
Photometric redshift estimation for precision cosmology

Markus Michael Rau



München 2017

Photometric redshift estimation for precision cosmology

Markus Michael Rau

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Markus Michael Rau
aus Sankt Wendel

München, den 07.09.2017

Erstgutachter: Prof. Dr. Ralf Bender
Zweitgutachter: Prof. Dr. Jochen Weller
Tag der mündlichen Prüfung: 04.10.2017

Contents

Zusammenfassung	vii
Abstract	ix
1 Introduction	1
1.1 Distance and Redshift	4
1.2 Expansion of the Universe	7
1.2.1 A catalog of the Universe	9
1.2.2 The Formation of Structure	14
1.2.3 The Power Spectrum	17
1.3 Cosmological Probes in Photometric Surveys	19
1.3.1 Angular Clustering	19
1.3.2 Gravitational Lensing	22
1.3.3 Weak Lensing by Large-scale structure	24
1.4 Photometric redshift estimation	30
1.4.1 Template Fitting	33
1.4.2 Machine Learning methods	34
2 Machine Learning	37
2.1 Basic probability theory	38
2.2 Regression	44
2.3 Classification	46
2.4 Model Complexity	48
2.4.1 The Bias-Variance Tradeoff	50
2.4.2 Evaluating Machine Learning Models	51
2.4.3 Bayesian model comparison	52
2.5 Decision trees and the Random Forest	53
2.5.1 Regression	54
2.5.2 Classification	55
2.5.3 The Random Forest	55
2.6 Sources of error in Machine Learning	56
2.6.1 Sample Selection Bias	56
2.6.2 Input noise	58
2.7 Density Estimation	59

2.7.1	Conditional Density estimation	61
3	Accurate photometric redshift PDF estimation	63
4	Fixing biases from estimating z-distributions	83
5	Accurate photometric redshift validation	101
5.1	Introduction	102
5.2	The Cosmos Validation Data	106
5.3	Variations in photometric noise	110
5.4	Spectroscopic Incompleteness	112
5.4.1	Analysis and Results	115
5.5	Compressing photometric redshift distributions	119
5.5.1	Compression Algorithms	120
5.5.2	Analysis and Results	121
5.6	Summary	122
6	Conclusions and Future Work	125
7	Additional Projects	129
7.1	Dark Energy Survey Year 1 Results	129
7.2	Anomaly detection for photo-z	132
7.3	Utility Optimization in Regression Tasks	135
8	Appendix	139
8.1	The CFHTLenS survey	139
8.2	The Dark Energy Survey	143
	Acknowledgements	167

List of Figures

1.1	Artists illustration of the evolution of the universe and the formation of structure	2
1.2	Overview over cosmologies in Ω_m - Ω_Λ plane	13
1.3	Galaxy distribution from the 2dF Galaxy Redshift Survey	15
1.4	Present day matter power spectrum	18
1.5	Geometry of Weak Gravitational Lensing	23
1.6	Weak Lensing by Large-scale structure	25
1.7	Effect of Weak Gravitational Lensing on the shape of a circular source	27
1.8	Illustration of the different shear components	28
1.9	Filter response functions of the DES survey plotted together with a SED template of a red galaxy	32
2.1	The normal distribution	42
2.2	Illustration of covariance ellipses	43
2.3	Illustration of a typical regression setting	45
2.4	Illustration of the impact of model complexity on the accuracy of Machine Learning models	49
2.5	Illustration of sample selection bias	57
2.6	Oversmoothing: effect of selecting a too large bandwidth on the accuracy of a density estimate	59
2.7	Undersmoothing: effect of selecting a too small bandwidth on the accuracy of a density estimate	60
5.1	Schematic illustration of the different sources of error in photometric redshift validation investigated in this work	103
5.2	Redshift distributions of the COSMOS multiband, narrow filter photometric sample and the matched spectroscopic sample	107
5.3	Magnitude distributions of the COSMOS multiband, narrow filter photometric sample and the matched spectroscopic sample	108
5.4	Distribution of i band flux error for the full DES science sample, the original DES photometry in the COSMOS field and the resampled DES photometry in the COSMOS field	109
5.5	Distribution of i band magnitude for the full DES science sample, the original DES photometry in the COSMOS field and the resampled DES photometry in the COSMOS field	110

5.6	Photometric redshift distributions obtained with the BPZ template fitting code on the resampled and degraded photometry, the corresponding COSMOS high-precision photometric redshifts and the spectroscopic redshift distributions . . .	111
5.7	The influence of photometric noise on photo- z predictions	112
5.8	The influence of photometric noise on the bias in the lensing convergence power spectrum	113
5.9	Weighted magnitude distribution of the COSMOS photo- z sample	115
5.10	Average relative bias in the convergence power spectrum produced by the weighted COSMOS high-precision photometric redshifts as a function of normalized effective sample size and the number of nearest neighbors selected during weighting	117
5.11	Scatter plot showing spectroscopic incompleteness of the full DES photometric color space in the COSMOS field	118
5.12	Photometric redshift error due to the inaccurate compression of photometric redshift distributions	119
7.1	Redshift distributions of the DES Y1 tomographic photometric redshift distributions	133
7.2	Photometric redshift error of the mean and median regression method as a function of the contamination rate hyper-parameter n_c	135
8.1	Constraint on the power spectrum amplitude σ_8 and Ω_m for a flat Λ -CDM model, obtained from the tomographic cosmic shear analysis presented in Heymans et al. (2013)	140
8.2	Filter bands of the CFHTLenS survey	141
8.3	Illustration of PSF anisotropy for the case of the m1m2 pointing in the W1 field	142
8.4	Footprint of the DES Y1A1 sample	144
8.5	Cosmological parameter constraints of the DES Y1 combined analyses of weak gravitational lensing and galaxy clustering	145
8.6	Filter functions of the Dark Energy Survey	146
8.7	PSF ellipticity in the DES Y1 data	147

List of Tables

7.1	Photometric redshift bias parameters Δz for the tomographic photometric redshift distributions estimated using BPZ. Adopted from Hoyle et al. (2017). . . .	130
8.1	Basic properties of the CFHTLens science data. Taken from Erben et al. (2013).	140
8.2	Properties of the Y1A1 GOLD dataset	144

Zusammenfassung

Die beobachtende Kosmologie trat in den letzten Jahren in eine Ära hoher statistischer Präzision ein. Beobachtungsprogramme wie das Dark Energy Survey erlauben die Messung der Position und Form von Galaxien in bislang unerreichter Genauigkeit und Tiefe. Die Eigenschaften des Dichtefelds der dunklen Materie und dessen zeitliche Entwicklung können anhand dieser Daten vorhergesagt werden, was in den nächsten Jahren unser Verständnis von dunkler Energie, dunkler Materie und kosmischer Strukturbildung stark verbessern wird. Genaue Messungen der Distanzen dieser Galaxien basierend auf ihrer Photometrie, die sogenannten photometrischen Rotverschiebungen, stellen eine der größten Herausforderungen für diese Programme dar. Deren hohe statistische Präzision benötigt daher eine genaue Kontrolle der systematischen Fehlerquellen in der Bestimmung dieser photometrischen Rotverschiebungen. Das Ziel dieser Arbeit ist daher die Entwicklung von Algorithmen und Methoden, welche die Qualität dieser photometrischen Rotverschiebungen verbessern, ihre Fehler genau quantifizieren und diese Systematiken in kosmologische Analysen miteinbeziehen.

Ich schlage einen neuen Algorithmus vor, der mithilfe der Methoden des Maschinellen Lernens die Modellierung von photometrischen Rotverschiebungen und deren Fehlern deutlich verbessert. In §3 demonstriere ich anhand öffentlicher Beobachtungsdaten aus dem CFHTLenS Programm, dass meine Methoden systematische Fehler in einer Reihe von wichtigen kosmologischen Messungen, verglichen mit etablierten Methoden, reduzieren können. Dazu gehören die Bestimmung der Masse von Galaxienhaufen mittels des Gravitationslinseneffektes, sowie die Modellierung von Winkel- und kosmischer Scherkorrelationsfunktionen. Ich schlage auch effiziente und genaue Kompressionstechniken vor, die den Speicherbedarf von photometrischen Beobachtungsprogrammen erheblich reduzieren können.

Um photometrische Rotverschiebungen präzise kalibrieren zu können, werden sehr genaue Referenzrotverschiebungen benötigt, die üblicherweise durch spektroskopische Beobachtungen bereitgestellt werden. Diese Messungen setzen vor allem bei lichtschwachen Galaxien, lange Belichtungszeiten voraus und sind daher kostenintensiv. Infolgedessen gibt es typischerweise wenig Spektren für lichtschwache Galaxienstichproben. Ich werde in §4 aufzeigen, dass dieser Mangel an Referenzdaten, in Kombination mit ungenau bestimmten photometrischen Rotverschiebungsverteilungen, die Genauigkeit der Messung von kosmologischen Parametern stark beeinträchtigen kann. Als Lösung schlage ich daher eine neuartige Methode vor, die es ermöglicht, genaue kosmologische Parameterschätzungen unter Verwendung einer kleinen Zahl wie etwa 5,000 repräsentativer spektroskopischer Referenzgalaxien vorzunehmen.

In §5 diskutiere ich Fehlerquellen, welche die Validierung von photometrischen Rotverschiebungen im Rahmen des Dark Energy Surveys beeinflussen können. Ich zeige, dass räumliche Variationen der Photometriegenauigkeit zu erheblichen Ungenauigkeiten in der Bestimmung von Rotverschiebungsfehlern führen können. Die bereits erwähnte Unvollständigkeit von spektroskopischen Referenzdaten kann die Validierung von photometrischen Rotverschiebungen zusätzlich erschweren. Ich untersuche daher, welche Galaxienrotverschiebungen sich nicht durch spektroskopische Daten kalibrieren lassen, und quantifiziere den erwarteten Fehler im Leistungsspektrum der Gravitationslinsenkonvergenz des projizierten Dichtefeldes. Darüber hinaus evaluiere ich die Genauigkeit von Kompressionstechniken für photometrische Rotverschiebungsverteilungen im Rahmen des Dark Energy Surveys.

Abstract

Observational cosmology has entered an era of high statistical precision in recent years. In particular, ongoing and future photometric surveys like the Dark Energy Survey observe the position and shape of galaxies to an unprecedented precision and depth. These measurements can then be used to estimate the properties of the dark matter density field and its evolution over time. The images of hundreds of millions of galaxies obtained by these programs are therefore expected to shed light on the greatest enigmas of modern cosmology and to refine our understanding of dark energy, dark matter and the growth of structure. One of the greatest challenges for these surveys are accurate measurements of distance, or redshift, from the galaxy photometry. As their high statistical power sets demanding requirements on the control of systematic errors, the development of methods and algorithms to control photometric redshift uncertainty is a vital prerequisite to enter the era of precision cosmology. The goal of this thesis is the development of algorithms and methods that improve the quality of photometric redshift estimates, accurately quantify their error and incorporate these systematics into cosmological analyses.

I propose a novel photometric redshift algorithm based on Machine Learning, that significantly improves the modelling of photometric redshift distributions. Using public imaging data from the CFHTLenS survey, I demonstrate in §3 that my approach reduces systematic biases in gravitational lensing cluster mass estimates, modeling of angular correlation functions, and modeling of cosmic shear correlation functions compared with results obtained using contemporary algorithms. I further propose efficient and accurate compression techniques, that can significantly reduce the storage requirements of large area photometric surveys.

Accurate photometric redshift estimation requires calibration samples of high-precision redshifts, that are usually provided by spectroscopic surveys. Taking spectra of faint galaxies requires long exposure times and is therefore costly. As a result, spectroscopic calibration data is typically not abundant at the faint end of the color-magnitude space. The lack of calibration data, in combination with inaccuracies in the estimated redshift distributions, can severely bias cosmological parameter constraints as shown in §4. I propose a novel strategy to correct these errors, which enables us to obtain unbiased cosmological parameter estimates using a small number of, e.g. 5,000, representative spectroscopic calibration galaxies per redshift distribution.

In §5, I analyse sources of error, that can bias the validation of photometric redshifts in the context of the Dark Energy Survey. I demonstrate that field-to-field variations in photometric noise can lead to a severe misestimation of photometric redshift error. The aforementioned incompleteness of spectroscopic samples of faint galaxies, can make photometric redshift validation difficult. I therefore investigate which galaxies cannot be well calibrated by spectroscopic data and quantify the expected error in terms of biases in the lensing convergence power spectrum. Furthermore I quantify the accuracy of compression techniques for photometric redshift distributions in the context of the Dark Energy Survey.

Chapter 1

Introduction

Cosmology is one of the most fundamental, exciting and ambitious branches of modern physics, having the goal to understand no less than the whole Universe and all its contents on all scales of length, temperature and time. Connecting the world at the largest scales with the properties of elementary particles is not only the final frontier of fundamental physics, but also perhaps one of the only opportunities to test its limitations.

In recent years, modern cosmology entered a golden age of progress, where the rapid advancements in observational techniques and instruments culminated in the development of the cosmological standard model (e.g. Fließbach, 1990; Dodelson, 2003; Carroll & Ostlie, 2007). This theory combines the concept of an isotropic and homogeneous universe with the ‘Big Bang picture’ (Fig. 1.1) that postulates that the Universe developed from a spacetime singularity, the ‘Big Bang’, where all energy and matter components were compressed into a singular point. The physical processes at the earliest moments until $t \approx 10^{-35}\text{s}$ after the Big Bang are still largely beyond our current knowledge of physics (see Fließbach, 1990, §54). It can however be speculated that the Universe underwent a phase of rapid expansion, the inflation era, that disconnected neighboring quantum fluctuations and put the Universe into a homogeneous and isotropic state (see e.g. Schneider, 2006a, §4.5). The continued expansion of the Universe rapidly decreased its temperature until we reach energy scales that are well within our current understanding of physics. About an hour after the Big Bang, nuclear fusion sets in, which led to the formation of ionized light elements like hydrogen, helium and a small amount of lithium (see Fließbach, 1990, §54). The Universe was then filled with free electrons and light ions, that constantly scattered with the hot photon gas, establishing a thermodynamic equilibrium between matter and radiation (see Fließbach, 1990, §54). While the universe cooled down, the electrons and ions assembled together to form neutral atoms. The thermodynamic equilibrium between matter and radiation was interrupted and the photons propagated freely through the now optically transparent Universe (see Fließbach, 1990, §54). The highly isotropic Cosmic Microwave Background (CMB) radiation, that is a relict of this era of recombination, can be observed today and is the most important currently observable signature of the young Universe (see e.g. Hu & Dodelson, 2002).

After this recombination era, our Universe was filled with cool neutral hydrogen and, except for the CMB, mostly devoid of radiation (Loeb, 2006). However even in these proverbial ‘dark ages’, the small amount of kinetic energy of the atoms in conjunction with some free electrons and the CMB radiation was enough to trigger the hyperfine structure transition of hydrogen.

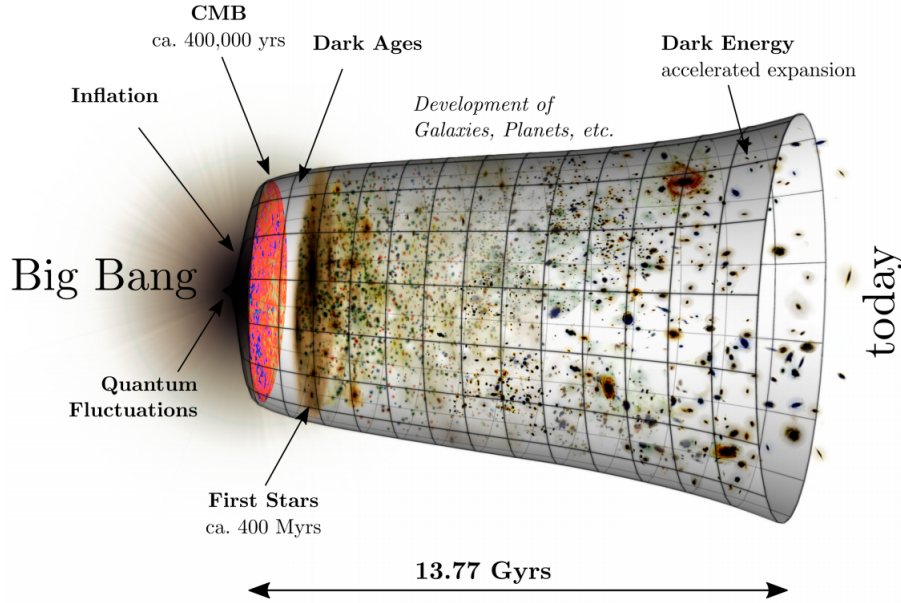


Figure 1.1: Artists illustration of the evolution of the universe and the formation of structure. Image credit: NASA/WMAP Science team <http://science.nasa.gov/missions/wmap/>, modified by Natascha Greisel and Anna Monna

This 21 cm line gives modern astrophysics a unique opportunity to study this very onset of structure formation (Loeb, 2006). About 100 Myrs after the Big Bang, the massive hydrogen clouds began to collapse under their own gravity to form the first generation of stars. Their light filled the universe with UV radiation and ionized the surrounding medium (Loeb, 2006).

The first generation of stars that formed during this era of ‘reionization’, where extremely massive and short lived (Schneider, 2006a, §9.4.1). They ended their lives in massive supernovae explosions that enriched the Universe with heavy elements, ultimately laying the foundation for the emergence of life (Schneider, 2006a, §9.4.1). Following the onset of structure formation, stars merged into galaxies, which then assembled to form galaxy clusters (see Carroll & Ostlie, 2007, §30.2). As the Universe continued to expand, its matter density decreased and dark energy became the dominant constituent (see e.g. Carroll & Ostlie, 2007, §29.3, Fig. 29.19). It is now widely accepted that the Universe is currently undergoing a state of accelerated expansion, driven by this dark energy content (e.g. Riess et al., 1998; Perlmutter et al., 1999; Frieman et al., 2008; Planck Collaboration et al., 2016). The exact nature of dark energy is currently one of the most active research areas in modern cosmology (Frieman et al., 2008). At the present time, the universe contains about $\approx 69.1\%$ dark energy, $\approx 25.9\%$ dark matter and only $\approx 4.9\%$ baryonic matter (taken from Planck Collaboration et al., 2016, Tab. 4, last column).

The groundbreaking discoveries that led to this picture of the Universe and its evolution, mark not only the biggest success of modern cosmology but also demanded measurements of drastically increased precision (Albrecht et al., 2006). Especially because the exact nature of dark energy and dark matter are largely unknown.

With the established goal to better understand the proverbial nature of this dark com-

ponent, large area sky surveys like the Sloan Digital Sky Survey (SDSS) (e.g. Blanton et al., 2017) have revolutionized the methodological aspects of cosmological research. These experiments are able to gather imaging information for hundreds of millions of galaxies and provide a chance to study the formation of structure to an unprecedented statistical precision. However the strong increase in statistical power gained by these surveys, requires the accurate control of systematic errors. One of the most pressing challenges in these surveys is to obtain accurate distance information for the observed galaxies. Inaccurate distance measurements distort the line-of-sight distribution of the cosmological density field in the modeling of cosmological probes. These errors can then propagate into systematic shifts in the derived quantities that parametrize the state and evolution of the Universe, like the fraction of dark matter or dark energy. To improve our current understanding of the Universe, it is therefore vital to control these sources of systematic error to high accuracy. The distance to a galaxy is commonly quantified by its cosmological redshift. When a light wave propagates through the expanding background Universe, its wavelength is ‘stretched’ and its spectrum therefore shifted to longer, redder wavelengths. This ‘redshift’ measures how long the light wave travelled through the expanding universe and therefore quantifies the distance to that galaxy. The traditional method to measure the redshift of galaxies compares the relative position of spectral emission and absorption lines to derive the global redshift of the spectrum. The instruments used in state-of-the-art spectroscopic surveys have a high spectral resolution and are able to resolve wavelength differences of up to $\Delta\lambda \sim 1\text{\AA} - 3\text{\AA}$ (e.g. Smee et al., 2013). Spectroscopic redshifts are therefore typically highly accurate, but require long exposure times especially for faint galaxies. Taking large samples of spectra is therefore not tractable for surveys, that observe faint galaxies in large areas of the sky. These large area photometric surveys instead image galaxies in different filters that typically cover the spectrum from the ultraviolet to the infrared, where the filter response functions have a width of $\Delta\lambda \sim 1000\text{\AA}$ (see Fig. 1.9). The redshift is then estimated using the flux of the galaxy in these optical and near infrared filters. These ‘photometric redshifts’ are generally less accurate than their spectroscopic counterparts due to the broad photometric filter kernel. On the other hand, this strategy is much more efficient at observing large numbers of galaxies, since light can be collected efficiently over a large field of view. In order to use these less accurate photometric redshift measurements for precision cosmology, it is therefore of vital importance to devise techniques that accurately quantify their error and incorporate them into the analysis.

This thesis contributes to this topic as follows. In §3 we propose techniques to accurately model the photometric redshift error and demonstrate that our methodology significantly improves over contemporary approaches. While these methods allow the accurate estimation of photometric redshifts, their statistical accuracy is limited by the amount of available calibration data. In §4 we therefore propose a method to incorporate this statistical error into the cosmological analysis. §5 then discusses possible sources of error in the context of the photometric redshift validation of the Dark Energy survey and §6 summarizes our findings and discusses prospects for future research.

The following sections provide the reader with a short introduction into the concepts of cosmology that are used in this work. To simplify the notation, we will use natural units $\hbar = c = k_B = 1$ in the presented equations and quote numerical values in SI, cgs or astrophysical units as appropriate.

1.1 Distance and Redshift

One of the foundations of modern cosmology, known as the ‘cosmological principle’, is the assumption that the matter distribution in the universe is isotropic and homogeneous at sufficiently large scales. An impressive experimental proof of isotropy is the Cosmic Microwave background radiation, a relic of the era of recombination, that shows remarkably small spatial temperature fluctuations of $\delta T/T \sim 10^{-5}$ (Banday et al., 1997; Rubakov & Vlasov, 2012). Isotropy implies homogeneity, if we further assume that we do not live at a special location in the Universe (Copernican Principle). While a direct proof of homogeneity is not possible, we can search for its violation using a variety of techniques reviewed in Maartens (2011). A metric that obeys this cosmological principle is the Friedmann-Lemaître-Robertson-Walker (FLRW) metric (cf. Bartelmann & Schneider, 2001; Kilbinger, 2015; Sánchez, 2017)

$$ds^2 = -dt^2 + a^2(t) (d\chi^2 + f_K(\chi)^2 [d\theta^2 + \sin^2(\theta)d\phi^2]) , \quad (1.1)$$

where

$$f_K(\chi) = \begin{cases} \frac{1}{\sqrt{K}} \sin(\sqrt{K} \chi) & \text{if } K > 0 \\ \chi & \text{if } K = 0 \\ \frac{1}{\sqrt{|K|}} \sinh(\sqrt{|K|} \chi) & \text{if } K < 0 \end{cases} . \quad (1.2)$$

Here χ is the radial coordinate and θ and ϕ denote the respective angular coordinates. The function $a(t)$ is the cosmological scale factor and describes the time dependent spatial expansion of the universe. The FLRW metric either describes a closed universe with positive spatial curvature $K > 0$, a flat universe $K = 0$ or an open universe $K < 0$ with negative spatial curvature. To draw an analogy with our three dimensional view, a closed universe would correspond to a sphere, a flat universe would correspond to a plane and an open universe would correspond to a hyperboloid (see e.g. Carroll & Ostlie, 2007, Fig. 29.17). The coordinate χ is fixed with respect to the expansion of the universe and is therefore called ‘comoving coordinate’. As a result, the comoving distance between two points that move with the expansion of the universe is also constant (Hogg, 1999). This notion of distance has to be distinguished from the distance one would measure with a ruler, the ‘proper distance’¹ d_p , which is related to the comoving distance as (cf. Hogg, 1999; Dodelson, 2003; Carroll & Ostlie, 2007)

$$d_p = a(t) \chi , \quad (1.3)$$

and therefore increases with the cosmic expansion.

Following Carroll & Ostlie (2007, §29.4), we now consider a lightwave travelling from a distant galaxy to us. Two successive wave crests separated by the light waves period Δt need to traverse the same distance until we receive the signal, where the line element of the FLRW metric can be written as

$$\chi = \int_{t_e}^{t_o} \frac{dt}{a(t)} , \quad (1.4)$$

where t_o denotes the time at which the light wave is observed and t_e denotes the time at which the light wave is emitted. Defining the period at the time of observation as Δt_o and at

¹I follow the notation in Dodelson (2003); Carroll & Ostlie (2007).

the time of emission as Δt_e , we note that both wave crests need to travel the same comoving distance, and therefore

$$\int_{t_e + \Delta t_e}^{t_o + \Delta t_o} \frac{dt}{a(t)} - \int_{t_e}^{t_o} \frac{dt}{a(t)} = 0. \quad (1.5)$$

After expanding the integral limits, we can rewrite this equation as

$$\int_{t_o}^{t_o + \Delta t_o} \frac{dt}{a(t)} - \int_{t_e}^{t_e + \Delta t_e} \frac{dt}{a(t)} = 0. \quad (1.6)$$

Assuming that the scale factor $a(t)$ doesn't change significantly during the period of the light wave, we can approximate these integrals as (cf. Fließbach, 1990)

$$\frac{\Delta t_o}{a(t_o)} = \frac{\Delta t_e}{a(t_e)}. \quad (1.7)$$

The cosmological redshift z_{cosmo} is then defined as the relative wavelength difference between the wavelength at the position of the emitter λ_e and the observer λ_o (cf. Fließbach, 1990)

$$z_{\text{cosmo}} = \frac{\lambda_o - \lambda_e}{\lambda_e}. \quad (1.8)$$

Using Eq. 1.7 we can express z_{cosmo} in terms of the cosmological scale factor $a(t)$ as (cf. Fließbach, 1990)

$$z_{\text{cosmo}} = \frac{a(t_o)}{a(t_e)} - 1, \quad (1.9)$$

or, by setting the scale factor at the present time to unity, $a(t_o) = 1$

$$a(z) = \frac{1}{z + 1}. \quad (1.10)$$

We note that the cosmological redshift defined in Eq. 1.9 originates purely from the expansion of the universe and has to be distinguished from local frequency shifts produced by the peculiar velocity that originates from gravitational interactions of galaxies with their neighbors. These motions relative to a fixed comoving position can cause an additional redshift by the Doppler effect. We also note that strong gravitational fields at the position of the source or the observer can also shift the spectrum. This work will however exclusively focus on the cosmological redshift and, in the following uses the term 'redshift' as its synonym.

Expanding the scale factor around the present time, we obtain a simple relationship between the redshift of a galaxy and its proper distance (cf. Fließbach, 1990)

$$z \approx H_0 d_p + \frac{(1 + q_0) H_0^2}{2} d_p^2, \quad (1.11)$$

where we introduced the 'Hubble constant' as (cf. Fließbach, 1990)

$$H_0 = \frac{\dot{a}(t_0)}{a(t_0)}, \quad (1.12)$$

and the deceleration parameter q_0 (cf. Fließbach, 1990)

$$q_0 = -\frac{\ddot{a}(t_0) a(t_0)}{\dot{a}(t_0)^2}. \quad (1.13)$$

The linear term in Eq. 1.11 is commonly known as the ‘Hubble law’ and was already established by Edwin Hubble in the 1920’s (Hubble, 1929). The Hubble constant² is proportional to the expansion speed and has been measured to high precision, e.g. using supernovae and Cepheid distances (e.g. Zhang et al., 2017), as

$$H_0 = 72.5 \pm 3.1 \text{ (stat)} \pm 0.77 \text{ (sys)} \frac{\text{km}}{\text{s Mpc}}, \quad (1.14)$$

where the errors quote the statistical and systematic uncertainties. The second term of Eq. 1.11 introduces corrections at higher redshift $z > 0.1$ and accounts for changes in the expansion rate. Measurements hint towards a negative value of q_0 , indicating that the expansion of the Universe is indeed accelerating (e.g. Riess et al., 1998; Perlmutter et al., 1999; Cattoën & Visser, 2008; Cai & Tuo, 2011).

The relation Eq. 1.11 implies that the basic strategy to constrain cosmic acceleration requires the measurement of the redshifts of sources and their proper distances.

In analogy to elementary trigonometry we define the distance d_A to an object of known physical size s as (cf. Dodelson, 2003)

$$d_A = \frac{s}{\tan(\theta)} \approx \frac{s}{\theta}, \quad (1.15)$$

where θ denotes the subtended angle. An object with known physical size is called a ‘standard ruler’ and its measured distance d_A is the ‘angular diameter distance’. The comoving size of the object can be derived as s/a and the angle θ can therefore be written in terms of the line-of-sight comoving distance χ (cf. Dodelson, 2003)

$$\theta = \frac{s/a}{\chi}. \quad (1.16)$$

Plugging Eq. 1.16 into Eq. 1.15 then relates the angular diameter distance d_A with the redshift and the line-of sight comoving distance χ (cf. Dodelson, 2003)

$$d_A = \frac{\chi}{1+z}. \quad (1.17)$$

We note that this relation is only valid for a flat universe $K = 0$. In the case of an open or closed universe we have to correct the comoving distance by the curvature of space as (see Hogg, 1999)

$$d_A = \frac{\chi_M}{1+z}, \quad (1.18)$$

where (see Hogg, 1999)

$$\chi_M = \begin{cases} \sqrt{|k|/H_0^3} \sin\left(\sqrt{|k|} H_0 \chi\right) & K > 0, \\ \chi & K = 0, \\ \sqrt{|k|/H_0^3} \sinh\left(\sqrt{|k|} H_0 \chi\right) & K < 0, \end{cases} \quad (1.19)$$

Instead of considering standard rulers, we can also determine the distance d_L to a source with known luminosity L , so called ‘standard candles’, by measuring the incoming flux f as (cf. Dodelson, 2003)

$$f = \frac{L}{4\pi d_L^2}. \quad (1.20)$$

²For historical reasons, it is sometimes expressed as a fraction of $100 \frac{\text{km}}{\text{s Mpc}}$, i.e. $H_0 = 100 h \frac{\text{km}}{\text{s Mpc}}$.

While this relation is defined for a static coordinate system, we can generalize it to an expanding universe as follows. Consider a comoving coordinate system centered on the source. The light emitted by that source would propagate outwards and we would observe it on a comoving shell with radius χ . In this coordinate system the observed flux then reads (cf. Dodelson, 2003)

$$f = \frac{L(\chi)}{4\pi\chi^2(a)}. \quad (1.21)$$

As the universe expands, the number of photons that cross our shell per unit time decreases proportional to a , since the physical distances that the photons need to travel increase over time. Similarly, the energy of the photons decreases due to the wavelength increase induced by the expanding background universe. In total the luminosity at shell radius χ scales as $L(\chi) = L a^2$, which can be combined with Eq. 1.21 to (cf. Dodelson, 2003)

$$f = \frac{L a^2}{4\pi\chi^2(a)}. \quad (1.22)$$

Comparing with Eq. 1.21, we see that the classical relation between luminosity and flux can be extended to an expanding universe, if we define the luminosity distance d_L as (cf. Dodelson, 2003)

$$d_L = \frac{\chi(a)}{a}. \quad (1.23)$$

Again this relation is only valid for a flat spacetime and generalizes to an open and closed universe as (cf. Dodelson, 2003)

$$d_L = \frac{\chi_M}{a} \quad (1.24)$$

where χ_M is defined in Eq. 1.19.

1.2 Expansion of the Universe

As discussed in the previous section, the FLRW metric describes a homogeneous and isotropic universe. To derive an equation that governs the dynamics of spacetime, i.e. the time evolution of the scale factor $a(t)$, we need to relate the energy content of the universe with its curvature. In general relativity this relation is given by Einsteins field equations (Einstein, 1915, 1916; Carroll, 1997; Sànchez, 2017)

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \quad (1.25)$$

where we will follow the notation in Sànchez (2017). The left hand side of Eq. 1.25 is the Einstein tensor that characterizes the curvature of spacetime, the right hand side describes the energy-momentum content of the universe with the energy-momentum tensor $T_{\mu\nu}$. The constant of proportionality that relates both sides is the gravitational constant G . We note that Einsteins equations obey energy-momentum conservation as (cf. Carroll, 1997; Sànchez, 2017)

$$\nabla^\mu T_{\mu\nu} = 0. \quad (1.26)$$

The following discussion will assume that the contents of the Universe, i.e. dark matter, baryonic matter, dark energy and radiation behave as perfect fluids, where their equation of state only depends on a single parameter w (cf. Carroll & Ostlie, 2007; Sànchez, 2017)

$$p = w \rho, \quad (1.27)$$

and is therefore completely described by their rest frame energy density ρ and their pressure p . The energy-momentum tensor of a perfect fluid then reads (cf. Dodelson, 2003; Sànchez, 2017)

$$T_{\mu}^{\nu} = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}. \quad (1.28)$$

Using these relations as well as the ansatz of the FLRW metric in Eq. 1.25 allows us to derive the Friedmann equations that describe the time evolution of the scale factor (cf. Sànchez, 2017)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \quad (1.29)$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p). \quad (1.30)$$

Here $a(t)$ is the cosmic scale factor, ρ is the density of the different matter and energy components of the universe and p is their pressure. The curvature parameter K was defined in the FLRW metric Eq. 1.1 and describes the curvature of space. Differentiating Eq. 1.29 and substituting Eq. 1.30 then leads to (cf. Sànchez, 2017)

$$\dot{\rho} + 3(\rho + p)\left(\frac{\dot{a}}{a}\right) = 0 \quad (1.31)$$

which can be seen as the first law of thermodynamics³ with $dQ = 0$ (cf. Sànchez, 2017)

$$dU + p dV = d(\rho a^3) + p d(a^3) = 0. \quad (1.32)$$

For a perfect fluid with an equation of state given by Eq. 1.27, we can solve the conservation equation by the ansatz (cf. Sànchez, 2017; Carroll & Ostlie, 2007)

$$\rho(a) = \frac{\rho_0}{a^{3(1+w)}}, \quad (1.33)$$

where ρ_0 is the present day density of the respective component.

Using this general scaling relation, we can rewrite Eq. 1.29 by summing over all N matter and energy components of the Universe (cf. Sànchez, 2017)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3} \sum_{i=1}^N \frac{\rho_{i,0}}{a^{3(1+w_i)}} - \frac{K}{a^2}, \quad (1.34)$$

where $\rho_{i,0}$ and w_i denote the present day density and equation of state parameter of the respective component i . It is convenient to normalize the density parameters $\rho_{i,0}$ by a reference density, which is the present day ‘critical density’ ρ_c of a flat universe

$$\rho_c = \frac{3H_0^2}{8\pi G}, \quad (1.35)$$

³We refer the interested reader to Carroll & Ostlie (2007, §29) for a very accessible derivation of Eq. 1.30 using the first law of thermodynamics in classical newtonian theory.

that is obtained by setting $K = 0$ in Eq. 1.29 and using Eq. 1.12. Normalizing the different density components $\rho_{i,0}$ by this critical density, conveniently expresses them as density parameters Ω_i

$$\Omega_i = \frac{\rho_{i,0}}{\rho_c} = \left(\frac{8\pi G}{3H_0^2} \right) \rho_{i,0}, \quad (1.36)$$

where we define the spatial curvature density parameter in analogy to Eq. 1.36 as (cf. Sánchez, 2017)

$$\Omega_K = -\frac{K}{H_0^2}. \quad (1.37)$$

With these definitions we can recast Eq. 1.34 into the more convenient form (cf. Sánchez, 2017)

$$\left(\frac{H}{H_0} \right)^2 = \left[\sum_{i=0}^N \Omega_i a^{-(1+3w_i)} + \Omega_K \right], \quad (1.38)$$

where we introduced the time dependent Hubble parameter $H = \dot{a}/a$. The inverse Hubble parameter (cf. Bartelmann & Schneider, 2001; Trodden & Carroll, 2004)

$$d_h \approx H(a)^{-1} \quad (1.39)$$

coincides, for flat universes that contain matter and radiation, to good accuracy with the size of the horizon, i.e. the distance photons can have travelled after the Big Bang (see Bartelmann & Schneider, 2001; Trodden & Carroll, 2004; Carroll & Ostlie, 2007).

1.2.1 A catalog of the Universe

The previous section discussed the dynamics of the expansion of the universe in terms of the respective density and equation of state parameters of its contents, i.e. matter, radiation and dark energy. In order to complement the dynamical picture of the universe, we now briefly discuss this cosmic inventory in the following subsection.

Matter

We start our overview over the different constituents of the universe in a self-centric manner with baryonic matter, the component that constitutes our own existence. Its most fundamental property in the context of astrophysics is its interaction with radiation, which makes the baryon density directly observable. This can be done for example by relating the observed fraction of Helium, Deuterium and Lithium with predictions of nucleosynthesis in the early universe (Walker et al., 1991) or by analyzing absorption features in the spectrum of high redshift quasars produced from hydrogen clouds along the line of sight (e.g. Rauch et al., 1997). The CMB also provides highly accurate measurements of the baryon density (Planck Collaboration et al., 2016). All these approaches provide striking evidence for a low value for the baryon density of around 5% of the critical density.

On the other hand measurements e.g. of the CMB, weak gravitational lensing or the clustering of galaxies (see §1.3) measure the total matter density to be much higher and around 30% of the critical density (e.g. Planck Collaboration et al., 2016; Hildebrandt et al., 2017; Salazar-Albornoz et al., 2017). This obvious mismatch is explained by the existence of a

matter component that interacts gravitationally, but does not interact with radiation⁴. This ‘dark matter’ component therefore constitutes the majority of matter in the Universe, but has not yet been detected in either direct (collider) or indirect searches. The two basic candidate particles for dark matter are hot dark matter and cold dark matter.

Hot dark matter particles, for which massive neutrinos are a likely, existing candidate, have a relativistic velocity dispersion that leads to a top-down structure formation. Hot dark matter does not form clumps, but instead fragments from large conglomerates into smaller structures like galaxy halos (cf. Carroll & Ostlie, 2007). The basic problem of the hot dark matter model is, that top-down structure formation would likely be too slow to explain e.g. observations of the highest redshift galaxies today (see e.g. Primack & Gross, 2001; Carroll & Ostlie, 2007).

An alternative theory that agrees much better with the current observations of structure formation is the slow moving cold dark matter paradigm. While there exist no direct detections of cold dark matter particles, the most likely candidate is the weakly interacting particle (WIMP). WIMPs interact only gravitationally and should have very high rest mass ranging from 10 GeV – TeV (see e.g. Olive et al., 2014). The low velocities of cold dark matter favours ‘bottom up’ structure formation. Smaller clumps merge together to form larger structures, where galaxies form preferentially in high density regions.

Since matter is pressureless ($w = 0$), its density decreases proportional to the volume increase of the Universe. Relating the matter density ρ_m with the cosmic scale factor using Eq. 1.33 we find (e.g. Dodelson, 2003)

$$\frac{\rho_m}{\rho_{\text{cr}}} = \Omega_m a^{-3}. \quad (1.40)$$

Radiation Besides dark matter and a fraction of baryonic matter, the Universe also contains radiation in the form of photons ρ_γ and neutrinos ρ_ν . We can estimate the photon density from the cosmic microwave background radiation using Plancks radiation law as (see e.g. Bartelmann & Schneider, 2001; Sánchez, 2017)

$$\rho_\gamma = \left(\frac{\pi^2}{15}\right) T_{\text{CMB}}^4. \quad (1.41)$$

Here T_{CMB} is the temperature of the cosmic microwave background radiation and ρ_γ is the radiation density of photons. Given a CMB temperature of $T_{\text{CMB}} = 2.725 \text{ K}$ (e.g. Fixsen, 2009) and the value for H_0 given in Eq. 1.14, we obtain for the radiation density parameter Ω_γ using Eq. 1.36, Eq. 1.35 and Eq. 1.41

$$\Omega_\gamma = 4.7 \cdot 10^{-5}. \quad (1.42)$$

Following Bartelmann & Schneider (2001), the density of the neutrino species can be estimated from Ω_γ by considering a time in the early universe where neutrinos and photons were in thermal equilibrium with the baryonic matter. When the temperature of the plasma dropped to 1 MeV, the expansion timescale of the universe exceeded the timescale of the weak interaction and the neutrinos decoupled from the hot baryon-photon plasma. At around 0.5 MeV

⁴The first observations, that suggested the existence of Dark Matter, were studies of cluster dynamics (Zwicky, 1933; Smith, 1936; Zwicky, 1937) and measurements of spiral galaxy rotation curves (Freeman, 1970; Bosma, 1978; Rubin et al., 1980; Bosma, 1981a,b).

positron-electron annihilation sets in, which heated up the medium. Since neutrinos decoupled earlier, they have a lower temperature than the generated photons. This difference can be easily calculated by considering entropy conservation before and after the electron-positron annihilation (see Bartelmann & Schneider, 2001)

$$(S_e + S_\gamma)_{\text{before}} = (S_\gamma)_{\text{after}} . \quad (1.43)$$

Following Bartelmann & Schneider (2001), the entropy per particle species scales with temperature as $S \propto g T^3$. Here, the factor g denotes the particles statistical weight, where bosons contribute $g = 1$ and fermions $g = 7/8$ for each spin state. In total the left hand side has a statistical weight of $g = 2 \cdot 1 + 4 \cdot 7/8 = 11/2$ and the right hand side $g = 2$. For the temperature increase we therefore obtain (cf. Bartelmann & Schneider, 2001; Dodelson, 2003)

$$T_\nu = \left(\frac{4}{11} \right)^{1/3} T_{\text{CMB}} . \quad (1.44)$$

We can now calculate the energy density of the neutrino species in terms of the photon energy density. As for photons, the energy density of the massless neutrinos is proportional to T^4 and since we have three massless neutrinos species we obtain⁵ (cf. Dodelson, 2003)

$$\rho_\nu = 3 \left(\frac{4}{11} \right)^{4/3} \rho_\gamma , \quad (1.45)$$

where we inserted Eq. 1.44 into Eq. 1.41. Using Eq. 1.35 and Eq. 1.42 we can then calculate the density parameter for the neutrino species as

$$\Omega_\nu \approx 3.7 \cdot 10^{-5} . \quad (1.46)$$

In total this then leads to a total critical density of the radiation component of

$$\Omega_{\text{RAD}} = \Omega_\gamma + \Omega_\nu \approx 8.4 \cdot 10^{-5} , \quad (1.47)$$

which is a negligible contribution to the critical density today. The early universe was however dominated by radiation, which implies that the radiation density has decreased rapidly with the expansion of the universe.

Using the equation of state parameter of photons $w_\gamma = 1/3$ (e.g. Dodelson, 2003; Sánchez, 2017) we find from Eq. 1.33

$$\frac{\rho_\gamma}{\rho_{\text{cr}}} = \Omega_\gamma a^{-4} . \quad (1.48)$$

Thus the radiation energy declines proportional to the volume increase of the universe with an additional factor of a^{-1} accounting for the wavelength increase due to its expansion (Carroll & Ostlie, 2007; Sánchez, 2017).

