}(1+z)^3+\Omega_{\Lambda,0}}}\times \\ \int\limits_{m_1}^{m_2} \frac{10^{\frac{2}{5}(M_*(z)-M(m,z,c_1,c_2))(\alpha(z)+1)}e^{-10^{\frac{2}{5}(M_*(z)-M(m,z,c_1,c_2))}}}{\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M_\mathrm{lim})}\right)}dmdz,\end{aligned} \tag{3.5}$$

where $B = 4\pi f(A)c/H_0$, $m_1 = \min\{m_{\lambda,\min}; m(M_\mathrm{lim}, z, c_1, c_2)\}$ and
$m_2 = \min\{m_{\lambda,\max}; m(M_\mathrm{lim}, z, c_1, c_2)\}$. Thus not only the integrand but also the limits depend on colour. Since K-corrections can be assumed to vary negligibly with m between bin edges, $\frac{dM}{dm} = 1$ within the limits of integration, meaning the analytic solution is identical, and we recover:

$$\begin{aligned}\widetilde{N}_{\mathrm{p},\lambda i c_1 c_2} = B \int\limits_{z_{i,\min}}^{z_{i,\max}} \frac{d_\mathrm{c}(z)^2\phi_*(z)}{\sqrt{\Omega_{\mathrm{m},0}(1+z)^3+\Omega_{\Lambda,0}}}\left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M_\mathrm{lim})}\right)\right]^{-1}\times \\ \left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_2,z,c_1,c_2))}\right)-\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_1,z,c_1,c_2))}\right)\right]dz.\end{aligned} \tag{3.6}$$

Here we calculate the integral over bin z_i by subdividing it in 10 subbins z_{i,sub_j} and summing over them.

It should be noted here that the role of the colour bins is to say what $\widetilde{N}_{\mathrm{p},\lambda i}$ would be for galaxies with colours c_1 and c_2, where the role of the magnitude and redshift bins were to integrate over them. Therefore the quantities in equation (3.6) can not simply be summed to find $\widetilde{N}_{\mathrm{p},\lambda}$, and thus we should weigh with the number of observed galaxies in a given bin of colour:

$$\begin{aligned}\widetilde{N}_{\mathrm{p},\lambda} &= \sum_i \widetilde{N}_{\mathrm{p},\lambda i} = \sum_i \sum_{c_1,c_2} \widetilde{N}_{\mathrm{p},\lambda i c_1 c_2}\frac{N_{\mathrm{p},c_1c_2}}{N_\mathrm{p,tot}} \\ &= B\sum_{c_1,c_2}\frac{N_{\mathrm{p},c_1c_2}}{N_\mathrm{p,tot}}\int_{z_\mathrm{min}}^{z_\mathrm{max}} \frac{d_\mathrm{c}(z)^2\phi_*(z)}{\sqrt{\Omega_{\mathrm{m},0}(1+z)^3+\Omega_{\Lambda,0}}}\left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M_\mathrm{lim})}\right)\right]^{-1}\times \\ &\left[\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_2,z,c_1,c_2))}\right)-\Gamma\left(\alpha(z)+1.10^{\frac{2}{5}(M_*(z)-M(m_1,z,c_1,c_2))}\right)\right]dz.\end{aligned} \tag{3.7}$$

where we remember a tilde denoted the model value, and thus $N_\mathrm{p,tot}$ and N_{p,c_1c_2} denoted the observed quantities.

Additionally our model for the clustering (equation (2.18)) changes due to the colour dependence of b_L, since calculating it involves the distance modulus. We could thus choose whether to keep colour information for clustering. However, since none of our model parameters changes the redshift distribution in a colour-dependent way, the model does not have enough freedom to for example increase the amount of red galaxies at higher redshift if it is observed there should be more red galaxies there. If the parameters change, they change for all colours equally. Furthermore splitting up our sample would severely degrade the clustering measurements and introduce more insufficient statistics, especially at the regions we already noticed to have problems with this. Additionally, we would have to add colour-dependent parameters, but since the model already has difficulty with degeneracies, this is unwanted. Therefore, we will average over the colour bins to find $\bar{b}_\mathrm{L}$. We thus set:

$$\bar{b}_\mathrm{L}(m_\lambda, z_i) = \sum_{c_1,c_2} \bar{b}_\mathrm{L}(m_\lambda, z_i, c_1, c_2) \frac{\widetilde{N}_{\mathrm{p},\lambda i c_1 c_2}}{\widetilde{N}_{\mathrm{p},\lambda i}}, \tag{3.8}$$

but since we have calculated K-corrections for all subbins in redshift we have to weigh $\bar{b}_\mathrm{L}(m_\lambda, z_i, c_1, c_2)$ over those already to find:

$$\begin{aligned}\bar{b}_\mathrm{L}(m_\lambda, z_i) &= \sum_{c_1,c_2} \frac{\widetilde{N}_{\mathrm{p},\lambda i c_1 c_2}}{\widetilde{N}_{\mathrm{p},\lambda i}} \sum_j b_\mathrm{L}(m_\lambda, z_{i,\mathrm{sub}_j}, c_1, c_2) \frac{\widetilde{N}_{\mathrm{p},\lambda i_j c_1 c_2}}{\widetilde{N}_{\mathrm{p},\lambda i c_1 c_2}} \\ &= \sum_{c_1,c_2,j} b_\mathrm{L}(m_\lambda, z_{i,\mathrm{sub}_j}, c_1, c_2) \frac{\widetilde{N}_{\mathrm{p},\lambda i_j c_1 c_2}}{\widetilde{N}_{\mathrm{p},\lambda i}},\end{aligned} \tag{3.9}$$

where:

$$b_\mathrm{L}(m_\lambda, z_{i,\mathrm{sub}_j}, c_1, c_2) = 1 + \frac{L(m_\lambda, z_{i,\mathrm{sub}_j}, c_1, c_2)}{L'} = 1 + 10^{-\frac{2}{5} M(m_\lambda, z_{i,\mathrm{sub}_j}, c_1, c_2) - M'}, \tag{3.10}$$

thus equation (2.18) becomes:

$$\begin{aligned}\widetilde{\bar{w}}_\mathrm{ps}(m_\lambda, z_i) &= \sum_k K \bar{b}_\mathrm{L}(m_\lambda, z_k) \frac{\widetilde{N}_{\mathrm{p},\lambda k}}{\widetilde{N}_{\mathrm{p},\lambda}} \bar{w}_\mathrm{ss}(z_i, z_k) \\ &= \sum_{k,c_1,c_2,j} K b_\mathrm{L}(m_\lambda, z_{k,\mathrm{sub}_j}, c_1, c_2) \frac{\widetilde{N}_{\mathrm{p},\lambda k_j c_1 c_2}}{\widetilde{N}_{\mathrm{p},\lambda}} \bar{w}_\mathrm{ss}(z_i, z_k)\end{aligned} \tag{3.11}$$

It should be noted here that implementation of a second Schechter function still goes analogous to equation (3.1), adding a term $\widetilde{N}'_{\mathrm{p},\lambda}$ to equation (3.7) equal to:

$$\begin{aligned}\widetilde{N}'_{\mathrm{p},\lambda} = B \sum_{c_1,c_2} \frac{N_{\mathrm{p},c_1 c_2}}{N_\mathrm{p,tot}} \int_{z_\mathrm{min}}^{z_\mathrm{max}} \frac{d_\mathrm{c}(z)^2 \phi'_*(z)}{\sqrt{\Omega_{\mathrm{m},0}(1+z)^3 + \Omega_{\Lambda,0}}} \left[\Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M_\mathrm{lim})}\right)\right]^{-1} \times \\ \left[\Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M(m_2, z, c_1, c_2))}\right) - \Gamma\left(\alpha'(z) + 1.10^{\frac{2}{5}(M'_*(z) - M(m_1, z, c_1, c_2))}\right)\right] dz.\end{aligned} \tag{3.12}$$

Again, the simple double Schechter model can be used, and M_* and M' could be made free parameters again.

3.3.2 Fitting

When fitting the K-corrected model to the data, we are still subject to the same constraints, $\bar{w}_{\mathrm{ps}}$ and $N_{\mathrm{p},\lambda}$, however, we have more information now and should instead use $N_{\mathrm{p},\lambda c_1 c_2}$ as a constraint, which can be found from the model and directly compared to observations, not losing any observed data, decreasing degeneracy. Still, note that K-corrections are imperfect and the error in them is unknown, which could bias our model outcome. Of course, this is also implicitly true when summing over colour, but the effect is more apparent without summing.

When using $N_{\mathrm{p},\lambda c_1 c_2}$ as a constraint, we update the likelihood (equation (2.29)) to include these quantities:

$$\begin{aligned}\ln \mathcal{L}(\boldsymbol{\theta}) = &-\frac{1}{2}\ln(|C(\boldsymbol{\theta})|) - \frac{1}{2}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}})^T C^{-1}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}}) \\ &+ \sum_{j=0}^{n_{\mathrm{m}}-1}\sum_{c_1=0}^{n_{r-i}-1}\sum_{c_2=0}^{n_{r-z}-1}\left[N_{\mathrm{p},jc_1c_2}\ln\left(\widetilde{N}_{\mathrm{p},jc_1c_2}(\boldsymbol{\theta})\right) - \widetilde{N}_{\mathrm{p},jc_1c_2}(\boldsymbol{\theta})\right].\end{aligned} \tag{3.13}$$

Clearly, only the second term is affected, and keeps its Poissonian nature, only being extended to include more data points.

In the next subsection we apply the model with incorporated K-corrections to the mock galaxy catalogue.

3.3.3 Results

Firstly, we extended the mock catalogue to now include more low magnitude galaxies, so we set $M_{\mathrm{lim}} = -14$. Still we use the outcome of the previous runs as starting point for the model, and first try to fit the dataset roughly by hand, again having one Schechter fit the low-luminosity slope and the high-luminosity cut-off, while the second Schechter adds the bump visible in the data, since, independent of which parameters were included, this was the structure of the result of the previous runs. Preliminary results of the implementation of models including equations (3.7) and (3.11) as constraints, using equation (3.12) when implementing a second Schechter and using (3.13) when fitting to data binned over colours are shown in the following figure 3.7.

As can be seen in this figure, the K-correction methodology (whether with or without extra binned constraints) has trouble fitting the extended mock galaxy catalogue. However, the figure in the bottom right shows this is not a problem inherent only to the K-correction methodology, since the original model also misfits this dataset. This presumably points to our likelihood surface having many local minima, and we tried different starting points, resulting in completely different results as well, confirming this. Therefore, either better fitting by trial-and-error must be done to find a better starting point, preferably better substantiated by theory, or we need to reconsider the parameters we want to include or exclude from the model. We therefore tried fitting the full double Schechter model without K-corrections, and it still returned comparable results. All results are collected in table 3.1, listing the starting point as well as best-fit parameters for the fiducial run (figure 2.1) as well as the runs discussed in this chapter.

Note here that the GLF may be highly accurately reproduced even for different parameters than the input (when this is fitted by trial-and-error) due to degeneracy with the normalization and realization noise.

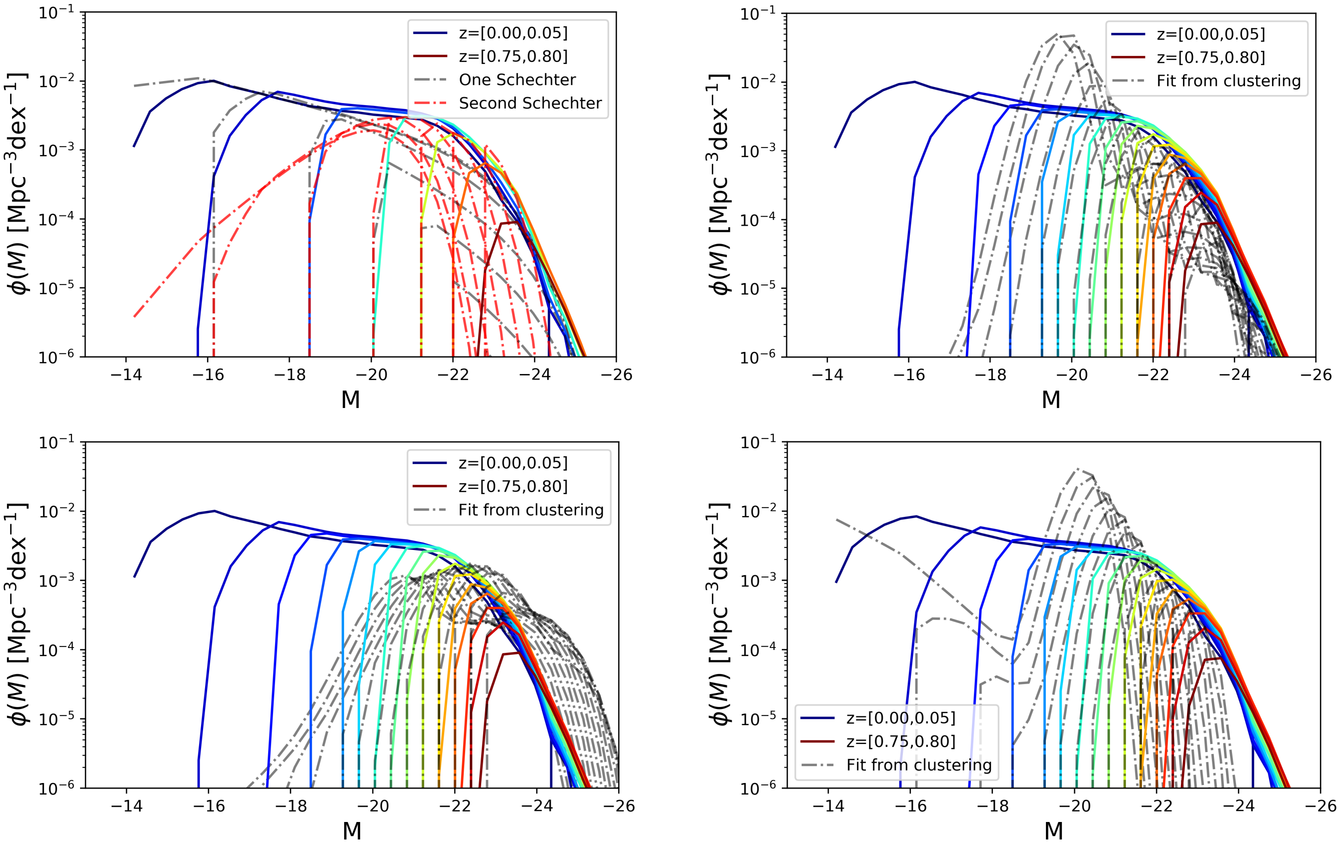

Figure 3.7: ***Top left:*** *The starting point of the model, fitted by hand to the extended mock galaxy catalogue.* ***Top right:*** *The result of the model using a simple double Schechter with M_* and M' free including K-corrections. Clearly the model has trouble fitting the data.* ***Bottom left:*** *Result of the same model as the top right, but now including constraints binned on colour, thus using likelihood from equation* (3.13)*. The extra constraints added by binning the number constraint on colours do not solve the problems from the top right figure.* ***Bottom right:*** *Result of the model without K-corrections (thus using equations* (3.1), (2.28), (2.18) *and* (2.29) *for a simple double Schechter with M_* and M' free) on the extended mock galaxy catalogue. Clearly, the problem does not lie solely with the K-correction methodology, since the original code also has trouble with the extended mock catalogue.*

To further discuss the results in figure 3.7, in the appendix we show the usual figures displaying N_m and N_z (as in 3.2 for example) for the four cases from figure 3.7 as well. There we see essentially that without using equation (3.13) one Schechter is made redundant for fitting $N_{\mathrm{p},\lambda}$, and that overestimation of high-redshift, high-luminosity galaxies is a continuing problem in the fit. Possible causes of the bad performance are discussed in the next section.