⁵Neutrinos are fermions and therefore follow the fermi-dirac statistic, which introduces a small correction of 7/8 to Eq. 1.45. Furthermore, they haven't fully decoupled during positron-electron annihilation, which slightly increases the prefactor $N_{\text{eff}} = 3.046 > 3$. We have neglected these aspects for simplicity, and refer to the literature (see Eq. (1) Planck Collaboration et al., 2014) for a more thorough treatment.

Dark Energy Since the early observations from Hubble and Slipher provided first hints for an expanding universe in 1929, more precise measurements (e.g. Perlmutter et al., 1999; Planck Collaboration et al., 2016) hint towards an accelerated expansion driven by a proverbial dark energy component. However relatively little is known about the exact properties of dark energy which is currently one of the most important active research areas in observational cosmology. The easiest model introduces an additional energy component to the universe as (see e.g. Einstein, 1917; Zel'Dovich, 1967; Carroll et al., 1992)

$$\Omega_\Lambda = \frac{\Lambda}{2H_0^2}, \quad (1.49)$$

that counteracts the influence of gravity. We note that the effect of introducing Λ to the Einstein equations is indistinguishable from the effect of zero point fluctuations of the vacuum, that curve spacetime even in the absense of matter (Carroll et al., 1992; Fließbach, 1990; Sánchez, 2017). While both effects are physically equivalent, their interpretation is fundamentally different, since vacuum energy would allow us to explain the observed accelerated expansion of the universe based on known quantum field theory (Carroll et al., 1992; Fließbach, 1990).

Since the vacuum ground state has to obey Lorentz invariance, the stress-energy-momentum tensor $\text{diag}(\rho, \mathbf{P})$ of dark energy has to be proportional to the Minkowski metric $\text{diag}(-1, 1, 1, 1)$, which implies $w_\Lambda = -1$ (Carroll et al., 1992), i.e. the dark energy density stays constant with the expansion of the universe. Unfortunately the vacuum energy density derived based on our knowledge of quantum field theory is about 120 orders of magnitude larger than the observational results from cosmology (Carroll et al., 1992). It is one of the great mysteries of cosmology, that the sum of the well motivated vacuum energy density and an unknown dark energy component therefore has to cancel almost exactly in order to explain these observations (Carroll et al., 1992). In the following discussion we will not explicitly use Λ in any equations and simply assume $w_\Lambda = -1$ as our fiducial model. This formalism is physically equivalent to introducing a cosmological constant to the Einstein equations, however provides more flexibility for the discussion of alternative equation of state parametrizations (Sánchez, 2017). We further denote the density parameter of dark energy as Ω_Λ .

The value of the dark energy equation of state parameter w has important consequences for the expansion history of the Universe, where we follow Caldwell et al. (2003) in this paragraph. If we consider a flat or an open universe without dark energy and $0 \leq \Omega_m \leq 1$, the scale factor increases at a smaller rate than the horizon. In this case, while the universe would become cooler and darker due to the continued expansion, more and more galaxies would enter the horizon and would therefore become observable. In contrast, if we consider the case of accelerated expansion for $-1 \leq w \leq -1/3$, the scale factor expands faster than the horizon. Thus, while the universe becomes cooler and darker, we could observe less galaxies in the future. For $w < -1$ the expansion of the universe is not only accelerating, but the scale factor also diverges in a finite time (e.g. Caldwell et al., 2003; Astashenok et al., 2012). The consequences are rather apocalyptic as gravitationally bound systems would be ripped apart as we come closer to this end-time. Ultimately this ‘Big Rip’ will also affect structures bound by the other elementary forces like molecules and even atoms. Constraining the exact value of w is therefore not only a considerable academic feat, but also answers the question about the ultimate fate of the Universe and the end of time.

Besides determining the value of the dark energy equation of state parameter, it is of particular interest to investigate a possible time dependence using the parametrization (Huterer

& Turner, 2001; Weller & Albrecht, 2002)

$$w(a) = w_0 + (1 - a) w_a . \quad (1.50)$$

The Λ -CDM model that describes dark energy as a cosmological constant and dark matter as ‘cold’ particles that trigger bottom-up structure formation, remains the state-of-the-art in modern cosmology that future experiments will seek to challenge. Following Carroll et al. (1992) and Carroll & Ostlie (2007), the following subsection will describe the different cosmologies that are possible under this Λ -CDM paradigm.

The dynamics of the Universe

The dynamics of Λ -CDM cosmologies are governed by the opposing forces of dark energy and dark matter. We established in the previous section, that the dark energy density is constant irrespective of the size of the universe, whereas the matter density decreases proportional to its volume. Thus if the universe reaches a certain size, dark energy will always drive the scale factor towards unlimited expansion.

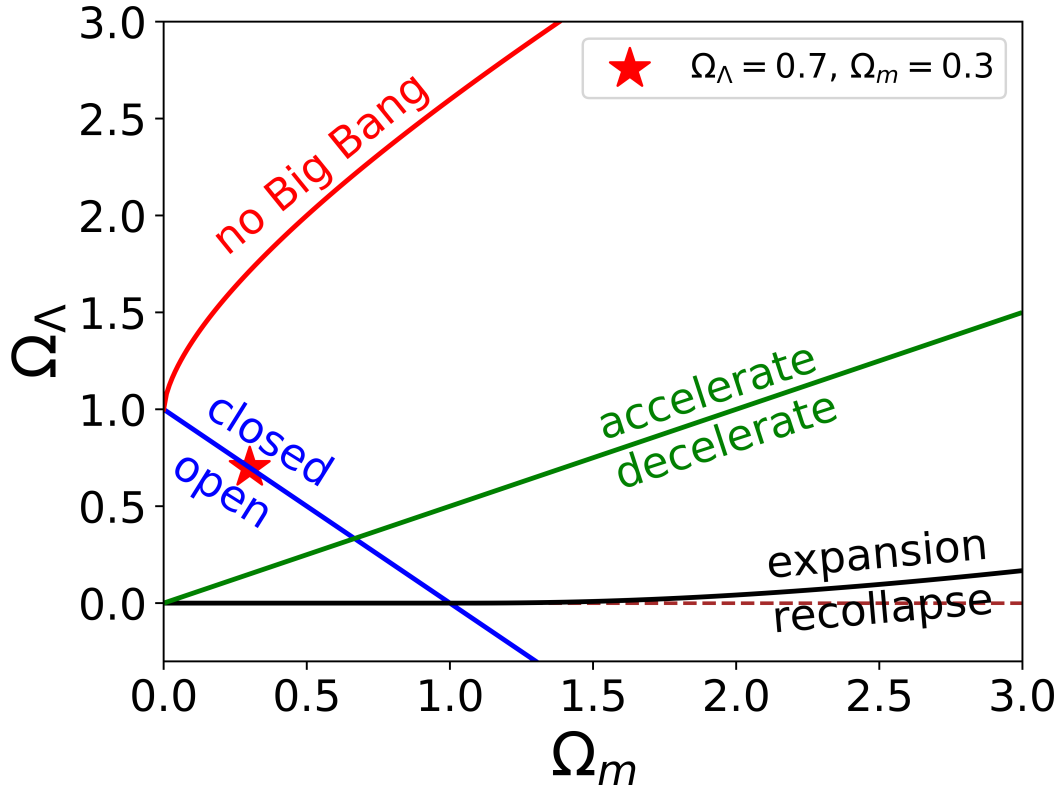


Figure 1.2: Cosmologies for different $\Omega_m - \Omega_\Lambda$ combinations. The red star indicates currently likely values for our Universe.

This will not happen only if the dark matter density is so high that the universe will recollapse before dark energy becomes dominant (cf. Carroll et al., 1992). Fig. 1.2 plots the diversity of different cosmologies in the $\Omega_m - \Omega_\Lambda$ plane. Unbounded expansion occurs for all

universes with (Carroll et al., 1992)

$$\Omega_\Lambda \geq \begin{cases} 0, & \text{if } 0 \leq \Omega_m \leq 1 \\ 4\Omega_m \left(\cos \left[\frac{1}{3} \arccos \left(\frac{1-\Omega_m}{\Omega_m} \right) + \frac{4\pi}{3} \right] \right)^3, & \text{if } \Omega_m > 1 \end{cases}. \quad (1.51)$$

If the dark energy component is abundant enough to avoid recollapse, the expansion can be either decelerated, uniform or accelerated. The critical line that separates both cases is obtained by setting $\ddot{a} = 0$ in Eq. 1.34, which leads to $\Omega_\Lambda = \frac{\Omega_m}{2}$ as shown by the green line. The blue line $\Omega_\Lambda = 1 - \Omega_m$ corresponds to flat universes with vanishing curvature density, separating closed and open cosmologies. For very large values of Ω_Λ , as illustrated by the region in the upper left corner, the dark energy component is so large that the universe cannot reach a singular state and there will never be a ‘Big Bang’. Instead these universes ‘bounce’ from an infinite size to a finite radius and then continue to expand forever. Universes that are barely outside this ‘bouncing region’, still have a Big Bang but stay at a constant size for some time. This ‘loitering phase’ is unstable and after some time they transition into exponential expansion. The boundary of universes that have no Big Bang is given by (Carroll et al., 1992)

$$\Omega_\Lambda \geq \begin{cases} 4\Omega_m \left(\cosh \left\{ \frac{1}{3} \operatorname{arccosh} \left(\frac{1-\Omega_m}{\Omega_m} \right) \right\} \right)^3, & \text{if } \Omega_m < \frac{1}{2} \\ 4\Omega_m \left(\cos \left\{ \frac{1}{3} \arccos \left(\frac{1-\Omega_m}{\Omega_m} \right) \right\} \right)^3, & \text{if } \Omega_m \geq \frac{1}{2} \end{cases}. \quad (1.52)$$

While the Friedman equations allow a large variety of different cosmologies, experimental results strongly hint towards a flat Λ -CDM universe with $(\Omega_m, \Omega_\Lambda) \approx (0.3, 0.7)$ as illustrated by the red star in Fig. 1.2.

The previous section discussed how the energy density of matter, radiation and dark energy scales with the size of the universe. At early times, i.e. for a small scale factor, the radiation term in Eq. 1.38 dominates over dark matter and dark energy, leading to $a(t) \approx t^{1/2}$ (cf. Carroll & Ostlie, 2007; Sánchez, 2017). After decoupling, the evolution of the universe becomes matter dominated, which leads to a faster expansion than during radiation domination $a(t) \approx t^{2/3}$ (cf. Carroll & Ostlie, 2007; Sánchez, 2017). Closely before our present time, dark energy began to dominate the cosmic expansion. The Universe is therefore currently transitioning into an era of exponential expansion $a(t) \propto \exp(H_0 t)$ (cf. Carroll & Ostlie, 2007; Sánchez, 2017).

1.2.2 The Formation of Structure

The previous section introduced the cosmological standard model, gave an overview over the cosmic inventory and reviewed the different cosmologies that result from the Friedmann-Lemaître equations. Up until now we have described the universe at large scales, i.e. larger than 260 Mpc/h (Yadav et al., 2010), where the universe can be assumed to be isotropic and homogeneous. At smaller scales the universe shows a plethora of structure and a fine, web-like pattern of galaxies. Fig. 1.3 shows a 4° slice from the 2dF Galaxy Redshift survey (Peacock et al., 2001) that illustrates the rich structure of the cosmic web.

The previous discussion introduced our Universe as being homogeneous, isotropic and flat. However it is a priori unclear, which physical process is responsible to generate such remarkably ‘simple’ initial conditions. The flatness of space observed today, requires that the total density of matter, radiation and dark energy $\rho_{\text{tot}} = \rho_m + \rho_{\text{RAD}} + \rho_\Lambda$ has to be finely tuned to the critical density, e.g. $|\rho_{\text{tot}}/\rho_c| \leq 10^{-15}$ for $z \sim 10^{10}$ (cf. Schneider, 2006a). Furthermore, in

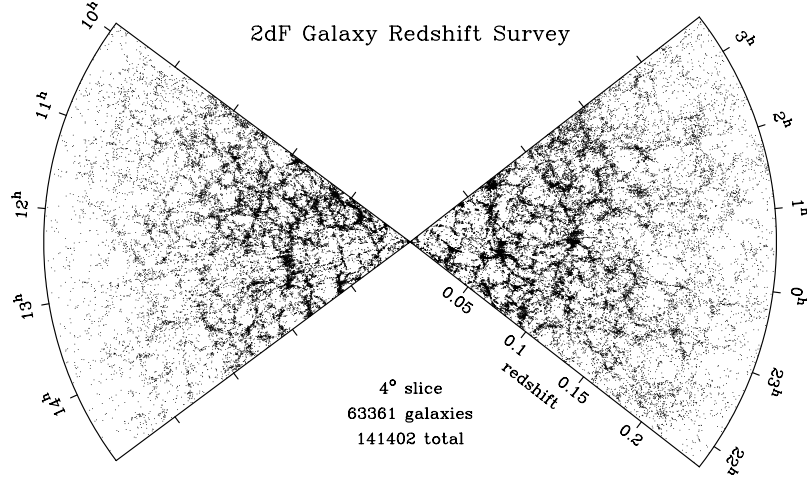


Figure 1.3: Plot of a 4° wide slice of the Galaxy distribution from the 2dF Galaxy Redshift Survey taken from Peacock et al. (2001). Image credit: Peacock et al. (2001)

order to obey the cosmological principle, the early Universe has to set highly isotropic initial conditions⁶(cf. Schneider, 2006a). Both these fundamental problems of the standard Big Bang picture can be resolved by an era of exponential expansion in the early Universe, the inflation era (Guth, 1981; Linde, 1982; Albrecht & Steinhardt, 1982). The rapid expansion of space during this period ‘washes out’ the curvature of space and allows small, causally connected regions to expand into sizes that equal the observable Universe (cf. Schneider, 2006a). Inflation also sets the initial conditions for the growth of structure, predicting that the early universe was filled with small, independent fluctuations that evolve under the influence of gravity to form larger structures (cf. Schneider, 2006a). Over time these initial density fluctuations grow and begin to interact with neighboring structures, which leads to nonlinear structure growth (cf. Schneider, 2006a). The expansion of the universe detaches neighboring density fluctuations and therefore counteracts the formation of structure. This preserves, at large scales, the form of the primordial density field as will be discussed in the following.

In order to describe the time evolution of these density fluctuations, we note that gravitational fields in the context of cosmology are typically rather weak and we can instead use a description based on classical physics within local regions of the universe (see e.g. Tegmark, 1994, §2.4). This description will be valid as long as the gravitational fields are weak ($\Phi \ll c^2$), the velocities are nonrelativistic ($v \ll c$) and the region we consider is much smaller than the Hubble horizon $d_H = 1/H_0$ (e.g. Coles, 2001).

The following description mainly follows Coles (2001), where I used Tegmark (1994); Peacock (1999); Mukhanov (2005); Schneider (2006a); Alborno (2016) as additional references.

We describe the evolution of density fluctuations using the Euler equations of a Newtonian fluid evolving under gravity. Using comoving coordinates χ and peculiar velocities $\mathbf{v}_p = a \dot{\chi}$,

⁶For example, CMB photons observed under a spatial separation of more than 1deg have been causally disconnected since their last matter interaction, due to the light speed limit (Schneider, 2006a).

the energy-momentum conservation equation then reads (Coles, 2001)

$$\frac{\partial (a\mathbf{v}_p)}{\partial t} + (\mathbf{v}_p \nabla_\chi) \mathbf{v}_p = -\frac{1}{\rho} \nabla_\chi p - \nabla_\chi \phi, \quad (1.53)$$

where ϕ is the peculiar gravitational potential, which describes fluctuations in the gravitational potential over a uniform background universe. Similarly we can formulate the mass conservation equation as

$$\frac{\partial \rho}{\partial t} + 3\frac{\dot{a}}{a}\rho + \frac{1}{a} \nabla_\chi (\rho \mathbf{v}_p) = 0 \quad (1.54)$$

and the poisson equation for Newtonian gravity as

$$\nabla_\chi^2 \phi = 4\pi G a^2 \rho_0 \delta. \quad (1.55)$$

Here we introduced the density contrast

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \rho_0}{\rho_0}. \quad (1.56)$$

with respect to the average density of the universe ρ_0 .

To provide an analytic solution to these equations, we can expand ρ and \mathbf{v}_p to first order assuming $|\mathbf{v}_p|, |\delta|, |\phi| \ll 1$ (Tegmark, 1994; Coles, 2001; Alborno, 2016). Ignoring pressure terms, the linearized Euler equations then read (cf. Tegmark, 1994; Peacock, 1999; Mukhanov, 2005; Schneider, 2006a; Alborno, 2016)

$$\frac{\partial \mathbf{v}_p}{\partial t} + \left(\frac{\dot{a}}{a}\right) \mathbf{v}_p + \frac{1}{a} \nabla_\chi \phi = 0 \quad (1.57)$$

$$\frac{\partial \delta}{\partial t} + \frac{1}{a} \nabla_\chi \cdot \mathbf{v}_p = 0 \quad (1.58)$$

$$\nabla_\chi^2 \phi - 4\pi G a^2 \rho_0 \delta = 0 \quad (1.59)$$

We can combine these equations by taking the time derivative of Eq. 1.58, the divergence of Eq. 1.57 and substitute Eq. 1.59, which then leads to (see e.g. Coles, 2001; Schneider, 2006a; Alborno, 2016)

$$\ddot{\delta}(\chi, t) + 2H(t)\dot{\delta}(\chi, t) - 4\pi G \rho_0 \delta = 0, \quad (1.60)$$

where we used $H(t) = \frac{\dot{a}(t)}{a(t)}$. This second order differential equation has the solution (see e.g. Tegmark, 1994; Alborno, 2016)

$$\delta(\chi, t) = A(\chi)D_+(t) + B(\chi)D_-(t), \quad (1.61)$$

where D_+ is a growing mode and D_- is the decaying mode. In the simple case $\Omega_m = 1$ these solutions scale as $D_+(t) \propto t^{2/3}$ and $D_-(t) \propto t^{-1}$ respectively (see e.g. Schneider, 2006a). As the initial fluctuations start out at very early times, the following discussion only considers the growing mode $D_+(t)$, where we simplify the notation by setting $D_+(t) = D(t)$. The generalization for universes that contain matter and a cosmological constant reads (Heath, 1977; Eisenstein, 1997; Hamilton, 2001)

$$D(a) = \frac{5\Omega_m}{2} H(a) \int_0^a \frac{da'}{a'^3 H(a')^3}, \quad (1.62)$$

where $D(a)$ is normalized such that $D(a) \rightarrow a$ as $a \rightarrow 0$ (cf. Eisenstein, 1997).

1.2.3 The Power Spectrum

In the previous section we reviewed how a single density fluctuation mode evolves with time. The evolution of the full density field $\delta(\mathbf{x})$, can be described in fourier space (cf. Coles, 2001)

$$\tilde{\delta}(\mathbf{k}) = \frac{1}{(2\pi)^3} \int d^3\mathbf{x} e^{-i\mathbf{k}\cdot\mathbf{x}} \delta(\mathbf{x}). \quad (1.63)$$

If $\delta(\mathbf{x})$ is a Gaussian random field, as predicted by inflation (Guth, 1981; Guth & Pi, 1982), its properties are completely determined by its variance. Assuming translational symmetry, reality and isotropy for the density field we obtain the matter power spectrum as (cf. Coles, 2001)

$$\langle \tilde{\delta}(k_1) \tilde{\delta}(k_2) \rangle = P(k_1) \delta^D(\mathbf{k}_1 + \mathbf{k}_2). \quad (1.64)$$

Inflationary models (Guth, 1981; Guth & Pi, 1982) predict that the primordial random field has a scale invariant ‘Harrison-Zeldovic’ spectrum of the form (Harrison, 1970; Zeldovich, 1972)

$$P_0(k) = A_s k^{n_s} \quad (1.65)$$

with a spectral index⁷ $n_s \approx 1$. We note that assuming that the fluctuations are small, each individual Fourier mode of the density field grows proportional to $D(t)$ and thus the power spectrum will scale proportional to $D(t)^2$.

Fluctuations that formed during the inflation era at different scales will reenter the horizon at different times during the expansion of the universe. As explained in e.g. Bartelmann & Schneider (2001), small scale fluctuations enter the horizon during radiation domination, which prevents them to grow until the matter density equals the radiation density. In contrast, large scale fluctuations that enter the horizon much later, i.e. after this time of matter-radiation equality, could grow unsuppressed and proportional to the scale factor. Fig. 1.4 shows the present time matter power spectrum for a Λ -CDM model. As can be seen, the matter power spectrum scales quite differently at small and large scales. At small scales, or large k , the suppression of growth during radiation domination produces a scaling of $P(k, t) \propto k^{-3}$ (e.g. Tegmark, 1994; Baugh, 2000; Trodden & Carroll, 2004; Schneider, 2006a; Sánchez, 2017). At large scales, or small k , the fluctuations were not suppressed and the matter power spectrum scales as $P(k, t) \propto k$, following its primordial shape (e.g. Tegmark, 1994; Baugh, 2000; Trodden & Carroll, 2004; Schneider, 2006a; Sánchez, 2017). Additional effects like acoustic sound waves also alter the shape of the matter power spectrum (see e.g. Eisenstein & Hu, 1998; Bassett & Hlozek, 2010). Incorporating all these effects into the modelling is a numerically challenging task that results in a prediction of the transfer function $T(k)$ that, together with the growth function, connects the primordial power spectrum with its present time shape (e.g. Tegmark, 1994; Coles, 2001; Albornoz, 2016)

$$P(k, t) = P_0(k) T^2(k) \frac{D^2(k, t)}{D^2(k, t_{\text{ini}})}, \quad (1.66)$$

where t_{ini} is set to be a very early reference time. In practise these predictions are obtained using dedicated codes like CLASS (Blas et al., 2011) or CAMB⁸ (Lewis et al., 2000).

⁷The results from CMB measurements, performed by the Planck-satellite (Planck Collaboration et al., 2016) measure $n_s = 0.9655 \pm 0.0062$, which slightly ‘tilts’ the Harrison-Zeldovic form.

⁸<http://camb.info/>

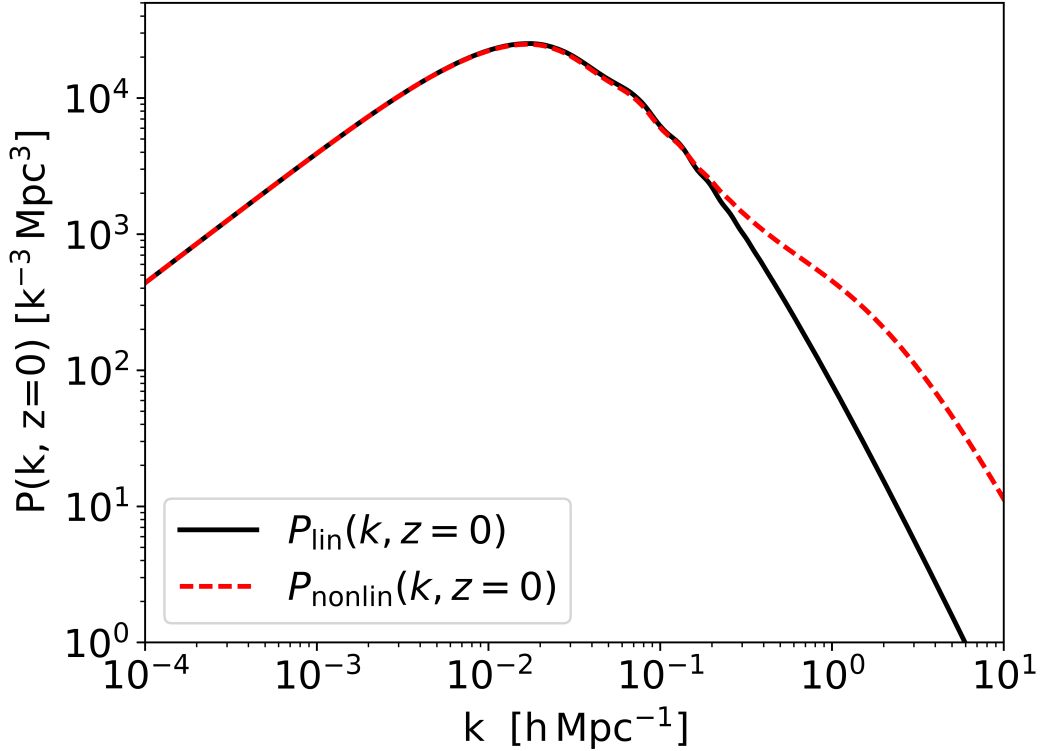


Figure 1.4: Present day ($z = 0$) power spectrum for a flat Λ -CDM model, estimated using the cosmosis (Zuntz et al., 2015) software.

The linear theory of structure formation we just described is only applicable on large scales and breaks down as soon as one enters scales of gravitational bound systems like galaxy clusters. This can be seen in Fig. 1.4 where the nonlinear power spectrum begins to deviate from the linear prediction at wavelengths of $\lambda \approx 10$ Mpc, which is about the size of a galaxy cluster. As an analytic treatment of the non-linear power spectrum is not practical, numerical simulations are used instead to derive suitable fitting formulas (Hamilton et al., 1991; Smith et al., 2003; Takahashi et al., 2012).

An alternative quantity that can be used to describe the density field in real space is the two point correlation function $\xi(r)$. The function $\xi(r)$ can be related to the probability density $dP_{1,2}$ to find two objects, e.g. galaxies, separated at distance r in the respective volume elements dV_1 and dV_2 as (Peacock, 1999; Baugh, 2000; Sánchez, 2017; Alborno, 2016)

$$dP_{1,2} = \rho_0 (1 + \xi(r)) dV_1 dV_2, \quad (1.67)$$

where ρ_0 is the average object density and the two point correlation function is therefore defined as the Fourier transform of the matter power spectrum

$$\langle \delta(x) \delta(x+r) \rangle = \xi(r). \quad (1.68)$$

We can therefore relate the matter power spectrum with the two point correlation function as (see e.g. Peebles, 1980)

$$\xi(r) = \frac{1}{2\pi^2} \int dk k^2 P(k) \left(\frac{\sin(kr)}{kr} \right), \quad (1.69)$$

1.3 Cosmological Probes in Photometric Surveys

The matter power spectrum and the two point correlation function are elegant and powerful tools to describe the three dimensional cosmic density field and its evolution. In order to directly measure them on data, one generally requires very accurate distance (i.e. redshift) information to trace the three dimensional structure of the cosmic web. Since redshifts provided by large area photometric surveys are obtained using photometry alone⁹, they are in general not accurate enough for this purpose. Instead large area photometric surveys use cosmological probes that consider projected density fields along the line of sight. Two of the most important probes will be described in the next subsection.

1.3.1 Angular Clustering

Starting in the 70s, redshift projects like the Lick (Seldner et al., 1977, and references therein), CFA (Tonry & Davis, 1979; Davis et al., 1982; Davis & Peebles, 1983) and APM (Maddox et al., 1990) surveys established galaxy correlation functions as an important cosmological probe in modern observational cosmology. Clustering measurements today mainly consider the two point correlation function of the three dimensional density field. As these measurements typically require a high redshift resolution, they are performed in the context of spectroscopic surveys like WiggleZ (Drinkwater et al., 2010; Hinton et al., 2017), BOSS (Dawson et al., 2013; Alam et al., 2017) and VVDS (Le Fèvre et al., 2005; Guzzo et al., 2008). However in recent years measurements of the angular correlation function, i.e. the correlation function of the projected galaxy density field along the line of sight, have been advocated (Asorey et al., 2012, 2014) and applied in spectroscopic (e.g. Salazar-Albornoz et al., 2017) as well as photometric surveys (Asorey et al., 2016). A major challenge in the modeling of angular correlation functions is redshift uncertainty. If the distribution of galaxies is inaccurately determined, the projection of the three dimensional density field is distorted, which propagates into biases in the derived cosmological parameters. Furthermore, these measurements consider correlation functions of the galaxy density field, which in general do not perfectly trace the underlying dark matter density field. In the following we will describe how this galaxy-dark matter bias can be incorporated into the modeling of clustering measurements.

Galaxy Bias

The power spectrum of the galaxy density field does not coincide with the dark matter power spectrum, but instead is offset, or biased, by a galaxy-dark matter bias function that in general depends on galaxy type, redshift and scale. We can expand the function that relates the galaxy density contrast $\delta_g(\mathbf{x})$ with the underlying dark matter contrast $\delta_m(\mathbf{x})$ locally in a Taylor series as (cf. Albornoz, 2016; Sánchez, 2017)

$$\delta_g(\mathbf{x}) = \sum_n \frac{b_n}{n!} \delta_m(\mathbf{x})^n, \quad (1.70)$$

where b_n are the galaxy-dark matter bias parameters. To first order we can therefore assume that the galaxy-dark matter bias function is a constant defined as

$$\delta_g(\mathbf{x}) = b \delta_m(\mathbf{x}), \quad (1.71)$$

⁹For an introduction into photometric redshift estimation we refer the reader to §1.4.

where b is the linear galaxy-dark matter bias parameter. Providing accurate models for the galaxy-dark matter bias requires a detailed knowledge about how galaxies form in dark matter halos, and there exists a large literature about more advanced models and prescriptions¹⁰. However in this work we assume a simple redshift dependent bias in the form of the Fry model (Fry, 1996; Clerkin et al., 2015). Following Clerkin et al. (2015), we assume that galaxies are stationary after forming at an initial redshift z_0 . They do not merge and their density field ‘passively’ follows the evolution of the underlying dark matter density field $\dot{\delta}_g = \dot{\delta}_m = \text{const.}$. We can then write

$$\delta_g(\mathbf{x}, z_0) - \delta_g(\mathbf{x}, z) = \delta_m(\mathbf{x}, z_0) - \delta_m(\mathbf{x}, z), \quad (1.72)$$

and divide this equation by $\delta_m(\mathbf{x}, z_0)$, which then leads to

$$\frac{\delta_g(\mathbf{x}, z_0)}{\delta_m(\mathbf{x}, z_0)} - \frac{\delta_g(\mathbf{x}, z)}{\delta_m(\mathbf{x}, z_0)} = 1 - \frac{\delta_m(\mathbf{x}, z)}{\delta_m(\mathbf{x}, z_0)}. \quad (1.73)$$

The first term quantifies the initial galaxy-dark matter bias b_0 , the second term can be written as $b(z)D(z)$ and the third term is the growth function $D(z)$. We therefore obtain for the redshift dependent galaxy-dark matter bias $b(z)$

$$b(z) = 1 + \frac{b_0 - 1}{D(z)}. \quad (1.74)$$

The Angular Correlation Power Spectrum

Measuring the galaxy correlation function of the three dimensional density field $\xi_g(r)$ or alternatively, its power spectrum $P_g(k)$, can be challenging in the context of large area photometric surveys, due to inaccurate redshift information. In the context of large area photometric surveys we instead typically consider the two dimensional projection of the three dimensional galaxy density field along the line of sight. We start by decomposing the two dimensional projected density contrast of galaxies on the sky into spherical harmonics $Y_{\ell m}$ (cf. Peebles, 1973; Crocce et al., 2011)

$$\delta(\hat{\mathbf{n}}) = \sum_{\ell \geq 0} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}), \quad (1.75)$$

where $\hat{\mathbf{n}}$ is a vector on the sky. The angular correlation power spectrum C_ℓ is defined as the variance of the coefficients $a_{\ell m}$ (cf. Crocce et al., 2011)

$$\langle a_{\ell m} a_{\ell' m'} \rangle = \delta_{\ell \ell'} \delta_{m m'} C_\ell, \quad (1.76)$$

where $\delta_{\ell, \ell'}$ and $\delta_{m, m'}$ are Kronecker delta functions.

Projecting the three dimensional density field $\delta_g(\mathbf{x}, z)$ along the line of sight direction $\hat{\mathbf{n}}$ leads to (cf. Crocce et al., 2011)

$$\delta(\hat{\mathbf{n}}) = \int dz n(z) \delta_g(\hat{\mathbf{n}}, z) \quad (1.77)$$

where $n(z)$ is the redshift distribution of galaxies and $\delta(\hat{\mathbf{n}})$ the projected density contrast.

¹⁰We refer the interested reader to the extensive review of Desjacques et al. (2016).

Performing the Fourier transformation of $\delta_g(\hat{\mathbf{n}}, z)$, using the spherical harmonics expansion of the plane wave

$$e^{i k \chi \hat{\mathbf{k}} \cdot \hat{\mathbf{n}}} = 4\pi \sum_{l \geq 0} \sum_{m=-l}^l i^l j_l(k\chi) Y_{lm}(\hat{\mathbf{k}}) Y_{lm}^*(\hat{\mathbf{n}}), \quad (1.78)$$

and comparing with Eq. 1.75 we obtain (cf. Crocce et al., 2011)

$$a_{\ell m} = 4\pi i^\ell \int dz n(z) \int \frac{d^3 k}{(2\pi)^3} \delta(\mathbf{k}, z) j_\ell(k\chi(z)) Y_{\ell m}^*(\hat{\mathbf{k}}). \quad (1.79)$$

With the orthogonality relation for spherical harmonics we can then derive the angular correlation power spectrum as (cf. Crocce et al., 2011; Kirk et al., 2015)

$$C_\ell = \frac{2}{\pi} \int dk k^2 P(k) \phi_\ell(k)^2, \quad (1.80)$$

where

$$\phi_\ell(k) = \int dz D(z) n(z) b(z) j_\ell(k\chi(z)). \quad (1.81)$$

Here $P(k)$ denotes the matter power spectrum today ($z = 0$), that is convolved with a window function $\phi_\ell(k)$. The window function of galaxy clustering contains an evolution factor $\propto D(z)$ and corrects for galaxy-dark matter bias $\propto b(z)$. The terms $j_\ell(k\chi(z))$ are spherical bessel functions and $n(z)$ represents the galaxy redshift distribution over which the three dimensional density field is projected.

In analogy to the three dimensional case considered earlier, the real space analog of the angular correlation power spectrum is given as the angular correlation function $w(\theta)$ (cf. Peebles, 1973; Crocce et al., 2011)

$$w(\theta) = \langle \delta_g(\hat{\mathbf{n}}) \delta_g(\hat{\mathbf{n}} + \hat{\theta}) \rangle. \quad (1.82)$$

The angular correlation function can then be related to the angular correlation power spectrum using Eq. 1.75, Eq. 1.76 and the addition theorem for spherical harmonics

$$P_\ell(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}') = \frac{4\pi}{2\ell + 1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\mathbf{n}}') Y_{\ell m}^*(\hat{\mathbf{n}}), \quad (1.83)$$

where $P_\ell(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}')$ are Legendre polynomials. After collecting terms we obtain (cf. Crocce et al., 2011)

$$w(\theta) = \sum_{\ell \geq 0} \left(\frac{2\ell + 1}{4\pi} \right) P_\ell(\cos(\theta)) C_\ell. \quad (1.84)$$

We reiterate, that angular correlation functions are less sensitive to photometric redshift uncertainty than the correlation function of the three dimensional density field. Galaxy clustering measurements in large area photometric surveys therefore use projected quantities like angular correlation functions or angular correlation power spectra.

1.3.2 Gravitational Lensing

The magnification and shear pattern of light, emitted by high redshift sources and lensed by the intervening matter distribution, provides another powerful tool to study the growth of structure. This gravitational lensing effect, i.e. the bending of the light trajectory under the influence of gravity, was first predicted around 1800 by Cavendish and Soldner (Valls-Gabaud, 2006). In 1919 Arthur Eddington, Frank Watson Dyson and colleagues (Dyson et al., 1920) then measured the change of the observed position of stars very close to the sun during a solar eclipse. The measured deflection angle was twice that of the prediction from Newtonian gravity, which provided the first experimental confirmation of General Relativity. Since these early measurements, modern Gravitational Lensing has evolved into a powerful tool for cosmology and astrophysics. Strong lenses like galaxy clusters are used e.g. as ‘cosmic telescopes’ to observe very faint galaxies that would otherwise not be detectable with current instruments (Goobar et al., 2009; Monna & Covone, 2012; Coe et al., 2013; Rydberg et al., 2015; Chan et al., 2017) or to measure their masses within the central regions (see e.g. Abdelsalam et al., 1998; Hoekstra et al., 2013; Monna et al., 2017). In a more cosmological context strong lensing also provides a method to measure the Hubble constant using gravitational time delays (Suyu et al., 2013; Hojjati & Linder, 2014; Treu & Marshall, 2016). The majority of cosmological applications however considers lensing by weak gravitational fields of the large scale mass distribution of the universe. This method has evolved into a powerful probe for dark matter and dark energy, especially in the context of large area photometric surveys like CFHTLenS (Heymans et al., 2013), KiDS (Hildebrandt et al., 2017) and DES (Abbott et al., 2016).

Fundamentals of Gravitational Lensing

Up until now we have considered light propagation through a homogeneous universe described by the FLRW metric. In the presence of gravitational fields produced by the matter distribution of the universe, the light trajectories have to be corrected for their influence. I follow Narayan & Bartelmann (1996, §2.1.1 - §2.1.2) in the next paragraphs. If the peculiar velocity of the lens is small and the gravitational potential weak, we can define the effective index of refraction produced by the gravitational potential of the lens $\Phi(\mathbf{x}, \chi')$ as

$$n_{\text{eff}}(\mathbf{x}, \chi') = 1 - 2 \Phi(\mathbf{x}, \chi'). \quad (1.85)$$

Here χ' is the transverse comoving distance to the lens along the line of sight and \mathbf{x} is the comoving separation vector. Using Fermat's principle we can derive the deflection angle as

$$\hat{\alpha} = - \int \nabla_{\perp} n_{\text{eff}}(\mathbf{x}, \chi') d\chi' = 2 \int \nabla_{\perp} \Phi(\mathbf{x}, \chi') d\chi', \quad (1.86)$$

where $\nabla_{\perp} = (\partial/\partial x_1, \partial/\partial x_2)$ is the transverse comoving gradient of the potential perpendicular to the light trajectory.

The most simplistic lensing system is a point mass, which the photons pass at an impact parameter b . The gravitational potential for this system then reads

$$\nabla_{\perp} \phi(b, z) = \frac{G M \mathbf{b}}{(b^2 + z^2)^{3/2}}, \quad (1.87)$$

where z is the position of the photon along the line-of-sight. From Eq. 1.86 we obtain a deflection angle of

$$\hat{\alpha} = \frac{4 G M}{b}. \quad (1.88)$$

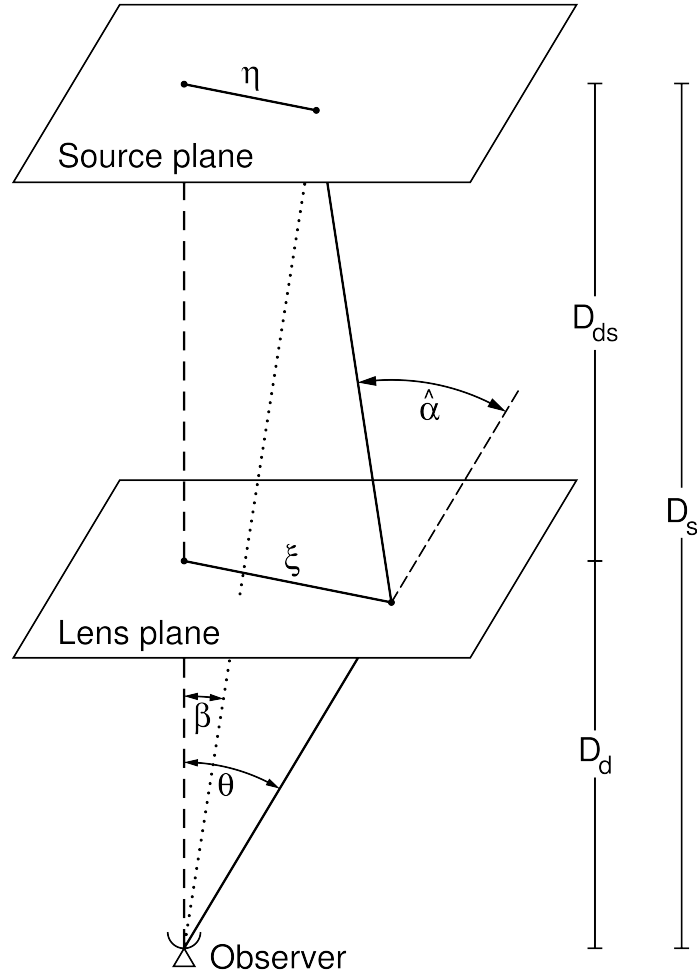


Figure 1.5: Geometry of Weak Gravitational Lensing. Image credit: (Bartelmann & Schneider, 2001)

For the more realistic case of an extended lens, we can assume that the lens is thin compared with the distance between lens and source and lens and observer. Considering the configuration shown in Fig. 1.5, we project the mass distribution $\rho(\xi, z)$ along the line-of-sight to yield the surface mass density

$$\Sigma(\xi) = \int \rho(\xi, z) dz. \quad (1.89)$$

Here ξ is a two dimensional vector and describes the position on the lens plane. The deflection angle $\hat{\alpha}$ can then be obtained as

$$\hat{\alpha}(\xi) = 4G \int \frac{(\xi - \xi') \Sigma(\xi')}{|\xi - \xi'|^2} d^2 \xi'. \quad (1.90)$$

From Fig. 1.5 we can relate the unperturbed deflection angle β and the perturbed angle

θ with the distance to the lens and source as (see Bartelmann & Schneider, 2001, §3.1.2)

$$\beta = \frac{\eta}{D_s} \quad (1.91)$$

$$\theta = \frac{\xi}{D_d}. \quad (1.92)$$

While these angles can be trivially defined in Euclidean space, it is nontrivial to see why they should be valid in curved spacetime. As discussed in §1.1, the distance measure where this relation holds by construction is the angular diameter distance Eq. 1.18, which we associate with $D_{d,s,ds}$.¹¹

Following Bartelmann & Schneider (2001, §3.1.2) we obtain for the deflection angle β

$$\beta = \theta - \frac{D_{ds}}{D_s} \hat{\alpha}(D_d \theta). \quad (1.93)$$

Introducing the scaled deflection angle

$$\alpha(\theta) = \frac{D_{ds}}{D_s} \hat{\alpha}(D_d \theta), \quad (1.94)$$

we can simplify the lens equation as

$$\beta = \theta - \alpha(\theta). \quad (1.95)$$

The scaled deflection angle α can then be expressed as a function of θ

$$\alpha(\theta) = \frac{1}{\pi} \int d^2 \theta' \kappa(\theta') \frac{\theta - \theta'}{|\theta - \theta'|^2}, \quad (1.96)$$

where we introduced the dimensionless surface mass density

$$\kappa(\theta) = \frac{\Sigma(D_d \theta)}{\Sigma_{\text{crit.}}} \quad (1.97)$$

The critical surface mass density

$$\Sigma_{\text{crit.}} = \frac{1}{4\pi G} \frac{D_s}{D_d D_{ds}}, \quad (1.98)$$

separates between the two regimes of gravitational lensing. In the strong lensing limit $\kappa > 1$, i.e. $\Sigma > \Sigma_{\text{crit.}}$ the lens can produce multiple images, which is not possible in the weak lensing regime $\kappa \ll 1$.

1.3.3 Weak Lensing by Large-scale structure

Previously we reviewed the basic geometry of gravitational lensing, assuming a singular object, e.g. a galaxy or a galaxy cluster, as the lens. Of particular interest for cosmology is the lensing effect by the three dimensional matter distribution along the line of sight. We will now see, that the density contrast of the matter distribution along the line of sight acts, to first order, as an

¹¹Specifically we note that $D_d + D_{ds} \neq D_s$. Here we changed the notation from d_A , which denoted the angular diameter distance in Eq. 1.18.

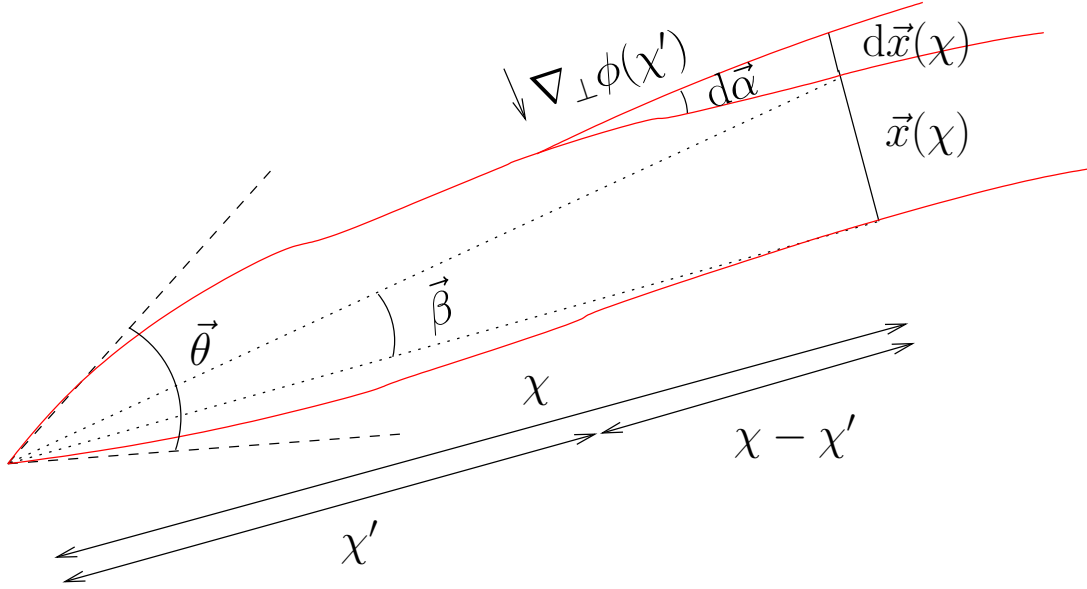


Figure 1.6: Bending of an incoming light wave by Weak Gravitational Lensing. The initial, undeflected separation of two parallel light waves $\mathbf{x}(\chi)$ would appear under the angle β as shown by the dotted lines. The gravitational potential $\nabla_{\perp} \Phi(\chi')$ bends the light by an angle of $d\alpha$. The observer then sees the light ray under the total deflection angle θ . The comoving distance χ denotes the distance from the observer to the source plane and χ' denotes the comoving distance to the lens plane. Image credit: (Kilbinger, 2015)

effective surface mass density κ for sources located at a certain redshift. The effective surface mass density produced by the matter distribution is the line of sight integral of the density contrast weighted by the geometrical factor $(D_d D_{ds})/D_s$ introduced earlier (see Schneider, 2006b, Eq. 95, page 359). Photometric redshift uncertainty enters this geometrical factor as a distortion of the redshift distribution of sources. While the mathematical treatment is rather technical, the fundamental equations that will be derived are very similar to the ones considered in the previous section. However we will use a slightly different notation for the lensing geometry as shown in Fig. 1.6, for consistency with the literature (e.g. Bartelmann & Schneider, 2001; Schneider, 2006b; Kilbinger, 2015). In the following discussion we will mainly follow Kilbinger (2015, §3.2).

Considering the lensing geometry shown in Fig. 1.6, we see that an infinitesimal deviation in the source plane $d\mathbf{x}$ induces a change in the deflection angle $d\alpha$ as

$$d\mathbf{x} = f_K(\chi - \chi') d\alpha, \quad (1.99)$$

where $f_K(\chi)$ was defined in the FLRW metric (Eq. 1.2) to correct for the effect of curvature. Using Eq. 1.86, we can then relate the infinitesimal shift in the deflection angle $d\alpha$ with an infinitesimal shift in the distance to the lens $d\chi'$ as

$$d\alpha = -2 \nabla_{\perp} \Phi(\mathbf{x}, \chi') d\chi' \quad (1.100)$$

Plugging Eq. 1.100 into Eq. 1.99 and integrating over the line of sight we obtain

$$\mathbf{x}(\chi, \theta) = f_K(\chi) \theta - 2 \int_0^{\chi} d\chi' f_K(\chi - \chi') \left[\nabla_{\perp} \Phi(\mathbf{x}(\theta, \chi'), \chi') - \nabla_{\perp} \Phi^{(0)}(\chi') \right], \quad (1.101)$$

where $\Phi^{(0)}(\chi')$ denotes the gravitational potential along the unperturbed ray. The apparent deflection angle $\boldsymbol{\theta}$ is then given by the lens equation Eq. 1.95

$$\boldsymbol{\theta} = \boldsymbol{\alpha} + \boldsymbol{\beta}. \quad (1.102)$$

Solving this equation for $\boldsymbol{\alpha}$ we obtain

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = 2 \int_0^\chi d\chi' \left(\frac{f_K(\chi - \chi')}{f_K(\chi)} \right) \left[\nabla_\perp \Phi(\mathbf{x}(\boldsymbol{\theta}, \chi'), \chi') - \nabla_\perp \Phi^{(0)}(\chi') \right]. \quad (1.103)$$

We can now connect the unlensed angle $\boldsymbol{\beta}$ with the lensed deflection angle $\boldsymbol{\theta}$ by the magnification matrix, or Jacobian matrix, $\mathbf{A} = \partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}$. Expanding \mathbf{A} to linear order in Φ we obtain (cf. Kilbinger, 2015)

$$A_{ij}(\boldsymbol{\theta}, \chi) = \delta_{ij} - \frac{\partial\alpha_i}{\partial\theta_j} \quad (1.104)$$

$$= \delta_{ij} - 2 \int_0^\chi d\chi' \frac{f_K(\chi - \chi') f_K(\chi')}{f_K(\chi)} \frac{\partial^2}{\partial x_i \partial x_j} \Phi(f_K(\chi')\boldsymbol{\theta}, \chi'). \quad (1.105)$$

In the case of weak gravitational lensing, \mathbf{A} is invertible, i.e. there exist no multiple images. The components of this matrix can be written more concisely

$$A_{ij} = \delta_{ij} - \partial_i \partial_j \psi, \quad (1.106)$$

where ∂_i denotes partial derivatives w.r.t. $\boldsymbol{\theta}$ and we define the lensing potential as

$$\psi(\boldsymbol{\theta}, \chi) = 2 \int_0^\chi d\chi' \left(\frac{f_K(\chi - \chi')}{f_K(\chi) f_K(\chi')} \right) \Phi(f_K(\chi')\boldsymbol{\theta}, \chi'). \quad (1.107)$$

The components of the magnification matrix then describe how the lensed image of a background source is distorted and magnified. We can parametrize \mathbf{A} as

$$\mathbf{A} = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix}, \quad (1.108)$$

where the respective components can be expressed in terms of the lensing potential

$$\kappa = \frac{1}{2} (\partial_1 \partial_1 + \partial_2 \partial_2) \psi \quad (1.109)$$

$$\gamma_1 = \frac{1}{2} (\partial_1 \partial_1 - \partial_2 \partial_2) \psi \quad (1.110)$$

$$\gamma_2 = \partial_1 \partial_2 \psi \quad (1.111)$$

As illustrated in Fig. 1.7, the convergence κ describes the magnification, and the shear coefficients γ_1 and γ_2 the distortion of the shape of the background source.

The two shear components parametrize the orientation of the sheared galaxy as shown in Fig. 1.8. We see that the two component shear vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ can be written as

$$\boldsymbol{\gamma} = |\boldsymbol{\gamma}| \exp(2i\phi), \quad (1.112)$$

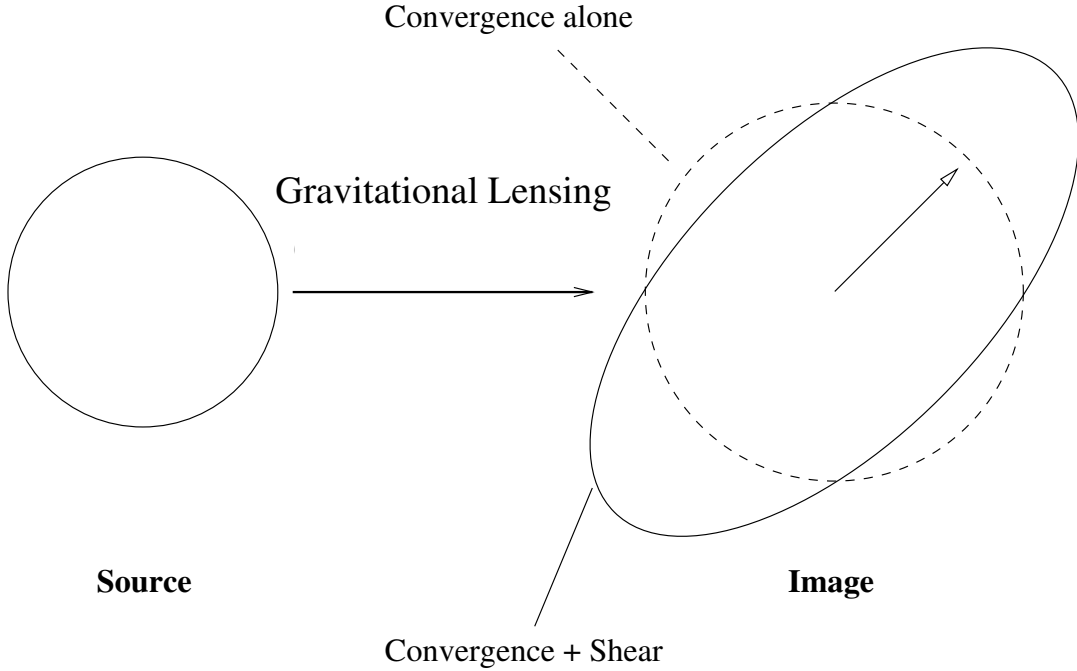


Figure 1.7: Effect of Weak Gravitational Lensing on the shape of a circular source. The pure convergence effect leads to a magnification of the circular source, whereas the shear makes the circular source more elliptical. Image credit: Umetsu (2010)

where ϕ is the polar angle between γ_1 and γ_2 . Since we recover the original orientation for $\phi = \pi$, shear transforms like a spin-2 quantity.

Before we consider the lensing effect on the full cosmic density field, we want to note that weak lensing is a powerful tool to measure the mass of e.g. galaxy clusters. The weak gravitational lensing effect of a single lens will align the shape of background galaxies tangentially relative to the lens. Following Hoekstra et al. (2013), we introduce the azimuthal angle ϕ wrt. the lensing galaxy, and define the tangential shear signal as

$$\gamma_T = -(\gamma_1 \cos(2\phi) + \gamma_2 \sin(2\phi)). \quad (1.113)$$

The average tangential shear signal $\langle \gamma_T \rangle$ (averaged over ϕ) can then be connected with the mass contrast that is defined wrt. the critical surface mass density Σ_{crit} (Eq. 1.98) as (Miralda-Escude, 1991; Kaiser, 1995)

$$\langle \gamma_T \rangle = \frac{\bar{\Sigma}(< r) - \bar{\Sigma}(r)}{\Sigma_{\text{crit}}}. \quad (1.114)$$

Here, $\bar{\Sigma}(< r)$ is the mean surface mass density within the radius r and $\bar{\Sigma}(r)$ is the surface mass density at the (border) radius r . We note that the critical surface mass density Σ_{crit} depends on the redshift of the background sources. Photometric redshift uncertainties of these background sources can therefore bias Eq. 1.114 and therefore the derived cluster mass estimate.

Coming back to our main goal, i.e. to use the weak lensing effect on the cosmological density field as a tool for cosmology (following Kilbinger, 2015, §3.2), we need to connect the convergence κ , with the underlying matter density contrast. This relation is found by the 3D

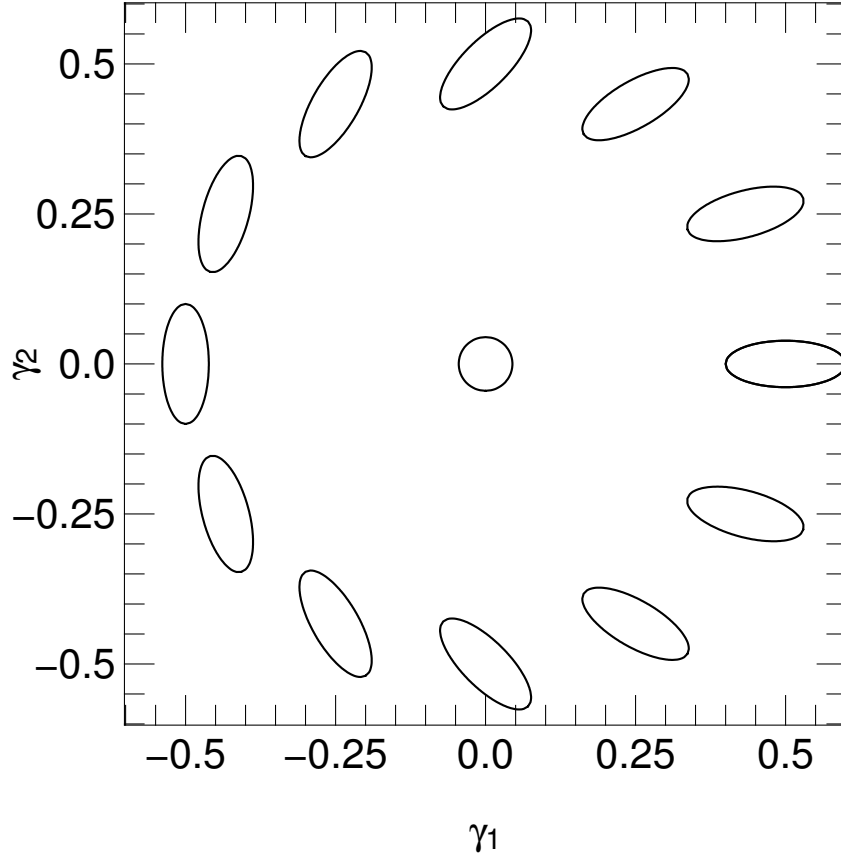


Figure 1.8: Schematic illustration of the influence of gravitational shear on a perfectly round galaxy shape, with respect to the two shear components γ_1 and γ_2 . Image credit: (Kilbinger, 2015)

Poisson equation in comoving coordinates (cf. Schneider, 2006b, §6.2)

$$\nabla^2 \Phi = \frac{3H_0^2 \Omega_m}{2a} \delta, \quad (1.115)$$

where Φ is the gravitational potential, δ is the density contrast, a is the cosmic scale factor and H_0 and Ω_m are the Hubble and matter density parameters respectively. Using Eq. 1.109 we can then express the lensing convergence as

$$\kappa(\boldsymbol{\theta}, \chi) = \frac{3H_0^2 \Omega_m}{2} \int_0^\chi d\chi' \frac{f_K(\chi') f_K(\chi - \chi')}{f_K(\chi)} \frac{\delta(f_K(\chi') \boldsymbol{\theta}, \chi')}{a(\chi')}. \quad (1.116)$$

We note that for a flat universe, the geometrical factor $\chi'(\chi - \chi')$ parametrizes the lensing efficiency. It is a parabola with a maximum at $\chi' = \frac{\chi}{2}$, which means that the lensing effect is strongest for lenses that are exactly between (in terms of χ) the observer and the source. As we are typically considering an ensemble of sources distributed along the line of sight following the redshift distribution $n(z)$, we need to average the convergence over this distribution of source redshifts as

$$\kappa(\boldsymbol{\theta}) = \int_0^{\chi_H} d\chi n(\chi) \kappa(\boldsymbol{\theta}, \chi). \quad (1.117)$$

The convergence can be written as

$$\kappa(\boldsymbol{\theta}) = \frac{3H_0^2\Omega_m}{2} \int_0^{\chi_H} d\chi q(\chi) f_K(\chi) \frac{\delta(f_K(\chi)\boldsymbol{\theta}, \chi)}{a(\chi)} \quad (1.118)$$

with the lensing weight function $q(\chi)$

$$q(\chi) = \int_{\chi}^{\chi_H} d\chi' n(\chi') \left(\frac{f_K(\chi' - \chi)}{f_K(\chi')} \right). \quad (1.119)$$

Here $\chi(a) \rightarrow \chi_H$ as $a \rightarrow 0$ (cf. Schneider, 2006b, §6.2). Note that we express the redshift distribution in terms of the comoving line-of-sight distance χ instead of the redshift z , where both parametrizations can be related as

$$n(z) dz = n(\chi) d\chi. \quad (1.120)$$

Comparing Eq. 1.116 with Eq. 1.97 we see that the lensing convergence is indeed the effective surface mass density of the density contrast, where the geometrical factor $\frac{f_K(\chi')f_K(\chi-\chi')}{f_K(\chi)}$ corresponds to the respective term in the inverse critical surface mass density Σ_{crit} .

In close analogy to the last section where we introduced the angular correlation power spectrum, we can relate the convergence in Fourier space to the convergence power spectrum, following Kilbinger (2015, §3.7), as

$$\langle \tilde{\kappa}(\ell) \tilde{\kappa}^*(\ell') \rangle = (2\pi)^2 \delta(\ell - \ell') C_\ell^\kappa(\ell) \quad (1.121)$$

where $\delta(\ell - \ell')$ is the Dirac delta function. The derivation that leads to the expression of $C_\ell^\kappa(\ell)$ closely follows the derivation of the angular correlation function and is presented in Kirk (2011, page 37-40). We expand the convergence into spherical harmonics, perform a fourier transformation of the density contrast, expand the plane wave into spherical harmonics and use the definition of the convergence power spectrum in Eq. 1.121. After collecting terms using the orthogonality relation of spherical harmonics, we use a simplification known as the Limber approximation (Limber, 1953; Kaiser, 1992; Simon, 2007; Kirk, 2011; Giannantonio et al., 2012) that omits correlations along the line of sight, assuming that these structures are sufficiently small, such that the projection ‘smears’ them out (see also Schneider, 2006b; Kirk, 2011; Kilbinger, 2015). After a lengthy calculation one arrives at (e.g. Blandford et al., 1991; Kaiser, 1992; Bartelmann & Schneider, 2001; Takada & Jain, 2004; Joachimi & Bridle, 2010; Kirk et al., 2015)

$$C_\ell^\kappa(\ell) = \frac{9}{4} \Omega_m^2 H_0^4 \int_0^{\chi_H} d\chi \frac{q^2(\chi)}{a^2(\chi)} P_\delta \left(k = \frac{\ell}{f_K(\chi)}, \chi \right). \quad (1.122)$$

The convergence power spectrum, similar to the angular correlation power spectrum, projects the matter power spectrum $P_\delta \left(k = \frac{\ell}{f_K(\chi)}, \chi \right)$ along the line of sight using the lensing weight function $q(\chi)$ defined in Eq. 1.119. As the lensing geometry critically depends on the photometric redshift distribution of sources, the accurate control of photometric redshift errors is a primary requirement to effectively use this probe for precision cosmology. The convergence power spectrum is directly sensitive to the matter density Ω_m and the Hubble parameter H_0 . Furthermore, the geometrical factor $q(\chi)$ depends, via Eq. 1.2, on the curvature of space. As

$C_\ell^\kappa(\ell)$ also depends on the power spectrum (Eq. 1.66) and the scale factor (Eq. 1.34), it is sensitive to the density parameters (Eq. 1.36), the power spectrum amplitude A_s and the dark energy equation of state parameters w_0 and w_a .

Instead of measuring the convergence power spectrum, weak gravitational lensing analyses usually instead use the cosmic shear correlation functions $\xi_\pm(\theta)$ (see e.g. Kaiser, 1992; Schneider, 2006b)

$$\xi_\pm(\theta) = \frac{1}{2\pi} \int_0^\infty d\ell C_\ell^\kappa(\ell) J_{0,4}(\ell\theta), \quad (1.123)$$

where J_0 (J_4) are the Bessel functions of the first kind that correspond to ξ_+ (ξ_-). These correlation functions have the decisive advantage that they relate to shear two point correlations functions that are easy to estimate by averaging the product of the measured shear of galaxy pairs as further discussed in Kilbinger (2015). In this work we will however mainly work with convergence power spectra.

In summary we note, that the modeling of both the angular correlation power spectra and the convergence power spectra depend on accurate redshift distributions. If these measurements are performed in the context of large area photometric surveys, we need to derive the redshift information using the photometry alone. The next section describes the photometric redshift methodologies that can be used to extract redshift information for photometrically observed galaxies.

1.4 Photometric redshift estimation

The radiation emitted by an astronomical source is measured as the incoming radiative energy per unit time and unit area (cf. Carroll & Ostlie, 2007; Greisel, 2015). As the incoming light of galaxies and stars is emitted in broad spectral ranges, we express its flux density either as a function of wavelength $[f_\lambda] = 1 \frac{\text{erg}}{\text{cm}^2 \text{s \AA}}$ or as a function of frequency $[f_\nu] = 1 \frac{\text{erg}}{\text{cm}^2 \text{s Hz}}$, where both parametrizations are related as (Hogg et al., 2002)

$$\nu f_\nu = \lambda f_\lambda. \quad (1.124)$$

Photometric surveys observe the galaxy through a set of photometric filters that are convolved with the SED of the galaxy to produce a flux measurement (see e.g. Hogg et al., 2002; Greisel, 2015)

$$f_i = \int \mathcal{B}_i(\lambda) f_\lambda(\lambda) d\lambda. \quad (1.125)$$

Here \mathcal{B}_i is the filter throughput in the photometric band and f_i is the measured flux of the galaxy in the filter. Instead of fluxes, photometry is typically quoted in terms of the apparent magnitudes (see e.g. Hogg et al., 2002; Greisel, 2015)

$$m_{\text{app,AB}} = -\frac{5}{2} \log_{10} \left(\frac{f}{f_{\text{ref}}} \right), \quad (1.126)$$

where f_{ref} is a reference flux. The choice of reference flux can vary, however it has become standard to use the AB system (Oke & Gunn, 1983), where $f_{\nu, \text{ref}} = 3631$ Jansky and $1 \text{ Jansky} = 10^{-23} \frac{\text{erg}}{\text{s Hz cm}^2}$.

The apparent magnitude of a galaxy depends on its distance, or redshift, and it is therefore customary to define the absolute magnitude M as the apparent magnitude that of the object

would have in 10 pc distance (see e.g. Hogg et al., 2002)

$$M = m_{\text{app}} - \text{DM} - K_{\text{corr}}, \quad (1.127)$$

where the distance modulus DM is defined as a function of the luminosity distance d_L (Eq. 1.24) as (see e.g. Hogg et al., 2002)

$$\text{DM} = 5 \log_{10} \left(\frac{d_L}{10 \text{ pc}} \right). \quad (1.128)$$

The measured flux of a galaxy in a certain filter band does not only depend on its distance, but also on the relative position between the filter response and the redshifted SED. This means that a certain position at the SED can fall into different filter bands, if we observe them in our restframe or 10 pc away from the galaxy. We therefore need to apply a correction term, the k-correction K_{corr} , to the absolute magnitude to account for this effect, which then requires knowledge, or an approximation, of the spectrum of the source (see e.g. Blanton et al., 2003; Hogg et al., 2002; Oke & Sandage, 1968). Instead of magnitudes, we frequently use galaxy colors c_{ij} , i.e. the difference between two magnitude measurements (see e.g. Greisel, 2015)

$$c_{ij} = m_i - m_j = -\frac{5}{2} \log_{10} \left(\frac{f_i}{f_j} \right), \quad (1.129)$$

to quantify the photometry of a galaxy¹².

If the source is redshifted, the luminosity L_ν that we observe changes as (cf. Hogg et al., 2002)

$$L_\nu(\nu_e) = \left(\frac{4\pi d_L^2}{1+z} \right) f_\nu(\nu_0) \quad (1.130)$$

$$\nu_e = (1+z)\nu_0, \quad (1.131)$$

where ν_e is the frequency at emission and ν_0 is the observed frequency. The photometric filter function will therefore traverse the redshifted SED and the measured flux in the filter bands therefore represent a good predictor for the redshift of the galaxy.

A typical example of a photometric filter system is the filter set of the Dark Energy survey (see §8.2) shown in Fig. 1.9, where we overplot the SED of an elliptical galaxy. When the spectrum is observed in the respective photometric filter, the full spectral information of the SED is reduced to a set of flux measurements. Methods that estimate photometric redshifts thus need to rely on prominent features in the galaxy's SED that propagate through the filter set to derive a redshift. This is illustrated in Fig. 1.9 for the example of an elliptical galaxy. As can be seen, the SED shows a characteristic break around $\lambda = 4000 \text{ \AA}$. The stellar population of ellipticals is old and therefore dominated by cool red giant stars and their SED therefore shows strong absorption features at the blue end of the spectrum (see e.g. Schneider, 2006a; Greisel, 2015). When the galaxy is redshifted, this 4000 \AA break shifts to higher wavelength and traverses through the filter set. The colors of the galaxy in the filters around this characteristic feature, are therefore excellent predictors for the galaxy's redshift.

¹²In combination with the magnitude of e.g. the first band, colors are in my experience better input features for Machine Learning algorithms (see §2), than the set of magnitudes.

¹⁴http://www.iasf-milano.inaf.it/~polletta/templates/swire_templates.html

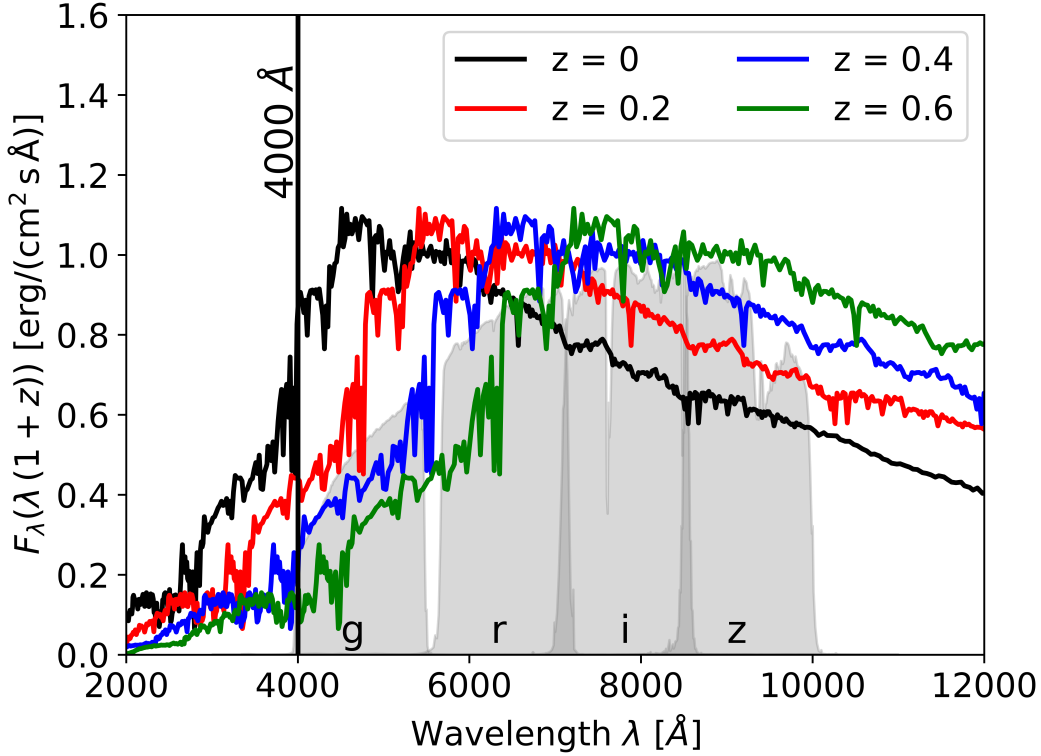


Figure 1.9: Filter response functions used in the DES Y1 analysis (see §8.2) plotted together with a SED template of a 2 Gyr old elliptical galaxy from the SWIRE library¹⁴(Silva et al., 1998; Polletta et al., 2007), at different redshifts. The SED was normalized at its value at 5500Å, the filter functions are normalized by the maximum i-band throughput. The solid vertical line shows the characteristic 4000 Å break at $z = 0$.

Not all galaxies show such a clear feature in their SED. For example spiral and irregular galaxies have a much less pronounced 4000 Å break, as their stellar population also contains young blue stars (see e.g. Greisel, 2015). It is therefore typically easier to derive photometric redshifts for red galaxies, compared with blue ones.

We distinguish two basic strategies to derive photometric redshifts from the available photometry. Template based methods fit model SEDs, or ‘SED templates’, to the observed photometry to derive a redshift and a spectral type. This is done by redshifting the SED template and then deriving synthetic photometry using the respective filter functions of the survey. We can then perform a χ^2 fit against the original photometry. The redshift of the template SED that best reproduces the original photometry, as measured by the χ^2 fit, is then reported as the final photometric redshift prediction.

The second class of photometric redshift algorithms are based on Machine Learning and use galaxies with available spectroscopic redshifts to ‘learn’ the mapping between the galaxies photometry and the redshift. In contrast to template fitting methods that model the SED of the galaxy, Machine Learning methods interpolate the photometry-redshift mapping using galaxies, for which both photometry as well as spectroscopic redshift measurements are available. The following section gives a brief overview over both methodologies and highlights their differences.

1.4.1 Template Fitting

The following subsection introduces the basic concepts of Template Fitting using concepts of Bayesian statistics that will be introduced in §2.1. The notation and derivations I present are similar, and consistent, with Benítez (2000); Feldmann et al. (2006a); Ilbert, O. et al. (2006); Greisel (2015), however I present a more explicit treatment of the different terms in view of §2.1. Template Fitting methods fit synthetic models of the galaxy's SED to the observed photometry to derive a redshift. Besides the redshift z of the galaxy, we would also like to predict its absolute magnitudes \mathbf{M} ; where we will summarize the target values as $\boldsymbol{\mu} = \{\mathbf{M}, z\}$ (see Greisel, 2015) and the input variables, i.e. the magnitudes and colors, as \mathbf{f} . Since not all galaxies have the same SED, the fitting routine uses a set of model SEDs that ideally is representative of all galaxy types in the respective catalog. These SED models, or SED 'templates', will be denoted as $\mathbf{T} = \{T_i, \text{ for } i \dots n_T\}$.

The goal of template fitting is to estimate the probability that a galaxy has a certain redshift and absolute magnitude $\boldsymbol{\mu}$ given its measured photometry \mathbf{f} . This information is expressed as the predictive distribution $p(\boldsymbol{\mu}|\mathbf{f})$, which can be decomposed as

$$p(\boldsymbol{\mu}|\mathbf{f}) = \sum_{i=1}^{n_T} \left[\frac{p(\mathbf{f}|\boldsymbol{\mu}, T_i)p(\boldsymbol{\mu}|T_i)p(T_i)}{p(\mathbf{f})} \right]. \quad (1.132)$$

Note that this equation has the form of Bayes theorem Eq. 2.4, where the left hand side is the posterior probability, or predictive distribution, that quantifies our knowledge about $\boldsymbol{\mu}$ of the galaxy after its photometry has been fitted by our model. The right hand side is a sum over all n template models T_i . For each individual model, we try to maximize the likelihood $p(\mathbf{f}|\boldsymbol{\mu}, T_i)$, i.e. the probability that the observed photometry \mathbf{f} is well reproduced by our model. This model is quantified by the respective template T_i and the redshift/absolute magnitude values $\boldsymbol{\mu}$. The posterior will however also depend on our prior knowledge about likely values for $\boldsymbol{\mu}$ given a particular galaxy template. This is quantified in the prior terms $p(\boldsymbol{\mu}|T_i)$ and $p(T_i)$, where $p(\boldsymbol{\mu}|T_i)$ quantifies our prior knowledge about reasonable redshift and absolute magnitude values for a given template galaxy, and $p(T_i)$ quantifies our prior knowledge about the correctness of the templates themselves.

The denominator $p(\mathbf{f})$ does not depend on our model and can therefore be seen as a normalization factor.

Assuming a uniform $p(T_i)$ and omitting the normalization $p(\mathbf{f})$, we can then simplify the equation as (cf. Benítez, 2000; Greisel, 2015)

$$p(\boldsymbol{\mu}|\mathbf{f}) \propto \sum_{T=1}^n p(\mathbf{f}|\boldsymbol{\mu}, T_i)p(\boldsymbol{\mu}|T_i). \quad (1.133)$$

The Gaussian likelihood is assumed to be of the typical χ^2 form (cf. Benítez, 2000; Ilbert, O. et al., 2006; Greisel, 2015)

$$p(\mathbf{f}|\boldsymbol{\mu}, T_i) \propto \exp(-\chi^2/2), \quad (1.134)$$

where the $\chi^2(z)$ function is given as (cf. Ilbert, O. et al., 2006; Greisel, 2015)

$$\chi^2(z) = \sum_{i \in \mathcal{F}} \left[\frac{f_{\text{obs},i} - \alpha f_{T_i,i}(z)}{\sigma_{\text{obs},i}} \right]^2. \quad (1.135)$$

Here \mathcal{F} denotes the set of photometric filters and α is a fitting parameter that scales the model flux to match the observations.

The accuracy of the derived photometric redshift posterior $p(\mu|\mathbf{f})$ depends on the availability of a template set \mathbf{T} that is representative of the galaxy SEDs in the full sample. Furthermore we note that the likelihood function in Eq. 1.134 assumes that the photometry and the photometric noise of each galaxy are known and well calibrated. As the prior term in Eq. 1.133 needs to be manually chosen, we should further require that the derived redshifts are relatively robust against these choices.

In practice however it can be difficult to guarantee the validity of these modeling assumptions without comparing the photometric redshift estimates with accurate spectroscopic reference data.

Despite these limitations, template fitting remains one of the primary methods for photometric redshift estimation. It has the decisive advantage, that it uses physically motivated models for the SED of galaxies and it does not directly require the availability of spectroscopic data during model fitting. Over the years, a variety of different template fitting methods have been developed, which are to a large degree, consistent with the framework we just discussed and differ mainly in the way priors and template sets are constructed. Among the more popular codes are ‘LePhare’ (Arnouts et al., 1999; Ilbert, O. et al., 2006), ‘BPZ’ (Benítez, 2000; Benítez et al., 2004; Coe et al., 2006), ‘HyperZ’ (Bolzonella et al., 2000), ‘PhotoZ’ (Bender et al., 2001), ‘ZEBRA’ (Feldmann et al., 2006b) and ‘EAZY’ (Brammer et al., 2008). More recent developments (Leistedt et al., 2016) aim towards an even more aggressive Bayesian modelling of photometric redshift distributions of samples of galaxies.

Due to the good performances in recent data challenges (e.g. Sánchez et al., 2014; Bonnett et al., 2016; Hildebrandt et al., 2010), Machine Learning based photometric redshift estimation was established as a powerful alternative to template fitting methods. The following section gives a brief overview over this class of methods, and §2 provides a more detailed description of Machine Learning.

1.4.2 Machine Learning methods

Machine Learning based photometric redshift estimation uses data from the overlapping regions between the photometric and the spectroscopic reference survey, to ‘learn’ the global mapping between the photometry and the redshift. This mapping can be derived using a wide variety of algorithms and methods like Neural Networks (e.g. Firth et al., 2003; Collister & Lahav, 2004; Bonnett, 2015), Support Vector Machines (Wadadekar, 2005; Wang et al., 2008), Self Organizing Maps (Carrasco Kind & Brunner, 2014a; Speagle & Eisenstein, 2015a,b), tree based methods (e.g. Carliles et al., 2010; Gerdes et al., 2010; Carrasco Kind & Brunner, 2013; Rau et al., 2015) or Diffusion Maps (Richards et al., 2009; Freeman et al., 2009).

While these methods differ in the way the mapping between photometry and redshift is derived, their basic application is model independent. Their underlying flexible model function is fit to a training dataset that provides measurements of both photometry and redshift. Once the model is optimized, it can then provide redshift predictions for new galaxies without spectral information. The Machine Learning approach therefore does not depend on choosing a set of representative templates to obtain a photometric redshift.

Instead Machine Learning derives, or ‘learns’, the mapping between photometry and redshift using a training dataset, that has to be representative of the sample of galaxies, for which photometric information needs to be obtained. If the training data does not accurately

resemble the full photometric catalog, i.e. because some regions of the training set color space are insufficiently populated by spectroscopic redshifts, we can expect large prediction errors for galaxies that reside in these regions. Since the training data of a photometric sample is typically obtained from spatially overlapping spectroscopic surveys, this situation is not the exception, but rather the norm. Taking spectra at the faint end of the color-magnitude space requires long exposure times and the target selection of spectroscopic surveys is therefore typically not well optimized to generate representative spectroscopic calibration data for large area photometric surveys. Instead these surveys often preferentially target certain galaxy types like e.g. emission line galaxies (e.g. Escoffier et al., 2013) or high redshift quasars.

While these effects are often discussed in the context of Machine Learning based photometric redshift estimation, they more importantly affect the process of validating photometric redshift estimates using spectroscopic reference data. Since every method of photometric redshift estimation makes implicit assumptions that can be violated in the data, accurate photometric redshift validation is a vital prerequisite to use photometric redshifts for applications in cosmology, irrespective of the chosen photometric redshift methodology. As discussed in more detail in §5.1, there exists a rich literature about different photometric redshift methods, but comparatively little work has been done on analyzing sources of error in photometric redshift estimation and validation. This thesis therefore contributes to the literature by introducing methods to more accurately quantify the error distribution in photometric redshift estimates by proposing accurate Machine Learning methods that significantly outperform more traditional methodologies in §3. In §4 we then propose resampling methods to incorporate the error in photometric redshift distributions into cosmological parameter constraints. In the final chapter §5 we study sources of error in the context of the DES photometric redshift validation pipeline and analyse their impact on cosmological observables.

Since the following chapters make frequent use of basic concepts of Machine Learning, we provide a brief introduction into its most fundamental aspects in the following chapter.

Chapter 2

Machine Learning

The process of extracting patterns and relationships from data is the essence of empirical sciences since the mid of the 16th century. Since then, data analysis has laid the foundation for modern physical sciences, induced drastic changes in society and ultimately contributed to the unprecedented standard of living in many parts of the world entering the 21th century. The methods and techniques developed by pioneers like Edmond Halley, studying the life expectancy in Breslau around 1693 (Senn, 2003), or Legendre and Gauss, who developed the concepts of regression to study planetary orbits (Hald, 2008) around 1809, are the foundations of data analysis to this day. The rapid development of computer science and Machine Learning (ML) is currently revolutionizing the well established field of data analysis. For the first time in history, tasks like classification or the derivation of regression functions can be performed by computer algorithms in a self-guided manner. If used correctly, and in conjunction with scientific intuition, these techniques can be an invaluable tool to efficiently and accurately process and analyze large amounts of scientific data.

Especially data intensive branches of astrophysics and cosmology benefit from the application of ML to a wide variety of problems like the classification of galaxies (e.g. Huertas-Company et al., 2011; Dieleman et al., 2015), variable stars (e.g. Richards et al., 2011; Sesar et al., 2017), supernovae (e.g. Kessler et al., 2010; Charnock & Moss, 2017) and photometric redshift estimation (Collister & Lahav, 2004; Rau et al., 2015; Hoyle et al., 2015a), just to name a few.¹ This chapter will give a brief introduction into Machine Learning, to prepare the reader for its specific application to the problem of photometric redshift estimation, presented in the following chapters. As Machine Learning is based on statistics, we start the discussion with a quick primer on basic probability theory.

References and Resources The following introduction in §2.1 is based on Sivia & Skilling (2006); Arnborg & Sjödin (2001); Horn (2003) and Bishop (2006), the subsequent sections §2.2, §2.3 and §2.4 largely follow Bishop (2006). For §2.5 there exist many good descriptions and resources, most notably Bishop (2006); Hastie et al. (2009). In my personal opinion, they are either too technical (Hastie et al., 2009) or too simplistic (Bishop, 2006) for my purposes. §2.5 therefore adapts aspects from both references. In the discussion about sample selection bias in §2.6.1, I do not follow a particular reference, however I recommend Quionero-Candela

¹We refer to Ball & Brunner (2010) for a detailed review over the various applications of Machine Learning in astronomy.

et al. (2009) as a general reference on this topic. The discussion on input noise in §2.6.2 follows Pischke (2007) with slightly adapted notation, however there exist numerous more general resources, where I recommend Carroll et al. (2006). The description on unconditional density estimation in §2.7 is based on Bishop (2006). References for the brief introduction into conditional density estimation presented in §2.7.1 are Rosenblatt (1969); De Gooijer & Zerom (2003); Holmes et al. (2012).

Nearly all analyses that I performed in this thesis use the following software packages: ‘python’ (Rossum, 1995), ‘SciPy’ (Jones et al., 2001–), ‘Numpy’ (van der Walt et al., 2011), ‘IPython’ (Perez & Granger, 2007), ‘Matplotlib’ (Hunter, 2007), ‘Pandas’ (McKinney, 2010), ‘scikit-learn’ (Pedregosa et al., 2011), ‘R’ (R Core Team, 2015).

For the preparation of the images I used the ‘Matplotlib’ package (Hunter, 2007). Fig. 2.2 was created incorporating a code snippet from Joe Kington². To create Fig. 2.4 and Fig. 2.5 I used the scikit-learn (Pedregosa et al., 2011) package.

2.1 Basic probability theory

The classical view on statistics defines the probability of an event as the frequency of its occurrence in a repeatable random experiment. However, in astrophysics and cosmology we often deal with scenarios, where this perspective seems to be unnatural. As an example consider the problem of measuring the distance to a supernovae using its lightcurve. The distance itself is just a fixed number and not a random variable and the experiment cannot be practically repeated, since the object under consideration exploded. While it is certainly possible to phrase this problem in terms of a random experiment, it seems more natural to regard the uncertainty in the derived distance as our degree-of-belief in the measurement process and our modeling. This ‘Bayesian’ interpretation of probability goes back to 19th century pioneers like Bernoulli, Bayes and Laplace, but was rejected by many of their colleagues at that time, because the definition of probability based on frequencies seemed to be the more rigorous approach (Sivia & Skilling, 2006). Even today, Bayesian statistics is often criticized as being subjective. As it interprets probability as a ‘degree-of-belief’, one can indeed correctly criticize that this notion might change from person to person. This is however not a criticism on the Bayesian approach itself, but on improper quantification and justification of the presumptions in its practical application (see e.g. Jaynes, 1989; Sivia & Skilling, 2006).

In 1946, Richard Cox derived rules³ that need to be fulfilled to ensure that our reasoning within Bayesian statistics is consistent and meaningful (Cox, 1946). As a first rule he required that our degree-of-belief in a statement should be measured in real numbers and has to depend on all prior information we have on the matter. If new information is obtained we have to update our assessment in a sensible way. Finally we should derive our belief in a statement in a consistent manner, i.e. if there exist more than one way to reason, all approaches have to yield the same result (see also Arnborg & Sjödin, 2001). It can be shown (see e.g. Horn, 2003; Sivia & Skilling, 2006), that Cox’s rules can be related to the basic rules of probability that we will discuss in the following.