3.3.4 Discussion

While the double Schechter model from section 3.1 (figure 3.3) fits the GLF visibly the closest, it seems further improvements to the model did not lead to better results. Firstly, adding M' as a free parameter is needed to correctly fit b_{L} to the dataset independent of the set itself, but did not lead to a better visual fit of the GLF. Most likely this is due to introduction of extra degeneracies, in this case mostly with K, but of course more indirectly also with the ϕ_* via equation (2.18). This results in more local minima in the likelihood surface, meaning the model is more dependent on the input

Run	α_0	M_{*0}	$\alpha_{\rm e}$	$M_{*{\rm e}}$	α'_0	M'_{*0}	$\alpha'_{\rm e}$	$M'_{*{\rm e}}$	M'
Input	**-1.01**	**-21.5**	**-0.15**	**-0.8**	**-**	**-**	**-**	**-**	**-**
Fiducial (vDW)	-1.050	-21.429	-0.155	-0.927	-	-	-	-	-
Input	**-0.57**	**-20.97**	**-0.72**	**-1.18**	**-1.36**	**-20.97**	**-0.19**	**-1.18**	**-**
DS	-0.211	-20.403	-0.745	-1.312	-1.469	-22.108	-0.318	-1.023	-
Simple DS	-0.370	-20.710	-1.199	-1.394	-1.285	-	-0.204	-	-
Input	**-1.36**	**-20.97**	**-0.19**	**-1.18**	**-0.60**	**-20.50**	**-0.04**	**-1.04**	**-23.3**
M_* free	-1.214	-21.856	-0.585	-1.340	-0.564	-20.562	-0.257	-1.131	-
M_* and M' free	-1.260	-21.830	-0.573	-1.334	-0.629	-20.490	-0.262	-1.189	-22.498
Input	**-1.25**	**-21.16**	**0.54**	**-0.50**	**0.92**	**-20.13**	**-0.37**	**-0.49**	**-22.29**
Kcorrdata	8.028	-17.700	4.899	-3.336	-2.188	-20.308	-1.416	-0.521	-25.571
Kcorrect	-3.000	-21.561	-1.239	-0.379	8.023	-17.504	4.672	-3.895	-24.347
Colour likelihood	0.339	-22.159	0.408	-2.116	3.244	-19.067	1.963	-3.238	-21.968

Table 3.1: *The best-fit GLF parameters derived from clustering data for each of the model runs above. A "-" denotes that the parameter is not included in the specific model run. "DS" stands for the double Schechter model from figure 3.1. "Simple DS" corresponds to figure 3.2. "M_* free" to figure 3.3 and "M_* and M' free" to figure 3.5. "Kcorrdata" denotes the run of the original model on the extended mock galaxy catalogue, corresponding to the bottom right figure of figure 3.7. The other panels of figure 3.7 correspond to the last "Input" line (top left), "Kcorrect" (top right) and "Colour likelihood" (bottom left).*

parameters.

This is also a possible cause of the bad performance of the K-corrected model. We would like the starting point to be explained by physical background, but this remains to be explored, since not enough is known on this topic yet ([Johnston, 2011]). Discerning between different types of galaxies (star-forming vs. quiescent) and different sources of mass-quenching in quiescent galaxies seems to be the best approach, although not knowing enough about M/L in these types of galaxies gets in the way of making conclusive statements. With these distinctions in the back of the mind, one could say a more logical starting point could be one Schechter fitting the cutoff, while another fits the low-luminosity slope, but that would make many parameters redundant (α of the first Schechter for example) and would also conflict with us wanting to constrain the universal GLF.

The fact that the model from section 3.1 also underperforms on the extended mock galaxy catalogue as seen in the bottom right panel of figure 3.7 also points to degeneracy being the main problem, more importantly because it is started at a different starting point than previous models. In table 3.1 we also see that the M_{*0}-parameter of the first and second Schechter interchange (Input parameters had $M_{*0} < M'_{*0}$ but the best-fit parameters had $M_{*0} > M'_{*0}$) for this run. We could bound the parameter space to the region where $M_{*0} < M'_{*0}$ to reduce at least one possibility of degeneracy.

Another possibility why the model starts to perform worse when including more low luminosity galaxies, especially for the low and high absolute magnitudes, exactly the limits where we have less data, could be due to insufficient sampling of the clustering in these regimes. That means that the assumption of a Gaussian likelihood breaks down in those regimes, since the Gaussian likelihood only holds in the limit of sufficient data, due to the law of large numbers. It is thus recommended to search for a more accurate form of likelihood to use in the regimes of statistically insufficient data.

The errors in `kcorrect` could be a source of error in these models, but as we see in figure 3.6, the correction always brings the naive distance modulus closer to the actual value while varying only slowly over the extent of the bins involved in our model. Therefore this is the least likely source of problems in the presented results. What could be a problem added to the model when implementing K-corrections is the extra bins we add for colours, by which we add more possibility for statistical artifacts to enter our analysis, although this is also partly counteracted by our ability to add more specific constraints to the model, and the model performs slightly better in overall shape when adding these constraints.

Lastly, this model certainly breaks down when reality can not be described by a Schechter function. Then we would need to find a different underlying model for the GLF that we want to fit to the data. The advantages of our current approach are mainly twofold. Firstly we avoid bias due to having spectroscopic samples being selected differently, since we do not use the properties of the spectroscopic galaxies to fit models to (as is done in SED-modelling or semi-analytic modelling as used in [Bates et al., 2019]). Secondly, we only have to make a simple assumption on the galaxy bias, namely the shape of the luminosity dependent galaxy bias factor, b_{L}, and incorporate the uncertainty about the galaxy bias as a free parameter in the model. Other results not able to do this will certainly be degenerate with the unknown redshift-dependent galaxy bias factor.

Summary and Conclusion

The main question of this thesis was to find out if improvements and extensions of the vDW method [van Daalen and White, 2018] would lead to better estimates of the universal GLF and its evolution through cosmic time. This model was constructed to be less dependent on spectroscopic data and the resulting biases when fitting galaxy evolution models to observations. In order to do that, we assert the assumptions made in their paper, as well as incorporate K-corrections in the conversion between apparent and absolute magnitudes. They do mention dust extinction and lensing magnification having an effect on the model, but these are outside the scope of this research.

Our research starts with a description, in chapter 1, of the general background of the GLF, most importantly the Schechter function, and theories of galaxy clustering. Most importantly we discussed the currently well-established theoretical concepts behind estimating the galaxy two-point correlation function. How galaxies are biased with respect to the underlying matter distribution, both due to halo bias and luminosity bias, is important in the context of the model.

Previous work on using cross-correlations to estimate the redshift distribution formed the basis and starting point of our research, and this is summarized in chapter 2. It presents how, using a spectroscopic sample of which we know the redshifts, we can assign redshifts to a photometric sample using cross-correlations between the samples. We describe how we average over multiple spatial scales and how we need to handle the difference in galaxy bias for the two samples, since objects good for spectroscopic measurements are more highly biased. Afterwards, we describe how we find a model value for $N_{\mathrm{p}}(m, z)$, the amount of photometric galaxies with apparent magnitude m and redshift z, given our parametric model dependent on a Schechter luminosity function. The way the model is fitted to data and the result of the model from vDW (their 'fiducial' model) were presented as background for chapter 3. Their model only serves as a conceptual proof, and should be extended to include effects that become important when applying this methodology to actual data.