²https://github.com/joferkington/oost_paper_code/blob/master/error_ellipse.py, Accessed: Mo 28 Aug 2017

³An excellent review over the Cox theorems can be found in Arnborg & Sjödin (2001); Sivia & Skilling (2006). Especially Sivia & Skilling (2006) provides an excellent introduction into Bayesian statistics on which some of the following discussion is based.

If we specify our confidence that a certain statement X is true, we automatically quantify our doubt in the reverse statement \bar{X} . Measuring our degree-of-belief as a real valued number in the interval $[0, 1]$, then immediately leads to the summation rule (see Sivia & Skilling, 2006)

$$\text{prob}(X) + \text{prob}(\bar{X}) = 1, \quad (2.1)$$

where \bar{X} denotes the negation of the statement X . Consider the case where we have confidence that the proposition Y is true. If we can further quantify our confidence that the statement X is true, given that our proposition Y is true, we have implicitly quantified our confidence that both X and Y are true. Introducing the conditional probability $\text{prob}(X|Y)$, i.e. the probability of X given Y is true, we can write (see Sivia & Skilling, 2006)

$$\text{prob}(X|Y) \times \text{prob}(Y) = \text{prob}(X, Y). \quad (2.2)$$

If we further require that the joint distribution $\text{prob}(X, Y)$ is symmetric $\text{prob}(X, Y) = \text{prob}(Y, X)$ we immediately obtain Bayes theorem (see Sivia & Skilling, 2006)

$$\text{prob}(X|Y) \text{prob}(Y) = \text{prob}(Y|X) \text{prob}(X). \quad (2.3)$$

Eq. 2.3 is central for applications in parameter inference, where we want to gain information about the validity of a theory given new data. From Bayes theorem we can write (see Sivia & Skilling, 2006)

$$\text{prob}(\text{hypothesis}|\text{data}) = \frac{\text{prob}(\text{data}|\text{hypothesis}) \text{prob}(\text{hypothesis})}{\text{prob}(\text{data})}. \quad (2.4)$$

Here $\text{prob}(\text{hypothesis}|\text{data})$ is the posterior probability and quantifies our degree-of-belief in our hypothesis, given the available data. The term $\text{prob}(\text{hypothesis})$ is the prior probability and specifies how much we trust our hypothesis before we considered the data. The prior is then modified by the ‘likelihood’ $\text{prob}(\text{data}|\text{hypothesis})$ that gauges how well the data is supported by our hypothesis. The ‘evidence’ $\text{prob}(\text{data})$ doesn’t play an important role in parameter estimation and can be regarded as a normalization factor (see Sivia & Skilling, 2006). However we will consider this term in more detail, when we discuss how different models can be compared in §2.4.3. The right hand side of Eq. 2.4 therefore reconciles our prior degree-of-belief in our hypothesis by the information gained from the new data via the likelihood. Instead of quantifying our prior confidence, or if we are largely oblivious about our hypothesis, we can instead simply adapt our hypothesis, such that the likelihood $\text{prob}(\text{data}|\text{hypothesis})$ is maximized for the current data. This ‘maximum likelihood’ approach (see Bishop, 2006; Sivia & Skilling, 2006) treats the optimized hypothesis as a fixed value, in contrast to the Bayesian approach, where the posterior is a probability function as explained shortly.

We would now like to generalize Eq. 2.1 to multiple statements by considering the sum over the joint probabilities (see Sivia & Skilling, 2006)

$$\text{prob}(X, Y) + \text{prob}(X, \bar{Y}), \quad (2.5)$$

which can be decomposed using Eq. 2.2 as (see Sivia & Skilling, 2006)

$$\text{prob}(X, Y) + \text{prob}(X, \bar{Y}) = (\text{prob}(Y|X) + \text{prob}(\bar{Y}|X)) \text{prob}(X). \quad (2.6)$$

The first term on the right hand side has to sum to unity following Eq. 2.1 and we obtain (see Sivia & Skilling, 2006)

$$\text{prob}(X, Y) + \text{prob}(X, \bar{Y}) = \text{prob}(X). \quad (2.7)$$

Thus if we sum over all possible states of Y , the result will no longer depend on Y , i.e. Y has been ‘marginalized’ over.

The previous discussion introduced the notion of probability in abstract terms as a degree-of-belief in a statement. In the following discussion we will make this notion more concrete and instead use the concept of a random variable X that represents a range of values distributed according to $p(x)$. A discrete random variable X can have multiple discrete states s , where we quantify the probability of each state by the probability mass function $p(X = s)$. In contrast, a continuous random variable can take on a continuous range of values x and is therefore described by a probability density function (pdf), denoted as $p(x)$. An example for a discrete random variable is the cube number in a gambling game, whereas a continuous variable could report the height of a random individual drawn from a large population.

Considering two discrete random variables X and Y , having n and m states respectively, the sum and marginalization rules then generalize as (see Sivia & Skilling, 2006)

$$\sum_k^n p(X = s_k) = 1 \quad (2.8)$$

$$\sum_k^m p(Y = s_k, X) = p(X). \quad (2.9)$$

In the case of a continuous random variable, the sum over discrete states is replaced by an integral over a continuum as (see Sivia & Skilling, 2006)

$$\int_{-\infty}^{\infty} dx p(x) = 1 \quad (2.10)$$

$$\int_{-\infty}^{\infty} dy p(x, y) = p(x). \quad (2.11)$$

The probability that e.g. the continuous random variable X takes on a value within the interval $x \in (a, b)$ is therefore given as (see Sivia & Skilling, 2006)

$$p(x \in (a, b)) = \int_a^b p(x) dx. \quad (2.12)$$

The Normal Distribution

In many applications of data analysis, random variables are distributed according to a normal, or Gaussian, distribution $\mathcal{N}(x|\mu, \sigma^2)$ where (see Bishop, 2006)

$$p(x) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x - \mu)^2\right). \quad (2.13)$$

The shape of this function is characterized by its mean μ and standard deviation σ (see Bishop, 2006)

$$\mu = \int_{-\infty}^{\infty} x p(x) dx \quad (2.14)$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx}, \quad (2.15)$$

as illustrated in Fig. 2.1. We note that the squared standard deviation $\sigma^2(X) = \text{Var}(X)$ is the variance of the random variable. If the input vector \mathbf{x} is D dimensional, the normal distribution $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ reads (see Bishop, 2006)

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right). \quad (2.16)$$

The mean μ is now a d dimensional vector and the variance generalizes to a $d \times d$ dimensional covariance matrix Σ

$$\Sigma_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) p(x_i, x_j) dx_i dx_j, \quad (2.17)$$

that is required to be symmetric and positive semi-definite. The diagonal entries of Σ quantify the variance of the individual components and the cross entries their interdependence.

This can be visualized for a two dimensional normal distribution by considering the elliptical region that corresponds to the 1σ interval in the one dimensional case. In 2D, the orientation and size of the elliptical 1σ region is determined by the entries of the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \quad (2.18)$$

To illustrate this, Fig. 2.1 plots the 1σ covariance ellipses for a variety of zero mean normal distributions, where

$$\Sigma_{1/2} = \begin{pmatrix} 1 & \pm 0.5 \\ \pm 0.5 & 1 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix} \quad \Sigma_{4/5} = \begin{pmatrix} 1.5/0.1 & 0 \\ 0 & 0.1/1.5 \end{pmatrix}. \quad (2.19)$$

We see that the diagonal covariance matrices Σ_3 , Σ_4 and Σ_5 quantify a different error in x_1 and x_2 , but are all aligned along the axes. In contrast the covariance ellipses for Σ_1 , Σ_2 show a correlation between the variables x_1 and x_2 . If the cross term Σ_{12} in Eq. 2.18 is positive, the variables are said to be correlated. This means that a large x_1 makes it likely that the corresponding x_2 variable will also be large. In contrast a negative cross term indicates anticorrelation, i.e. a large x_1 makes it likely that the corresponding x_2 will be small. We further note, that the ellipses that correspond to $\Sigma_{1,2}$ have a different area than those that correspond to $\Sigma_{4,5}$ and Σ_3 . If x_1 and x_2 are interpreted as parameters that need to be estimated, covariance ellipses with different areas therefore describe measurements that have a different total accuracy. We will regularly encounter these covariance ellipses in §4, when we discuss how photometric redshift uncertainties affect cosmological parameter constraints.

We now consider the problem of estimating the parameters in Eq. 2.13 using a sample of values that was drawn independently from the same normal distribution⁴. Denoting the

⁴Such a sample is called independent and identically distributed, which we will always assume to be the case, if not stated otherwise.

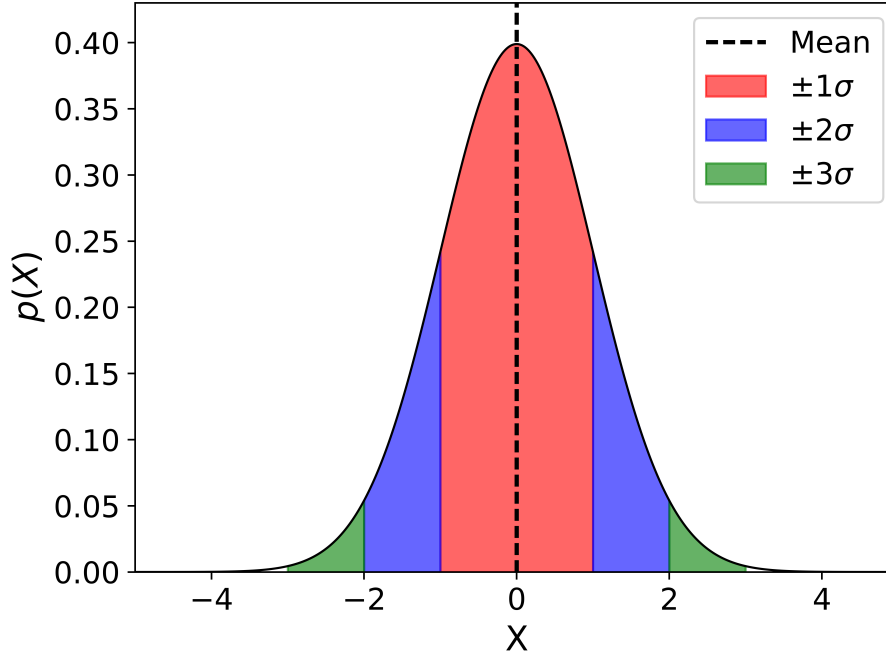


Figure 2.1: A normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1.0$. The colored regions show the $1\sigma - 3\sigma$ intervals, that contain 68.27%, 95.45% and 99.72% of the area under the curve, as measured from the mean.

gaussian model as $\mathcal{N}(x|\mu, \sigma^2)$, we can formulate the likelihood, i.e $p(\text{data}|\text{hypothesis})$, as (see Bishop, 2006)

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2) . \quad (2.20)$$

If we place the Gaussian at a wrong position in parameter space, its likelihood will be small. In contrast, if we guess the parameters correctly, we will obtain a large likelihood value, since the majority of samples will fall into a region where the probability density function is large, as can be seen in Fig. 2.1. Maximizing the likelihood is therefore a very reasonable and natural criterion to select the model parameters. Since the product over normal distributions is a bit difficult to handle both analytically and numerically, we instead maximize the logarithm of the likelihood, or log-likelihood (see Bishop, 2006)

$$\log p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log (2\pi) . \quad (2.21)$$

In general the log-likelihood is maximized simultaneously for all parameters in the model, however for the special case of a normal distribution this optimization can be done separately for the mean and the standard deviation. Taking the derivative with respect to the log-likelihood in Eq. 2.21, we can easily obtain the optimal solutions ⁵ for μ and σ^2 as (e.g.

⁵Note that the maximum likelihood estimate for the variance is biased, because it is estimated with respect to the sample mean μ_{ML} instead of the true mean. A better estimate can be obtained using $\sigma_{\text{corr}}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$. This underestimation of variance is a general problem of the maximum likelihood approach and we refer to the literature (e.g. Bishop, 2006) for further details.

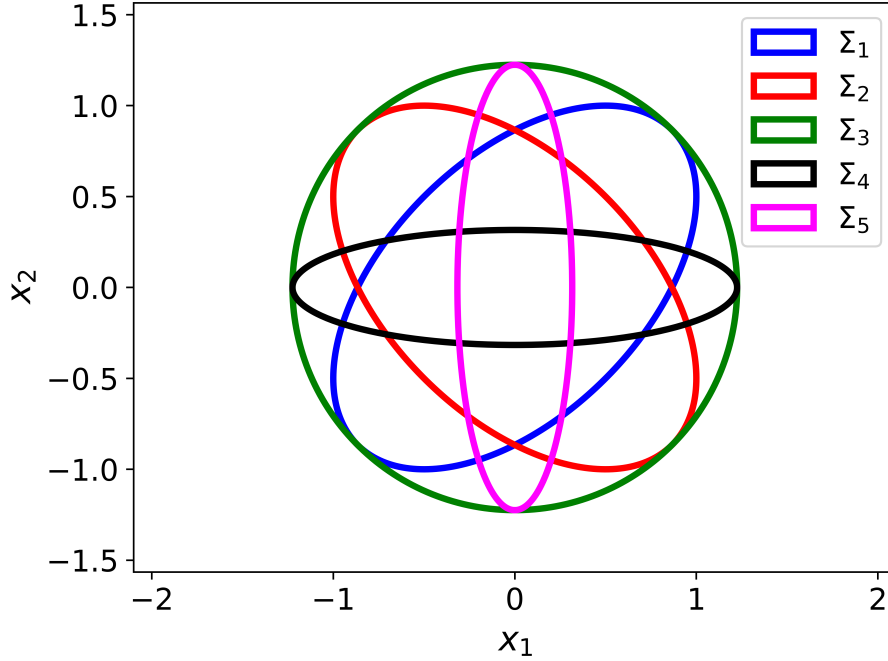


Figure 2.2: Elliptical 1σ regions described by the covariance matrices defined in Eqs. 2.19.

Bishop, 2006)

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad (2.22)$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2. \quad (2.23)$$

In this example we considered the task of estimating the parameters of a normal distribution. However, in Machine Learning applications we want to infer a functional mapping between a set of inputs and a target variable. In photometric redshift estimation for example, we predict the redshift of a galaxy given its photometry. The redshift cannot be predicted perfectly based on the photometry in a limited number of filter bands. As a result, the photometric redshift predicted for a certain galaxy will be distributed according to a probability density function. Since the photometric redshift z_{phot} of the galaxy depends on its photometry \mathbf{f} , this ‘predictive distribution’ is a conditional probability density function $p(z_{\text{phot}}|\mathbf{f})$.

If we assume the conditional pdf to be Gaussian, its shape will be completely determined by its mean and variance. In contrast to the unconditional case discussed previously, the conditional pdf depends on a set of inputs, e.g. the photometry \mathbf{f} . As a result, the two parameters that describe its shape, will be functions of these input variables. This problem of ‘regression’ will be discussed in a broader context in §2.2.

The Bernoulli Distribution

Besides continuous probability density functions, many random processes have discrete states and are therefore described by discrete random variables. The classical example of a discrete

random variable in the context of Machine Learning is the class of an object like the morphological type of a galaxy, or the type of a variable star. One of the simplest discrete random experiments is the coin toss, where a potentially biased coin is thrown to yield either heads or tails. If we parametrize the possible states as $\mathbf{s} = \{0, 1\}$, the probabilities for the individual outcomes $p(X = 1) = \mu$ and $p(X = 0) = 1 - \mu$ depend only on the ‘success rate’ μ (e.g. Bishop, 2006). The corresponding probability distribution for this process is then given as (e.g. Bishop, 2006)

$$p(X = s) = \mu^s (1 - \mu)^{1-s} . \quad (2.24)$$

For a dataset $\mathcal{D} = \{x_1, \dots, x_N\}$, the likelihood of the Bernoulli distribution can be obtained as (see Bishop, 2006)

$$p(\mathcal{D}|\mu) = \prod_{i=1}^N \mu^{x_i} (1 - \mu)^{1-x_i} , \quad (2.25)$$

and the value for μ that maximizes the log-likelihood function (see Bishop, 2006)

$$\log p(\mathcal{D}|\mu) = \sum_{i=1}^N (x_i \log \mu + (1 - x_i) \log (1 - \mu)) , \quad (2.26)$$

is given by the sample mean

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i . \quad (2.27)$$

With the introduction of the normal distribution to describe continuous random variables and the Bernoulli distribution for discrete binary random variables, we have laid the foundation for the following discussion on Regression and Classification. Up until now, these probability functions have been unconditional, i.e. they depend on fixed valued parameters. In the following sections, we will instead describe e.g. the mean or the success rate, using functional models that depend on input variables \mathbf{x} . This simple modification will make the Normal distribution and Bernoulli distribution conditional on these sets of inputs \mathbf{x} , and enables us to use them for Regression and Classification tasks respectively.

2.2 Regression

The basic task of regression is to estimate the functional relationship between a given set of inputs x and a target variable t . In Fig. 2.3 we illustrate a typical regression setting, where the blue points correspond to the data that we want to model. We see that these datapoints do not follow an exact functional relationship, but instead have an intrinsic scatter around the true function shown in red. The modelling is therefore not restricted to determining this function, but also includes the estimation of the predictive distribution $p(t|x)$ of the target variable t given the input x . We follow Bishop (2006) in this section.

In analogy to the discussion in the previous section, we model the predictive distribution as a Gaussian, where the mean is given by a polynomial function $y(\mathbf{x}, \mathbf{w})$ and the variance β is assumed to be independent of \mathbf{x}

$$p(t|\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \sigma^2) . \quad (2.28)$$

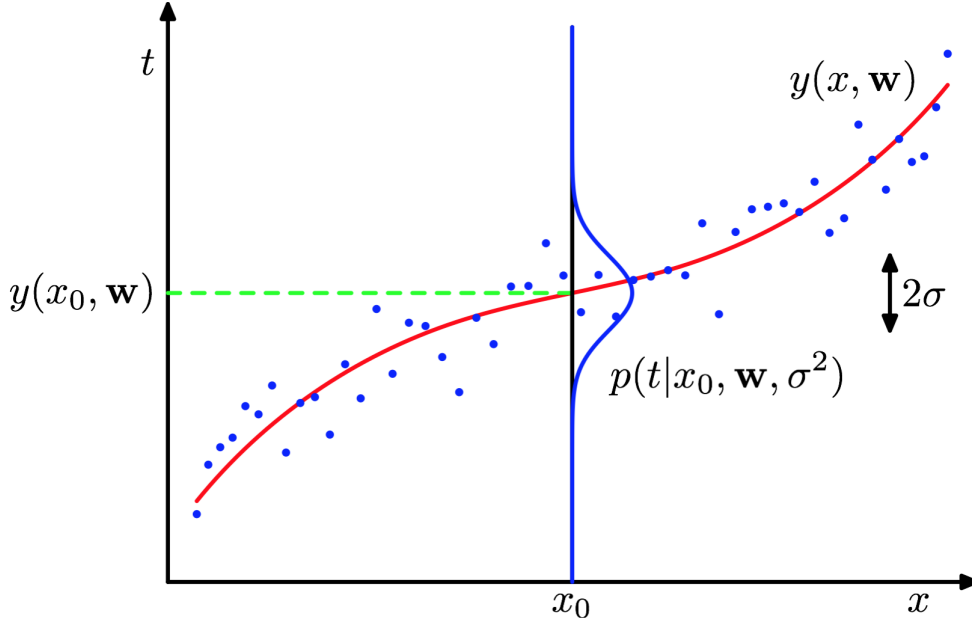


Figure 2.3: Illustration of a typical regression setting. The blue points illustrate the data, that needs to be modeled by the predictive distribution $p(t|x, \mathbf{w}, \sigma^2)$. The function $y(x, \mathbf{w})$ parametrizes its mean and depends on the model parameters \mathbf{w} . The parameter σ^2 is the variance of the predictive distribution and is assumed to be independent of the input x . Image credit: Christopher M. Bishop (Bishop, 2006), modified.

If we do not have any prior information about the functional form of $y(\mathbf{x}, \mathbf{w})$ we can assume a very complex function that will adapt to the data. A simple choice could be a polynomial model or, more general a linear basis function model

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \quad (2.29)$$

where $\mathbf{w} = (w_0, \dots, w_{M-1})^T$ is a M dimensional vector of free model parameters, or weights. In the corresponding basis function vector $\boldsymbol{\phi} = (\phi_0, \dots, \phi_{M-1})^T$ we define $\phi_0 = 1$ such, that the first weight is an offset to the function. The basis functions $\boldsymbol{\phi}$ themselves can be e.g. linear combinations, polynomial terms or Gaussians that depend on the input \mathbf{x} . To give an example, consider the case of a single input variable x , where we obtain a polynomial model by setting $\phi_j(x) = x^j$. The choice of suitable basis functions will strongly depend on the data at hand and has to be carefully chosen in advance to reflect our intuition about the data. With a linear basis function model for the mean of the predictive distribution, we obtain the likelihood as

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{i=1}^N \mathcal{N}(t_i|y(\mathbf{x}_i, \mathbf{w}), \sigma^2), \quad (2.30)$$

where the data \mathcal{D} consists of a set of N input variables $\mathbf{X} = \{\mathbf{x}_i \text{ for } i = 1, \dots, N\}$ with their associated target values $\mathbf{t} = \{t_i \text{ for } i = 1, \dots, N\}$. In analogy to the previous section, we

jointly optimize the log-likelihood with respect to the free parameters \mathbf{w} of the mean function $y(\mathbf{x}, \mathbf{w})$ and the variance⁶ σ^2 (e.g. Bishop, 2006)

$$\log(p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n, \mathbf{w}))^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi). \quad (2.31)$$

We can first optimize the log-likelihood with respect to the weights vector \mathbf{w} by minimizing the sum-of-squares error function

$$\text{SSE} = \frac{1}{2} \sum_{i=1}^N [t_i - y(\mathbf{x}_i, \mathbf{w})]^2. \quad (2.32)$$

For the linear basis function model this optimization can be performed analytically to yield the maximum likelihood estimate (see Bishop, 2006)

$$\mathbf{w}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}, \quad (2.33)$$

where the components of the matrix Φ are given as $\Phi_{nj} = \phi_j(\mathbf{x}_n)$ and \mathbf{t} denotes the column vector of the target observations $\{t_i \text{ for } i = 1, \dots, N\}$. In analogy to the previous section, the variance can then be estimated using the optimized model for the mean as

$$\sigma_{\text{opt.}}^2 = \frac{1}{N} \sum_{i=1}^N [t_i - y(\mathbf{x}_i, \mathbf{w}_{\text{opt.}})]^2. \quad (2.34)$$

We finally note that the sum-of-squares loss function Eq. 2.32 can also be used as a metric to evaluate the performance of the regression model, as discussed in more detail in §2.4.2. While the linear basis function model has an analytic solution for the optimal weights vector, this is in general not the case for other Machine Learning architectures and model functions, as we will see in the next section. In these cases, we have to resort to numerical optimization methods and efficient algorithms to calculate the necessary gradients. Furthermore, instead of a simple Gaussian, we can consider more complex distributions or mixtures of distributions, that may be advantageous in some applications. However besides changes in the parametrization of the predictive distribution and the optimization routine, all concepts that have been developed in this section remain valid.

2.3 Classification

In contrast to the regression setting, where the goal was to predict a continuous target variable, a classifier assigns discrete classes to the objects. In this section, we focus on binary classification for simplicity. However we will discuss in §3, how an ensemble of binary classifiers can be combined to produce multiclass outputs.

In analogy to the previous discussion on the regression task, we start by formulating the predictive distribution for binary classification. In §2.1 we introduced the Bernoulli distribution of a coin toss experiment, that depends only on a single free parameter μ which is the probability of the positive result, i.e. the success rate. Associating the states of a Bernoulli

⁶Here we assume that the variance is independent of \mathbf{x} for the sake of simplicity. In more complex scenarios, we can parametrize the variance as a function, just as we did for the mean.

distributed random variable with the target class values, the class probabilities for an object with features \mathbf{x} are given as $p(\text{class } 1|\mathbf{x}) = \mu$ and $p(\text{class } 2|\mathbf{x}) = 1 - p(\text{class } 1|\mathbf{x})$ (see Bishop, 2006). We can use the basis function model Eq. 2.29 to parametrize $p(\text{class } 1|\mathbf{x})$, if we scale its output such, that it represents probabilities (see Bishop, 2006).

This can be done by the sigmoid function, that scales the output of the linear basis function model to an interval $[0, 1]$ (see Bishop, 2006)

$$p(\text{class } 1|\mathbf{x}) = y(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))}, \quad (2.35)$$

where we defined the weights \mathbf{w} and the basis functions $\boldsymbol{\phi}(\mathbf{x})$ in analogy to Eq. 2.29.

Considering a dataset \mathcal{D} that consists of a set of N inputs $\mathbf{X} = \{\mathbf{x}_i \text{ for } i = 1, \dots, N\}$ and associated target class variables $\mathbf{t} = \{t_i \text{ for } i = 1, \dots, N\}$ with $t_i \in [0, 1]$, we can derive the Bernoulli likelihood as (see Bishop, 2006)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N y(\mathbf{x}_i, \mathbf{w})^{t_i} (1 - y(\mathbf{x}_i, \mathbf{w}))^{1-t_i}. \quad (2.36)$$

Taking the logarithm of Eq. 2.36 then leads to the so called cross-entropy error function defined as (see Bishop, 2006)

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \sum_{i=1}^N (t_i \log(y(\mathbf{x}_i, \mathbf{w})) + (1 - t_i) \log(1 - y(\mathbf{x}_i, \mathbf{w}))), \quad (2.37)$$

which is maximal⁷ if the prediction is perfect $y(\mathbf{x}_i, \mathbf{w}) = t_i$ and penalized if the classifier predicts the wrong class. In contrast to the linear basis function model, it is not possible to derive an analytical solution for the weights \mathbf{w} that optimizes the log-likelihood, due to the nonlinearity introduced by the sigmoid (see Bishop, 2006). However the optimization can be performed efficiently by the Newton-Raphson method as detailed in the literature (see Bishop, 2006). While the discussion in this section focussed on the two-class problem, the generalization to multiple classes is streight-forward, and the interested reader is referred to the standard literature on the topic (e.g. Bishop, 2006; Hastie et al., 2009)

We finally note that the evaluation of classifiers is more complex than in the regression setting. Consider for example a very simple metric that measures classification accuracy as the average fraction of incorrect classifications. While this approach is sensible, if we can guarantee an equal number of objects in each class and balanced classification performance for all classes, it turns out to be suboptimal in the majority of applications, where this is not the case. This can be easily understood by considering the hypothetical example of automatically detecting diseases, like HIV or Hepatitis, in a blood donation sample. In general only a small fraction of the active blood donors will have one of these diseases, due to the extensive screening process before the donation. Let us for instance assume that only 0.1% of the blood donors are infected and thus a ‘classifier’ that always predicts a clean blood sample, will only make mistakes 0.1% of the time. A naive interpretation would lead us to believe that we have a very accurate classifier, whereas in reality the classifier performs catastrophically. On the other hand we don’t want to mistakenly diagnose too many clean blood samples as being infected, due to

⁷Instead of maximizing the log-likelihood as done here, it is sometimes customary to minimize the negative log-likelihood in accordance with the notion of a ‘loss function’ that needs to be minimized.

the enormous psychological stress imposed on the blood donors. Clearly the problem arises because the dataset we are considering is extremely imbalanced, which makes loss functions like the misclassification rate or the cross entropy error suboptimal choices to interpret the accuracy of the predictions.

Let us denote the subset of infected blood samples as the ‘true class’ and the subset of clean blood samples as the ‘false class’. When a classifier is applied to this problem, its predictions can be partitioned into four subsets. The ‘true positive’ (tp) sample contains all infected samples that have been correctly classified as being infected. Accordingly the ‘true negative’ (tn) sample will contain all samples that have been correctly identified as being clean. If these two subsamples already constitute the full sample, the classifier provides perfect predictions.

In the presence of classification error, we will encounter a fraction of objects that are incorrectly classified as being infected, but are in reality clean. We call this subsample the false positives (fp). Equivalently the false negatives (fn) are those samples, that are incorrectly classified as being clean, but are actually infected.

Counting the number of samples that fall into each category, we can define two performance metrics that correctly account for class imbalance. The ‘purity’ metric measures which fraction of those samples that have been classified as being infected, are actually infected (e.g. Witten et al., 2011)

$$\text{purity} = \frac{tp}{tp + fp}, \quad (2.38)$$

and the completeness of a sample measures how many of the total number of infected samples have been correctly identified as such (e.g. Witten et al., 2011)

$$\text{completeness} = \frac{tp}{tp + fn}. \quad (2.39)$$

To obtain a summary statistic we define the F1-score as the harmonic mean between purity and completeness as (e.g. Witten et al., 2011)

$$F1 = 2 \left(\frac{\text{purity} \cdot \text{completeness}}{\text{purity} + \text{completeness}} \right). \quad (2.40)$$

The classification performance in terms of the F1-score is worst if $F1 = 0$ and best if $F1 = 1$. We finally note that in the Machine Learning literature, purity is usually referred to as the precision and completeness is referred to as recall.

This concludes our brief introduction into the basic concepts of regression and classification. In the following section we will introduce the reader to sources of error that have to be considered when Machine Learning algorithms are applied in practise.

2.4 Model Complexity

A critical question in Machine Learning is how complex of a model should be in order to fit the data appropriately. In the basis function model described in the previous section, one can increase the complexity by adding additional basis function terms. If we do not have a clear idea about the underlying structure of the data, a naive approach would select a very complex model in order to guarantee a perfect fit to the data. However this methodology runs a severe risk of overfitting, where the model fits the intrinsic noise of the data, rather than the underlying structure.

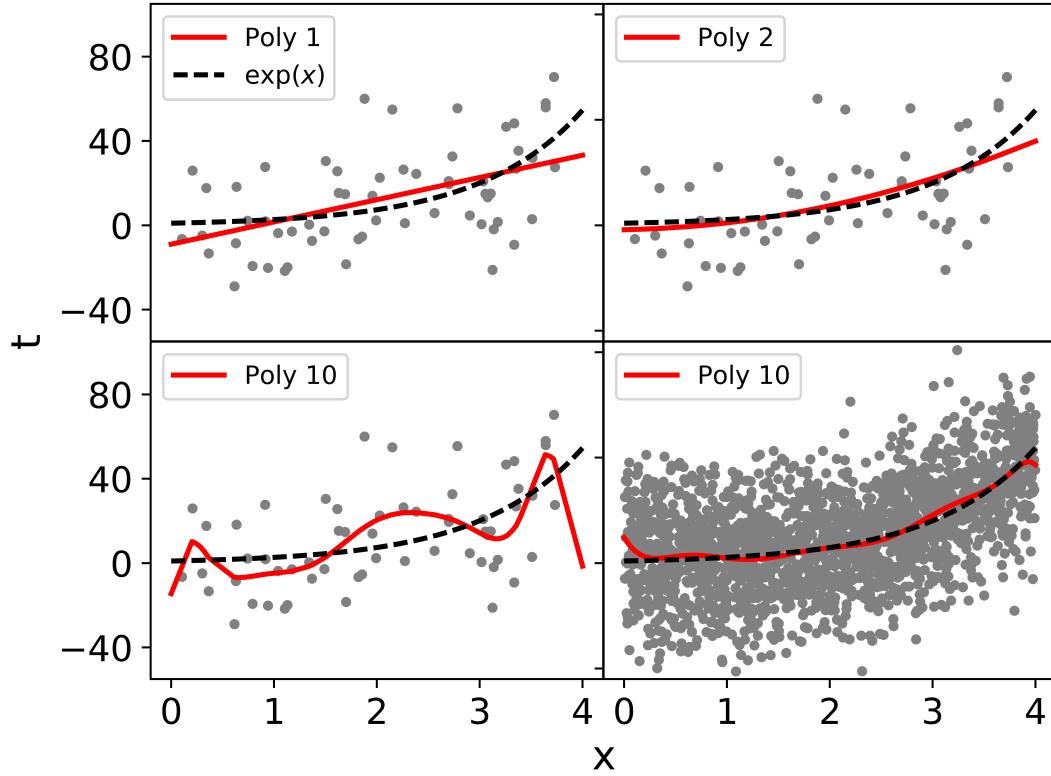


Figure 2.4: Illustration of the impact of model complexity on the accuracy of Machine Learning models. The black dashed line shows the exponential function from which $N = 60$ grey data points have been generated by adding zero mean Gaussian noise with $\sigma = 20$. The red curves show the polynomial models of order M that have been fitted to this data set. The bottom right panel shows the performance of a polynomial model with $M = 10$ fitted on a larger sample with $N = 2000$.

This issue is illustrated in Fig. 2.4, where we estimate an exponential function from the underlying data shown in grey using a polynomial model. While the simple linear model (‘Poly 1’) is not complex enough to capture the nonlinear structure in the data, adding higher order polynomial terms does not always improve the quality of the recovered model, but can instead lead to overfitting as shown in the lower left panel for a 10th order polynomial (‘Poly 10’). Here the complex model is adapting too tightly to the small amount of available data and thus strongly fluctuates around the truth. For this small sample, a quadratic model (‘Poly 2’) produces the best fit. Increasing the data sample to $N = 2000$ points, significantly stabilizes the high order polynomial, which now perfectly recovers the underlying exponential function. This simple experiment demonstrates, that the complexity of the model needs to be balanced with the nonlinearity of the data, as well as the number of available data samples. If we have more data available, we can increase the model complexity to more accurately capture the complex structure in the data. In contrast, if data is sparse, increasing the complexity of the fit, can even lead to greater errors than a more simplistic model.

2.4.1 The Bias-Variance Tradeoff

In order to gain more insights into the tradeoff between model complexity and fitting accuracy, I follow Bishop (2006) in this subsection, and consider the sum of squared error function (Eq. 2.32) for a sample that is distributed according to $p(\mathbf{x}, t)$

$$E[L] = \int \int [y(\mathbf{x}) - t]^2 p(\mathbf{x}, t) d\mathbf{x} dt, \quad (2.41)$$

where $y(\mathbf{x})$ is assumed to be a very flexible model function. We can determine the function that minimizes Eq. 2.41 by setting the variational derivative of $E[L]$ with respect to $y(\mathbf{x})$ to zero (see Bishop, 2006)

$$\frac{\delta E[L]}{\delta y(\mathbf{x})} = 2 \int [y(\mathbf{x}) - t] p(\mathbf{x}, t) d\mathbf{x} dt = 0. \quad (2.42)$$

Rearranging terms using Eq. 2.11 and Eq. 2.2 we obtain

$$y(\mathbf{x}) = \int t p(t|\mathbf{x}) dt. \quad (2.43)$$

Thus the function that minimizes the sum of squared error is given by the conditional mean of the predictive distribution $p(t|\mathbf{x})$ irrespective of its shape. This generalizes the result we obtained earlier, where we derived the sum-of-squares error function for the specific model of a Gaussian predictive distribution. We note that this derivation implicitly assumes knowledge of the joint distribution $p(\mathbf{x}, t)$ and imposes no restrictions on the form of $y(\mathbf{x})$. The function Eq. 2.43 can therefore be interpreted as the best performing model that would be obtained, if a very complex model is fitted to an infinitely large dataset.

If the dataset is limited, we will in general not obtain this perfect solution, which we show by evaluating the expected sum-of-squared error for our estimated, potentially suboptimal, model $\hat{y}(\mathbf{x})$

$$E[L] = \int \int [\hat{y}(\mathbf{x}) - t]^2 p(\mathbf{x}, t) d\mathbf{x} dt, \quad (2.44)$$

and expanding the loss function using the optimal prediction $y(\mathbf{x})$

$$E[L] = \int (\hat{y}(\mathbf{x}) - y(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} + \int \int (y(\mathbf{x}) - t)^2 p(\mathbf{x}, y) d\mathbf{x} dt. \quad (2.45)$$

The first term on the right hand side characterizes the deviation of our estimate from the optimal solution, i.e. the conditional mean. The second term describes the intrinsic, irreducible noise in the data that does not dependent on our estimate $\hat{y}(\mathbf{x})$.

Obtaining near optimal regression functions in real world applications is primarily complicated by the limited amount of data available to construct an estimate $\hat{y}(\mathbf{x})$ for the function $y(\mathbf{x})$. As discussed in the last section, this leads to two forms of error that largely depend on the complexity of the model. If we select a too simplistic model, we will not be able to fit the complex underlying structure of the data. In contrast, if we select a too flexible model, we will overfit the intrinsic noise in the data. These two forms of error can be disentangled by taking a frequentist position and assuming that we fit a certain model to a large number of datasets, each generated by the same underlying distribution. If we take the expectation

value with respect to this ensemble of datasets \mathcal{D} , we can decompose the total expected loss as

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise} \quad (2.46)$$

where (see Bishop, 2006)

$$(\text{bias})^2 = \int \{E_{\mathcal{D}}[\hat{y}(\mathbf{x}; \mathcal{D})] - y(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} \quad (2.47)$$

$$\text{variance} = \int E_{\mathcal{D}} \left[\{\hat{y}(\mathbf{x}, \mathcal{D}) - E_{\mathcal{D}}[\hat{y}(\mathbf{x}; \mathcal{D})]\}^2 \right] p(\mathbf{x}) d\mathbf{x} \quad (2.48)$$

$$\text{noise} = \int \int \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt. \quad (2.49)$$

The bias term (Eq. 2.47) quantifies how well our estimate on average coincides with the conditional mean. This term is large, if we select a model that is not complex enough to capture the underlying structure in the data and it is small, if we consider complex models. The variance (Eq. 2.48) does not depend on the true function $y(\mathbf{x})$ and quantifies how strongly the fitted models vary between the generated datasets. This variance term is in general large, if a too complex model is fitted to a relatively small dataset as shown in the lower left panel of Fig. 2.4. If the variance contributes significantly to the total error, it is advisable to consider less complex models that are more stable and thus have a smaller variance or, even better, collect more data. The intrinsic noise (Eq. 2.49) does not depend on our estimate $\hat{y}(\mathbf{x}; \mathcal{D})$ and represents the irreducible error in the data. We conclude, that in order to optimize the complexity of the model, we have to trade-off the bias and variance terms that both contribute to the expected sum of squared loss for regression. This can be achieved by varying the complexity of the model and subsequently estimating the expected loss on unseen data using the methods described in the following section.

2.4.2 Evaluating Machine Learning Models

The bias-variance tradeoff (Eq. 2.46) decomposes the error of a Machine Learning model into three additive effects, i.e. the squared bias, the variance and the intrinsic error of the data. Especially the variance component makes the tuning and evaluation of Machine Learning models challenging. Since the error functions described so far, always favour more complex models, we have to perform training, evaluation and testing on three disjunct datasets. Otherwise we will select a too complex model during tuning, or more importantly underestimate the error of our model on unseen data. For instance consider fitting a highly complex model that simply connects all datapoints. This model will perform perfectly on this particular dataset, but very poorly on unseen data, drawn from the same parent distribution. We therefore split the complete data randomly into three disjunctive parts, the training data, the validation data and the test data. The model is trained on the training data, tuned on the validation data and, after all these steps have been finished, applied to the test data, to evaluate the performance of the model. We want to highlight, that it is very important to hold out and blind the test data from all prior steps of data processing and model tuning. While there are no strict rules which fraction of the data should reside in each of these catalogs, I suggest based on my own experience that using a 50%/25%/25% split for the respective training/validation/test samples yields good results.

If there is not enough data available to obtain model and performance estimates with reasonable statistical accuracy, by simply splitting the available data, we can instead use the

k -fold cross-validation technique (e.g. Bishop, 2006; Hastie et al., 2009; Witten et al., 2011). The data is randomly split into k disjunctive parts. We train the model on $k - 1$ of them and apply the trained model to the single remaining subset to obtain a performance estimate. We then continue to hold out the next subsample and train the model on the remaining parts until we have used each of the k subsamples once for testing purposes. The result of this algorithm are then k performance estimates that allow us to obtain a mean performance as well as a quantification of its statistical uncertainty, where a typical number of cross validation folds is $k = 10$. We note that if we want to incorporate a model tuning step into this algorithm, we can apply the k -fold cross validation technique during each training step on the respective training data. This effectively nests two k -fold cross validation runs, where the outer one is used to obtain k performance estimates of the model performance, and the inner ones are applied to each of the respective training sets to select the best model. This ensures that the outer test sets are never used during their respective training runs.

Besides estimating the performance of a Machine Learning algorithm on unseen data, we would also like to obtain confidence intervals to report the error of our model. A popular method to derive these confidence intervals is the bootstrap technique (Hastie et al., 2009). From the training sample, we randomly select N elements with replacement, where N is the sample size of the training set. This sampling process is then repeated until we reach a large enough number of resampled datasets. We then fit our model function to each of these datasets to generate a model ensemble. We can then apply the test sample to each model in the ensemble to generate a variety of possible model predictions for each element in the test set. The distribution of predictions generated by this ensemble of models then quantifies our uncertainty about the true model. We want to highlight that these confidence intervals are not estimates of the width of the predictive distribution, but of the accuracy of the model itself. In the case of regression, where we parametrized the mean of a gaussian predictive distribution as a flexible basis function model, these intervals would for example correspond to our uncertainty w.r.t. the true conditional mean. We finally note, that the bootstrap technique can also be used in the same manner as k -fold cross validation to estimate the model performance on unseen data. Here the training set contains all objects that have been selected during resampling and the test set consist of the disjunctive set of samples, that have not been selected. In this way we ensure that training and test set don't contain the same objects, which is a vital requirement for accurate model validation as noted earlier.

The k -fold cross-validation and bootstrap techniques are inherently frequentistic. While these methods are very powerful and almost always used in practise, we can gain further insights into the problem of model selection by considering an elegant Bayesian framework following Bishop (2006) in the next subsection.