We implemented three improvements to the vDW fiducial model to apply it to realistic data. The theory and results of these improvements are presented in chapter 3. First is the addition of a second Schechter function since actual data is often better fitted with two Schechter functions as proposed for instance in [Peng et al., 2010]. We describe how the choice of which parameters to include influences the location of local minima on the likelihood surface and conclude the model in figure 3.3 provides the current best fit to a simulated mock galaxy catalogue from the Planck Millennium all-sky lightcones simulation [Henriques et al., 2015]. This model implements a second Schechter function of which the normalization is assumed to evolve equally to the normalization of the first Schechter and only differs by a relative factor, while other parameters are free. Afterwards, we describe how the luminosity dependent galaxy bias factor, b_{L} needs to be modelled or fitted to data, and decide it is better to include an additional parameter for fitting b_{L} to actual datasets, since otherwise we could only fit it to the spectroscopic set and thus introduce the bias we wanted to avoid. Lastly we discussed how `kcorrect` by [Blanton and Roweis, 2007] calculates K-corrections to the distance modulus given three observed apparent magnitudes, and how to incorporate these corrections into the model by binning over two observed colours. We show that the K-corrections vary negligibly over the bins. However, the preliminary results of this addition are shown to be less than ideal.

Still, the extensions of the model show promising results for when we ultimately want to apply the model to actual data, since the overall shape of the GLF is fitted to a very good extent by the model from figure 3.3. We therefore laid foundations for further work extending the model and accounting for the problems encountered during this research.

Therefore we advise future research to further study the following:

- Most importantly try to get a better understanding of the likelihood surface and its structure of minima. Of course this is quite complicated seeing the dimensionality of the problem. Therefore a good start would be to gain a better theoretical background on the GLF, thereby understanding even better the dependencies of its parameters, and thus which parameters need to be included, or need to be subject to what constraints (i.e. constrain $M_{*0} < M'_{*0}$). This would also enable us to find a more substantiated starting point for the two Schechter components for the model. Since much previous work is about the Schechter mass function [Johnston, 2011], this needs to be extended to the GLF. Extending it naively brings us to the current assessment, but assumes a constant M/L-ratio, which is well-established to depend strongly on specific galaxy properties and environments which were outside the scope of this work.

- The assessment of the errors and the variability of calculated K-corrections from `kcorrect` could be extended to find an even more optimal database of K-corrections to use. By this we mean to reduce the statistical artifacts introduced by subdividing all bins over colour as well as reducing the bias introduced due to K-corrections being derived from spectral models.

- The mathematics behind the likelihood for the correlation function $\bar{w}_{\mathrm{ps}}$ could lead to better insights of how to fit the model. Since we experience sampling noise at both the high redshift, high luminosity end (due to only the brightest objects being observed at those distances) and the low redshift end (due to cosmic variance), the assumption of a Gaussian likelihood breaks down, and should be altered. The exact mathematical expression is an area that should be explored.

- Lastly, before ultimately applying the model to actual data, the possible effects of lensing magnification as well as dust extinction need to be implemented to reduce their influence on the result. Since these effects concern only higher-order effects, their effect should be small, but important when we want to extract the exact parameters describing the galaxy population. We note that lensing magnification effects are not present in the mock galaxy catalogue, but will be in actual data.

Acknowledgements

Although I of course, since it is a custom, have to start by saying thanks to Dr. Marcel van Daalen for giving me the opportunity to work on this exciting project, that does not express enough thanks for the help I received during work on this project, but not concerning the project. I had not done a presentation of my own work for about a year when starting this project, but he helped me not only get my confidence back on track, but also start my path to a further scientific career by helping with and encouraging me in my search for PhD positions. Now concerning the project, we had meetings sometimes lasting so long I forgot to have lunch completely, since he was always ready to explain the intricacies of his model until I (at least thought I) understood every part, staying patient when the next meeting I had to ask again. Additionally this project has skyrocketed my understanding and ability of coding (which is apparent since while this thesis is the shortest I have written during my studies, I think the amount of lines of code written throughout more than makes up for this), so I feel the the opportunity to do this project has benefited me enormously.

Thanks also to Koenraad Kuijken for taking over supervision, albeit for only a short time.

Then, as has become a custom as well, I would like to thank my father for proofreading my thesis and giving feedback, even in the short timespan left before the deadline. This short timespan of course mostly being a result from the extensive house-hunting we needed to do in the same weeks accommodating my move to the outer territories.

Due to the (here we go again with the customs) weird times this year, it occasionally became hard to focus on studying, but my friends were always ready for emotional support, especially Merel. I would also like to thank Toon, Iris, Thomas, Emiel and Matthijs for occasional relaxing conversations and activities.

Bibliography

[Bates et al., 2019] Bates, D. J., Tojeiro, R., Newman, J. A., Gonzalez-Perez, V., Comparat, J., Schneider, D. P., Lima, M., and Streblyanska, A. (2019). Mass functions, luminosity functions, and completeness measurements from clustering redshifts. *MNRAS*, 486(3):3059–3077.

[Beare et al., 2014] Beare, R., Brown, M. J. I., and Pimbblet, K. (2014). An Accurate New Method of Calculating Absolute Magnitudes and K-corrections Applied to the Sloan Filter Set. *ApJ*, 797(2):104.

[Benoist et al., 1996] Benoist, C., Maurogordato, S., da Costa, L. N., Cappi, A., and Schaeffer, R. (1996). Biasing in the Galaxy Distribution. *ApJ*, 472:452.

[Bezanson et al., 2016] Bezanson, R., Wake, D. A., Brammer, G. B., van Dokkum, P. G., Franx, M., Labbé, I., Leja, J., Momcheva, I. G., Nelson, E. J., Quadri, R. F., Skelton, R. E., Weiner, B. J., and Whitaker, K. E. (2016). LEVERAGING 3d-HST GRISM REDSHIFTS TO QUANTIFY PHOTOMETRIC REDSHIFT PERFORMANCE. *The Astrophysical Journal*, 822(1):30.

[Binney and Tremaine, 2008] Binney, J. and Tremaine, S. (2008). *Galactic Dynamics: Second Edition*.

[Blanton et al., 2005] Blanton, M. R., Lupton, R. H., Schlegel, D. J., Strauss, M. A., Brinkmann, J., Fukugita, M., and Loveday, J. (2005). The Properties and Luminosity Function of Extremely Low Luminosity Galaxies. *ApJ*, 631(1):208–230.

[Blanton and Roweis, 2007] Blanton, M. R. and Roweis, S. (2007). K-Corrections and Filter Transformations in the Ultraviolet, Optical, and Near-Infrared. *AJ*, 133(2):734–754.

[Bolzonella et al., 2000] Bolzonella, M., Miralles, J. M., and Pelló, R. (2000). Photometric redshifts based on standard SED fitting procedures. *A&A*, 363:476–492.

[Bonne et al., 2015] Bonne, N. J., Brown, M. J. I., Jones, H., and Pimbblet, K. A. (2015). THE INFLUENCE OF RED SPIRAL GALAXIES ON THE SHAPE OF THE LOCALK-BAND LUMINOSITY FUNCTION. *The Astrophysical Journal*, 799(2):160.

[Chilingarian et al., 2010] Chilingarian, I. V., Melchior, A.-L., and Zolotukhin, I. Y. (2010). Analytical approximations of K-corrections in optical and near-infrared bands. *MNRAS*, 405(3):1409–1420.

[Choi et al., 2016] Choi, A., Heymans, C., Blake, C., Hildebrandt, H., Duncan, C. A. J., Erben, T., Nakajima, R., Van Waerbeke, L., and Viola, M. (2016). CFHTLenS and RCSLenS: testing photometric redshift distributions using angular cross-correlations with spectroscopic galaxy surveys. *MNRAS*, 463(4):3737–3754.