2.4.3 Bayesian model comparison

In the previous sections, we always applied a single model to a dataset and derived a maximum likelihood solution for its parameter vector. As discussed in § 2.1 a fully Bayesian treatment would specify our confidence in the model via the prior and use the likelihood to derive the full posterior distribution. In the context of multiple models \mathcal{M}_i that are applied to the same dataset \mathcal{D} , we can calculate the posterior as (see Bishop, 2006)

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i). \quad (2.50)$$

Here $p(\mathcal{M}_i)$ is the prior model probability and $p(\mathcal{D}|\mathcal{M}_i)$ is the model evidence that quantifies how well the data supports this particular model. If we have no prior preference for a certain model, i.e. $p(\mathcal{M}_i)$ is uniform, we should select the model, that maximizes $p(\mathcal{D}|\mathcal{M}_i)$. Denoting the parameters of the respective model \mathcal{M}_i as \mathbf{w}_i , we can marginalize over the set of parameters as (see Bishop, 2006)

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}. \quad (2.51)$$

This integral can be approximated analytically, under a couple of simplifying assumptions. We assume that the posterior distribution $p(\mathcal{M}_i|\mathcal{D})$ is sharply peaked around the optimal weight vector \mathbf{w}_{opt} and has a width of $\Delta w_{\text{posterior}}$ and further, that the prior for each parameter has the form $p(\mathbf{w}|\mathcal{M}_i) = 1/\Delta w_{\text{prior}}$. For an M parameter model we can then obtain an approximation for the logarithm of the model evidence as (see Bishop, 2006)

$$\log p(\mathcal{D}|\mathcal{M}_i) \approx \log p(\mathcal{D}|\mathbf{w}_{\text{opt}}) + M \log \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right), \quad (2.52)$$

where we note that the ratio $\log(\Delta w_{\text{posterior}}/\Delta w_{\text{prior}})$ is negative as $\Delta w_{\text{posterior}} < \Delta w_{\text{prior}}$. To develop intuition about Eq. 2.52, consider a scenario where we fit a very complex model to a low number of data samples. In this case, the width of the prior would be large in comparison to the width of the posterior, as the model would be very tightly fit to the data. The likelihood term $p(\mathcal{D}|\mathbf{w}_{\text{opt}})$ would then be penalized by this ratio proportional to the number of parameters M in the model. In contrast a too simplistic model would be penalized less strongly by the second term, but would also produce a worse fit, which decreases the first term. Thus Eq. 2.52 reflects our basic scientific intuition, that a good model should not only fit the data well, but also minimize the number of free parameters.

2.5 Decision trees and the Random Forest

Machine Learning algorithms based on decision trees are one of the most important techniques used in Data Mining up to date (Wu et al., 2007)⁸. The simple idea behind these architectures is to separate the input space of the training set in such a way, that each cell contains objects of the same class or with similar target values. The tree based architecture then allows the algorithm to efficiently find the cell that corresponds to a new object. The training set objects that reside in the selected cell then constitute a reference sample of similar objects, from which we can derive a prediction. Despite their simplicity, tree based Machine Learning architectures often provide very accurate predictions; and visualizations of the tree structure can be an invaluable information for domain experts.

In the following discussion we will first describe the general workings of the decision tree algorithm and then move to their specific application to regression and classification. We denote the input space of our training set as \mathbf{x}_T and the target variables as t_T . The goal of the decision tree is then to partition the input space of the training set \mathbf{x}_T such, that the

⁸Even though other architectures like Deep Convolutional Neural Networks (see e.g. Schmidhuber, 2014) became more relevant in applications like image (e.g. Cun et al., 1990; Gu et al., 2015; Rawat & Wang, 2017) or speech recognition (Hinton et al., 2012), the Random Forest can still be regarded as one of the most influential Machine Learning algorithms, especially if input variables can be designed based on human intuition.

elements in each partition have similar target values, where classification and regression use different ‘measures of similarity’ as explained shortly.

The tree is grown recursively on the training data in the following manner:

1. start by considering the whole input space \mathbf{x}_T which represents the root node of the tree
2. perform an exhaustive search for the split in the variable i at position j that makes the elements in the two resulting partitions most similar, as quantified by the minimum of the ‘split selection function’ $\text{SSF}(i, j)$

$$\min_{ij} (\text{SSF}(i, j)) \quad (2.53)$$

If the globally best split is identified, we branch the input space into the respective regions, which then constitute the new child nodes of the tree.

3. Step (2) is recursively performed, until a stopping criterion is reached. In our context the tree building process is stopped, if the leaf nodes contain a minimum number of elements.

If we want to predict the target value for a new element, we can query it down the tree to its corresponding leaf node, which will then contain similar training set elements. Using their known target values, we can then estimate the target value for the new element.

2.5.1 Regression

In the regression setting, the splits are selected based on the sum of squared error function defined in Eq. 2.32. More concretely we consider a split on the i th variable at position j , that branches the tree into a left and a right leaf node. The region that corresponds to one of these leaf nodes l is then referred to as $\mathcal{R}_{l,ij}$, where $l \in \{\text{left}, \text{right}\}$. We then calculate the sum of squared error (e.g. Hastie et al., 2009) for the training set samples that reside in region $\mathcal{R}_{l,ij}$ as

$$\text{SSE}(\mathcal{R}_{l,ij}) = \sum_{\mathbf{x}_n^T \in \mathcal{R}_{l,ij}} (t_n - E[t]_{\mathcal{R}_{l,ij}})^2 \quad (2.54)$$

where

$$E[t]_{\mathcal{R}_{l,ij}} = \left[\frac{1}{N_{\mathcal{R}_{l,ij}}} \sum_{\mathbf{x}_n^T \in \mathcal{R}_{l,ij}} t_n \right]. \quad (2.55)$$

Here $N_{\mathcal{R}_{l,ij}}$ denotes the number of training set elements in region $\mathcal{R}_{l,ij}$.

After evaluating Eq. 2.54 for the left ($l = \text{left}$) and the right ($l = \text{right}$) leaf nodes, corresponding to a split in the i th variable at the position j , we assess the split selection function $\text{SSF}(i, j)$ as

$$\text{SSF}(i, j) = \text{SSE}(\mathcal{R}_{\text{left},ij}) + \text{SSE}(\mathcal{R}_{\text{right},ij}). \quad (2.56)$$

The tree constructed in such a manner can then be used to make predictions for new elements that are queried down the tree. Denoting the region into which the new element with input value \mathbf{x} is placed as $\mathcal{R}(\mathbf{x})$, we obtain the final prediction as

$$y(\mathbf{x}) = \frac{1}{N_{\mathcal{R}(\mathbf{x})}} \sum_{\mathbf{x}_n^T \in \mathcal{R}(\mathbf{x})} t_n, \quad (2.57)$$

where the sum goes over all $N_{\mathcal{R}(\mathbf{x})}$ training set elements in the region $\mathcal{R}(\mathbf{x})$.

2.5.2 Classification

In the context of classification, the general structure of the algorithm remains the same, but the split selection, as well as the prediction function are different.

In §2.3 we introduced the cross-entropy loss function for binary classification (Eq. 2.37), which we will use to select the input parameter splits for a decision tree in the classification context. The split selection function $\text{SSF}(i, j)$ evaluated for a split at the i th feature at position j reads then

$$\text{SSF}(i, j) = \text{CE}(\mathcal{R}_{\text{left}, ij}) + \text{CE}(\mathcal{R}_{\text{right}, ij}). \quad (2.58)$$

We define the fraction of elements $p_{\mathcal{R}_{1, ij}}(k)$ within a certain region $\mathcal{R}_{1, ij}$ that are member of class k as

$$p_{\mathcal{R}_{1, ij}}(k) = \frac{1}{N_{\mathcal{R}_{1, ij}}} \sum_{\mathbf{x}_n^T \in \mathcal{R}_{1, ij}} \delta_{t_n=k}, \quad (2.59)$$

where $N_{\mathcal{R}_{1, ij}}$ denotes the number of training samples in region $\mathcal{R}_{1, ij}$ and $\delta_{t_i=k}$ is the Kronecker delta function, that is 1 for all samples of class k and 0 otherwise. We can then define the cross entropy (e.g. Hastie et al., 2009) by summing over all K classes as

$$\text{CE}_{\mathcal{R}_{1, ij}} = - \sum_{k=1}^K p_{\mathcal{R}_{1, ij}}(k) \log \{p_{\mathcal{R}_{1, ij}}(k)\}, \quad (2.60)$$

which is just the generalization of Eq. 2.37 to multiple classes.

If a new sample with input feature \mathbf{x} is queried down the tree, we predict the majority class of all training samples that reside in the respective region $\mathcal{R}_{1, ij}$ as

$$y(\mathbf{x}) = \underset{k}{\text{majority}} \{t_{\mathbf{x}_i^T \in \mathcal{R}(\mathbf{x})} \text{ for } i = 1, \dots, N_{\mathcal{R}(\mathbf{x})}\}, \quad (2.61)$$

where $N_{\mathcal{R}(\mathbf{x})}$ denotes the number of training samples in region $\mathcal{R}(\mathbf{x})$.

2.5.3 The Random Forest

While the decision tree is a powerful tool to generate predictions in regression and classification, it is typically quite sensitive to the noise in the training set (see e.g. Hastie et al., 2009, Sec. 15.2, page 588). This means that small fluctuations can lead to very different, and potentially suboptimal, split selections, whose effect can propagate down into the subsequent nodes. As a result, decision trees tend to have a high variance. The Random Forest is an attempt to stabilize the decision tree by bootstrap aggregation. In this method we generate a set of bootstrap realizations of the training set and grow a separate tree on each individual dataset in the ensemble. The prediction for a new instance is then the average (regression) or majority vote (classification) of the tree predictions in the ensemble. The averaging process reduces the variance of the individual trees in the ensemble prediction. Optionally we can randomly select a subset of features to be considered before each split selection, which can further improve the performance of the Random Forest method.

In summary the procedure then reads

1. generate b bootstrap realizations of the training data
2. grow a tree on each bootstrap realization of the training data as follows

- (a) Randomly select p input variables from the full set of input variables
 - (b) Select the best attribute/split combination from these p input variables
 - (c) Perform the selected split and recursively continue to grow the tree while iterating all prior steps (a) - (c)
3. After all trees in the ensemble have been grown, return the ensemble.

The Random Forest is a powerful, practical Machine Learning algorithm, that provides highly accurate predictions and requires, based on my own experience, a very small amount of tuning compared with other contemporary techniques like e.g. Neural Networks (see e.g. Bishop, 1995; Haykin, 2009) or Support Vector Machines (see e.g. Hearst et al., 1998; Bishop, 2006).

2.6 Sources of error in Machine Learning

The previous section gave a brief introduction into the basic concepts of Machine Learning and presented the Random Forest algorithm, a state-of-the art practical Machine Learning method. Besides the aforementioned sources of error, i.e. bias and variance, the accuracy of a Machine Learning system also strongly depends on the quality of the training set. Specifically we require, that the training set is a good representation of the test data. The following discussion therefore considers scenarios, where differences between the training and test data can lead to biases in the final prediction.

2.6.1 Sample Selection Bias

The training data used in a Machine Learning algorithm is representative by construction, if it was drawn from the same parent distribution as the test data. In many practical applications however, this cannot be guaranteed and we have to consider several effects, that can bias the training data.

Training samples with input variables \mathbf{x}_T and target values t_T can be biased either in terms of the marginal distribution $p(\mathbf{x}_T)$ or in terms of their conditional distribution $p(t_T|\mathbf{x}_T)$. The former simply means that some areas of the full test set attribute space are not well populated by training samples. Biases in the distribution of target values are more subtle and appear e.g. if the target values and the input variables have been obtained from inconsistent data sources. The source that provides the target values might use unknown selection criteria, or can be subject to systematic biases, that cannot be reproduced in the test set. As a result, some target values might be incorrect or systematically missing, where we give an example of this below. In the following discussion we will refer to biases in the input variables as ‘covariate shifts’ and biases in the target direction as ‘target shifts’.

We illustrate the effect of these types of sample selection biases in Fig. 2.5, where a dataset (x, t) is generated using an exponential function with additive Gaussian noise

$$t = \exp(x) + \epsilon. \quad (2.62)$$

Here ϵ is a zero mean random variable with $\sigma_\epsilon = 20$ and we generate 4 separate datasets, three training sets and one test set. The test set is obtained by drawing 2000 values x from a uniform distribution $\mathcal{U}(0, 5)$ between 0 and 5. We then obtain an unbiased training set and

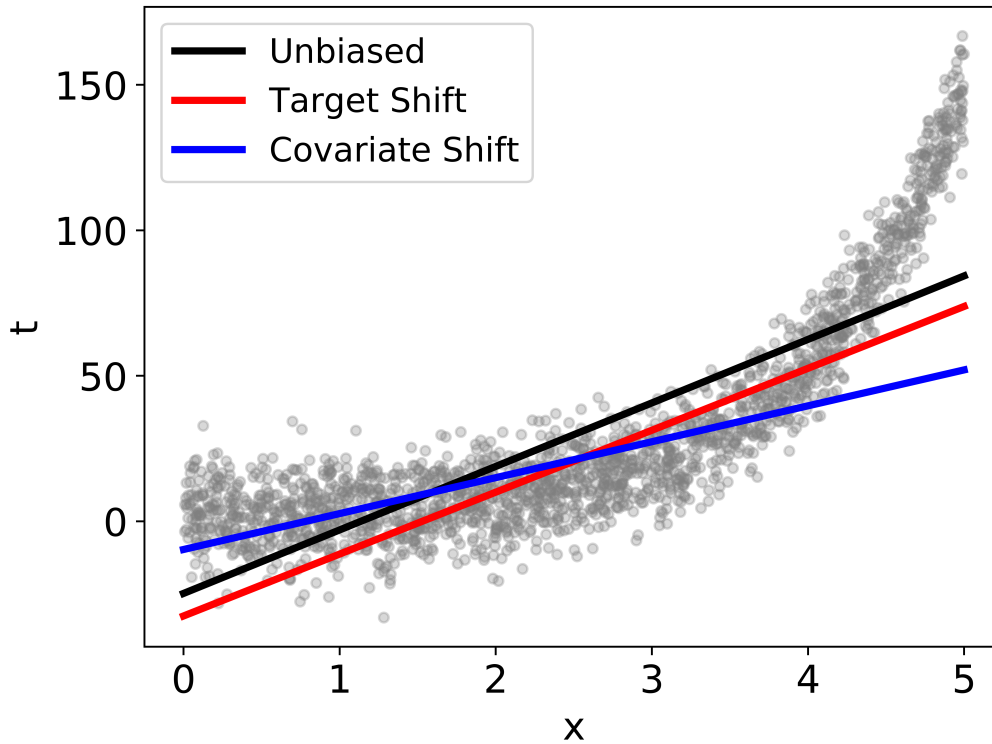


Figure 2.5: Illustration of several sources of sample selection bias. The grey points illustrate the test data, that was generated by adding zero mean Gaussian noise with $\sigma = 20$ to an exponential function evaluated on uniformly distributed points. The black line shows the linear regression model obtained using an unbiased training set with $N = 2000$ data points. The red and blue lines show the corresponding result obtained using a training set with a target shift or covariate shift as explained in the text.

two training sets that are subject to a covariate shift and a target shift respectively. We then fit a linear regression function to each of these training sets. The training set that is subject to a covariate shift is generated by reducing the number of samples for $x > 2.5$ to 10% of its original size. As can be seen by comparing the black and the blue curves, the covariate shift focusses the regression curve on the flat part of the exponential function. As a result, the blue curve performs very poorly for $x > 2.5$, where the exponential function steeply rises.

The effect of a target shift is illustrated by the red line in Fig. 2.5, where we systematically bias the mean of the target noise by $\delta\epsilon = -8$, which biases the function to lower t values. It is notable, that the target shift cannot be detected based on the distribution of x , which is the same as for the unbiased data. In contrast, covariate shifts can be detected relatively easily and corrected either by a weighting scheme or using a culling method⁹. We will provide a more detailed discussion of sample selection biases and the methods that can be used to correct them in §5.4.

⁹Culling methods are Machine Learning methods that detect outliers in a dataset. A simple culling method could be obtained by fitting a multidimensional normal distribution to a dataset, flagging all datapoints that lie in its tails.

2.6.2 Input noise

Another common source of error in the practical application of Machine Learning algorithms arises, if the noise properties of the input variables in the training data are not representative of those in the test set. For example, in many astrophysical applications a Machine Learning architecture is trained on simulations or high quality data, and subsequently applied to data that is subject to larger measurement errors. If this measurement error is not properly incorporated into either the training set or the model, the derived predictions will be biased. I follow Pischke (2007) in this subsection.

While the effect of measurement errors can become very complex depending on the Machine Learning model, the problem domain and the noise model, we can gain a basic understanding by considering the simple linear regression setting

$$t = ax + \epsilon \quad (2.63)$$

where t is the response variable and x is the attribute. Here we assume that the target noise ϵ is a zero mean random variable with standard deviation σ_ϵ . The optimal solution for a in this simple regression setting is given as

$$a_{\text{opt}} = \frac{\text{Cov}[x, t]}{\text{Var}[x]}. \quad (2.64)$$

We now assume that the attribute x is subject to a zero mean error u with standard deviation σ_u , where u is uncorrelated with x , t and ϵ . Under the new attribute error we do not measure the true attribute x but instead the noisy version (see Pischke, 2007)

$$w = x + u, \quad (2.65)$$

and the estimate for the model parameter \tilde{a} that is subject to the covariate noise σ_u can then be obtained as (see Pischke, 2007)

$$\tilde{a}_{\text{opt}} = \frac{\text{Cov}[x + u, t]}{\text{Var}[x + u]} = \frac{\text{Cov}[x, t]}{\text{Var}[x] + \text{Var}[u]} = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \right) a_{\text{opt}} = \lambda a_{\text{opt}}, \quad (2.66)$$

where we used that u is uncorrelated with t and substituted Eq. 2.64. The slope of the regression curve is therefore tilted and the corresponding target error (see Pischke, 2007)

$$\tilde{\epsilon} = t - \tilde{a}_{\text{opt}}w = \epsilon - (t - a_{\text{opt}}x) + t - \tilde{a}_{\text{opt}}w = \epsilon + (a_{\text{opt}} - \tilde{a}_{\text{opt}})x - \tilde{a}_{\text{opt}}u, \quad (2.67)$$

will also increase as a result of the measurement error as (see Pischke, 2007)

$$\tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + (1 - \lambda)^2 a_{\text{opt}}^2 \sigma_x^2 + \lambda^2 a_{\text{opt}}^2 \sigma_u^2. \quad (2.68)$$

We conclude, that the measurement error not only biases the parameters in the regression function, but also leads to a significantly larger residual error. As the shape of the nonlinear regression function can change in a complex manner, as a result of additional covariate noise, we cannot directly generalize these simple considerations. However we will demonstrate in §5 that similar errors arise in the context of photometric redshift validation. A general method to alleviate these errors is to match the attribute noise in the training set and the test set by artificially degrading the one that has the lower noise level.

2.7 Density Estimation

In the previous sections we have always assumed, that the predictive distribution is a Gaussian, where its mean is given by the regression function and the standard deviation is assumed to be a constant. In practical applications, this assumption will not always hold, and it is desirable to introduce techniques that estimate densities in a model independent way.

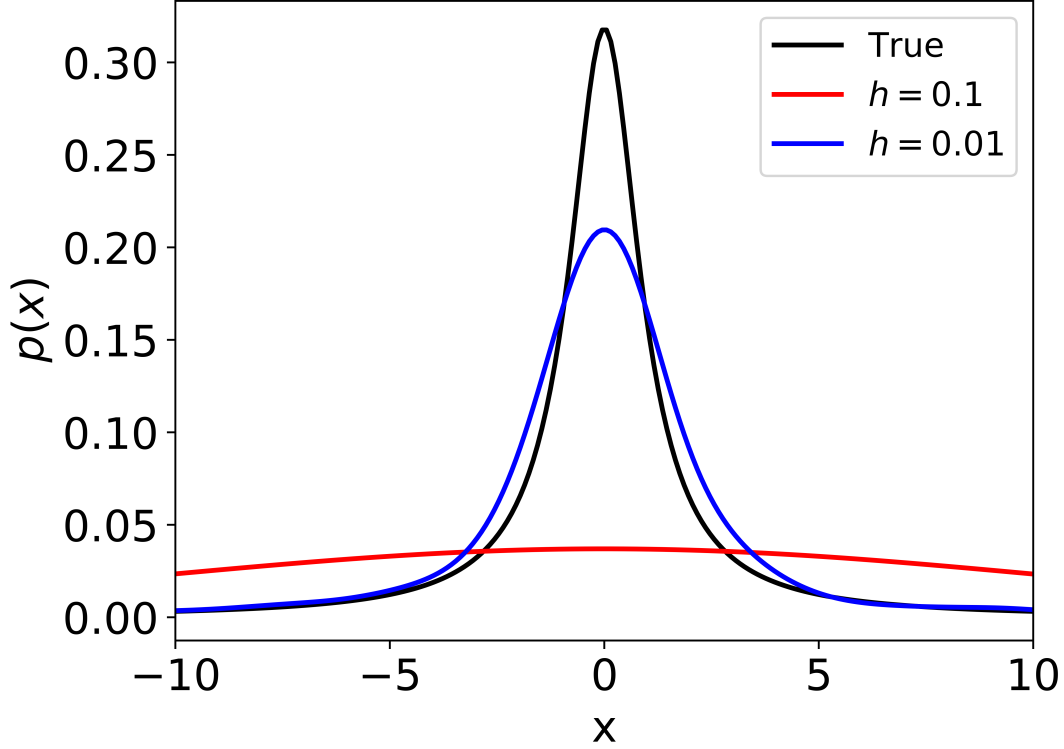


Figure 2.6: The effect of selecting a too large bandwidth h on the accuracy of the recovered distribution. We show the recovered distribution using kernel density estimates Eq. 2.75 with a Gaussian kernel and a bandwidth of $h = 0.1$ (red) and $h = 0.01$ (blue) respectively. The density estimates are applied to a sample of size $N = 1000$. The black line shows the original, true distribution.

Following Bishop (2006) we start by considering the probability $P_{\mathcal{R}}$ to find a sample within a region \mathcal{R} of the feature space \mathbf{x} . Assuming that we knew the true pdf of the underlying distribution that generated the sample, we can write

$$P_{\mathcal{R}} = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}. \quad (2.69)$$

Provided a sufficiently large sample of size N , the number of objects k that fall into region \mathcal{R} are then distributed according to a binomial distribution¹⁰, and we can estimate k as

$$k = NP_{\mathcal{R}}. \quad (2.70)$$

¹⁰The Binomial distribution quantifies how likely it is, to find a certain number of successes in a coin toss, or Bernoulli experiment. We refer to the statistics literature (see also Bishop, 2006) for a further introduction.

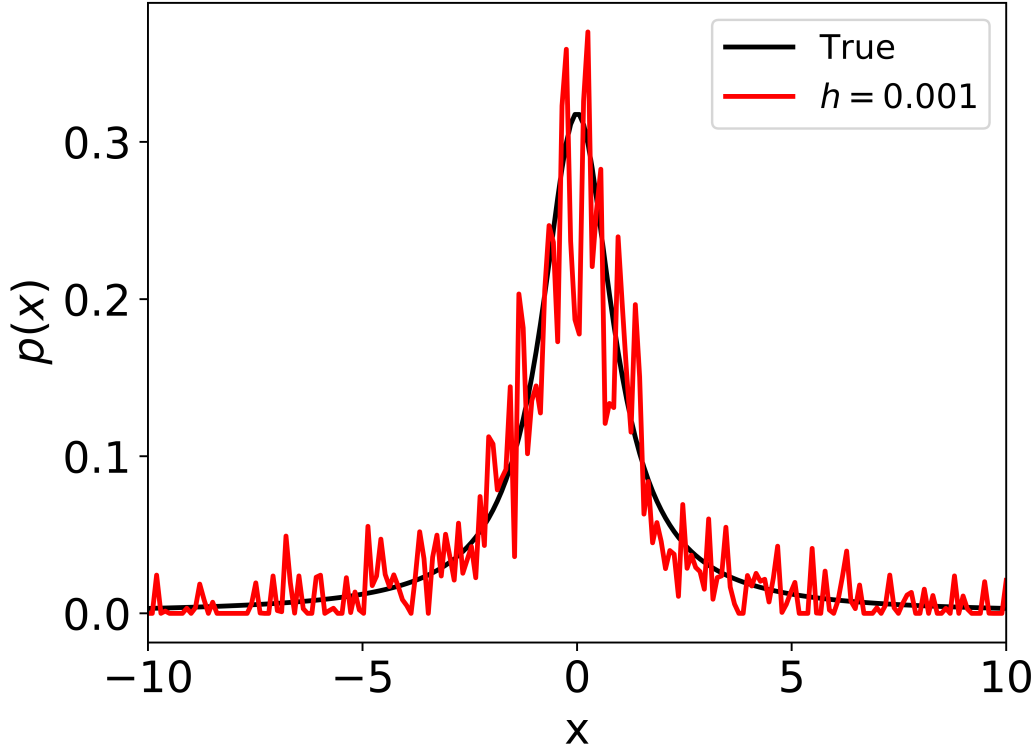


Figure 2.7: The effect of selecting a too small bandwidth h on the accuracy of the recovered distribution. We show the recovered distribution using a kernel density estimate Eq. 2.75 with a Gaussian kernel and a bandwidth of $h = 0.001$ (red). The density estimate is applied to a sample of size $N = 1000$. The black line shows the original, true distribution.

If \mathcal{R} is small enough, such that the density $p(\mathbf{x})$ doesn't change much across this region with volume V , we can write

$$p(\mathbf{x}) = \frac{P_{\mathcal{R}}}{V}. \quad (2.71)$$

If we combine Eq. 2.70 and Eq. 2.71 the density $p(\mathbf{x})$ is given as

$$p(\mathbf{x}) = \frac{k}{NV}. \quad (2.72)$$

We see that Eq. 2.72 is determined by two free parameters, the number of objects k within the region and its volume V . To estimate the density $p(\mathbf{x})$ we therefore either fix the number of points k and constrain V from the data, or we fix the volume V and count the number of objects k within it. Fixing k leads to the k -nearest neighbor estimate that will be considered in §5, the latter option motivates the kernel density estimate introduced in the following.

We consider a hypercube in input space of side length h . The total number of objects k that fall into this box centered at the point \mathbf{x} is

$$k = \sum_{i=1}^N \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (2.73)$$

In the following we will refer to $\mathcal{K}(\mathbf{a})$ as the kernel function, where our kernel is given by a d dimensional hypercube

$$\mathcal{K}(\mathbf{a}) = \begin{cases} 1, & |a_i| \leq 1/2, \text{ for } i = 1, \dots, d \\ 0, & \text{otherwise} \end{cases}. \quad (2.74)$$

Filling the volume of the input space \mathbf{x} with hypercubes leads to an estimate for the density as (see Bishop, 2006)

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh^d} \sum_{i=1}^N \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (2.75)$$

Instead of using hypercubes, which have discontinuous boundaries, we can instead use smooth functions like normal distributions as kernel functions. In the multidimensional case the gaussian kernel density estimate reads (see Scott & Sain, 2005)

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \frac{1}{(2\pi h^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2h^2}\right), \quad (2.76)$$

where the standard deviation of the gaussian σ is associated with the ‘bandwidth’ h of the kernel function.

We note that we can in principle use any kernel function that is positive for all inputs $\mathcal{K}(\mathbf{a}) \geq 0$ and that integrates to unity (see Bishop, 2006).

One of the main challenges in the application of kernel density estimates is the selection of a suitable bandwidth h as illustrated in Fig. 2.6 and Fig. 2.7. Selecting a too large bandwidth h leads to a density estimate that can oversmooth a distribution as shown in Fig. 2.6. This is especially problematic for peaked or multimodal distributions as illustrated here for the case of a student-t distribution¹¹ with $r = 1$ degrees of freedom. In contrast, selecting a too small bandwidth will correctly capture the peak, but will also produce a very noisy estimate as shown in Fig. 2.7. This again is an example of the trade-off between bias and variance that we considered previously in the context of regression. Applied to the problem of density estimation, an estimate with a large bias would have a too large bandwidth. Similarly a high variance estimate would be constructed on a too small bandwidth. The trade-off between bias and variance in density estimation will be discussed in more detail in §4.

Up until now we have reviewed how a kernel density estimate can be applied to estimate unconditional distributions. In the next section we extend this method to conditional distributions, generalizing the previous discussion on regression, that was restricted to Gaussian predictive distributions.

2.7.1 Conditional Density estimation

In order to estimate the conditional density of the target variable t given the input variables \mathbf{x} , we can use the density ratio estimate

$$\hat{p}(t|\mathbf{x}) = \frac{\hat{p}(t, \mathbf{x})}{\hat{p}(\mathbf{x})}, \quad (2.77)$$

¹¹We refer to the statistics literature (e.g. Bulmer, 1967) for a detailed description of the student-t distribution.

between the joint distribution of the target-attribute space $\hat{p}(t, \mathbf{x})$ and the marginal distribution \mathbf{x} .

The joint density $\hat{p}(t, \mathbf{x})$ can be estimated by the product kernel (see Scott, 1992)

$$\hat{p}(t, \mathbf{x}) = \frac{1}{N h_t h_x^d} \sum_{i=1}^N \mathcal{K} \left(\frac{\mathbf{x} - \mathbf{x}_i}{h_x} \right) \mathcal{K} \left(\frac{t - t_i}{h_t} \right) \quad (2.78)$$

where N is the sample size, d is the dimension of \mathbf{x} , and the parameters h_x , h_t are the bandwidths for the attribute and target dimensions. Estimating the function $\hat{p}(\mathbf{x})$ using Eq. 2.75, we can write the conditional density estimate of the predictive distribution as (see Rosenblatt, 1969; De Gooijer & Zerom, 2003; Holmes et al., 2012)

$$p(t|\mathbf{x}) = \sum_{i=1}^N w_i(\mathbf{x}) \mathcal{K} \left(\frac{t - t_i}{h_t} \right), \quad (2.79)$$

where we define the weights as (see De Gooijer & Zerom, 2003)

$$w_i(\mathbf{x}) = \frac{1}{N h_t} \frac{\mathcal{K}((\mathbf{x} - \mathbf{x}_i)/h_x)}{\sum_{i=1}^N \mathcal{K}((\mathbf{x} - \mathbf{x}_i)/h_x)}. \quad (2.80)$$

While Eq. 2.79 provides a simple way to estimate conditional densities, its performance will in general strongly depend on the selection of suitable bandwidth values.

The concepts introduced in this chapter are the basis for the techniques that will be discussed in the rest of this thesis. In the next chapter, we will discuss a method for conditional density estimation, that is based on the highly accurate Random Forest algorithm. In §4 we will use the bootstrap method to quantify the accuracy of density estimates, and §5 will discuss the effect of sample selection biases and field-to-field variations in photometric noise on the accuracy of photometric redshift distributions. We will also use a nearest neighbor density estimate to investigate which regions of the photometric input space are sufficiently populated by spectra.

Accurate photometric redshift PDF estimation

In the following I attach the preprint version of my paper ‘Accurate photometric redshift probability density estimation - method comparison and application’, published in the ‘Monthly Notices of the Royal Astronomical Society, Volume 452, Issue 4, p.3710-3725’. The authors of this paper are: Markus Michael Rau, Stella Seitz, Fabrice Brimiouille, Eibe Frank, Oliver Friedrich, Daniel Gruen, Ben Hoyle. This preprint version was updated to match the published version.

Scientific Context Accurate photometric redshift estimation is a vital requirement in order to fully exploit large area photometric surveys like CFHTLenS (see §8.1) or DES (see §8.2). In this paper we present a variety of novel approaches to estimate photometric redshift distributions of individual galaxies and samples of galaxies and demonstrate the effectiveness of these methods in a variety of applications, that are relevant for cosmology, like the modeling of angular correlation power spectra, cosmic shear correlation functions and cluster mass estimates.

As mentioned in §1.4.2 there exist a large variety of different photometric redshift codes to estimate photometric redshift distributions. Most notably, the methods proposed by Bonnett (2015) and Carrasco Kind & Brunner (2013) use a classification based approach, that is structurally similar to the method used in my paper to reconstruct the photometric redshift distribution. In contrast to their work, our approach additionally incorporates the ordering in consecutive redshift bins into the prediction process, which improves the accuracy of the photometric redshift distributions, as will be demonstrated in §5.1 of the paper. This ordinal classification approach was previously proposed in Frank & Hall (2001), however we complement this method with efficient bandwidth selection schemes for Kernel density estimates and a Gaussian mixture model estimator. The former technique significantly improves the accuracy of the reconstructed photometric redshift distributions, while the latter enable the efficient parametrization of photometric redshift distributions as mixtures of Gaussians. For many applications in cosmology, estimates of the photometric redshift distribution of an individual galaxy is not required, and we are instead mainly interested in the photometric redshift distribution of samples of galaxies. For these specific applications we propose a method that allows the estimation of photometric redshift distributions of samples of galaxies, using only

a single floating point number per galaxy. In contrast to other approaches that draw a single redshift value from the photometric redshift distribution of an individual galaxy (Carrasco Kind & Brunner, 2014b), the ‘Highest Weight Element’ estimator proposed in this work does not require this sampling step and is therefore structurally more efficient.

We apply these methods to data from the CFHTLenS survey and demonstrate that our methods are able to significantly reduce photometric redshift induced biases in cluster mass estimates, the modelling of angular correlation power spectra and the modelling of cosmic shear correlation functions.

Challenges During the course of this project it became apparent, that the accuracy of individual galaxy redshift distributions is more sensitive to the selection of an accurate bandwidth for the kernel density estimate, than to the hyperparameters of the Random Forest algorithm. The photometric redshift distribution of an individual galaxy is typically estimated using a small number of galaxies, that are strongly upweighted. This then leads to a very unstable estimate that will have a high error. This was the ultimate motivation to investigate methods to derive accurate error intervals on these density estimates, leading to the paper discussed in the next section. During my work to investigate efficient bandwidth selection techniques, I compared a variety of bandwidth selection techniques (see Jones et al. 1996 for a review), e.g. cross validation bandwidth selection methods or plug-in methods. While these more advanced techniques seemed to perform better with respect to the negative log-likelihood loss (MNLP) defined in Eq. 15 of the paper, they are computationally expensive, especially if they are applied to a large number, e.g. 100M, galaxies. After visual inspection of the individual density estimates I had the impression, that kernel density estimates selected by the MNLP loss tend to be undersmoothed, which agrees with the statistics literature (see Zambom & Dias, 2013). Since elaborated bandwidth selection schemes didn’t seem practical, I decided to choose a starting bandwidth based on the Scott’s rule (see Scott, 1992) and choose a global prefactor based on the MNLL criterion to adapt the smoothing based on the ‘average’ shape of the redshift distributions in the sample. This method produced good results, while still remaining computationally tractable. A possible alternative is the usage of Gaussian mixture models, which are analytical functions and therefore potentially a more appropriate choice, if a photometric redshift distribution has to be fitted on a small number of samples. However even with Gaussian mixture models one needs to select the number of components in the model. We tested resampling/cross validation methods like the bootstrap (Eibe Frank personal communication). However again I found these methods computationally too expensive to be applicable to large galaxy samples. I therefore decided to select the number of Gaussian components based on the negative entropy criterion, which I found to be much faster compared with the usage of contemporary cross validation techniques, while still providing highly accurate estimates of the individual galaxy photometric redshift distribution.

We note that these issues are not unique to Machine Learning based photometric redshift estimation, but are also problematic in the more general context of photometric redshift validation, if the sample sizes are small as discussed in §4.

We also found, that the accuracy of photometric redshift distributions of samples of galaxies does not always translate into a lower bias in terms of the cosmic shear correlation function, as discussed in the Appendix of the paper. This was not the case for the angular correlation power spectrum, where improvements in the overall shape of the photometric redshift distribution also led to a reduction in the angular correlation function modeling bias. This

strongly suggests, that the angular correlation power spectrum and the cosmic shear correlation function have a different sensitivity to the moments of the estimated photometric redshift distribution, where we refer to §5 for a further discussion.

Contributions and Acknowledgements I performed the majority of the analyses, matched the datasets, performed the reweighting of the catalog, wrote the text and prepared the figures. Furthermore I implemented all Machine Learning algorithms, except the ANNz code that was used from Collister & Lahav (2004). My Quantile Regression Forest (QRF) implementation is based on the original package by Nicolai Meinshausen¹ (Meinshausen, 2006), that I modified to return a weighted training set instead of the conditional quantiles. My implementation of the Ordinal Class PDF (OCP) method uses some advanced statistics functionality from the R language (R Core Team, 2015) like the preimplemented ‘isoreg’ function for isotonic regression, and the R implementation of the RandomForest method (Liaw & Wiener, 2002). Furthermore, my implementation of the OCP method was adapted and added to the UBL package by Paula Branco² (Branco et al., 2016). I used the Rmixmod package (Biernacki et al., 1999; Auder et al., 2014) and preimplemented statistics functions in R to implement the Gaussian mixture model and the cross-validation bandwidth selection extensions to the OCP method. I also implemented the HWE method based on the QRF package (Meinshausen, 2006) that I modified, and performed the analyses in §5.1 and §5.2. I also prepared the distributions and performed all tasks of data processing, including the re-weighting of the spectrophotometric data to the shape catalog, for the analyses in §5.3.

FB provided the catalogs used in this work and wrote the dataset description (§4). EF contributed useful discussions and helped in the verification of the algorithms via code comparisons. OF provided the modelling of the shear correlations functions, the critical surface mass density estimates, provided the results that I plotted in Fig. 19 and helped in the analysis presented in the appendix. BH contributed the modelling of the angular correlation functions. All authors contributed useful discussions and helped in the writing of the text.

¹<https://cran.r-project.org/web/packages/quantregForest/quantregForest.pdf>

²<https://cran.r-project.org/web/packages/UBL/UBL.pdf>

Accurate photometric redshift probability density estimation - method comparison and application

Markus Michael Rau^{1,2}, Stella Seitz^{1,2}, Fabrice Brimiouille³, Eibe Frank⁴,
Oliver Friedrich^{1,2}, Daniel Gruen^{1,2}, Ben Hoyle^{1,5}

¹*Ludwig-Maximilians-Universität München, Universitäts-Sternwarte, Scheinerstr. 1, D-81679 Munich, Germany*

²*Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, 85748 Garching, Germany*

³*Observatório Nacional, Rua Gal. José Cristino, 20921-400, Rio de Janeiro, Brazil*

⁴*Department of Computer Science, University of Waikato, Hamilton, New Zealand*

⁵*Excellence Cluster Universe, Boltzmannstr. 2, D-85748 Garching, Germany*

E-mail: mmrau@usm.lmu.de

Accepted —. Received —; in original form —.

ABSTRACT

We introduce an ordinal classification algorithm for photometric redshift estimation, which significantly improves the reconstruction of photometric redshift probability density functions (PDFs) for individual galaxies and galaxy samples. As a use case we apply our method to CFHTLS galaxies. The ordinal classification algorithm treats distinct redshift bins as ordered values, which improves the quality of photometric redshift PDFs, compared with non-ordinal classification architectures. We also propose a new single value point estimate of the galaxy redshift, that can be used to estimate the full redshift PDF of a galaxy sample. This method is competitive in terms of accuracy with contemporary algorithms, which stack the full redshift PDFs of all galaxies in the sample, but requires orders of magnitudes less storage space.

The methods described in this paper greatly improve the log-likelihood of individual object redshift PDFs, when compared with a popular Neural Network code (ANNz). In our use case, this improvement reaches 50% for high redshift objects ($z \geq 0.75$).

We show that using these more accurate photometric redshift PDFs will lead to a reduction in the systematic biases by up to a factor of four, when compared with less accurate PDFs obtained from commonly used methods. The cosmological analyses we examine and find improvement upon are the following: gravitational lensing cluster mass estimates, modelling of angular correlation functions, and modelling of cosmic shear correlation functions.

Key words: galaxies: distances and redshifts, catalogues, surveys.

1 INTRODUCTION

The determination of distance, or redshift, estimates to galaxies is a vital requirement before using large scale photometric galaxy surveys for many cosmological analyses. Large scale surveys, such as the SDSS (York et al. 2000a), PanSTARRS (Tonry et al. 2012), DES (Flaugher 2005) and LSST (Tyson et al. 2003) rely on a combination of photometric and more accurate spectroscopic redshifts when providing distance estimates to photometrically identified galaxies.

Photometric redshifts are used throughout astrophysics and cosmology, for example in large scale structure analyses (Staniszewski et al. 2009; de Simoni et al. 2013), in galaxy

cluster weak lensing analyses (Gruen et al. 2013), and in galaxy-galaxy lensing analyses (Brimioulle et al. 2013). Photometric redshifts are obtained using either machine learning methods or template fitting techniques (see e.g., Benítez 2000; Csabai et al. 2000; Bender et al. 2001; Ilbert et al. 2006; Feldmann et al. 2006; Greisel et al. 2013). Machine learning techniques range from early works employing artificial Neural Networks (Firth, Lahav & Somerville 2003; Collister & Lahav 2004) as photometric point predictors, to recent developments that estimate the full photometric redshift PDF of the galaxy (Lima et al. 2008; Cunha et al. 2009; Carrasco Kind & Brunner 2013; Bonnett 2015). For

detailed reviews and comparisons of different photometric redshift techniques we refer the reader to Sánchez et al. (2014); Hildebrandt et al. (2010); Dahlen et al. (2013). This work focuses on machine learning methods for photometric redshift PDF estimation for samples of galaxies (hereafter sample PDF) as well as individual galaxies (hereafter individual PDFs). We apply the results to a range of analyses in weak gravitational lensing, cosmic shear and large scale structure.

In general, machine learning algorithms learn a mapping between the photometry of an object and the spectroscopic redshift. To train the machine learning models to learn this mapping, one typically identifies spectrophotometric data that overlaps with the photometric feature space of the final data sample for which one would like to estimate redshifts. However, recent work shows that machine learning can also be performed with spectroscopic reference data that is brighter than the photometric sample (Hoyle et al. 2015a). Many photometric surveys include a dedicated spectroscopic follow up program, which allows such a machine learning system to be built, e.g., SDSS-I/II (York et al. 2000b), 2dF (Colless et al. 2001), VVDS (Le Fèvre et al. 2005), WiggleZ (Drinkwater et al. 2010).

The mapping obtained with machine learning is only approximate: the redshift of an object cannot be exactly determined by its corresponding photometry. Moreover, most machine learning methods produce a point estimate, which reduces the individual PDF to one number. The point estimate only predicts the most likely value of the redshift, irrespective of the quality of the photometry, and the shape of the distribution. In order to enter the era of precision cosmology, one must be able to incorporate the uncertainty in the redshift estimate into the cosmological analysis. This means that the use of single point redshift predictions is no longer sufficient. To achieve precision cosmology, we are required to incorporate the full redshift uncertainty using the individual PDFs.

We can obtain a sample PDF by stacking the individual PDFs. This distribution describes the probability that a randomly sampled galaxy has a certain redshift. The accurate estimation of the redshift distribution of the full sample is important for many cosmological analyses, e.g., in large scale structure, weak gravitational lensing, and cosmic shear.

However, effectively estimating and storing the photometric redshift PDF instead of the point estimate, for each object in a large astronomical dataset, is a challenging task. This process requires efficient and accurate photometric estimation algorithms, and scalable data storage solutions. These algorithms must be benchmarked using carefully constructed performance metrics to be useful for the next generation large scale structure photometric surveys (e.g., Laureijs et al. 2011).

We discuss such metrics to quantify performance of photometric redshift PDF estimation in §2. We describe the Ordinal Class PDF (OCP) algorithm in §3.2, which improves the estimation accuracy over commonly used non-ordinal classification architectures. We continue in §3.4 by showing how the OCP method can become more storage efficient, by combining it with the Gaussian mixture model. This enables the storage of the PDFs of individual galaxies even within massive datasets without significant demands on disc space.

Many applications in cosmology require an estimation

of the sample PDF. We propose a single point estimator for this quantity in §3.5, and show how this single floating point number can be computed very efficiently, and achieves good performance when compared with algorithms that stack individual PDFs. The performance of the proposed techniques is demonstrated and analysed in a method comparison in §5.1 and §5.2 using a spectrophotometric dataset (§4) obtained from the public CFHTLS WIDE survey.