[Coles and Lucchin, 2002] Coles, P. and Lucchin, F. (2002). *Cosmology: The Origin and Evolution of Cosmic Structure, Second Edition.*

[Cucciati et al., 2016] Cucciati, O., Marulli, F., Cimatti, A., Merson, A. I., Norberg, P., Pozzetti, L., Baugh, C. M., and Branchini, E. (2016). Measuring galaxy environment with the synergy of future photometric and spectroscopic surveys. *MNRAS*, 462(2):1786–1801.

[Cunha et al., 2009] Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., and Lin, H. (2009). Estimating the redshift distribution of photometric galaxy samples - II. Applications and tests of a new method. *MNRAS*, 396(4):2379–2398.

[Henriques et al., 2015] Henriques, B. M. B., White, S. D. M., Thomas, P. A., Angulo, R., Guo, Q., Lemson, G., Springel, V., and Overzier, R. (2015). Galaxy formation in the Planck cosmology - I. Matching the observed evolution of star formation rates, colours and stellar masses. *MNRAS*, 451(3):2663–2680.

[Hogg et al., 2002] Hogg, D. W., Baldry, I. K., Blanton, M. R., and Eisenstein, D. J. (2002). The K correction. *arXiv e-prints*, pages astro–ph/0210394.

[Johnston, 2011] Johnston, R. (2011). Shedding light on the galaxy luminosity function. *The Astronomy and Astrophysics Review*, 19(1).

[Kerscher et al., 2000] Kerscher, M., Szapudi, I., and Szalay, A. S. (2000). A Comparison of Estimators for the Two-Point Correlation Function. *ApJ*, 535(1):L13–L16.

[Landy and Szalay, 1993] Landy, S. D. and Szalay, A. S. (1993). Bias and Variance of Angular Correlation Functions. *ApJ*, 412:64.

[Lima et al., 2008] Lima, M., Cunha, C. E., Oyaizu, H., Frieman, J., Lin, H., and Sheldon, E. S. (2008). Estimating the redshift distribution of photometric galaxy samples. *MNRAS*, 390(1):118–130.

[Loveday et al., 2012] Loveday, J., Norberg, P., Baldry, I. K., Driver, S. P., Hopkins, A. M., Peacock, J. A., Bamford, S. P., Liske, J., Bland-Hawthorn, J., Brough, S., Brown, M. J. I., Cameron, E., Conselice, C. J., Croom, S. M., Frenk, C. S., Gunawardhana, M., Hill, D. T., Jones, D. H., Kelvin, L. S., Kuijken, K., Nichol, R. C., Parkinson, H. R., Phillipps, S., Pimbblet, K. A., Popescu, C. C., Prescott, M., Robotham, A. S. G., Sharp, R. G., Sutherland, W. J., Taylor, E. N., Thomas, D., Tuffs, R. J., van Kampen, E., and Wijesinghe, D. (2012). Galaxy and Mass Assembly (GAMA): ugriz galaxy luminosity functions. *MNRAS*, 420(2):1239–1262.

[Matthews and Newman, 2010] Matthews, D. J. and Newman, J. A. (2010). RECONSTRUCTING REDSHIFT DISTRIBUTIONS WITH CROSS-CORRELATIONS: TESTS AND AN OPTIMIZED RECIPE. *The Astrophysical Journal*, 721(1):456–468.

[McQuinn and White, 2013] McQuinn, M. and White, M. (2013). On using angular cross-correlations to determine source redshift distributions. *Monthly Notices of the Royal Astronomical Society*, 433(4):2857–2883.

[Ménard et al., 2013] Ménard, B., Scranton, R., Schmidt, S., Morrison, C., Jeong, D., Budavari, T., and Rahman, M. (2013). Clustering-based redshift estimation: method and application to data. *arXiv e-prints*, page arXiv:1303.4722.

[Mo et al., 2010] Mo, H., van den Bosch, F. C., and White, S. (2010). *Galaxy Formation and Evolution*.

[Morrison et al., 2017] Morrison, C. B., Hildebrandt, H., Schmidt, S. J., Baldry, I. K., Bilicki, M., Choi, A., Erben, T., and Schneider, P. (2017). the-wizz: clustering redshift estimation for everyone. *MNRAS*, 467(3):3576–3589.

[Muzzin et al., 2013] Muzzin, A., Marchesini, D., Stefanon, M., Franx, M., McCracken, H. J., Milvang-Jensen, B., Dunlop, J. S., Fynbo, J. P. U., Brammer, G., Labbé, I., and van Dokkum, P. G. (2013). THE EVOLUTION OF THE STELLAR MASS FUNCTIONS OF STAR-FORMING AND QUIESCENT GALAXIES TOz= 4 FROM THE COSMOS/UltraVISTA SURVEY. *The Astrophysical Journal*, 777(1):18.

[Norberg et al., 2001] Norberg, P., Baugh, C. M., Hawkins, E., Maddox, S., Peacock, J. A., Cole, S., Frenk, C. S., Bland-Hawthorn, J., Bridges, T., Cannon, R., Colless, M., Collins, C., Couch, W., Dalton, G., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Madgwick, D., Peterson, B. A., Sutherland, W., and Taylor, K. (2001). The 2dF Galaxy Redshift Survey: luminosity dependence of galaxy clustering. *MNRAS*, 328(1):64–70.

[O'Mill et al., 2011] O'Mill, A. L., Duplancic, F., García Lambas, D., and Sodré, Laerte, J. (2011). Photometric redshifts and k-corrections for the Sloan Digital Sky Survey Data Release 7. *MNRAS*, 413(2):1395–1408.

[Padmanabhan et al., 2007] Padmanabhan, N., Schlegel, D. J., Seljak, U., Makarov, A., Bahcall, N. A., Blanton, M. R., Brinkmann, J., Eisenstein, D. J., Finkbeiner, D. P., Gunn, J. E., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Loveday, J., Lupton, R. H., Nichol, R. C., Schneider, D. P., Strauss, M. A., Tegmark, M., and York, D. G. (2007). The clustering of luminous red galaxies in the Sloan Digital Sky Survey imaging data. *MNRAS*, 378(3):852–872.

[Peacock et al., 2001] Peacock, J. A., Cole, S., Norberg, P., Baugh, C. M., Bland-Hawthorn, J., Bridges, T., Cannon, R. D., Colless, M., Collins, C., Couch, W., Dalton, G., Deeley, K., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Percival, W. J., Peterson, B. A., Price, I., Sutherland, W., and Taylor, K. (2001). A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey. *Nature*, 410(6825):169–173.

[Peng et al., 2010] Peng, Y.-j., Lilly, S. J., Kovač, K., Bolzonella, M., Pozzetti, L., Renzini, A., Zamorani, G., Ilbert, O., Knobel, C., Iovino, A., Maier, C., Cucciati, O., Tasca, L., Carollo, C. M., Silverman, J., Kampczyk, P., de Ravel, L., Sanders, D., Scoville, N., Contini, T., Mainieri, V., Scodeggio, M., Kneib, J.-P., Le Fèvre, O., Bardelli, S., Bongiorno, A., Caputi, K., Coppa, G., de la Torre, S., Franzetti, P., Garilli, B., Lamareille, F., Le Borgne, J.-F., Le Brun, V., Mignoli, M., Perez Montero, E., Pello, R., Ricciardelli, E., Tanaka, M., Tresse, L., Vergani, D., Welikala, N., Zucca, E., Oesch, P., Abbas, U., Barnes, L., Bordoloi, R., Bottini, D., Cappi, A., Cassata, P., Cimatti, A., Fumana, M., Hasinger, G., Koekemoer, A., Leauthaud, A., Maccagni, D., Marinoni, C., McCracken, H., Memeo, P., Meneux, B., Nair, P., Porciani, C., Presotto, V., and Scaramella, R. (2010). Mass and Environment as Drivers of Galaxy Evolution in SDSS and zCOSMOS and the Origin of the Schechter Function. *ApJ*, 721(1):193–221.