Finally, we demonstrate in §5.3 that the methods introduced in this work improve the precision of gravitational lensing cluster mass estimates, measurements of angular correlation functions, and analyses of cosmic shear correlation functions, when compared with results obtained using a common Neural Network code. We conclude and summarize in §6.

2 FUNDAMENTAL CONCEPTS

The following section gives a brief review of important statistical concepts needed in this work. We start with a short introduction to density estimation, introduce metrics to quantify the performance of density estimators and finally describe a scheme to assess the performance of a machine learning model.

2.1 Kernel Density Estimation

The goal of kernel density estimation is to find a good estimator¹ $\hat{p}(\mathbf{x})$ for the probability density function $p(\mathbf{x})$ of a random variable \mathbf{X} using N samples \mathbf{x}_i . Consider a small region \mathcal{R} centred on a point \mathbf{x} . We can then assume that $p(\mathbf{x})$ is approximately constant across \mathcal{R} . Based on this assumption we can estimate the density at point \mathbf{x} as

$$\hat{p}(\mathbf{x}) = \frac{k}{NV_{\mathcal{R}}} . \quad (1)$$

The number of objects² k in Eq. 1 can be estimated by considering a D dimensional hyper cube with volume

$$V_{\mathcal{R}} = h^D \quad (2)$$

centred on the point \mathbf{x} with side length h . Using Eq. 1, we obtain k as

$$k = \sum_{i=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) , \quad (3)$$

where

$$K(\mathbf{d}) = \begin{cases} 1, & |d_i| \leq 1/2, \quad 1 \leq i \leq D \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

is an example of a kernel function. Note that this kernel has discontinuities at the boundaries. The bandwidth h determines how much the kernel density estimate interpolates (or smoothes) between the given data points. A bandwidth that is too large oversmooths important structures in the

¹ In the following we will mark the estimator for a quantity with a hat.

² Fixing the number of points k that fall into \mathcal{R} and estimating the volume $V_{\mathcal{R}}$ leads to the k nearest neighbour density estimation technique (see e.g. Scott 1992).

density whereas one that is too small leads to a noisy density estimate. The density estimate $\hat{p}(\mathbf{x})$ can then be written as

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h^D} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) =: \frac{1}{N} \sum_{i=1}^N \tilde{K}(\mathbf{x}, \mathbf{x}_i, h). \quad (5)$$

Instead of using Eq. 4, which has discontinuities at the boundaries, we can alternatively use smooth and symmetric functions, for example, a Gaussian.

The estimation of photometric redshift PDFs for individual objects (individual PDFs) is an application of conditional probability density function estimation, since the individual PDF $p(z|\mathbf{f})$ is conditional on the objects photometry \mathbf{f} . The estimation of conditional probability density functions can be formulated in close analogy to Eq. 5. We can estimate the individual PDF $p(z|\mathbf{f})$ as a weighted kernel density estimate in redshift space of the form

$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N_{\text{tr}}} w_i(\mathbf{f}) K(z, z_i^{\text{spec}}, h), \quad (6)$$

using a dataset, the so called training set, containing N_{tr} objects. $K(z, z_i^{\text{spec}}, h)$ denotes a kernel function with bandwidth h centred on the spectroscopic redshift values z_i^{spec} . The weights $w_i(\mathbf{f})$ sum to unity and depend on the photometry \mathbf{f} of the object.

The conditional cumulative distribution function $F(z|\mathbf{f})$ defined as

$$F(z|\mathbf{f}) = \int_{-\infty}^z p(z'|\mathbf{f}) dz' \quad (7)$$

can be estimated (Meinshausen 2006) as

$$\hat{F}(z|\mathbf{f}) = \sum_{i=1}^{N_{\text{tr}}} w_i(\mathbf{f}) I(z_i^{\text{spec}} \leq z). \quad (8)$$

$I(z_i^{\text{spec}} \leq z)$ equates to unity if $z_i^{\text{spec}} \leq z$ and to zero otherwise.

The redshift distribution $\hat{p}(z)$ of a sample (sample PDF) containing N objects can be estimated by stacking the individual PDFs

$$\hat{p}(z) = \sum_{i=1}^N w_{\text{stack},i} \hat{p}(z|\mathbf{f}_i). \quad (9)$$

The normed weights $w_{\text{stack},i}$ can be set to $1/N$ or chosen to give more weight to certain sub populations. For example, we can favour certain redshift intervals $z \in [a, b]$ by defining weights as

$$w_{\text{stack}} = \int_a^b p(z|\mathbf{f}) dz = \hat{F}(b|\mathbf{f}) - \hat{F}(a|\mathbf{f}), \quad (10)$$

and we show an example of such a weighting in §5.2. The above weights are normalized afterwards to sum to unity.

2.2 The Gaussian Mixture Model

In this paper, we consider kernel density estimators and Gaussian mixture models for density estimation. A Gaussian mixture model (see, for example, Bishop 2006) for the probability density function $p(x)$ of a random variable X is a linear combination of K normal densities defined as

$$p(x) = \sum_{i=1}^K \alpha_i \mathcal{N}(x, \mu_i, \sigma_i) \quad (11)$$

where α_i is the amplitude, μ_i is the mean, and σ_i is the standard deviation of the mixture component i .

We define the weight proportion $\gamma_k(x)$ of component k as

$$\gamma_k(x) = \frac{\alpha_k \mathcal{N}(x, \mu_k, \sigma_k)}{\sum_{j=1}^K \alpha_j \mathcal{N}(x, \mu_j, \sigma_j)}, \quad (12)$$

where $\gamma_k(x)$ determines how much a certain component of the Gaussian mixture model contributes to the total density at point x .

2.3 Evaluation Metrics

Consider an estimate $\hat{p}(\mathbf{x})$ of the true probability density function $p(\mathbf{x})$ describing the distribution of the random variable \mathbf{X} . We can measure the quality of the estimate $\hat{p}(\mathbf{x})$ by its distance $D(\hat{p}(\mathbf{x})||p(\mathbf{x}))$ to the true distribution $p(\mathbf{x})$, which is generally unknown. The Kullback-Leibler divergence between the true density $p(\mathbf{x})$ and the estimate $\hat{p}(\mathbf{x})$ is defined using the natural logarithm as,

$$D(p||\hat{p}) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{\hat{p}(\mathbf{x})} \right) d\mathbf{x}. \quad (13)$$

A good estimate \hat{p} for p should minimize $D(p||\hat{p})$. Rewriting the logarithm we obtain

$$D(p||\hat{p}) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log(p(\mathbf{x})) d\mathbf{x} - \int_{-\infty}^{\infty} p(\mathbf{x}) \log(\hat{p}(\mathbf{x})) d\mathbf{x}, \quad (14)$$

and we note that the first term is a constant that does not depend on the model parameters, for example bandwidth, kernel or shape of kernel function. Thus, the second term in Eq. 14 can be used as a relative measure of the accuracy of $\hat{p}(\mathbf{x})$. If we use the sample mean to estimate the expectation with respect to $p(\mathbf{x})$, we obtain the mean negative log likelihood loss, hereafter MNLL, (Habbema, Hermans & Van den Broek 1974; Duin 1976)

$$MNLL = -\frac{1}{N} \sum_{i=1}^N \log(\hat{p}(\mathbf{x}_i) + \epsilon), \quad (15)$$

where we set $\epsilon = 10^{-6}$ to avoid floating point underflow. The Kullback-Leibler Divergence is a distance and thus non negative and it is smallest if the MNLL is smallest.

A suitable loss function for individual PDFs can be defined analogously (see e.g., Takeuchi, Nomura & Kanamori 2009; Frank & Bouckaert 2009; Sugiyama et al. 2010). We estimate $p(z|\mathbf{f}_i)$ for each of the N objects in the sample, in order to establish performance using a sample of objects for which spectroscopic redshift values have been observed. We then evaluate $\hat{p}(z|\mathbf{f}_i)$ at the object's observed spectroscopic redshift $\hat{p}(z = z_{\text{spec},i}|\mathbf{f}_i)$. In the rest of the paper, the abbreviation MNLL refers to the mean negative log-likelihood loss evaluated for individual PDFs.

2.4 Model Training

We randomly sample three non-overlapping datasets without replacement from the available data: the training set, the validation set and the test set. The model is trained on the training set and the model parameters are chosen by testing the performance of the trained model with different parameter settings on the validation set.

The validation set is used during model tuning and therefore it does not provide a good estimate of the performance on unseen data. We measure this generalization performance on a test set that is not used during training and tuning.

To evaluate the machine learning algorithms, we construct a training set containing 9000 objects, a validation set containing 3000 objects and a test set containing 22072 objects. After the validation set has been used to determine the best combination of model parameters, we merge the training set and the validation set and train the respective model again with this best setup. In this way, we make optimal use of the available data to build the final model. All results described in §5 are obtained on the test set, which, we reiterate, was not used in all prior steps of model training and tuning.

In this work we choose to use the aperture magnitudes of the CFHTLS WIDE five band photometry as input attributes. Other photometric features may be used, for example see Hoyle et al. (2015b) for a feature importance analysis.

3 ALGORITHMS

We have introduced the estimator for the photometric redshift PDF of individual objects (individual PDF) in Eq. 6 as a weighted kernel density estimate that depends on the weights $w(\mathbf{f})$. The following section discusses two algorithms that can be used to estimate these weights.

3.1 Quantile Regression Forest (QRF)

The Quantile Regression Forest³ (Meinshausen 2006) is a generalization of the Random Forest (Breiman 2001) that can be used to reconstruct individual PDFs, an algorithm known as TPzreg (Carrasco Kind & Brunner 2013) in astrophysics.

A regression/classification tree partitions the input space and returns the mean/majority vote of the response values (i.e., the redshift values) of the training set objects in each partition as the final prediction for new objects falling into that partition. The tree partitions the input data such that the training set objects in each partition are most similar with respect to their response values. In regression, we measure similarity using the sum of squares loss function *SSE*, defined as

$$SSE = \sum_{\tau=1}^l \sum_{\mathbf{f}_i \in \mathcal{R}_\tau} (z_{\text{spec},i} - \langle z_{\text{spec},\tau} \rangle)^2. \quad (16)$$

The sum runs over all l leaf nodes of the tree $1 \leq \tau \leq l$, which each represents a certain partition \mathcal{R}_τ in input space, and over all objects in the training set $(\mathbf{f}_i, z_{\text{spec},i})$ with attribute values \mathbf{f}_i that fall into \mathcal{R}_τ . The term $\langle z_{\text{spec},\tau} \rangle$ denotes the mean spectroscopic redshift of all training set objects that fall into \mathcal{R}_τ .

³ The method was originally developed to estimate conditional quantiles hence the name Quantile Regression Forest.

The binary tree is recursively grown by choosing a splitting attribute and split point for each region using brute-force search such that the SSE is minimized.

The Random Forest algorithm combines several trees by bootstrap aggregation which is described as follows. New training sets are drawn from the original training set with replacement, which is also known as bootstrapping. We train a tree model on each of these bootstrapped training sets, to obtain an ensemble of trees. Combining the estimates from all trees in the ensemble reduces variance. In addition, the Random Forest algorithm makes the resulting models even more diverse by modifying the way each tree is grown. Before each split selection, the routine randomly selects a certain number of attributes, as specified by the ‘mtry’ parameter, on which the algorithm can perform the split.

The complexity of the tree model is governed by the size of the leaves of the tree. We stop the recursive tree building process when a specified minimum number of objects in each leaf, denoted as ‘nodesize’ is reached. If the nodesize is small, very complex trees are grown and the tree might overadapt to the training set. This is an example of overfitting. The prediction from the Random Forest is the mean, in regression, or the majority vote, in classification, of the predictions from the ensemble of trees.

A single tree in the Random Forest splits the space spanned by the input attributes derived from the photometry of the objects into partitions which are represented by the tree leaves. Each leaf defined in this manner is associated with the mean spectroscopic redshift value of the training set objects in this leaf. The tree therefore approximates the underlying smooth function by a step function. If a new object is queried, it will be placed in a leaf containing objects with similar photometry. Following the formulation by Meinshausen (2006), we can write the photometric redshift prediction

$$z_{\text{phot}}(\mathbf{f}) = \sum_{i=1}^{N_{\text{tr}}} w_i(\mathbf{f}) z_{\text{spec},i} \quad (17)$$

as a weighted sum over the spectroscopic redshift values $z_{\text{spec},i}$ of the N_{tr} training set objects. In order to distinguish the different trees in the ensemble, which are characterized by different split selections, we introduce a parameter θ , which characterizes each tree. All training set objects with photometry \mathbf{f}_i^{tr} that are located in the same region $\mathcal{R}_{l(\mathbf{f},\theta)}$ (defined by the leaf $l(\mathbf{f},\theta)$) as the newly queried object with photometry \mathbf{f} , get a constant weight, and all other training set objects get zero weight. This can be written as

$$w_i(\mathbf{f}, \theta) = \frac{I(\mathbf{f}_i^{\text{tr}} \in \mathcal{R}_{l(\mathbf{f},\theta)})}{\sum_{j=1}^{N_{\text{tr}}} I(\mathbf{f}_j^{\text{tr}} \in \mathcal{R}_{l(\mathbf{f},\theta)})}, \quad (18)$$

where the weights are normalized such that they sum to unity.

The same concept holds for the Random Forest prediction, in which the weights associated with each training set object are averaged over k trees, each grown on different bootstrapped datasets, and therefore each described by a different parameter θ_b :

$$w_i(\mathbf{f}) = \frac{1}{k} \sum_{b=1}^k w_i(\mathbf{f}, \theta_b). \quad (19)$$

The weights can be used to estimate the individual PDF and

	QRF/HWE	OCP	NOCP	OCP GMM
nodesize	3,5,7,10	1,2,3,5,7,9	1,2,3,5	1,2,3,5,7,9
mtry	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5
BW mod	0.5,0.6,...,1.8,...,3.0	0.5,0.6,...,2.5,...,3.0	0.5,0.6,...,2.0,...,3.0	-
Gauss Comp.	-	-	-	1,2,3

Table 1. Model parameters of the Quantile Regression Forest (QRF), the classification based PDF estimation algorithms (OCP/NOCP) and the OCP algorithm used with the Gaussian mixture model OCP GMM. ‘nodesize’ and ‘mtry’ are model parameters of the Random Forest described in §3. ‘BW mod’ is the bandwidth modification factor employed in the Scott’s rule (Eq. 24) and Gauss Comp. denotes the maximum number of components allowed in the Gaussian mixture model. The best parameter configuration for each algorithm picked on the validation set during model tuning (§2.4) is marked in bold type.

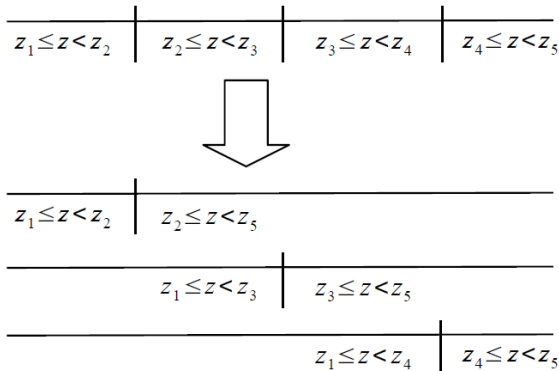


Figure 1. An illustrative example of a nominal classification problem with four redshift bins. These bins can be reformulated into three binary classification problems by merging neighbouring bins. The class probabilities from the binary classification problems can be recombined to incorporate the ordering between the redshift bins (see text) into the final classification.

corresponding statistics like the conditional mean, the conditional cumulative distribution function or the conditional standard deviation defined as

$$\hat{\sigma}^2(z|\mathbf{f}) = \sum_{i=1}^{N_{tr}} w(\mathbf{f}_i) (z_{\text{spec},i} - z_{\text{phot}}(\mathbf{f}_i))^2. \quad (20)$$

The following section introduces an alternative way of estimating the weights in Eq. 6, using a classification scheme.

3.2 Ordinal Class PDF (OCP) estimation

The basic idea of classification-based PDF estimation is to bin the spectroscopic data by redshift and use a classification algorithm that outputs probabilities for bin membership to reconstruct the PDF. Bin membership is viewed as an ordinal variable. Ordinal scale variables, in contrast to nominal ones, exhibit an intrinsic order. If the classes in a classification problem are ordinal, we can use this information to improve the classification (Frank & Hall 2001).

Current classification-based PDF estimation methods in the astrophysics literature (e.g., Bonnett 2015; Carrasco Kind & Brunner 2013) treat redshift bins as nominal classes. In the following, we will refer to the latter as the non-ordinal class PDF (NOCP) algorithm. The ordinal class PDF (OCP) algorithm trains a separate classifier that estimates the probability $p(z \geq z_i)$ that a new object has redshift z above a certain threshold z_i given by the edge of the respective redshift bin. This scheme is illustrated in Fig. 1. The probability

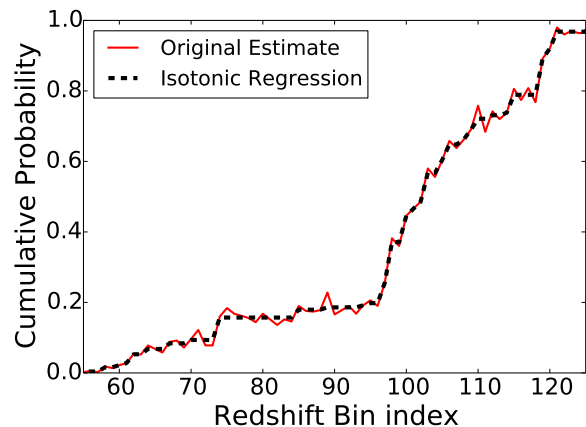


Figure 2. Ordinal classification can result in non monotonic cumulative distribution functions. We calibrate them using isotonic regression. Isotonic Regression (black) approximates the original estimate (red) as a monotonically increasing step function.

that the redshift of an object resides in the original bins is then calculated from these separate classification models as (Frank & Hall 2001)

- 1 $p(z \in [z_1, z_2]) = 1 - p(z \geq z_2)$
- 2 $p(z \in [z_{i-1}, z_i]) = p(z \geq z_{i-1}) - p(z \geq z_i)$
- 3 $p(z \in [z_{k-1}, z_k]) = p(z \geq z_{k-1}) - p(z \geq z_k)$, $1 < i < k$.

The reconstruction of the class probabilities $p(z_i)$ has the idealistic assumption that each of the classifiers used to estimate the probability $p(z \geq z_i)$ outputs perfect probabilities. In practice, this will not be the case and the recovered cumulative distribution function, which is a monotonically increasing function, has to be calibrated. Schapire et al. (2002); Frank & Bouckaert (2009) use a heuristic approach to ensure this monotonicity requirement. Alternatively, we use the ‘isotonic’ regression technique to calibrate the class probabilities. Isotonic regression is synonymous for monotonically increasing regression and is a technique for which efficient implementations are available (de Leeuw, Hornik & Mair 2009). For increasing bin index, isotonic regression optimizes the mean squared error between the original function values and the isotonic fit such that the fit is a monotonic increasing step function as shown in Fig. 2.

We use bins of fixed size $\Delta z = 0.01$ in the range between the minimum and the maximum spectroscopic redshift value in the training set, since we found that equal frequency binning degrades photometric redshift accuracy for catalogues with long-tailed sample PDF. The weights sum to unity and

are calculated using

$$w_i(\mathbf{f}) = \frac{\hat{p}(b_i|\mathbf{f})}{n_{b_i}}, \quad (21)$$

where n_{b_i} is the number of training objects with a redshift value in bin b_i . The quantity $\hat{p}(b_i|\mathbf{f})$ is an estimate for the class probability that a newly queried object with photometry \mathbf{f} has a spectroscopic redshift inside the bin b_i . The method used to obtain the class probability estimates $\hat{p}(b_i|\mathbf{f})$ is interchangeable (e.g., using Neural Networks Bonnett 2015, or the Random Forest Frank & Bouckaert 2009; Carrasco Kind & Brunner 2013). We use the Random Forest algorithm for consistency with the Quantile Regression Forest and implemented the OCP algorithm using the ‘randomForest’ (Liaw & Wiener 2002) package for the R programming language (R Core Team 2013).

The original paper by Schapire et al. (2002) used the histogram estimator defined in Frank & Bouckaert (2009) as

$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N_{\text{tr}}} w_i(\mathbf{f}) \frac{I(b_{z_i} = b_z)}{r_{b_z}}. \quad (22)$$

Here b_z is defined as an index denoting the bin in which z is located and r_{b_z} denotes the corresponding bin width. We can interpret this histogram as a weighted kernel density estimate with value $r_{b_z}^{-1}$ for all training set objects in a bin specified by b_z and zero outside. Frank & Bouckaert (2009) improved the algorithm using a Gaussian kernel function and demonstrated its superiority over the histogram kernel in numerical experiments on machine learning benchmark datasets that are unrelated to the photometric redshift problem.

3.3 Bandwidth Selection

The algorithms we use to obtain PDFs for individual objects require the selection of an appropriate bandwidth for the weighted kernel density estimator (Eq. 6). This section proposes a bandwidth selection scheme that selects the bandwidth for the Gaussian kernel during model tuning using the MNLL.

The choice of a proper bandwidth depends on the shape of the underlying distribution and the number of objects available to construct the estimator. Assuming a normal distribution and a Gaussian kernel function one can obtain the optimal bandwidth as

$$\sigma_{\text{scott}} = 1.06 \frac{\hat{\sigma}}{N^{1/5}}, \quad (23)$$

where $\hat{\sigma}$ is the sample standard deviation and N denotes the number of objects. This so-called ‘Scott’s rule’ is commonly used in the machine learning and statistics literature (e.g. Takeuchi, Nomura & Kanamori 2009; Wang & Wang 2007). To apply this bandwidth selection rule to weighted data, we need to calculate the weighted standard deviation from the weighted training set using Eq. 20. Scott’s rule gives a good first estimate of a suitable bandwidth for distributions which are approximately normal.

Photometric redshift PDFs are in general not Normal distributions and Eq. 23 can pick a non-optimal bandwidth.

Thus, we modify Eq. 23

$$\sigma_{\text{scott}} = a \frac{\hat{\sigma}}{N^{1/5}}, \quad (24)$$

with a pre-factor a that is chosen to minimize the MNLL on the validation set. We can stack the N_{te} individual PDFs in the test set using an individual bandwidth σ_a for each object

$$\hat{p}(z) = \sum_{a=1}^{N_{\text{te}}} w_{\text{stack},a} \sum_{i=1}^{N_{\text{tr}}} w_i(\mathbf{f}_a) \mathcal{N}(z, z_i, \sigma_a), \quad (25)$$

or we can use a global bandwidth $\sigma_a = \sigma$.

3.4 The Gaussian Mixture Model Estimator

Storing the individual PDFs obtained by weighted kernel density estimation for every element in the test set requires a large amount of storage. Carrasco Kind & Brunner (2014) proposed several different methods, including a Gaussian mixture model, to more efficiently store a previously obtained estimate. The authors store individual PDFs using 10 - 20 numbers compared with 200 used previously. Instead of giving a previously estimated individual PDF a sparse representation, we fit the Gaussian mixture model directly to the weighted spectroscopic redshift values in the training set and ensure model sparsity by penalizing the model likelihood dependent on the number of components in the mixture model.

More specifically, we fit the Gaussian mixture to the weighted spectroscopic data with the expectation maximization algorithm (for an introduction see Chen & Gupta 2010) as implemented in the Rmixmap package (Biernacki et al. 2006; Auder et al. 2014). In §5.1 and §5.2 during the analysis using CFHTLS, we select the number of Gaussian components for each object in the test set using the normalized entropy criterion (Celeux & Soromenho 1996; Biernacki, Celeux & Govaert 1999), abbreviated as NEC in the following. The maximum number of Gaussian components that can be included in the mixture model is a parameter that is selected during model tuning as described in §2.4.

For a K -component Gaussian mixture model fitted on the weighted training data, the NEC criterion reads

$$NEC(K) = \frac{E(K)}{L(K) - L(1)} \quad (26)$$

where $L(K)$ denotes the maximum weighted log-likelihood

$$L(K) = \sum_{i=1}^N w(\mathbf{f}_i) \log \left(\sum_{k=1}^K \alpha_k \mathcal{N}(z_{\text{spec},i}, \mu_i, \sigma_i) \right) \quad (27)$$

for the K component Gaussian mixture model. The entropy $E(K)$ is defined as

$$E(K) = - \sum_{k=1}^K \sum_{i=1}^N w(\mathbf{f}_i) \gamma_k(z_{\text{spec},i}) \log(\gamma_k(z_{\text{spec},i})) \leq 0, \quad (28)$$

where the definition of the component weight proportions, following Eq. 12, is used. We pick the number of components K such that the NEC criterion is minimized, where $NEC(1) = 1$ (Biernacki, Celeux & Govaert 1999).

The NEC criterion normalizes the entropy by the maximum weighted log-likelihood, in which the offset for a one-component mixture is subtracted. There are two reasons

(Celeux & Soromenho 1996) why we cannot use the entropy $E(K)$ directly. The entropy for $K = 1$ provides a lower bound

$$E(K) \geq E(1) \quad \forall K > 1 \quad (29)$$

and the maximum weighted log-likelihood function is an increasing function of K , which makes $E(K)$ unequal for different values of K . The entropy term $E(K)$ measures how much overlap there is between the different components of the Gaussian mixture model. In the case where the components in the model fit completely separated data clusters, the entropy term approaches zero. If we select too many components, the quantity $E(K)$ will increase because the components will overlap strongly. This can be compensated by the higher likelihood of the more complex model. In this way, we can efficiently determine a suitable number of components to include into the mixture.

3.5 Highest Weight Element

A common application for individual PDFs is the estimation of the sample PDF. Storing and processing individual PDFs is computationally expensive. We propose the Highest Weight Element (hereafter HWE), a single point estimate for each object from which we can accurately reconstruct the sample PDF. We first run the QRF algorithm to determine weights as for individual PDF estimation. Instead of using the individual PDF, we select the spectroscopic redshift value that is associated with the maximum weight. If more than one spectroscopic redshift value has the same maximum weight, we randomly select one of those values.

4 DATASET

We use photometric imaging data from the CFHTLS Wide survey using the following bands u^* , g' , r' , i' and z' -band as obtained from the public CFHTLenS data release (Erben & CFHTLenS Collaboration 2012)⁴. We obtain the photometry analogously to Brimiouille et al. (2013), i.e., we degrade all images to match the band with the worst seeing, and use the unconvolved i' -band as the detection band and the convolved frames as the extraction band. Then we correct for the remaining zeropoint calibration uncertainties and varying galactic extinction by comparing the measured star colours from the catalogues with predictions of the Pickles star library (Pickles 1998). In this way, we eliminate possibly remaining field-to-field variations in the photometric calibration.

We then match our photometric catalogues to public spectroscopic redshift samples. These samples are the Visible Multiobject Spectrograph (VIMOS) VLT Deep Survey (VVDS) (Le Fèvre et al. 2004; Garilli et al. 2008), VVDS-F22, the Deep Extragalactic Evolutionary Probe-2 (DEEP-2) program (Davis et al. 2007; Vogt et al. 2005; Weiner et al. 2005) and the VIMOS Public Extragalactic Redshift Survey (VIPERS) (Garilli et al. 2014; Guzzo et al. 2014). We only make use of spectroscopic redshifts with confidence values

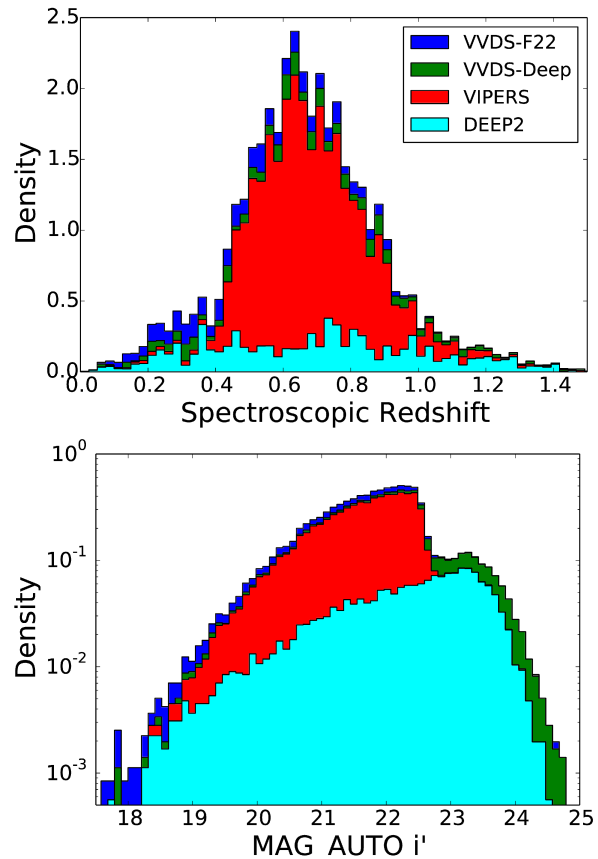


Figure 3. Spectroscopic redshift and $MAG_AUTO\ i'$ distributions of the compiled dataset described in §4. Objects matched from different spectroscopic surveys are indicated by different colors. We limit the spectroscopic redshift range to $z_{\text{spec}} < 1.5$ in the plots excluding 34 objects with higher redshift.

of at least 95% and only use pointings where the i' -data is available and where the i' -band serves as detection band.

This produces a total sample of 28159 objects with $i' \leq 22.5$ and additional 5893 objects with $22.5 < i' \leq 24.5$ with spectroscopic redshifts and five band photometry.

5 RESULTS

Future large area photometric surveys will produce large amounts of photometric data for which we need to obtain redshift information. Efficiency in terms of runtime and disk space will be important in order to use algorithms for photometric redshift estimation effectively on these large datasets. Additionally we are required to produce high quality photometric redshift PDFs in order to obtain accurate constraints on, for example, cosmological parameters or cluster masses.

We use the public CFHTLS data described in §4, to compare the accuracy of photometric redshift PDFs estimated by the algorithms described in §3. We show that these methods improve the modelling of angular correlation functions, cluster mass estimates, and the modelling of shear correlation functions compared to results obtained with the Neural Network code ANNz (Collister & Lahav 2004) commonly used in the literature (e.g., Sheldon et al.

⁴ <http://www.cfhtlens.org>

	η	$\sigma(\Delta z)$	$\langle \Delta z \rangle$	σ_{68}
ANNz	1.23%	0.092	-0.001	0.044
PhotoZ	2.27%	0.129	-0.008	0.050

Table 2. Point prediction performance of the Neural Network code ANNz and the template fitting code PhotoZ quantified by the metrics described in §5.1.

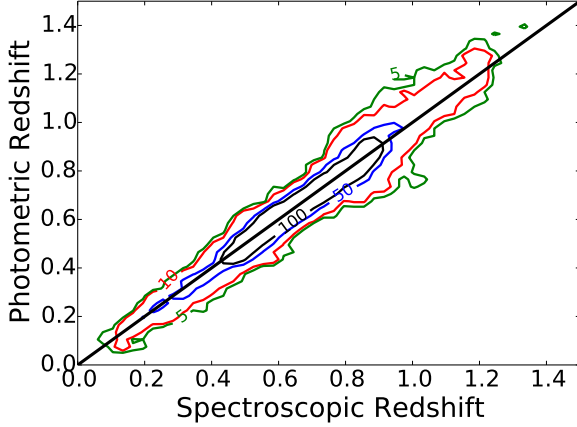


Figure 4. Density contours of photometric redshift estimates from ANNz against the spectroscopic redshift.

2009; Williamson et al. 2011; Smith et al. 2012; Planck Collaboration et al. 2015).

5.1 Comparison with ANNz

We train an ensemble of 20 Neural Networks with two hidden layers, each consisting of 12 nodes, following the methodology described in §2.4.

The photometric redshift estimates obtained from ANNz are competitive compared to those obtained with the template fitting code PhotoZ (Bender et al. 2001; Brimiouille et al. 2008; Greisel et al. 2013) in terms of common photometric redshift performance metrics. As shown in Table 2, ANNz improves upon the photometric redshift performance obtained with PhotoZ by 46%, 29%, 88% and 12% in terms of outlier rate, scatter, bias and spread of the residuals. The outlier rate η is defined as the fraction of objects with $|z_{\text{spec}} - z_{\text{phot}}| > 0.15$. The bias $\langle \Delta z \rangle$ and scatter $\sigma(\Delta z)$ are the mean and standard deviation of the distribution of the residuals $\Delta z = z_{\text{phot}} - z_{\text{spec}}$. The spread of the residual distribution is measured by the σ_{68} metric which is defined as half the difference between the 16% and 84% quantile.

The quality of the photometric redshifts obtained with ANNz is illustrated in Fig. 4. It shows a tightly aligned correlation between photometric and spectroscopic redshift. We estimate sample PDFs from the ANNz point predictions and the stacked Normal densities constructed from the ANNz error estimates, in the following referred to as ‘ANNz-stack’. While showing excellent point prediction performance, ANNz and ANNz-stack do not accurately estimate the sample PDF as shown in Fig. 5. The sample PDF constructed from ANNz-stack deviates from the true spectroscopic redshift distribution in the central redshift

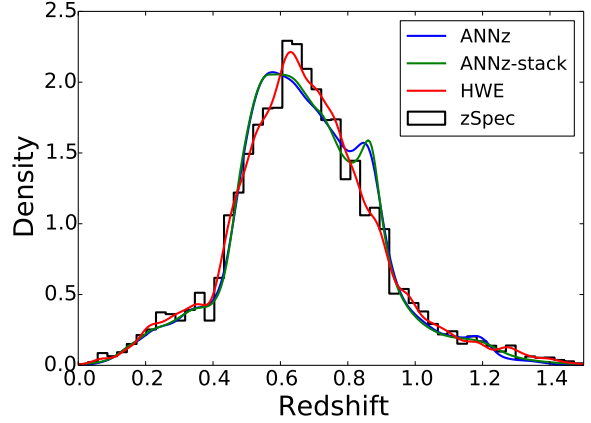


Figure 5. Sample PDF estimated using ANNz and the Highest Weight Element. The histogram shows the true spectroscopic redshift distribution.

	Total	[0, 0.585[[0.585, 0.7488[[0.7488, 3.818[
OCP	-1.3577	-1.3905	-1.6432	-1.0395
NOCP	-1.2847	-1.3029	-1.5880	-0.9648
QRF	-1.3483	-1.3627	-1.6470	-1.0347
GMM	-1.3181	-1.3591	-1.5606	-1.0354
ANNz	-1.1588	-1.3138	-1.4891	-0.6731

Table 3. MNLL of the Quantile Regression Forest (QRF), the classification based PDF estimation algorithms (OCP/NOCP) and the OCP algorithm used with the Gaussian mixture model (GMM). The values are evaluated over the full spectroscopic redshift range and in three bins. The result is illustrated in Fig. 6.

range [0.45, 0.85]. We will show in the following sections that these deviations from the true spectroscopic redshift PDF introduce a systematic bias in several important analyses in cosmology. To compare the quality of photometric redshift PDFs of individual objects (individual PDFs), we evaluate the MNLL (Eq. 15) of the four discussed algorithms (QRF, NOCP, OCP and OCP GMM) on the full range of redshift values and in three redshift bins ([0, 0.585[, [0.585, 0.7488[and [0.7488, 3.818[). The results are shown in Table 3 and illustrated in Fig. 6. QRF, NOCP and OCP employ the weighted kernel density estimate. OCP GMM denotes the Gaussian mixture model applied in combination with weights determined using the ordinal classification method described in §3.2.

We illustrate the relative improvement MNLL_{rel} gained by applying these algorithms compared with ANNz-stack

$$\text{MNLL}_{\text{rel}} = \left(\frac{\text{MNLL}_{\text{ANNz}} - \text{MNLL}_{\text{alg.}}}{|\text{MNLL}_{\text{ANNz}}|} \right) \quad (30)$$

in Fig. 6. A high value in terms of MNLL_{rel} translates into an improvement in the log-likelihood of the individual PDFs over those obtained with ANNz-stack. The boundaries of the redshift intervals are picked such that they contain approximately the same number of test set objects. All discussed methods improve over ANNz-stack. For the highest redshift objects, our methods show improvement of up to 50%. The OCP routine performs the best and improves the NOCP routine. This verifies the superiority of the ordinal classi-

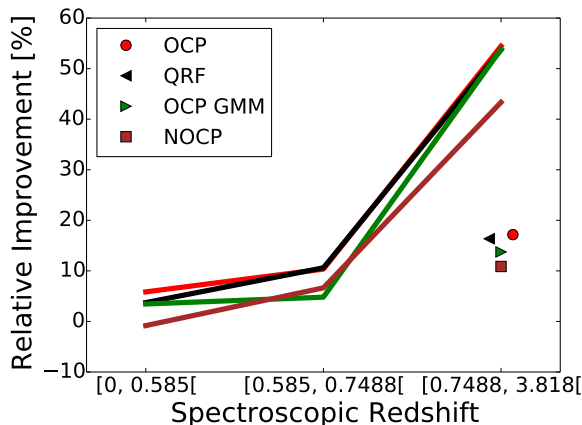


Figure 6. Relative improvement in MNLL over the performance of ANNz-stack. We compare the classification-based PDF estimators (OCP, NOCP), the ordinal classification PDF estimator combined with a Gaussian mixture model (OCP GMM) and the Quantile Regression Forest (QRF) in three spectroscopic redshift bins. The plotted points show the average improvement over the full spectroscopic redshift range.

cation technique. The Quantile Regression Forest performs on par with OCP. OCP GMM shows mediocre results, but provides the most efficient parametrization using a single Normal density per object.

5.2 Stacked photometric redshift distribution

Applications like shear tomography require the photometric selection of objects in a certain redshift range. We stack the individual PDFs compared in §5.1 using weights that quantify their overlap with a certain redshift interval using Eq. 10. These estimates are compared with the weighted kernel density estimate obtained from the spectroscopic redshift values using the same weights. The weights are determined using each of the OCP, NOCP, OCP GMM and QRF methods individually. The Highest Weight Element (HWE) uses the weighted kernel density estimate with weights determined using the QRF algorithm. We use Scott’s rule to choose the bandwidth for the weighted kernel density estimates of the HWE and the spectroscopic redshift values. The sample PDFs obtained with the OCP, NOCP and QRF algorithms are very similar. We therefore restrict the following discussion to the OCP method.

The results shown in Fig. 7 compare the weighted sample PDFs obtained with the HWE, OCP and OCP GMM methods in the redshift intervals $[0, 0.585[$, $[0.585, 0.7488[$ and $[0.7488, 3.818[$. They differ mainly in the amount of smoothing present in the estimate. Notably the OCP GMM method oversmooths features in the density estimate. This is because a single Gaussian was selected during model tuning based on the performance of individual object PDFs. Allowing more components reduces the amount of smoothing. The HWE is competitive with methods that estimate the individual PDFs, with the advantage that the HWE is extremely efficient to calculate and, being a point estimate, requires storing only a single floating point number per object.

The weighted distributions of all methods have tails that extend outside the desired redshift range. We can reduce these tails by stacking only the objects with the highest weight in the respective redshift bin as demonstrated in the lower right panel of Fig. 7. We estimate the sample PDF from the HWE predictions of the objects with the 5000 highest weights in the respective redshift bin. The estimated weighted sample PDF of these objects has less overlap with neighbouring redshift bins, compared with the estimate that incorporates all objects. Furthermore it agrees well with the equally weighted spectroscopic redshift distribution of the corresponding objects.

Instead of weighting the objects in the respective redshift range, we can select objects based on a photometric redshift point estimate in analogy to Benjamin et al. (2013). We perform the same cut in $\text{MAG_AUTO } i' < 23.0$ and estimate the sample PDF in the same photometric redshift intervals selected after our ANNz estimate. The results for the HWE are shown in Fig. 8 and agree well with the spectroscopic redshift distribution. The agreement is better in the central bins, which contain more objects, because the histogram approximates the underlying distribution better.

5.3 Applications to Cosmology

We now investigate how the previously discussed methods can be used to improve analyses that use photometric redshifts. We estimate the sample PDF using the Highest Weight Element and ANNz. We use kernel density estimates with bandwidths selected using Scott’s rule.

Where required, we impose a flat Λ -CDM cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $n_s = 0.96$, $H = 0.7$, $\sigma_8 = 0.79$.

5.3.1 The Angular Power Spectrum

The angular power spectrum measures the clustering of galaxies and is an important tool to constrain cosmological models.

In the following we adopt the notation of Thomas, Abdalla & Lahav (2010). Consider the line-of-sight projection of the 3D mass distribution in the universe, δ_{2D} . The harmonic modes of δ_{2D} are given by

$$\delta_\ell = i^\ell \int \frac{d^3k}{(2\pi)^3} \delta(\mathbf{k}) W_\ell(k), \quad (31)$$

where the window function $W_\ell(k)$ is sensitive to the sample PDF of light sources, $p(z)$, and can be computed by the integral

$$W_\ell(k) = \int p(z) D(z) \left(\frac{dz}{dx} \right) j_\ell(kz) dz. \quad (32)$$

Here $D(z)$ is the linear growth factor, $j_\ell(kz)$ are the Bessel functions and $\left(\frac{dz}{dx} \right)$ relates the redshift to the radial comoving coordinate x .

The angular power spectrum C_ℓ , is the variance of the modes δ_ℓ^5 ,

$$C_\ell = \langle \delta_\ell \delta_\ell^* \rangle = 4\pi \int \Delta^2(k) W_\ell^2(k) \frac{dk}{k}, \quad (33)$$

⁵ In our analysis, we are assuming a galaxy-dark matter bias equal to one.