[Planck Collaboration, 2016] Planck Collaboration (2016). Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13.

[Rahman et al., 2016] Rahman, M., Mendez, A. J., Ménard, B., Scranton, R., Schmidt, S. J., Morrison, C. B., and Budavári, T. (2016). Exploring the SDSS photometric galaxies with clustering redshifts. *Monthly Notices of the Royal Astronomical Society*, 460(1):163–174.

[Schmidt et al., 2013] Schmidt, S. J., Ménard, B., Scranton, R., Morrison, C., and McBride, C. K. (2013). Recovering redshift distributions with cross-correlations: pushing the boundaries. *Monthly Notices of the Royal Astronomical Society*, 431(4):3307–3318.

[Schulz, 2010] Schulz, A. E. (2010). CALIBRATING PHOTOMETRIC REDSHIFT DISTRIBUTIONS WITH CROSS-CORRELATIONS. *The Astrophysical Journal*, 724(2):1305–1315.

[Tomczak et al., 2014] Tomczak, A. R., Quadri, R. F., Tran, K.-V. H., Labbé, I., Straatman, C. M. S., Papovich, C., Glazebrook, K., Allen, R., Brammer, G. B., Kacprzak, G. G., Kawinwanichakij, L., Kelson, D. D., McCarthy, P. J., Mehrtens, N., Monson, A. J., Persson, S. E., Spitler, L. R., Tilvi, V., and van Dokkum, P. (2014). GALAXY STELLAR MASS FUNCTIONS FROM ZFOURGE/-CANDELS: AN EXCESS OF LOW-MASS GALAXIES SINCEz= 2 AND THE RAPID BUILDUP OF QUIESCENT GALAXIES. *The Astrophysical Journal*, 783(2):85.

[van Daalen and White, 2018] van Daalen, M. P. and White, M. (2018). A cross-correlation-based estimate of the galaxy luminosity function. *Mon. Not. Roy. Astron. Soc.*, 476(4):4649–4661.

Appendix A

Additional Figures and Derivations

A.1 Chapter 1

A.1.1 Section 1.2.2

Another, more natural way of seeing that the linear bias model introduced in section 1.2.2. is viable is as follows (from [Mo et al., 2010, pp. 679-681]): When splitting the two-point correlation function of galaxies into a one-halo and two-halo term, dependent on masses of those respective halos, the two-point correlation function of the halos can be found from the galaxy-galaxy two-halo two-point correlation function. Since dark matter halos are correlated in space, the joint probability of finding a halo of mass M_1 at $\mathbf{x}$ and one of mass M_2 at $\mathbf{x}'$ is proportional to:

$$n(M_1)dM_1 n(M_2)dM_2[1+\xi_{\mathrm{hh}}(\mathbf{x}-\mathbf{x}'|M_1,M_2)]d^3\mathbf{x}d^3\mathbf{x}', \tag{A.1}$$

where ξ_{hh} is the two-point correlation function of the halos, given their masses. Thus, the probability of having an interhalo galaxy pair, separated by a vector $\mathbf{r}$, hosted by an $M_1 - M_2$ pair separated by $\mathbf{x}-\mathbf{x}'$ is equal to the product of the number of interhalo galaxy pairs given the separation and masses M_i and the joint probability from equation (A.1).

The average number of interhalo galaxy pairs per volume given purely their separation is then an integral over the halo masses M_1 and M_2 and locations $\mathbf{x}$ and $\mathbf{x}'$. Then we can write:

$$\xi_{\mathrm{gg}}(\mathbf{r}) = \frac{[\mathrm{GG}^{1\mathrm{h}}(\mathbf{r}) + \mathrm{GG}^{2\mathrm{h}}(\mathbf{r})]dV_1 dV_2}{\mathrm{RR}(\mathbf{r})dV_1 dV_2} - 1, \tag{A.2}$$

where $\mathrm{GG}^{i\mathrm{h}}$ is the number of galaxy-galaxy pairs per volume squared in a single halo (1h) or interhalo (2h), $\mathrm{RR}(\mathbf{r})dV_1dV_2 = \overline{n}_{\mathrm{g}}^2 dV_1 dV_2$, with $\overline{n}_{\mathrm{g}} = \int n(M)\langle N|M\rangle\, dM$ the mean number density of galaxies, is the expected number of pairs in the absence of clustering. The number $\mathrm{GG}^{1\mathrm{h}}$ of course depends on the average spatial distribution of galaxies in a halo of mass M, $u(\mathbf{x}|M)$. Therefore, when ξ_{hh} can be obtained, the two-point correlation function of galaxies is determined by the first two moments of the halo occupation distribution $P(N|M)$ having moments $\langle N^k|M\rangle \equiv \sum_N N^k P(N|M)$ and the function $u(\mathbf{x}|M)$. On larger scales, where the individual halo sizes can be neglected, however, the halo correlation function is related to that of the matter, $\xi_{\mathrm{mm}}(r)$, by the linear bias relation $\xi_{\mathrm{hh}}(r|M_1, M_2) = b_{\mathrm{h}}(M_1)b_{\mathrm{h}}(M_2)\xi_{\mathrm{mm}}(r)$, so we recover:

$$\xi_{\mathrm{gg}}(r) \approx b_{\mathrm{g}}^2 \xi_{\mathrm{mm}}(r); \qquad b_{\mathrm{g}} = \int dM n(M) b_{\mathrm{h}}(M) \frac{\langle N|M\rangle}{\overline{n}_{\mathrm{g}}}. \tag{A.3}$$

A.2 Chapter 2

A.2.1 Section 2.2.1

Equation (2.29) follows from a general likelihood where we combine a Gaussian likelihood for the spatial correlations and a Poissonian likelihood for the number of galaxies in each bin. We also set $n = n_{\mathrm{z}} \times n_{\mathrm{m}}$ and $\boldsymbol{\theta}$ to be the vector of parameters:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \exp\left[-\frac{1}{2}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}})^T C^{-1}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}})\right] \times \prod_{j=0}^{n_{\mathrm{m}}-1} \frac{\widetilde{N}_{\mathrm{p},j}^{N_{\mathrm{p},j}} \exp[-\widetilde{N}_{\mathrm{p},j}]}{N_{\mathrm{p},j}!}, \tag{A.4}$$

thus:

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta}) = &-\frac{n_{\mathrm{m}} n_{\mathrm{z}}}{2} \ln(2\pi) - \frac{1}{2}\ln(|C(\boldsymbol{\theta})|) - \frac{1}{2}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}})^T C^{-1}(\overline{\mathbf{w}}_{\mathrm{ps}} - \widetilde{\overline{\mathbf{w}}}_{\mathrm{ps}}) \\ &+ \sum_{j=0}^{n_{\mathrm{m}}-1}\left[N_{\mathrm{p},j} \ln\left(\widetilde{N}_{\mathrm{p},j}(\boldsymbol{\theta})\right) - \widetilde{N}_{\mathrm{p},j}(\boldsymbol{\theta}) - \sum_{k=1}^{N_{\mathrm{p},j}} \ln(k)\right]. \end{aligned} \tag{A.5}$$

Here we can leave out constants to get to equation (2.29).

Note that the assumption of a Gaussian likelihood only holds due to assumption of enough sampling that the law of large numbers holds. Where this does not hold, however (which could for instance happen in the low luminosity and high luminosity high redshift regimes) this assumption breaks down and we need a different expression for the likelihood. The exact expression needed should be explored in further research.

A.2.2 Section 2.3

Here we discuss in some more detail what is shown in Figure 2.1.