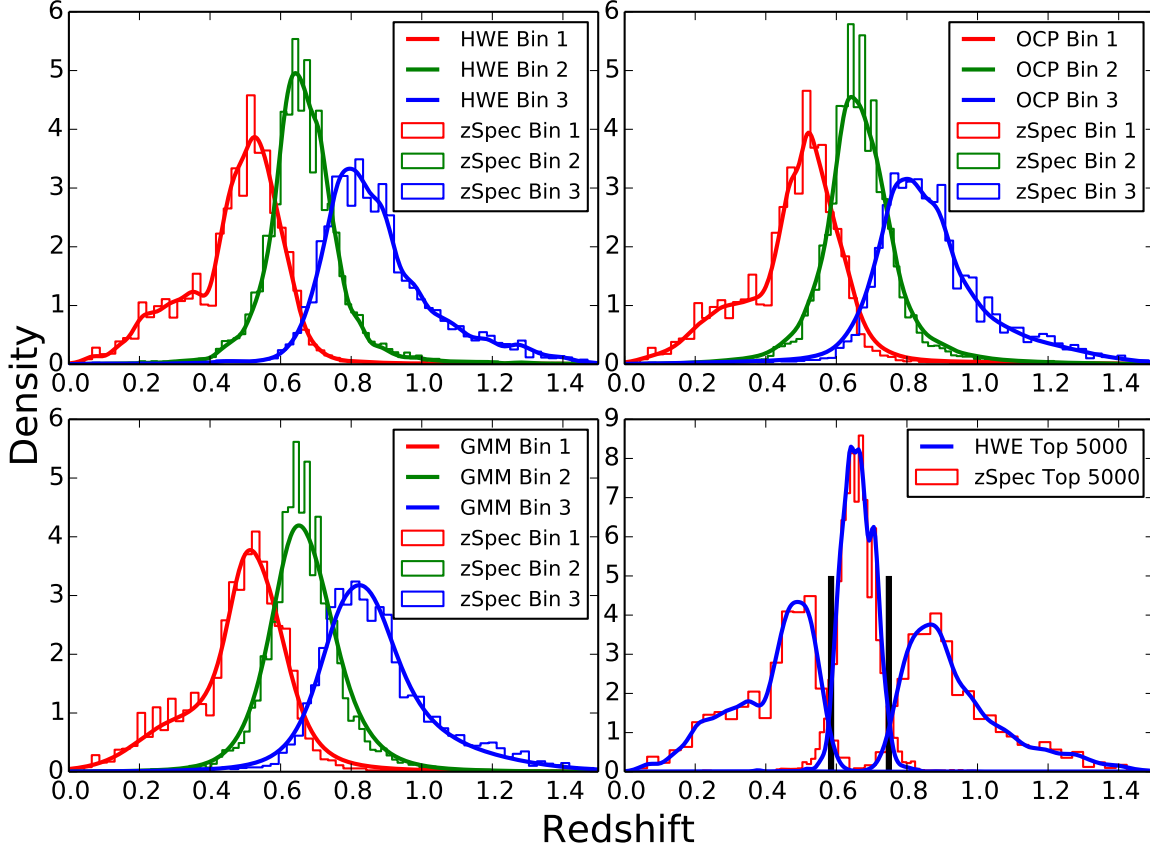


Figure 7. Sample PDFs weighted in three redshift intervals $[0, 0.585]$, $[0.585, 0.7488]$ and $[0.7488, 3.818]$. The PDFs are obtained using the Highest Weight Element (upper left), the ordinal classification PDF estimator (upper right) and the ordinal classification PDF estimator combined with a Gaussian mixture model (lower left). The histograms show the weighted spectroscopic redshift distribution using weights determined using the respective algorithms. The lower right panel shows the weighted distribution of the HWE predictions for the objects with the 5000 highest weights in the three intervals (blue) and the corresponding weighted histogram of spectroscopic redshifts (red).

where the dimensionless 3D power spectrum $\Delta^2(k)$ is given in terms of the usual 3D matter power spectrum $P_\delta(k)$ as

$$\Delta^2(k) = \frac{4\pi k^3 P_\delta(k)}{(2\pi)^3}. \quad (34)$$

From Eq. 32 it can be seen, that the modelling of C_ℓ depends highly on the assumed sample PDF of the data. We use the distributions shown in Fig. 5 to model the angular correlation power spectrum with the CLASS software package (Blas, Lesgourgues & Tram 2011). We define the bias introduced by the C_ℓ^{phot} of the angular correlation function estimated using photometric redshifts, as the relative difference to the results based on the PDF of spectroscopic redshifts C_ℓ^{spec} :

$$\text{Bias}_{C_\ell} = \left(\frac{C_\ell^{\text{phot}} - C_\ell^{\text{spec}}}{C_\ell^{\text{spec}}} \right). \quad (35)$$

The resulting biases are shown in Fig. 9. We find that the results obtained with the Highest Weight Element have a lower systematic bias in C_ℓ by a factor of four compared to the ANNz results and that the improvement is almost independent of ℓ .

5.3.2 Gravitational Lensing

We investigate two important applications in gravitational lensing: quantifying cluster masses by the light deflection from background sources, and obtaining cosmic shear correlation functions. In contrast to the previously considered analysis of the angular correlation function, applications in gravitational lensing require careful selection of sources with successfully measured shapes. Since the spectrophotometric dataset used previously is not representative for datasets generally used in gravitational lensing analyses, we first weight our catalogue such that it mimics a CFHTLS shape catalogue. To do this, we obtain a photometric shape catalogue from public CFHTLS data, which is then used as the reference to weight the spectrophotometric dataset.

5.3.3 Catalogue Creation and Weighting

Whether the shape of an object can be measured, depends primarily on its intrinsic size and magnitude in the respective band. We therefore reweight our spectrophotometric catalogue such that it resembles a CFHTLS shape catalogue

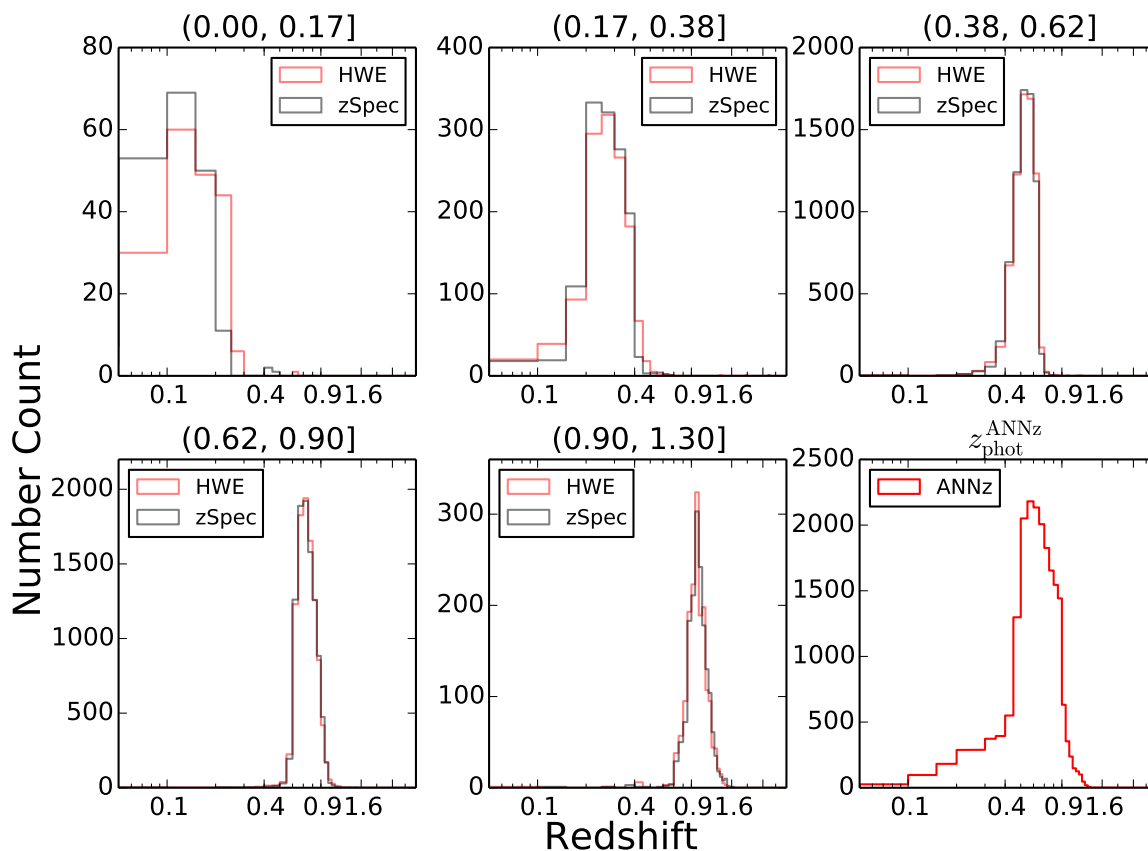


Figure 8. Sample PDFs estimated using the HWE for subsamples selected in analogy to Benjamin et al. (2013, Fig. 1) using a cut at $\text{MAG_AUTO } i' < 23.0$. The subsamples are selected using the photometric redshift estimates from ANNz in intervals shown in the subfigure titles.

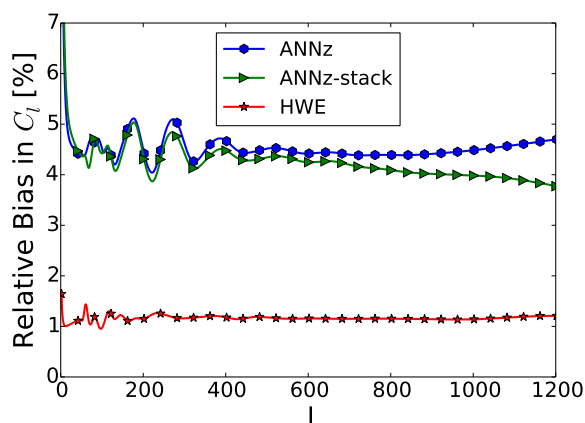


Figure 9. Bias in the angular correlation power spectrum obtained for different estimates for the sample PDF. We restrict the comparison to $\ell < 1200$.

in terms of these properties. We obtain the shape catalogue in analogy to Brimiouille et al. (2013) for the full CFHTLS survey region. Intrinsic sizes s_{intr} are calculated for each object from the measured $\text{FWHM}_{\text{image}}$ and corrected for seeing

as follows

$$s_{\text{intr}} = \sqrt{\text{FWHM}_{\text{image}}^2 - \langle \text{FWHM}_{\text{psf}} \rangle^2}, \quad (36)$$

where $\langle \text{FWHM}_{\text{psf}} \rangle$ is the average size of the point spread function for the respective chip⁶.

In this way, we obtain s_{intr} and $\text{MAG_AUTO } i'$ entries for each object in the shape and spectrophotometric catalogue. We now determine weights for the spectrophotometric catalogue such that, after weighting, it matches the size and magnitude distribution of the shape catalogue. Furthermore, the results obtained with the reweighted spectrophotometric catalogue have to be robust against the removal of the objects with the highest weights (Sánchez et al. 2014). Since we do not have enough spectroscopically observed objects to mimic the shape catalogue at the faint end, we have to employ a magnitude cut in order to fulfill both requirements. For the analyses presented in §5.3.4 and §5.3.5, we employ a magnitude cut at $\text{MAG_AUTO } i' < 23.5$ and $\text{MAG_AUTO } i' < 23.0$, respectively. We give a detailed discussion of these cuts in the appendix.

⁶ We work on image stacks, but (as in Brimiouille et al. 2013) only consider objects, for which all images contribute to the stack from the same CCD-chip.

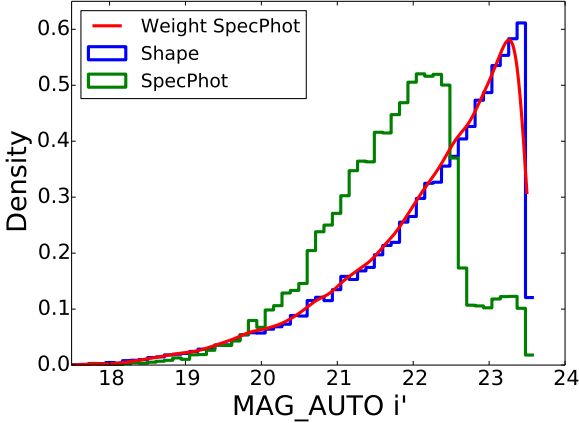


Figure 10. Distributions in $\text{MAG_AUTO } i'$ band for the original spectrophotometric dataset, the re-weighted spectrophotometric dataset and the shape catalogue for $\text{MAG_AUTO } i' < 23.5$.

We combine bootstrap re-sampling with the k nearest neighbour estimator to determine weights for the elements in the spectrophotometric catalogue, such that the weighted catalogue mimics the distribution of the shape catalogue in the two dimensional space spanned by the intrinsic size of the objects and their magnitude $\text{MAG_AUTO } i'$. To this end, we draw bootstrap samples from the shear catalogue and find the k nearest neighbours in the spectrophotometric catalogue. The nearest neighbour of an object in the shear catalogue with the lowest Euclidean distance to this object. Accordingly, the k nearest neighbour algorithm selects the k nearest objects with respect to the Euclidean distance. The number of times an object in the spectrophotometric catalogue is selected as one of the k nearest neighbours corresponds to its weight. This process is similar to previous work done by Lima et al. (2008), which employs a nearest neighbour based approach to determine weights for objects in a spectroscopic sample to estimate the sample PDF of the photometric data. In contrast to our method, which is based on bootstrap re-sampling, they calculate the density ratio between the distributions characterizing the two catalogues using a nearest neighbour approach. For the data at hand, we draw 10^6 bootstrap samples and consider three nearest neighbours $k = 3$. This method accurately weights the spectrophotometric data to mimic the size and i-band magnitude distributions of the shape catalogue, as shown in Figs. 10 and 11. The following analysis uses the estimated weights to weight the sample PDF of ANNz, ANNz-stack, the Highest Weight Element and the spectroscopic data as shown in Fig. 12.

5.3.4 Cluster Mass Measurement

Galaxy Clusters are one of the primary tools to probe the Λ -CDM model (for a review, see e.g. Allen, Evrard & Mantz 2011). Cluster masses can be determined by measuring the

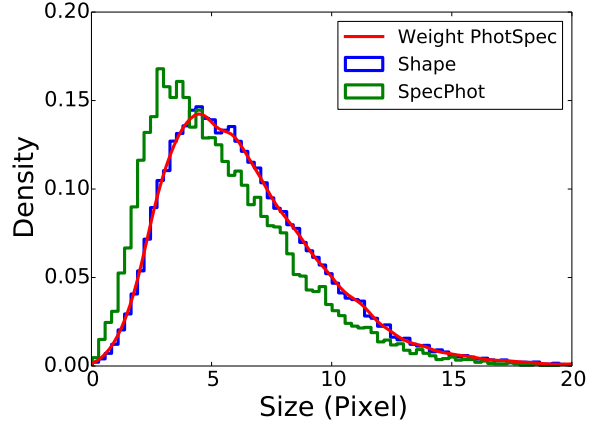


Figure 11. Distributions in intrinsic size (Eq. 36) for the original spectrophotometric dataset, the re-weighted spectrophotometric dataset and the shape catalogue for $\text{MAG_AUTO } i' < 23.5$.

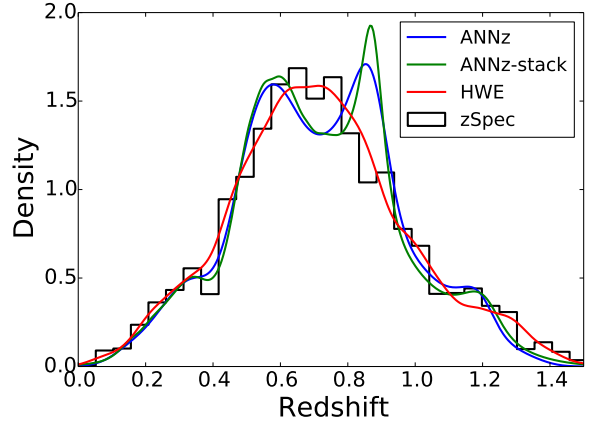


Figure 12. Weighted stacked sample PDF estimated using ANNz, ANNz-stack and the Highest Weight Element. The histogram shows the weighted spectroscopic redshift distribution. We use a cut on $\text{MAG_AUTO } i' < 23.5$. The used weights and cuts are described in §5.3.3.

tangential alignment of gravitationally lensed galaxies⁷ located behind the clusters. The accuracy of these weak lensing mass estimates suffers from uncertainties in the photometric redshift of the lensed sources. In combination with other effects such as cluster mass profile variances, they can introduce systematics at the 5% to 10% level (see e.g. Applegate et al. 2014). In the following, we will only consider uncertainties due to errors in photometric redshift estimates (Seitz & Schneider 1997; Mandelbaum et al. 2008; Dawson et al. 2012; Gruen et al. 2013, 2014; Applegate et al. 2014). The excess surface density inside radius R

$$\langle \Sigma(r) \rangle_{r < R} - \bar{\Sigma}(R) = \Sigma_{\text{crit}} \gamma_{\text{tan}}(R) \quad (37)$$

⁷ For a introduction into gravitational lensing we refer to Bartelmann & Schneider (2001).

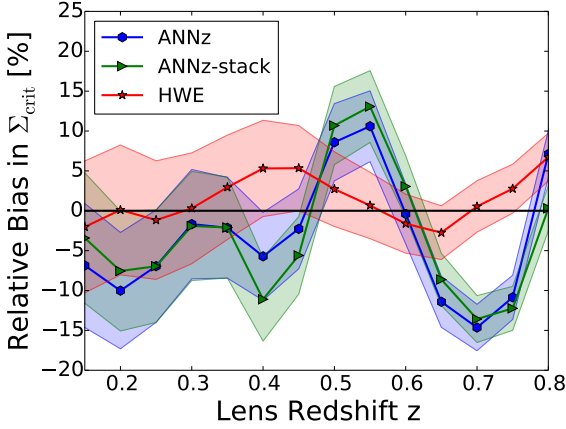


Figure 13. Relative bias in the mean critical surface density (Eq. 39) for different lens redshifts obtained using different estimates for the sample PDF. The filled area shows the 1σ error interval.

is proportional to the critical surface density

$$\langle \Sigma_{\text{cr}} \rangle \propto \int_{z_{\text{Lens}}}^{\infty} dz p(z) \left(\frac{D_d(z_{\text{Lens}}) D_{ds}(z_{\text{Lens}}, z)}{D_s(z)} \right) \quad (38)$$

of the lens at redshift z_{Lens} . Here D_d , D_s and D_{ds} denote the angular diameter distance to the lens, the source and between the lens and the source respectively. Uncertainties in the sample PDF of background sources $p(z)$ will propagate into systematic errors in the determination of the critical surface density. This introduces systematic errors in the excess surface density and therefore in the cluster mass estimate.

We quantify the systematic bias of the critical surface density as

$$\text{Bias}_{\langle \Sigma_{\text{cr}} \rangle} = \left(\frac{\langle \Sigma_{\text{cr}} \rangle_{\text{photo}} - \langle \Sigma_{\text{cr}} \rangle_{\text{spec}}}{\langle \Sigma_{\text{cr}} \rangle_{\text{spec}}} \right), \quad (39)$$

where $\langle \Sigma_{\text{cr}} \rangle_{\text{photo}}$ is estimated from the photometry of the objects (e.g., using machine learning) and $\langle \Sigma_{\text{cr}} \rangle_{\text{spec}}$ from the spectroscopic redshifts. We estimate the error σ on this bias with respect to our test set containing N objects as

$$\sigma^2 = \left(\frac{\sigma_{\text{photo}}(\Sigma_{\text{cr}})}{\sqrt{N} \langle \Sigma_{\text{cr}} \rangle_{\text{spec}}} \right)^2. \quad (40)$$

The mean and standard deviation of the distribution of Σ_{cr} are estimated using the probability density function estimates obtained from ANNz and the Highest Weight Element and we present the results in Fig. 13.

The Highest Weight Element estimate for the sample PDF reduces the systematic bias in the critical surface density compared with ANNz by a factor of four for medium lens redshifts $z \in [0.45, 0.6]$. The systematic bias in $\langle \Sigma_{\text{cr}} \rangle$ obtained from the HWE is consistent with zero for lens redshifts $z < 0.7$ and, in general, outperforms the results obtained with ANNz. Higher lens redshifts are however unrealistic for current survey depths.

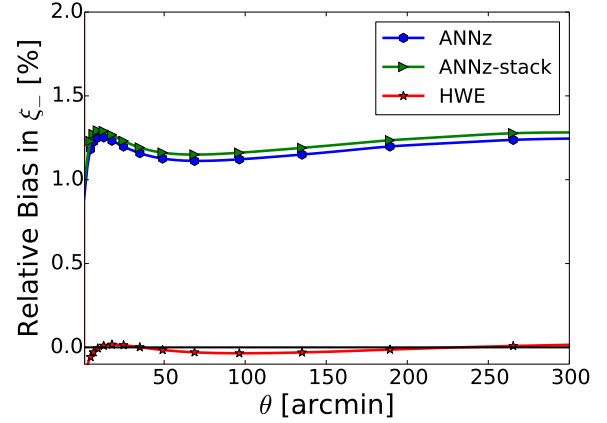


Figure 14. Relative bias in the shear correlation function estimate for ξ_- (Eq. 43) obtained using different estimates for the sample PDF.

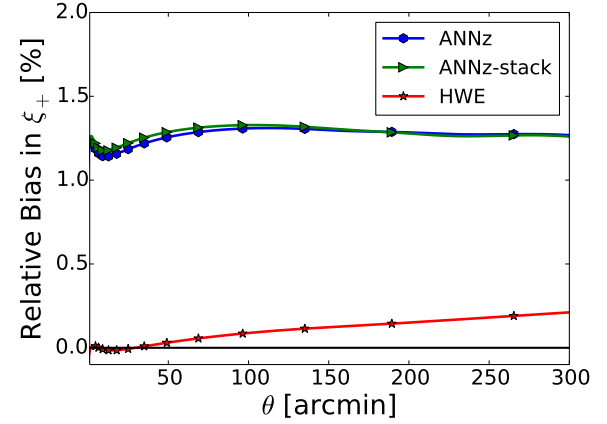


Figure 15. Relative bias in the shear correlation function estimate for ξ_+ (Eq. 43) obtained using different estimates for the sample PDF.

5.3.5 Cosmic Shear

Cosmic shear is the weak lensing effect generated by the inhomogeneous matter distribution of the universe and has become an important tool to constrain cosmological parameters (see, e.g., Kilbinger et al. 2013, and references therein). Similar to our discussion of the angular correlation function, we derive a power spectrum $P_{\kappa}(\ell)$ of the lensing convergence κ , which is the source of the lensing potential, defined with respect to the radial co-moving coordinate x

$$P_{\kappa}(\ell) = \int_0^{\infty} dx \left(\frac{q^2(x)}{x^2} \right) P_{\delta} \left(\frac{\ell}{x}, x \right). \quad (41)$$

We calculate the power spectrum $P_{\delta} \left(\frac{\ell}{x}, x \right)$ using the halo fit formula from Smith et al. (2003). The lensing efficiency $q(x)$ quantifies how strongly the objects in an infinitesimal shell of radial comoving coordinates deflect the light coming from background sources. Since the radial comoving coordinates of the objects are related to their redshifts, the lensing efficiency $q(x)$ depends on the sample PDF $p(z)$. From the lensing convergence power spectrum, we can obtain the two

shear correlation functions (Kaiser 1992) as

$$\xi_{\pm}(\theta) = \frac{1}{2\pi} \int_0^{\infty} d\ell \ell P_{\kappa}(\ell) J_{0,4}(\ell\theta), \quad (42)$$

where the Bessel function J_0 (J_4) corresponds to the ξ_+ (ξ_-) correlation function. In analogy to the previous sections we quantify the bias in the shear correlation functions obtained from photometric data ξ_{\pm}^{phot} by their relative error with respect to the results obtained from the spectroscopic data ξ_{\pm}^{spec} ,

$$\text{Bias}_{\xi_{\pm}} = \left(\frac{\xi_{\pm}^{\text{photo}} - \xi_{\pm}^{\text{spec}}}{\xi_{\pm}^{\text{spec}}} \right). \quad (43)$$

The results are presented in Figs. 14 and 15. We reduce the bias in the shear correlation function estimates, using the Highest Weight Element estimate instead of the photometric redshift estimates from ANNz, by a factor of 12 for ξ_- and a factor of 6 for ξ_+ .

6 SUMMARY AND CONCLUSIONS

The next generation photometric surveys will measure the positions on the sky of thousands of millions of galaxies. We must be able to reliably estimate the distance to, or the redshift of, each photometrically identified galaxy before we can use these galaxies in analyses to derive the values of cosmological parameters. Furthermore to maximize the precision and accuracy of any derived parameters, we require a complete understanding of the full shape of the photometric redshift probability density function (hereafter PDF) for both each individual object, and the entire galaxy sample.

In this work we develop and discuss methods drawn from machine learning, to accurately estimate photometric redshift PDFs, which will meet both the future storage demands of large surveys, and the precision demands for cosmological parameter estimation.

As a working example, we apply these algorithms to a sample of galaxies selected from the CFHTLS survey for a set of cosmological analyses. We demonstrate that these methods reduce the biases in all of the analyses examined. We also show that these biases result from the mishandling of the full shape of the photometric redshift PDFs.

This advancement is quantified by comparing several accurate methods to estimate photometric redshift PDFs for individual objects (hereafter individual PDFs). We estimate individual PDFs using a classification scheme that classifies objects into redshift bins and thereby constructs the PDF using the probabilities for bin membership. In contrast to the classification-based PDF estimation methodology commonly used in the astrophysics literature, we incorporate the order of consecutive redshift bins into the classification framework. This produces more accurate individual PDFs. We quantify the performance of the methods by measuring the average log-likelihood of all PDF estimates in a test sample. Our method outperforms other non-ordinal classification and regression schemes, for example classification trees and Neural Networks. Specifically, for high redshift objects, our method reaches performance gains of over 50% in average log-likelihood when compared with the results obtained using the common Neural Network code ANNz. We construct the individual PDFs using kernel density estimation

which inherently requires the selection of a suitable bandwidth to govern the smoothing scale. We propose an efficient method to choose the smoothing scale on an object by object basis. We further discuss a Gaussian mixture model, whose complexity is adaptively selected for each individual object, using a criterion that penalizes model complexity. This method shows solid performance compared with kernel density estimates, while providing a more efficient parametrization of individual PDFs.

Many cosmological analyses require an accurate knowledge of the full shape of the galaxy sample PDF, instead of estimates for the individual PDFs of each galaxy. Sample PDFs are typically obtained by stacking the PDFs of individual galaxies, and so their estimation and storage is required. This reconstruction of the individual PDF typically requires the storage of several hundred floating point numbers. Complex post processing algorithms can reduce this number to 10 - 20 floating point numbers per object at the expense of additional computation time. However in this work, we propose a new single point estimator for each galaxy, called Highest Weight Element (HWE), which can be used to accurately reconstruct the full sample PDF. This leads to a significant reduction in the storage requirements of future photometric surveys. Furthermore, we note that reconstructing the full sample PDF using the point estimator method described in this paper requires orders of magnitude less computation time than using other common redshift codes.

Applications such as shear tomography require the accurate photometric selection of objects in redshift bins. We weight photometrically observed galaxies such that their sample PDF lies within the predefined redshift range. The weights are estimated from the overlap between the individual redshift PDFs and the redshift selection interval. We further use these weights to improve the selection of a sample of galaxies, such that their sample redshift PDF is more accurately confined to be within the predefined redshift bin.

We now return our attention to the specific use case highlighted above using CFHTLS galaxies. In particular we examine the following cosmological analyses: the estimation of cluster masses using weak gravitational lensing, the modelling of galaxy angular correlation functions, and the modelling of cosmic shear correlation functions. In each case we compare the results, and estimate biases, using results obtained with ANNz.

For lensing clusters within the redshift interval $0.45 < z < 0.6$, we show that our methods reduce the relative bias in the cluster mass reconstruction by up to a factor of 4. Furthermore our methods improve the relative biases in the modelling of the explored large scale structure, and cosmic shear correlation functions by similar values.

In this paper we have shown that the usual point estimate of a photometric redshift is a poor estimator when used to reconstruct the full sample redshift PDF. We note that these point estimates are still used in many recent analyses, and we have shown that their continued use can lead to large biases in cosmological analysis. By using the new HWE point estimator method, highlighted in this paper, we show that the full shape of the sample PDF can be estimated more accurately and that this reduces the biases incurred by mis-estimating the sample PDF.

The results discussed in this paper have been obtained

under the idealized assumption that the data used to train the models is completely representative of the test data. In applications where this is not the case, data augmentation techniques (Hoyle et al. 2015a) can be used to artificially populate regions of color-magnitude space, that are not fully covered by spectroscopy. These techniques assume a model for the data distribution and can be seen as a form of extrapolation. Weighting methods (§5.3.3) are in some cases an alternative to data augmentation. If all relevant attributes are included, these algorithms can be used to determine weights, such that the weighted dataset resembles a reference dataset.

To aid the common adoption of these tools and techniques we will make the source code of all algorithms publicly available on the homepage of the first author.

7 ACKNOWLEDGEMENTS

S. Seitz thanks Ofer Lahav for drawing her attention towards machine learning techniques several years ago, and for inspiring discussions on machine learning vs. template fitting during mutual visits.

M. M. Rau especially thanks Jolanta Krzyszowska and Natascha Greisel for useful discussions.

This work was supported by SFB-Transregio 33 ‘The Dark Universe’ by the Deutsche Forschungsgemeinschaft (DFG) and the DFG cluster of excellence ‘Origin and Structure of the Universe’.

APPENDIX: TESTS OF WEIGHTING SCHEME

The analyses in §5.3.5, §5.3.4 have been carried out by weighting the photospectroscopic dataset such that it resembles a shape catalogue. If only a few objects in the reweighted catalogue are given high weights, the analyses can strongly depend on these objects. We lack spectroscopically observed objects at the faint end of the shape catalogue and therefore employ a magnitude cut to avoid giving large weight to the faint, unrepresentative part of the spectrophotometric catalogue. In analogy to Sánchez et al. (2014), we test the robustness of our weighting scheme with respect to the considered applications by excluding the top 5% of the objects that are given the highest weights.

The bias in the critical surface density is robust against the exclusion of the highest weighted objects for a magnitude cut at $\text{MAG_AUTO } i' < 23.5$ as shown in Fig. 16. The results improve for all algorithms if these objects are removed. The conclusions of the analysis, i.e. that the Highest Weight Element (HWE) leads to a lower bias compared with ANNz, remain valid.

The analysis of the biases incurred in estimates of the cosmic shear correlation functions requires a more conservative cut at $\text{MAG_AUTO } i' < 23.0$, to be robust against the removal of a small number of highly weighted objects, as can be seen in Fig. 17 and Fig. 18. For a magnitude cut at $\text{MAG_AUTO } i' < 23.5$, ANNz⁸ gives a better overall result compared with the HWE, while the opposite is true if the 5% objects with the highest weight are left out.

⁸ The results for ANNz-stack are very similar. Therefore we do not show them here.

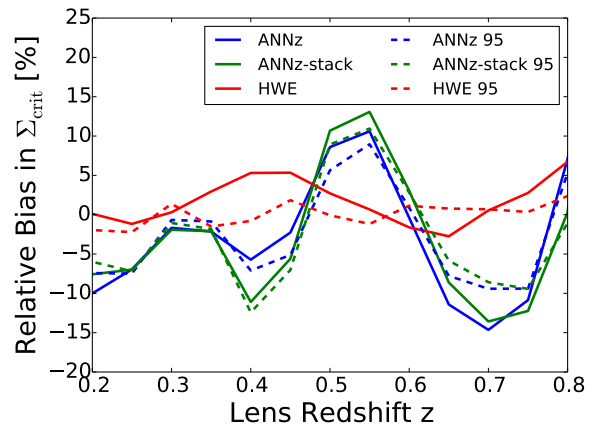


Figure 16. Relative bias in the mean critical surface density (Eq. 39) for different lens redshifts obtained using different estimates for the sample PDF. We show the relative biases obtained for the weighted dataset cut at $\text{MAG_AUTO } i' < 23.5$ in solid lines, and the corresponding results with the 5% highest weighted objects removed in dashed lines.

Note that this is *not* because the $p(z)$ reconstruction of ANNz is superior at faint magnitudes. Instead this can be explained by considering the bias in the integrand in Eq. 42 with respect to the spectroscopic result given as

$$\text{Bias} = \frac{\ell}{2\pi} J_{0,4}(\ell\theta) \left(P_{\kappa}^{\text{phot}}(\ell) - P_{\kappa}^{\text{spec}}(\ell) \right). \quad (44)$$

As shown in Fig. 19, ANNz both partly underestimates and overestimates the true spectroscopic integrand at different redshift values such that these two effects compensate each other. Since the lensing efficiency is dominated by the high redshift tail of the stacked PDF, the peculiar shape of the ANNz reconstruction in this range happens to outperform the otherwise superior HWE method. The shape of the high redshift tail strongly depends on a small number of faint objects, which are given a high weight. Accordingly, this artifact is no longer present if the top 5% of the objects with the highest weights are left out. For a more conservative cut at $\text{MAG_AUTO } i' < 23.0$, the analysis is no longer dominated by a few highly weighted objects at the faint end of our spectrophotometric catalogue, the ANNz analysis does not outperform the HWE, and the interpretation does not depend on the removal of the objects with the highest weights.

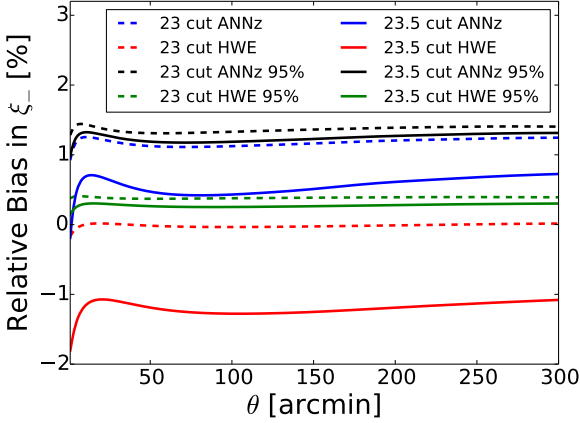


Figure 17. Relative bias in the shear correlation function estimate for ξ_- (Eq. 43) obtained using different estimates for the sample PDF. We show the relative biases obtained for the weighted dataset cut at $\text{MAG_AUTO } i' < 23.5$ in solid lines and $\text{MAG_AUTO } i' < 23.0$ in dashed lines and the corresponding results with the 5% highest weighted objects removed.

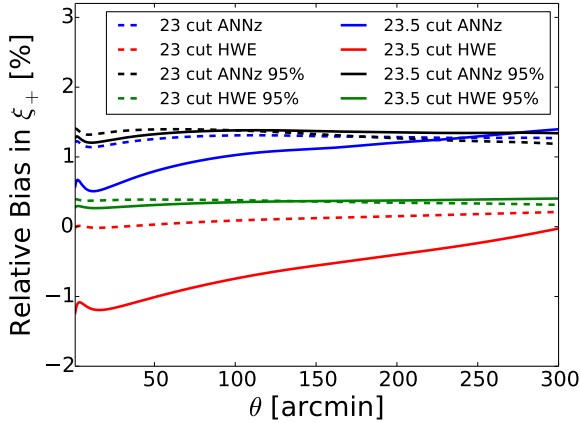


Figure 18. Relative bias in the shear correlation function estimate for ξ_+ (Eq. 43) obtained using different estimates for the sample PDF. We show the relative biases obtained for the weighted dataset cut at $\text{MAG_AUTO } i' < 23.5$ in solid lines and $\text{MAG_AUTO } i' < 23.0$ in dashed lines and the corresponding results with the 5% highest weighted objects removed.

REFERENCES

- Allen S. W., Evrard A. E., Mantz A. B., 2011, *Annu. Rev. Astron. Astrophys.*, 49, 409
 Applegate D. E. et al., 2014, *MNRAS*, 439, 48
 Auder B., Lebrete R., Iovleff S., Langrognat F., 2014, *Rmixmod: An interface for MIXMOD*. R package version 2.0.2
 Bartelmann M., Schneider P., 2001, *Physics Reports*, 340, 291
 Bender R. et al., 2001, in *Deep Fields*, Cristiani S., Renzini A., Williams R. E., eds., p. 96
 Benítez N., 2000, *The Astrophysical Journal*, 536, 571
 Benjamin J. et al., 2013, *MNRAS*, 431, 1547
 Biernacki C., Celeux G., Govaert G., 1999, *Pattern Recognition Letters*, 20, 267

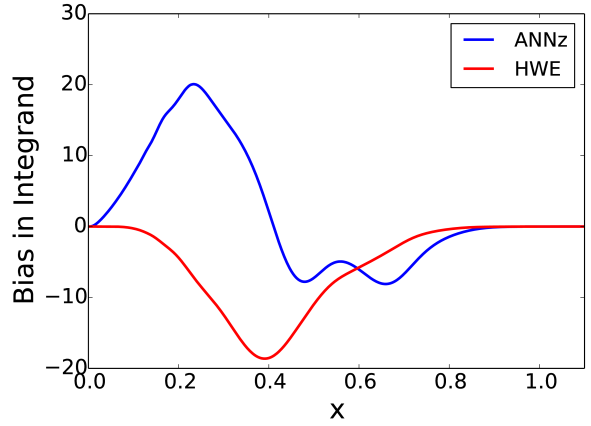
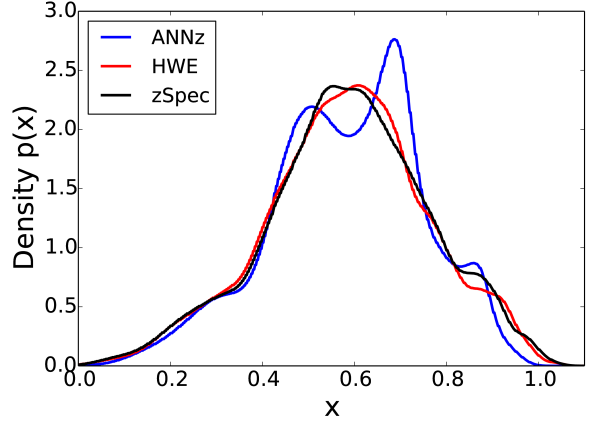


Figure 19. The sample PDFs for a cut at $\text{MAG_AUTO } i' < 23.5$ expressed in radial comoving coordinates x for the spectroscopic, ANNz and HWE reconstructions (top panel). The bias (Eq. 44) obtained with the sample PDFs from ANNz and the HWE for the example $l = 100$ (bottom panel).

- Biernacki C., Celeux G., Govaert G., Langrognat F., 2006, *Computational Statistics & Data Analysis*, 51, 587
 Bishop C. M., 2006, *Pattern Recognition and Machine Learning*. Springer
 Blas D., Lesgourgues J., Tram T., 2011, *JCAP*, 7, 34
 Bonnett C., 2015, *MNRAS*, 449, 1043
 Breiman L., 2001, *Mach. Learn.*, 45, 5
 Brimiouille F., Lerchster M., Seitz S., Bender R., Snigula J., 2008, preprint (arXiv:0811.3211)
 Brimiouille F., Seitz S., Lerchster M., Bender R., Snigula J., 2013, *MNRAS*, 432, 1046
 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, 441, 3550
 Celeux G., Soromenho G., 1996, *Journal of Classification*, 13, 195
 Chen Y., Gupta M. R., 2010, *Em demystified: An expectation-maximization tutorial*
 Colless M. et al., 2001, *MNRAS*, 328, 1039
 Collister A. A., Lahav O., 2004, *PASP*, 116, 345
 Csabai I., Connolly A. J., Szalay A. S., Budavári T., 2000, *AJ*, 119, 69
 Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, *MNRAS*, 396, 2379

- Dahlen T. et al., 2013, *ApJ*, 775, 93
- Davis M. et al., 2007, *ApJL*, 660, L1
- Dawson W. A. et al., 2012, *ApJL*, 747, L42
- de Leeuw J., Hornik K., Mair P., 2009, *Journal of Statistical Software*, 32, 1
- de Simoni F. et al., 2013, *MNRAS*, 435, 3017
- Drinkwater M. J. et al., 2010, *MNRAS*, 401, 1429
- Duin R., 1976, *IEEE Transactions on Computers*, 25, 1175
- Erben T., CFHTLenS Collaboration, 2012, in *American Astronomical Society Meeting Abstracts #219*, American Astronomical Society Meeting Abstracts
- Feldmann R. et al., 2006, *MNRAS*, 372, 565
- Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, 339, 1195
- Flaugher B., 2005, *International Journal of Modern Physics A*, 20, 3121
- Frank E., Bouckaert R., 2009, *Springer Proceedings of First Asian Conference on Machine Learning*, ACML, Nanjing China, 2-4
- Frank E., Hall M., 2001, in *Lecture Notes in Computer Science*, Vol. 2167, *Machine Learning: ECML 2001*, De Raedt L., Flach P., eds., Springer Berlin Heidelberg, pp. 145–156
- Garilli B. et al., 2014, *Astron. & Astrophys.*, 562, A23
- Garilli B. et al., 2008, *Astron. & Astrophys.*, 486, 683
- Greisel N., Seitz S., Drory N., Bender R., Saglia R. P., Snigula J., 2013, *ApJ*, 768, 117
- Gruen D. et al., 2013, *MNRAS*, 432, 1455
- Gruen D. et al., 2014, *MNRAS*, 442, 1507
- Guzzo L. et al., 2014, *Astron. & Astrophys.*, 566, A108
- Habbema, Hermans, Van den Broek, 1974, A stepwise discrimination analysis program using density estimation. *Compstat: Proceedings in Computations Statistics*
- Hildebrandt H. et al., 2010, *Astron. & Astrophys.*, 523, A31
- Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015a, *MNRAS*, 450, 305
- Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015b, *MNRAS*, 449, 1275
- Ilbert O. et al., 2006, *Astron. & Astrophys.*, 457, 841
- Kaiser N., 1992, *ApJ*, 388, 272
- Kilbinger M. et al., 2013, *MNRAS*, 430, 2200
- Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
- Le Fèvre O. et al., 2004, *Astron. & Astrophys.*, 417, 839
- Le Fèvre O. et al., 2005, *Astron. & Astrophys.*, 439, 845
- Liaw A., Wiener M., 2002, *R News*, 2, 18
- Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, *MNRAS*, 390, 118
- Mandelbaum R. et al., 2008, *MNRAS*, 386, 781
- Meinshausen N., 2006, *Journal of Machine Learning Research*, 7, 983999
- Pickles A. J., 1998, *PASP*, 110, 863
- Planck Collaboration et al., 2015, preprint (arXiv:1502.01595)
- R Core Team, 2013, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria
- Sánchez C. et al., 2014, *MNRAS*, 445, 1482
- Schapire R. E., Stone P., McAllester D. A., Littman M. L., Csirik J. A., 2002, in *ICML*, Sammut C., Hoffmann A. G., eds., Morgan Kaufmann, pp. 546–553
- Scott D. W., 1992, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 1st edn. Wiley
- Seitz C., Schneider P., 1997, *Astron. & Astrophys.*, 318, 687
- Sheldon E. S. et al., 2009, *ApJ*, 703, 2217
- Smith D. J. B. et al., 2012, *MNRAS*, 427, 703
- Smith R. E. et al., 2003, *MNRAS*, 341, 1311
- Staniszewski Z. et al., 2009, *ApJ*, 701, 32
- Sugiyama M., Takeuchi I., Suzuki T., Kanamori T., Hachiya H., Okanohara D., 2010, *IEICE Transactions on Information and Systems*, E93-D, 583
- Takeuchi I., Nomura K., Kanamori T., 2009, *Neural Computation*, 21, 533
- Thomas S. A., Abdalla F. B., Lahav O., 2010, preprint (arXiv:1011.2448)
- Tonry J. L. et al., 2012, *ApJ*, 750, 99
- Tyson J. A., Wittman D. M., Hennawi J. F., Spergel D. N., 2003, *Nuclear Physics B Proceedings Supplements*, 124, 21
- Vogt N. P. et al., 2005, *ApJS*, 159, 41
- Wang B., Wang X., 2007, preprint (arXiv:0709.1616)
- Weiner B. J. et al., 2005, *ApJ*, 620, 595
- Williamson R. et al., 2011, *ApJ*, 738, 139
- York D. G. et al., 2000a, *AJ*, 120, 1579
- York D. G. et al., 2000b, *AJ*, 120, 1579

Fixing biases from estimating z-distributions

In the following I attach the preprint version of my paper ‘Correcting cosmological parameter biases for all redshift surveys induced by estimating and reweighting redshift distributions’, published in the ‘Monthly Notices of the Royal Astronomical Society, Volume 466, Issue 3, p.2927-2938’. The authors of this paper are: Markus Michael Rau, Ben Hoyle, Kerstin Paech, Stella Seitz. This preprint version was updated to match the published version.