The top-left panel shows the GLFs as a function of absolute magnitude in each redshift bin. Solid lines are the GLF as measured from the catalogue directly. The fit by the model clearly overestimates the amount of dim galaxies in low-redshift bins, largely due to cosmic variance. At high redshift too, the number is overestimated slightly. Still, in general, the best-fit distribution reproduces the GLF very accurately, including the bright and dim-end dropoffs due to the rarity of high-luminosity galaxies and due to the cut-off in apparent magnitude shifting to brighter galaxies at higher redshifts.

The top-right shows the distribution over apparent magnitude, and contains the black line of the total number of galaxies per apparent magnitude, which is $N_{\mathrm{p},\lambda}$, one of the constraints of the model. Again, at low redshifts the model overestimates the number of dim galaxies, where the effect of cosmic variance is largest, and the clustering signal has a large relative uncertainty.

The bottom-left shows the redshift distribution in each bin of apparent magnitude, where we see an excellent reproduction of the data by the model, especially, again, at intermediate redshifts.

In figure 3 from vDW (which was reproduced in figure 2.1), shaded bands for cosmic variance are plotted which have been calculated from a thousand randomly placed surveys from the same simulation, with the same sky area as the fiducial survey area. Thus, it can be seen that the lowest redshift bin contains significantly less objects than average, while the second redshift bin contains significantly more. This is the main reason the model has difficulty to match the low-redshift end of the GLF.

Lastly, the sharp downturn at high redshifts means that the bin is not captured in its fullest by the model, causing the model to overestimate the number of galaxies in that bin. To show the mismatch at low redshift is mostly due to cosmic variance, a fit directly to absolute magnitude and redshifts is done (both of which are unknown to the model), and results are shown to be extremely closely comparable between the two methods.

A.3 Chapter 3

A.3.1 Section 3.1

For completeness, we show the result of the fiducial model on the unaltered mock galaxy catalogue in figure A.1.

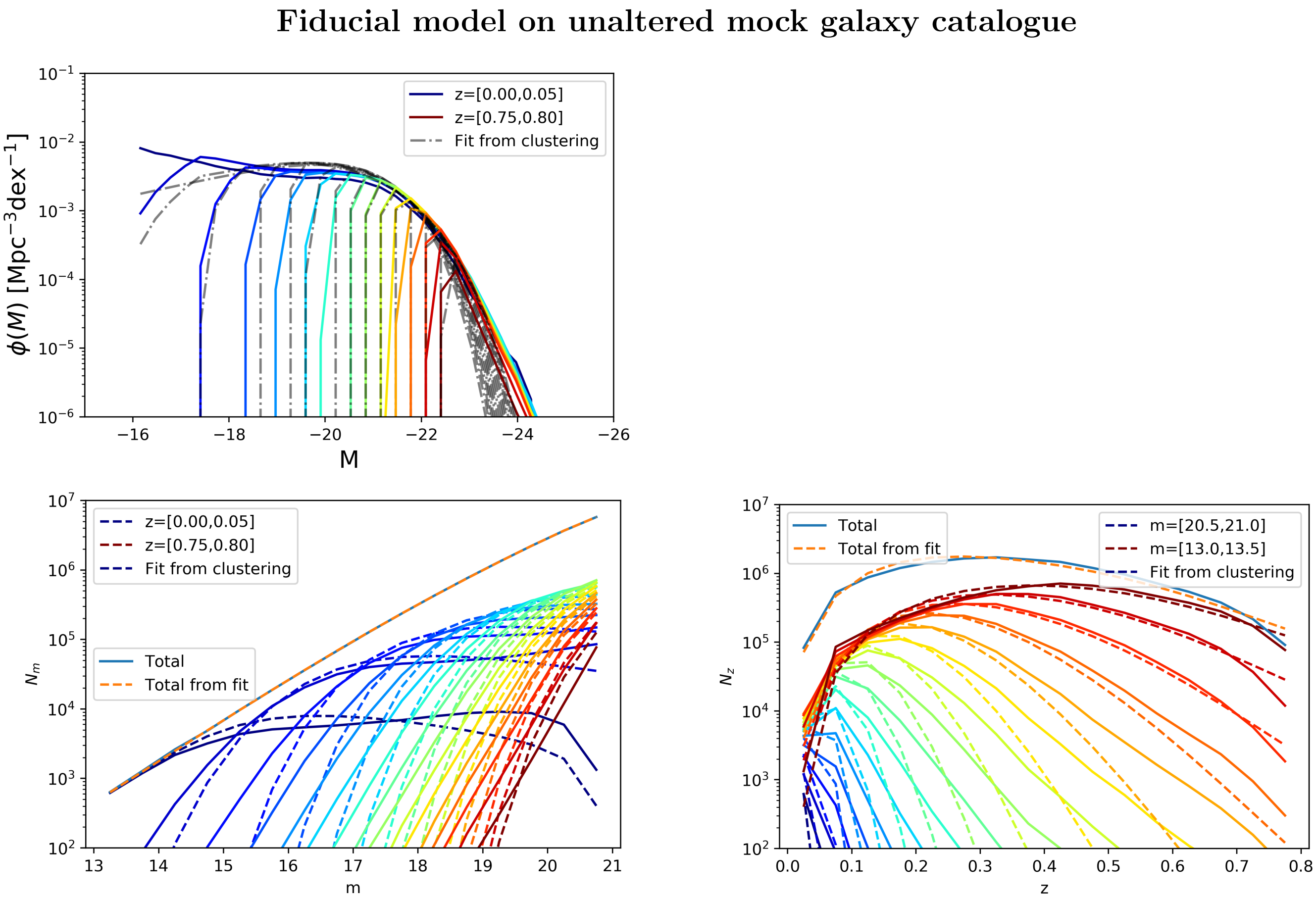

Figure A.1: *As figure 2.1 but now for the unaltered catalogue with the same model. Starting and best-fit parameters are shown in table A.1. Clear underestimation in multiple regimes is seen.*

When comparing these results to those in figures 3.1, 3.3 and 3.5, we already see that in all cases, a double Schechter function manages to make a better fit of the data, confirming our belief that we can better fit this data with a double Schechter model. The starting and best-fit parameters are presented in table A.1.

Input	**-0.57**	**-20.97**	**-0.72**	**-1.18**
Fiducial on unaltered catalogue	-0.572	-20.790	-1.037	-1.309

Table A.1: *Input and best-fit parameter for the fiducial model on unaltered data. Input was the same as for the double Schechter model in table 3.1.*

A.3.2 Section 3.3.3

Here we show the N_m and N_z counterparts to the models shown in figure 3.7 as was done in for example figure 3.2 too.

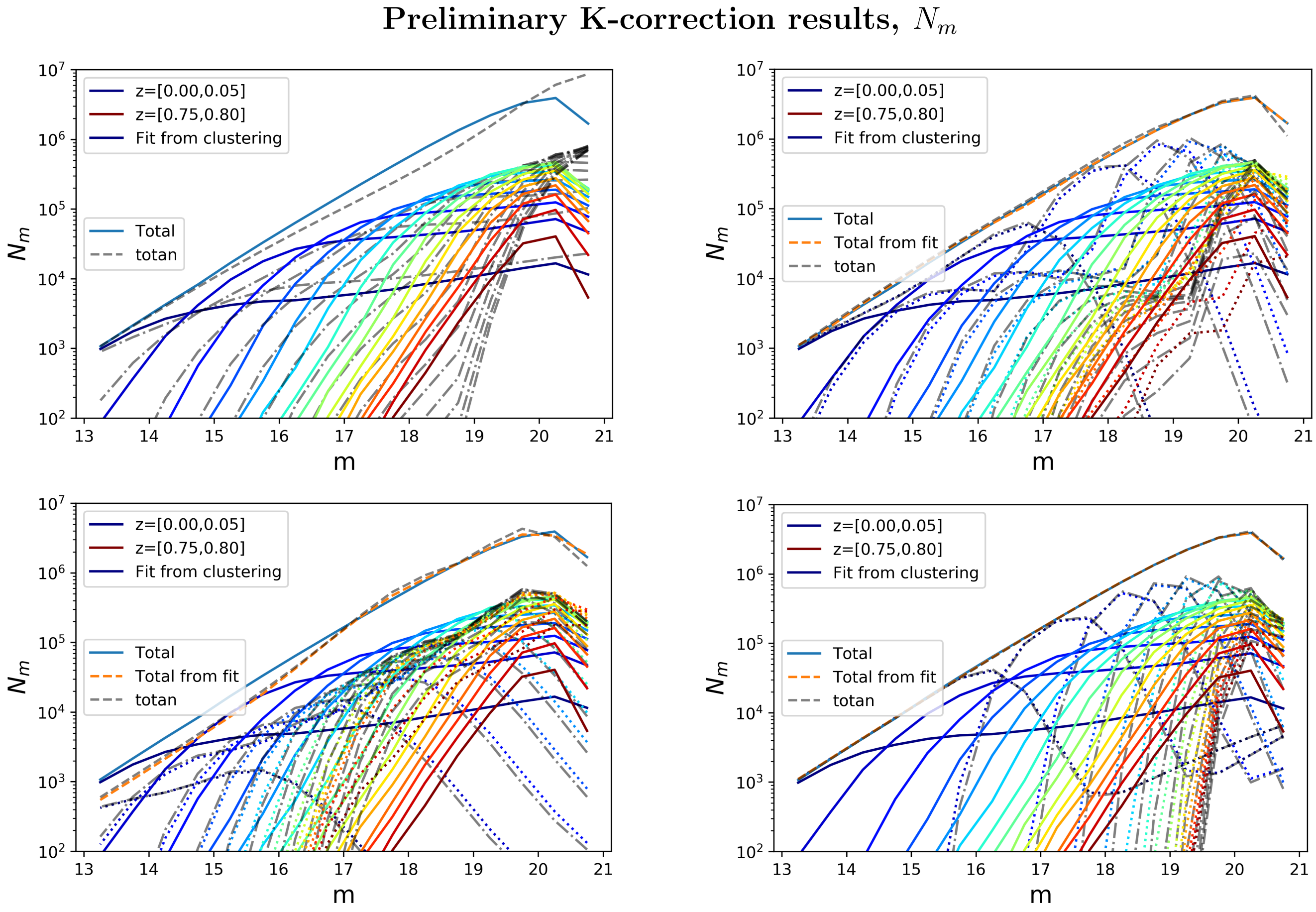

Figure A.2: *As figure 3.7 but now N_m is shown. Dotted coloured lines show the actual fit from the model, while the dot-dashed lines show the analytical fit given the best-fit parameters. Note that, while the constraint in the top left figure is not fit well, it is fitted perfectly by both the top and bottom right models, and is underestimated by the bottom left model. The binned data is however misfitted completely in the results.*

We notice here that, while the starting point is not the best, it does use both Schechters to build up the contributions to both the total constraint and the binned data.

In the top right panel, we see this structure broken down, and the two Schechters show two clear maxima, using mostly the second maximum to fit the constraint. This structure is also seen in the

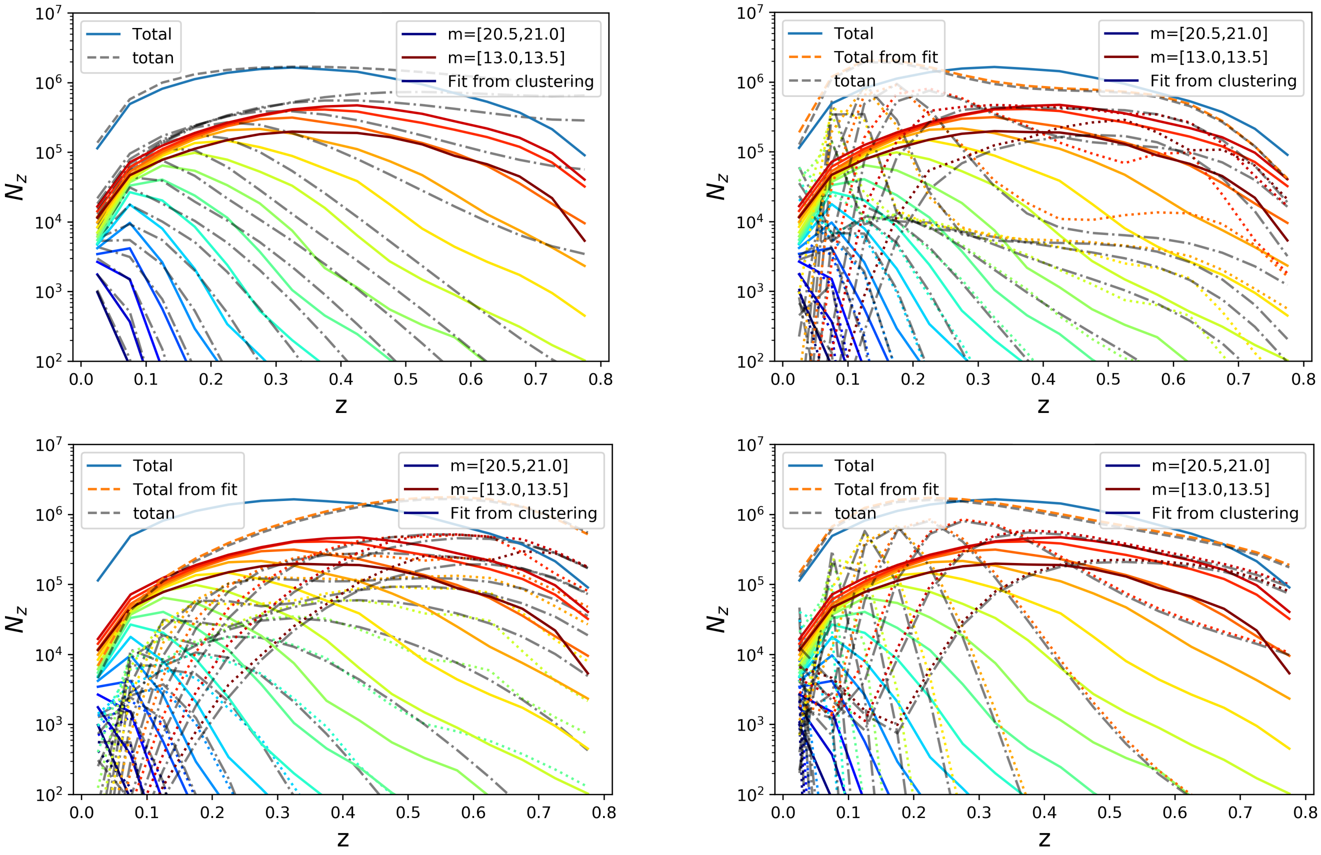

Figure A.3: *As figure A.2 but now N_z is shown. Note the apparent overestimation for higher redshift, high luminosity galaxies in the top left panel. The original model (bottom right) compensates the underestimation of low-luminosity galaxies with overestimation of high-luminosity, high-redshift galaxies. This is also apparent in the top right. The bottom left panel again is closer the the overall shape of the data, but still misfits horribly.*

bottom right, where it seems like one Schechter function is actually almost made redundant by the model.

The bottom left panel of course looks different since the shown total is not used as a constraint but the data is additionally binned on colour here. The result seems to follow the shape of the data better, but does not match the data at all.

As already seen in the original runs (for example figure 3.3), the high-redshift, high-luminosity galaxies are overestimated by the fit made by trial-and-error, meaning this is possibly unavoidable when assuming the structure of the two Schechter functions we use. This starting point then seems to retain its bad placement during the fitting of the model, as is apparent in all other figures. For the bottom left panel, the overall shape comes closest to what we want to have, but it is still clearly wrong.

For further discussion, see section 3.3.3 and on.