Scientific Context We demonstrated in the last section, how Machine Learning can be used to estimate accurate photometric redshift distributions for individual galaxies and samples of galaxies. However, we did not answer the question, how one can derive error intervals for these estimates. Photometric redshift distributions can be biased by a variety of effects, e.g. due to cosmic variance, errors in the photometry or selection effects introduced by the spectroscopic targeting strategy. This paper exclusively focuses on the intrinsic error in density estimates like e.g. histograms, kernel density estimates or Machine Learning based conditional density estimation. We perform a galaxy clustering forecast to demonstrate, that this intrinsic error can be large enough to propagate into significant biases in cosmological parameter constraints. We resolve these errors using a modification of the classical bootstrap method. This modified bootstrap, or smoothed bootstrap, improves upon the classical bootstrap technique, that is typically applied in this context (e.g. Bonnett et al., 2016), since it is able to estimate the bias in the density estimate. This bias is produced by the oversmoothing of the redshift distribution estimate, due to the selection of a too large bandwidth; an effect that cannot be quantified using the classical bootstrap approach. As a result, we can first correct the parameter contours for the bias in the density estimate and then marginalize over its statistical uncertainty, i.e. its variance. When both corrections are applied, we are able to recover the original, unbiased cosmological parameter constraints to sufficient accuracy. We want to highlight, that even in the absence of other sources of error, photometric redshift accuracy will be limited by the intrinsic accuracy of density estimates. This also defines requirements on the minimum number of unbiased spectroscopic calibration redshifts, that are necessary to constrain cosmological parameters with a given statistical precision.

We note that resampling methods have been applied in previous work, not only to correct for the variance in the photometric redshift distribution, but also to quantify the effect of large-scale structure correlations between neighboring photometric redshift bins (e.g. Mandelbaum et al., 2008). We want to reiterate, that the purpose of this work is to correct for the intrinsic

error in a density estimate, that is different from large-scale structure induced redshift biases.

To the best of my knowledge the resampling method proposed in the following paper was not yet applied in the context of photometric redshift estimation, nor was the effect of inaccurate selection of bin width explicitly studied in the photometric redshift literature. Finally I would like to note, that the smoothed bootstrap method, while previously proposed in the statistics literature (see e.g. Scott, 1992), is to the best of my knowledge not yet established in the Machine Learning literature on conditional density estimation. Furthermore I found no mention of my approach to apply the smoothed bootstrap to weighted samples in the literature.

In the following paragraph I briefly comment on challenges that arose during the analyses.

Challenges In the following paper we use a Fisher forecast to propagate biases in the photometric redshift distribution into biases in the cosmological parameter constraints. This choice is motivated by the immense computational workload of a full MCMC approach. Fisher forecasts require considerable tuning to ensure an accurate numerical calculation of the derivatives. I start with a large parameter difference in the calculation of the numerical derivative, that I successively decrease. If this parameter difference is too small, the calculated derivative begins to fluctuate, since I hit the numerical accuracy of the angular correlation power spectrum calculation.¹ In contrast, if I select a difference that is too large, I obtain an insufficient numerical accuracy. The goal of my tuning procedure is therefore to select the smallest possible parameter difference step, while still ensuring numerical stability. I repeated my tuning for all cosmological parameters in turn, which led to stable results. I would also like to highlight, that I tested ‘rule of thumb’ approaches of choosing the difference size equal to the expected 1σ error of the particular parameter. Neither of these approaches yielded overall satisfactory results. In my opinion, careful tuning of the derivative in each individual parameter with subsequent consistency checks using the full parameter set, is the only viable option to ensure sufficient numerical accuracy.

I also extensively checked for numerical errors in my calculations, which I suspected e.g. in the matrix inversion of the covariance matrix.

In this project the usage of Fisher forecasts was necessary due to computational limitations. Due to the aforementioned problems, I would highly discourage their use, if a full MCMC approach is feasible.

Contributions and Acknowledgements I prepared the text and performed the analyses as detailed in the following. I implemented the Fisher forecast code including a module that propagates the photometric redshift uncertainties into the cosmological parameter constraints. For the calculation of the angular correlation power spectrum I used the *cosmosis* software (Zuntz et al., 2015). I further prepared the tomographic redshift distributions, where I used the mock catalog from Jouvel et al. (2009). To bin the tomographic redshift distributions I obtained photometric redshifts using the *randomForest* package (Liaw & Wiener, 2002). To estimate the histograms and kernel density estimates I used the following packages: *astroML* (Vanderplas et al., 2012), *SciPy* (Jones et al., 2001–), *Numpy* (van der Walt et al., 2011), *statsmodels* (Seabold & Perktold, 2010). I implemented the smoothed bootstrap methods, wrote the whole analysis pipeline and prepared the figures.

¹I use the *cosmosis* software (Zuntz et al., 2015) in this work.

BH helped in the early stages of the project by cross checking some results I obtained using his modification of the CLASS software. BH, KP, SS provided useful discussion during the project and proof read the text.

Correcting cosmological parameter biases for all redshift surveys induced by estimating and reweighting redshift distributions

Markus Michael Rau^{1,2}, Ben Hoyle¹, Kerstin Paech^{1,3}, Stella Seitz^{1,2}

¹*Ludwig-Maximilians-Universität München, Universitäts-Sternwarte, Scheinerstr. 1, D-81679 Munich, Germany*

²*Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, 85748 Garching, Germany*

³*Excellence Cluster Universe, Boltzmannstr. 2, D-85748 Garching, Germany*

E-mail: mmrau@usm.lmu.de

Accepted —. Received —; in original form —.

ABSTRACT

Photometric redshift uncertainties are a major source of systematic error for ongoing and future photometric surveys. We study different sources of redshift error caused by choosing a suboptimal redshift histogram bin width and propose methods to resolve them. The selection of a too large bin width is shown to oversmooth small scale structure of the radial distribution of galaxies. This systematic error can significantly shift cosmological parameter constraints by up to 6σ for the dark energy equation of state parameter w . Careful selection of bin width can reduce this systematic by a factor of up to 6 as compared with commonly used current binning approaches. We further discuss a generalised resampling method that can correct systematic and statistical errors in cosmological parameter constraints caused by uncertainties in the redshift distribution. This can be achieved without any prior assumptions about the shape of the distribution or the form of the redshift error. Our methodology allows photometric surveys to obtain unbiased cosmological parameter constraints using a minimum number of spectroscopic calibration data. For a DES-like galaxy clustering forecast we obtain unbiased results with respect to errors caused by suboptimal histogram bin width selection, using only 5k representative spectroscopic calibration objects per tomographic redshift bin.

Key words: galaxies: distances and redshifts, catalogues, surveys.

1 INTRODUCTION

Ongoing and future photometric surveys such as DES (Flaugher 2005), KIDS (de Jong et al. 2013) and Euclid (Laureijs et al. 2011) will photometrically observe hundreds of millions of galaxies. With this rapid increase in statistical power, comes the need to control systematic uncertainties with even higher accuracy, if we wish to remain in the era of precision cosmology. One of the dominant sources of systematic error in these broad band photometric surveys is our ability to obtain accurate distance information characterised by the photometric redshift for the observed galaxies. The accuracy in the photometric redshift distribution for a selected galaxy sample is particularly important, since it enters into the modelling of a wide variety of measurements. Examples of these include projected two point statistics like angular correlation power spectra, or estimates of the crit-

ical surface density of a cluster that is important for weak lensing cluster mass measurements (Rau et al. 2015; Bonnett et al. 2016). Misestimating the photometric redshift distribution will introduce biases in the respective theoretical models that will cause errors in the modelling of the signal and hence lead to biased estimates for e.g. cosmological parameters or cluster masses.

The main goal of these large area photometric surveys is to improve our understanding of dark energy and the growth of structure. A particularly important probe for this are accurate measurements of two point statistics which are, as mentioned, quite sensitive to errors in the photometric redshift distribution. Since the lack of accuracy in photometric redshift estimates already challenges current multiband photometric surveys like CFHTLenS (Choi et al. 2016; Kitching et al. 2016), DES (Sánchez et al. 2014; Bonnett et al. 2016) or KIDS (Hildebrandt et al. 2016), it will likely retain its

relevance in the next decade where Euclid will probe the Universe to even fainter magnitudes.

The primary methods to obtain photometric redshift point estimates and estimates of the redshift distribution for individual galaxies are template fitting methods (e.g. Koo 1985; Benítez 2000; Bender et al. 2001; Leistedt, Mortlock & Peiris 2016), empirical methods based on machine learning (e.g. Carrasco Kind & Brunner 2013; Bonnett 2015; Rau et al. 2015) and combinations of the two (Hoyle et al. 2015; Speagle et al. 2016; Speagle & Eisenstein 2015b,a; Beck et al. 2016). Alternatively one can use cross correlations between photometric and spatially overlapping spectroscopic samples to obtain photometric redshift information (Newman 2008; Ménard et al. 2013). While these cross correlation techniques show great promise and are already applied to photometric data sets (Hildebrandt et al. 2016; Rahman et al. 2016), the aforementioned established methods remain the state of the art in photometric redshift estimation and are therefore the main focus of this paper.

Template fitting uses models of the spectral energy distribution (SED) of the different types of galaxies and fits them against the measured photometry to constrain their redshift. As a limited number of broad photometric bands only provides limited information about the SED, the color space spanned by these templates is typically degenerate. This means that the galaxy photometry can be represented by several SED templates and redshifts. If the wrong template is fit to the photometry, large photometric redshift errors can occur that can shift cosmological constraints (Hearin et al. 2010). Empirical methods using machine learning have recently become a popular and accurate method for photometric redshift estimation that often outperform contemporary template fitting techniques (Sánchez et al. 2014). Instead of using theoretical SED templates to model the mapping between photometry and redshift, these methods ‘learn’ it directly from spectroscopic calibration data. This data is taken from the spatially overlapping regions between a spectroscopic and the photometric survey and thus provides both photometric and spectroscopic information. Flexible machine learning methods can then use this data to mimic the mapping between photometry and redshift. The result of this fitting process is a model that can provide photometric redshifts for all galaxies in the photometric data set. The process of fitting these flexible models to the color-redshift space of the calibration data can also be supported by extending the available calibration data using artificial galaxies from simulations or SED templates (Hoyle et al. 2015). In this way we can incorporate our understanding of galaxy evolution and the shape of galaxy SEDs into an otherwise completely data driven process. However empirical methods assume, that the calibration data is representative of the true photometric science sample. If the calibration data is not representative of the full science sample, the algorithm can produce biased photometric redshifts, since the model is forced to extrapolate into unknown regions of color-magnitude space.

While empirical methods based on machine learning naturally dependent on representative spectroscopic data, both methods require them to verify their results. These calibration data sets are usually much smaller than the photometric catalogues for which they provide redshift calibration. The main reason for this is the lack of accurate spectro-

scopic redshift measurements for faint galaxies. Overlapping spectroscopic surveys are not able to completely cover the faint end of the color-magnitude distribution of the photometric survey, because taking spectra at high magnitudes is extremely expensive and requires long exposure times. As a result, the photometric redshifts of significant portions of the faint photometric science sample cannot be verified using accurate spectroscopic redshifts (e.g. Bonnett et al. 2016). Photometric data from these regions is unreliable for usage in cosmological analyses. Thus it needs to be removed (Cunha et al. 2014; Bonnett et al. 2016; Hildebrandt et al. 2016) or small samples of spectroscopic redshifts need to be upweighted to obtain a representative validation catalogue (Sánchez et al. 2014; Rau et al. 2015). The spectroscopic redshift distributions constructed on these weighted spectroscopic validation catalogues can then be used to test the quality of the photometric redshift distributions. However as these spectroscopic redshift distributions are constructed with a limited number of data that is strongly upweighted in the sparsely populated high redshift tail, they will be noisy and thus show a large error. This limits our ability to accurately validate photometric redshift distributions and thus contributes to the total error of the final measurement.

We note that this source of error is relevant independent of the method used to generate photometric redshift estimates. Even methods that fit SED templates and do not directly use spectroscopic galaxies during training, also need to be validated on spectroscopic data. The spectroscopic redshift distributions that are constructed during the validation process are then subject to the aforementioned sources of error. This in turn limits the accuracy of photometric redshift validation.

In this paper we study how this statistical error propagates into cosmological parameter shifts in a DES like galaxy clustering forecast. The goal is to explore how the error in the redshift distribution can be reduced and how the remaining uncertainty can be incorporated into the parameter likelihood. In §4.1 we will show that the selection of a too large histogram bin width can shift cosmological parameter constraints. Subsequently we compare several different binning strategies to reduce this error and provide guidelines for their successful application. The following section §4.2 then considers how cosmological parameter constraints are affected by introducing weights to a sample. Some previous work has been done to incorporate errors in the photometric redshift distribution into the parameter likelihood. Most notably the recent work by Bonnett et al. (2016) uses an analytical model for the bias in the tomographic redshift bins and selects a prior on this parameter by comparing several photometric redshift codes. In contrast we study the application of bootstrap techniques to incorporate the uncertainty in photometric redshifts into the parameter likelihood without imposing a specific model. The bootstrap was used in the work from Cunha et al. (2009); Sánchez et al. (2014) and more recently Bonnett et al. (2016); Hildebrandt et al. (2016) to quantify photometric redshift uncertainty from statistical shot noise. In §5 we improve upon the bootstrap by studying the ‘smoothed bootstrap’, a modification of this popular resampling method. In addition to the accurate treatment of the statistical shot noise, this new method is also able to correct for systematic shifts in the parameter likelihoods caused by the selection of a too large bin width.

In §6 we conclude with a discussion on how these methods can be best applied in ongoing and future large area photometric surveys.

2 DATA

To mimic the typical shape of photometric redshift distributions, we use the public galaxy mock catalogue from Jouvel et al. (2009), that resembles the color-redshift space of future imaging surveys like DES or Euclid. We remove data with large magnitude errors $\sigma_{\text{mag}} > 0.1$ to produce a catalogue that contains 13k objects with a median value in the i band magnitude of 21. Spectroscopic surveys measure accurate redshifts for their observed galaxies. In contrast, the photometric redshifts available to imaging surveys have a higher redshift uncertainty and photometric tomographic redshift bins can therefore have broad, non Gaussian, or even multimodal shapes (e.g. Benjamin et al. 2013; Sánchez et al. 2014; Becker et al. 2015; Bonnett et al. 2016).

If the shape of a distribution shows non Gaussian structure, it becomes increasingly hard to estimate this more complicated function with a limited number of samples. In order to obtain the shape of realistic photometric tomographic redshift bins, we need to simulate the way samples of galaxies are selected by their photometric redshift. We therefore first estimate photometric redshift predictions for our sample using the Random Forest (Liaw & Wiener 2002) algorithm. This method is a popular algorithm for photometric redshift estimation (Carrasco Kind & Brunner 2013; Sánchez et al. 2014; Rau et al. 2015) and was adopted as one of the standard photometric redshift codes within the DES collaboration.

Using 5 band photometry in g , r , i , z and Y , we obtain a photometric redshift performance with a mean and scatter on the residuals $\Delta z = z_{\text{phot}} - z_{\text{spec}}$ of $\langle \Delta z \rangle = -0.0010$ and $\sigma(\Delta z) = 0.095$. This is comparable with the photometric redshift performance obtained for the DES science verification data as reported in Sánchez et al. (2014). Using these predictions, we split the sample into 5 photometric tomographic redshift bins, such that each tomographic bin contains approximately the same number of objects. The true redshifts of these galaxy samples is then used to estimate the distribution of the tomographic redshift bins. We define these five tomographic redshift distributions as the true underlying redshift distributions. In the remaining paper we will draw new catalogues of varying sizes from these tomographic redshift distributions. This allows us to compare various estimators for the redshift distribution on these samples. Their accuracy can then be compared with the true underlying redshift distributions. We note that it is therefore important for this analysis to use simulated datasets, as the true underlying redshift distribution has to be known. In real data this underlying truth is never perfectly known. As will be shown in §4.1 biases can persist even in the presence of a very large number of calibration galaxies. In using the true redshift, we implicitly assume that the algorithm used to produce the photometric redshift distributions does not contribute to further biases in the redshift distribution. This is an optimistic assumption as there is typically a noticeable disagreement between codes (e.g. Sánchez et al. 2014; Bonnett et al. 2016) that estimate photometric redshift dis-

tributions and this will further contribute to the total error budget. Accordingly the total error of the estimated redshift distributions used in the next sections will likely be higher in practise than assumed here.

3 METHODOLOGY

In the following section we give a brief introduction into density estimation and describe the methods used to select the smoothing scale, i. e. the bin width, in the density estimate. We discuss the different sources of error in density estimates and describe a resampling method to estimate and correct these errors. Finally we describe the Fisher formalism used to propagate the error in the redshift distribution into shifts in the cosmological parameters. To avoid confusion between the ‘bias’ in redshift distributions and the ‘bias’ in cosmological parameters, we will refer to biases on cosmological parameters as cosmological parameter ‘shifts’.

3.1 Density Estimation

The modelling of cosmological observables like angular correlation functions depends on the accurate modelling of the redshift distribution of the tracers. The most common estimator for these distributions is the histogram. To obtain a smooth function that allows for accurate integration, the density at the midpoints is interpolated using spline interpolation. An alternative estimator to obtain smooth density estimates is the kernel density estimate (KDE) that interpolates the density by placing Gaussians on each sample point.¹ Each of these estimators $\hat{p}(z)$ only approximate the underlying distribution $p(z)$, and their mean squared error can be decomposed into a bias and variance component as

$$E [\hat{p}(z) - p(z)]^2 = \text{Var} \{\hat{p}(z)\} + \text{Bias}^2 \{\hat{p}(z)\}, \quad (1)$$

where the variance and bias terms are defined as

$$\text{Var} \{\hat{p}(z)\} = E [(\hat{p}(z) - E[\hat{p}(z)])^2] \quad (2)$$

and

$$\text{Bias} \{\hat{p}(z)\} = E[\hat{p}(z)] - p(z). \quad (3)$$

The bias and variance of a density estimate as defined in Eq. 2 and 3 are functions of redshift and quantify the error from the full shape of the redshift distribution. The bias of a density estimate determines how closely the model fits the data. Picking a small bin width leads to a very noisy density estimate. It has a low bias as the small scale features of the particular sample are closely fit. However the variance in this estimate will be large, since some of its bumps may be spurious and not actual features of the underlying distribution to be estimated. Thus the same density estimate, e.g. a histogram with the same fine grained binning, will look quite different for multiple samples independently drawn from the same parent distribution. In contrast, picking a large bin width leads to very smooth functions with a low variance.

¹ Instead of Gaussians other so called kernel functions can be used instead. These are non-negative real-valued integrable functions that integrate to unity and exhibit axis symmetry.

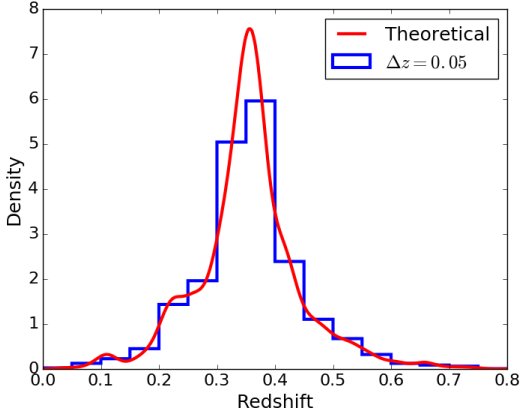


Figure 1. Histogram applied to 16000 objects sampled from the second tomographic bin shown in Fig. 2. The bin width $\Delta z = 0.05$ was chosen in analogy to Benjamin et al. (2013).

The density estimate is very stable but can oversmooth important small scale structure in the underlying distribution as shown in Fig. 1. We randomly sample 16000 galaxies from the second tomographic bin shown in Fig. 2 and apply a histogram with the same bin width used in Benjamin et al. (2013). The estimate significantly underestimates the peak of the true t-shaped distribution. Increasing the number of samples will not significantly improve upon this result, as the fixed bin width does not allow the histogram to further adapt to the narrow peak. In order to gain improvement, it is necessary to select a smaller bin size, which will also reduce the risk of oversmoothing. As illustrated in this example, the accuracy of a density estimate like the histogram will primarily depend on the amount of smoothing, i.e. the bin- or kernel width, in the estimate². The optimal bin width depends on the number of data samples as well as the shape of the underlying distribution to be estimated. Larger samples allow for smaller bins without producing noisy histograms. Narrow distributions also require narrower bins, to properly estimate the structure of the peak. The selected bin width affects the bias and variance terms in Eq. 1 in opposite ways and the tradeoff between both types of error needs to be balanced in any real density estimate to produce the lowest possible total error. Thus we need to adapt the bin size as a function of the number of samples and the shape of the distribution. In the following section we will discuss methods that optimize the size of the bins for this purpose.

3.1.1 The Histogram Estimate

The histogram is a density estimate that approximates the underlying density as a step function. The height of each step is proportional to the number of objects falling into a particular grid cell and the smoothing scale of the histogram is determined by the width of these bins. As the modelling of cosmological observables contains integrals over

² To simplify the notation, we refer to the bin size as the parameter that governs the smoothing of all density estimates that will be discussed in this work.

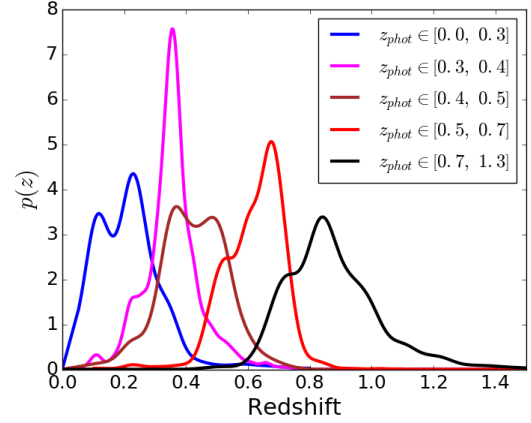


Figure 2. Tomographic redshift distributions generated from the mock catalogue as discussed in §2. The legend shows the photometric redshift bins used to generate the distributions. These tomographic bins are used as reference densities to generate new mock catalogues in a Monte Carlo experiment as explained in §4.

the redshift distribution, software packages like *cosmosis* (Zuntz et al. 2015) interpolate the midpoints of the histograms with splines to perform this integration accurately and efficiently. Following the *cosmosis* software package we use the Akima spline interpolation scheme (Akima 1970). This method minimizes spurious wiggles for low density values that otherwise pose a problem when using cubic spline interpolation. As the density estimate can still be negative due to numerical errors, we set all negative density values to zero and renormalize. We want to ensure the same numerical accuracy for all considered density estimates, irrespective of the selected bin width. Thus we evaluate each interpolated histogram on a fixed grid with 1000 equally spaced grid-points over the whole redshift range $z \in (0.0, 1.5)$. For the following analysis we compare a histogram evaluated on a fixed grid using a bin width of $\Delta z = 0.05$ in analogy to Benjamin et al. (2013), with a bin width selection scheme that adapts the size of the bin as a function of the number of samples and the shape of the distribution. Assuming a Gaussian reference distribution, one can show (Scott 1992) that the optimal bin width for a linearly interpolated histogram is

$$h = 2.15 \hat{\sigma} n^{-1/5}, \quad (4)$$

where $\hat{\sigma}$ is the estimator for the standard deviation and n is the number of galaxies.³ Even though this rule was derived for the case of a linearly interpolated histogram, it performs well for the case of an Akima spline interpolation as shown later in §4.1. Note that Eq. 4 also depends on the shape of the distribution as parametrized by the sample standard deviation. This ensures that the density estimate favours smaller bin widths for strongly peaked distributions which reduces the chance of ‘oversmoothing’ peaks in the density estimate. In the following we will refer to the histogram bin width selection rule defined in Eq. 4 as the adaptive bin width selection rule.

³ The bin width relates to the number of bins as $k = \lceil (\max z - \min z)/h \rceil$.

3.1.2 Kernel Density Estimate

The kernel density estimate (KDE) approximates the underlying distribution as a sum of Gaussians centered on the sample points. More formally, the density $p(z)$ is determined by interpolating the density between the weighted sample points z_i with weights w_i using Gaussians $\mathcal{N}(z|z_i, \sigma)$ centred on each sample point z_i

$$\hat{p}(z) = \sum_{i=1}^n w_i \mathcal{N}(z|z_i, \sigma). \quad (5)$$

The standard deviation σ of the Gaussians determines how smooth the resulting estimate will be. Broad Gaussians over-smooth the small scale structure of the underlying density and have a similar effect to selecting a too large histogram bin width. In contrast, a standard deviation that is too small can lead to spurious wiggles in the resulting estimate. For simplicity, the following discussion will globally refer to the bin width as the smoothing parameter of kernel density estimates. Assuming that the underlying distribution is Gaussian or close to Gaussian, the Scott rule selects a bin width that minimizes the error in the density estimate. To estimate the underlying density using a sample of size n , the optimal value for σ is given as

$$\sigma_{\text{scott}} = 1.06 \hat{\sigma} n^{-1/5}, \quad (6)$$

where $\hat{\sigma}$ denotes the standard deviation of the sample.

3.1.3 The Knuth Rule

In the previous sections we considered simple rules for bin width selection, that assume parent distributions of close to Gaussian shape. This has computational advantages and also allows for an easy application to weighted data. The next section compares these simple methods with a more advanced method developed in Knuth (2006), that uses Bayesian inference to fix the number of bins in the histogram. The idea is to maximize the posterior probability for the number of bins M given the data vector \mathbf{d}

$$\hat{M} = \arg \max \{ \log p(M|\mathbf{d}) \}. \quad (7)$$

Using bayes theorem this posterior is constructed as a nested integral over $\underline{\pi}$ that denotes the vector of probabilities that samples are drawn from each of the M histogram bins

$$p(M|\mathbf{d}) \propto \int d\underline{\pi} p(\underline{\pi}|M) p(M) p(\mathbf{d}|\underline{\pi}, M), \quad (8)$$

where the data likelihood $p(\mathbf{d}|\underline{\pi}, M)$ takes the form of a multinomial distribution, however with a different normalization factor. They continue by choosing a noninformative prior for the bin probabilities $p(\underline{\pi}|M)$, known as the Jeffreys's prior for the multinomial likelihood (Jeffreys 1961) and a uniform prior for the number of bins $p(M)$. We use the implementation in the astroML package (Vanderplas et al. 2012) which at the time of this work does not support the application to weighted data. The Knuth rule selects bins of equal width. We also tested the Bayesian Blocks method (Scargle et al. 2013) which adapts the width of the individual bins, again using the implementation in the astroML package. However the results we obtained using Bayesian Blocks were much worse compared with all algorithms considered in this work.

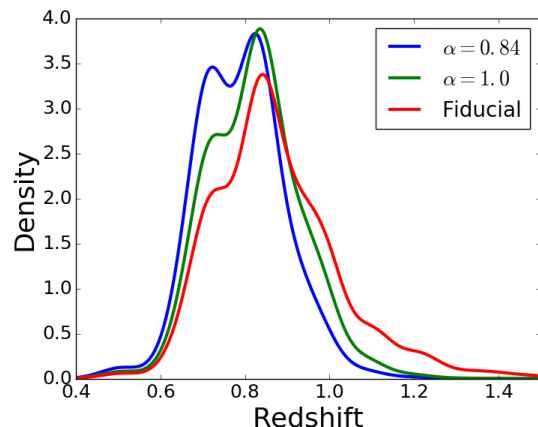


Figure 3. Weighting scheme applied to the highest tomographic redshift bin to mimic the lack of spectroscopic calibration data at high redshift. The weighted distributions are generated by multiplying the fiducial analytical distribution of the highest tomographic redshift bin with the sigmoid weighting function defined in Eq. 9. The distributions are normalized afterwards and the parameter α parametrizes the position of the sigmoid used to penalize the high redshift tail.

3.1.4 Weighting

As already described in the introduction, some weighting schemes are usually applied to the galaxy sample, when computing photometric redshifts. These weights are often the result of empirical photometric redshift codes that interpolate the photometric redshift of a large number of galaxies using a small number of spectroscopic calibration data. The high redshift tail of a photometric redshift distribution is then obtained by giving large weight to a small number of spectroscopic training objects. We mimic this depletion of available spectroscopic objects in high tomographic redshift bins by multiplying a sigmoid weighting function to the highest tomographic redshift bin

$$w(z|\alpha) = [1 + \exp(10(z - \alpha))]^{-1}, \quad (9)$$

where α is a parameter that parametrizes the redshift position of the sigmoid. This weighting scheme is illustrated in Fig. 3 where the fiducial density (red) is penalized at the high redshift tail by the weighting functions $w(z|\alpha = 1.0)$ (green) and $w(z|\alpha = 0.84)$ (blue). The high redshift tail is suppressed after applying the weights which mimics the decrease in the number density of spectroscopically observed galaxies beyond $z > 1.0$. During photometric redshift estimation the penalized distributions shown in green and blue would then be remapped onto the red distribution by the introduction of weights. These weights will give more weight to the few galaxies drawn from the high redshift tail of the blue and green curves and downweight the bulk of the distribution at lower redshift. As there are only few objects at high redshift, the histogram constructed on this weighted data set can be quite noisy. In the analysis in §4.2 we study how shifts in cosmological parameters are affected by this increase in noise after applying weights to a sample. This is studied in a Monte Carlo experiment by first generating samples from the penalized distribution with $w(z|\alpha = 0.84)$ (blue line in Fig. 3). For each of the drawn samples we then use the inverse of the weighting function $1/w(z)$ to attach a

weight that maps the penalized distribution back onto the original red one. Weighting a sample introduces a correlation between the individual samples, which reduces the statistical power of the full sample. In order to compensate for this, we use the effective sample size (Gray 1969)

$$n_{\text{eff}} = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}, \quad (10)$$

that replaces n in Eqs. 4 and 6. The mean and the standard deviation of a weighted sample are then computed using equations

$$\langle z \rangle = \sum_{i=1}^n w_i z_i \quad (11)$$

and

$$\sigma^2 = \sum_{i=1}^n w_i (z_i - \langle z \rangle)^2 \quad (12)$$

respectively.

3.2 The Smoothed Bootstrap

As discussed previously, the total error of density estimates can be split into a bias component and a variance component. The bias quantifies how closely the obtained density estimate fits the underlying distribution. The variance measures the noise in the estimate. Drawing an analogy to a similar cosmological effect, it can be viewed as the shot noise across bins. In practise these contributions need to be estimated using a limited number of samples. This can be done using resampling techniques like the bootstrap. A commonly used way to incorporate the variance in photometric redshift predictions into the analysis is the popular bootstrap method, here ‘normal bootstrap’, as used more recently in Bonnett et al. (2016); Hildebrandt et al. (2016).

We generate new normal bootstrap samples from our available calibration data set by sampling with replacement new data sets of the same size as the original catalogue. In this work we generate 100 of these bootstrap samples and apply the respective density estimate to each of them. Following Scott (1992) the point wise error bands generated by these bootstrapped density estimates reflect the variance of the histogram.

While the normal bootstrap is able to estimate the variance contribution to the total error, it does not quantify the bias generated by oversmoothing the histogram by picking a too large bin width. A resampling method that properly reflects this bias is the smoothed bootstrap (Scott 1992). The basic goal is to estimate both the bias and the variance of a given density estimate. We generate the new smoothed bootstrap data sets by sampling from this density estimate, instead of drawing from the original data set with replacement as done in the normal bootstrap. This smears out the bootstrap samples on the same smoothing scale used to construct the estimate. As for the normal bootstrap, the density estimate is then reapplied to the generated samples and the bias and variance can be measured. This can be seen as a form of Monte Carlo experiment, where our density estimate approximates the true underlying distribution. We reiterate that the only difference between the normal bootstrap and

the smoothed bootstrap lies in the way the bootstrap samples are generated. The normal bootstrap generates them by sampling from the data with replacement, the smoothed bootstrap samples from the density estimate whose bias and variance needs to be estimated.

In the following we consider a kernel density estimate (KDE) of the form given by Eq. 5 from which it is particularly easy to draw samples. The KDE can be seen as a Gaussian mixture model, where the Gaussians are centered on the sample points. In order to generate a single smoothed bootstrap realisation of a particular density estimate, we first draw n_{eff} samples from the estimate and then reapply the density estimate to this newly generated sample. For the aforementioned case of a kernel density estimate constructed on a weighted sample (w_j, z_j) with $j \in \{1, \dots, n\}$ of size n , sampling from the density estimate is done as follows:

- (i) Randomly pick one component j with replacement
- (ii) Draw a sample z^* from $\mathcal{N}(z|z_j, \sigma)$
- (iii) Return z^* and the weight (z^*, w_j) attributed to z_j

In this way we obtain an ensemble of density estimates $\hat{p}^*(z)$. If the original density estimate $\hat{p}(z)$ is a proper approximation of the underlying distribution $p(z)$, then samples generated from them should have comparable statistical properties. The bias between the original density estimate $\hat{p}(z)$ and its smoothed bootstrap realisations $\hat{p}^*(z)$ should thus be similar to the bias between the true unknown density $p(z)$ and the original density estimate $\hat{p}(z)$. This fact will later allow us to correct for the shifts in the cosmological parameters. The variance of the smoothed bootstrap realizations will also be similar to the variance in the original density estimate. Note that the normal bootstrap is the special case of the smoothed bootstrap where $\sigma \rightarrow 0$. In this work we will construct estimates using 100 bootstrap samples. In the following section, we briefly describe the Fisher forecast formalism (e.g. Knox, Scoccimarro & Dodelson 1998; Huterer 2002; Joachimi & Schneider 2009) used to propagate the errors in the photometric redshift distribution into shifts in the cosmological parameters.

3.3 Forecasting the parameter shifts

In this work, we focus on biases introduced into the modelling of the angular clustering of galaxies, where the corresponding angular correlation power spectrum for a combination of tomographic bins (i, j) is defined as

$$C_{i,j}(\ell) = \frac{2}{\pi} \int W_i(\ell, k) W_j(\ell, k) k^2 P(k) dk. \quad (13)$$

Here $P(k)$ is the matter power spectrum, k is the wavevector and the galaxy clustering window functions for galaxy clustering are defined as

$$W_i(\ell, k) = \int b_g(k, z) p_i(z) j_\ell[k\chi(z)] D(z) dz. \quad (14)$$

The modelling of $W_i(\ell, k)$ depends on the galaxy-dark matter bias $b_g(k, z)$, the redshift distribution of the galaxy sample $p_i(z)$, the comoving distance $\chi(z)$ and the linear growth factor $D(z)$.

The offset in the angular correlation power spectrum

caused by the inaccurate estimation of the redshift distribution is defined as

$$\Delta_{\mathbf{C}\mathbf{l}} = \mathbf{C}(\ell)_{\text{bias}} - \mathbf{C}(\ell)_{\text{fid}}, \quad (15)$$

where $\mathbf{C}(\ell)_{\text{bias}}$ denotes the vector of angular correlation power spectra estimated using the non optimal estimator for the tomographic redshift distributions and $\mathbf{C}(\ell)_{\text{fid}}$ the fiducial, unbiased, vector of angular correlation power spectra obtained from the tomographic distributions in Fig. 2.

The shift in cosmological parameters $\Delta_{\mathbf{p}}$ caused by this systematic error in the cosmological observable is given to linear order as

$$\Delta_{\mathbf{p}} = \mathbf{F}^{-1} \mathbf{D} \Sigma^{-1} \Delta_{\mathbf{C}\mathbf{l}}, \quad (16)$$

where \mathbf{F} is the Fisher matrix defined in Eq. 18, Σ is the corresponding covariance matrix and \mathbf{D} contains the derivatives with respect to the cosmological parameters p_{α}

$$D_{\alpha,\beta} = \frac{\partial C_{\beta}(\ell)}{\partial p_{\alpha}}. \quad (17)$$

Here the index β runs over all elements in the data vector $\mathbf{C}(\ell)$, i.e. over all auto and cross correlation power spectra. The Fisher matrix \mathbf{F} is estimated from the data covariance matrix Σ as

$$F_{\alpha,\beta} = \sum_{\ell=\ell_{\min}}^{\ell_{\max}} \sum_{(i,j),(m,n)} \frac{\partial C_{i,j}(\ell)}{\partial p_{\alpha}} \Sigma^{-1} \frac{\partial C(\ell)_{m,n}}{\partial p_{\beta}}. \quad (18)$$

The covariance Σ is modelled as

$$\Sigma_{(i,j)}^{(k,l)}(\ell) = A(\ell) \left(\bar{C}^{(i,k)}(\ell) \bar{C}^{(j,l)}(\ell) + \bar{C}^{(i,l)}(\ell) \bar{C}^{(j,k)}(\ell) \right), \quad (19)$$

where

$$A(\ell) = \frac{\delta_{\ell,\ell'}}{(2\ell+1)f_{\text{sky}}} \quad (20)$$

weights the covariance by the inverse fractional sky coverage f_{sky} and $\bar{C}^{(i,j)}(\ell)$ denotes the angular power spectra estimates, including the shot noise contribution

$$\bar{C}^{(i,j)}(\ell) = C^{(i,j)}(\ell) + \frac{\delta_{i,j}}{\bar{n}_g^i}. \quad (21)$$

Here \bar{n}_g^i is the number of galaxies per steradian in the respective sample. The angular correlation power spectra are estimated using the cosmosis software (Zuntz et al. 2015). In this work we follow Huterer, Cunha & Fang (2013) and Shafer & Huterer (2015) by using a five dimensional fiducial parameter space $(\Omega_m, w_0, w_a, A_s, n_s)$ in our forecast. This neglects the uncertainty in the parameters $\Omega_b = 0.04$ and $h = 0.72$ which we fix to default values used in the cosmosis software package. Huterer, Cunha & Fang (2013) and Shafer & Huterer (2015) justify this simplification by arguing that these remaining parameters are well constrained by other probes like *Planck*. Additionally we marginalize over a multiplicative galaxy-dark matter bias and include modes from $[\ell_{\min}, \ell_{\max}] = [10, 1000]$ assuming a fractional sky coverage of $f_{\text{sky}} = 0.12$ with a number density of 2 arcmin^{-2} for each of the five tomographic redshift bins shown in Fig. 2. We use the redshift dependent bias model by Fry (1996)

$$b_g(z) = 1 + \frac{b_g - 1}{D(z)}, \quad (22)$$

where $D(z)$ is the linear growth function and b_g is the Fry

Table 1. Fiducial cosmological parameter values and corresponding cosmological constraints for our DES like galaxy clustering forecast.

Parameter p	Error σ_p	Fiducial value
Ω_m	0.013	0.3
w_0	0.093	-1.0
w_a	0.42	0.0
A_s	$1.6 \cdot 10^{-10}$	$2.1 \cdot 10^{-9}$
n_s	0.023	0.96
b_g	0.021	1.0

parameter that we set to the fiducial value of $b_g = 1$, such that the fiducial galaxy-dark matter bias model coincides with a constant value $b_g = 1$. We summarize the fiducial cosmological parameter values and constraints in Table 1.

4 COSMOLOGICAL BIASES

A density estimate has two sources of error which contribute to its total mean squared error, as shown in Eq. 1. The bias of the density estimate increases if we introduce more smoothing by choosing a larger bin width. This stabilizes the density estimate but can oversmooth the density thereby destroying its small scale structure. The second contribution to the total error is the variance of the density estimate. This statistical error occurs, since we use a limited number of spectroscopic calibration data to obtain the density estimate, which leads to errors across bins. As a result the same density estimate applied to multiple catalogues independently drawn from the same parent distribution will be different. This effect is larger when the sample size of the catalogue is small. The introduction of weights to the sample can decrease its effective sample size, which further increases the variance of the density estimate.

In the next subsection we will study how errors in the redshift distribution propagate into shifts in the cosmological parameters in a Monte Carlo (MC) experiment. We note that performing a simulation is necessary as the parent distribution from which real spectroscopic data sets are drawn is unknown. A particular catalogue corresponds only to a single realization of the MC experiment. This makes it impossible to estimate the true cosmological parameter shift with respect to this true unknown distribution using real data.

We generate 100 samples from the distributions shown in Fig. 2 and subsequently apply the density estimates discussed in §3.1 to these samples. For each method, we end up with an ensemble of 100 density estimates. If the density estimate would be perfect, each of the obtained distributions would coincide with the (theoretical) parent distributions shown in Fig. 2. However as we have a limited amount of data available to construct the estimate, it is not possible to obtain perfect estimates of the redshift distribution. As discussed in Bonnett et al. (2016), errors on the mean and the width of the redshift distribution are expected to be the dominant source of photometric redshift error for weak gravitational lensing. However it can be expected that higher order statistics like skewness also contribute to the total error budget. This can be especially important for galaxy clustering and cross correlations like galaxy-galaxy lensing. We therefore take a more general approach and include the full shape of the redshift distribution into our analysis without

making any assumptions about the shape of the photometric redshift error. We use the Fisher forecast formalism described in §3.3 to estimate the shift in cosmological parameters with respect to the values obtained using the theoretical parent distributions. In order to study the distribution of parameter shifts composed of the 100 Monte Carlo experiments as a function of the number of galaxies available to construct the estimate, we repeat the MC experiment using a variety of different sample sizes. We investigate sample sizes per tomographic redshift bin ranging from $N_{\text{bin}} = 5000$ to $N_{\text{bin}} = 30000$, which results in a fully representative calibration sample of 25k - 150k galaxies distributed in 5 tomographic redshift bins. These numbers have to be compared with the 50k spectroscopic validation objects that are currently used in DES (Bonnett et al. 2016) and the 25k used in KIDS (Hildebrandt et al. 2016, Tab. 2). However we also discuss the large sample limit of $N_{\text{bin}} = 13 \cdot 10^6$ per tomographic redshift bin, which results in $65 \cdot 10^6$ for the full sample.

In §4.1 we compare the various density estimates using unweighted data, §4.2 then investigates the effect of introducing weights. In the following analyses we will normalize the cosmological parameter shifts Δ_p defined in Eq. 16 by the fiducial constraint σ_p for the respective parameter \mathbf{p} quoted in Table 1:

$$\mathbf{p}_{\text{norm}} = \Delta_p / \sigma_p. \quad (23)$$

In the following we will refer to the normalized quantities \mathbf{p}_{norm} defined in Eq. 23 as the cosmological parameter shifts.

4.1 Oversmoothing Errors

Fig. 4 shows the parameter shift distributions in the set of four parameters (Ω_m, b_g, w_0, A_s) obtained using the different density estimates introduced in §3.1. The results for w_a are not shown, because they are very similar to those obtained for w_0 due to the intrinsic correlation between w_0 and w_a . For each parameter we plot the distribution of parameter shifts on the vertical axis as a function of the number of galaxies N_{bin} per tomographic bin in units of 1000 objects on the horizontal axis. We showcase these distributions by the respective mean parameter shift and the standard deviation of parameter shifts. The dashed regions denote the $\pm 1\sigma$ regions of this distribution, the points denote the mean values. The following discussion refers to the mean and standard deviation of the distribution of parameter shifts as the mean parameter shift and the parameter shift scatter respectively. The horizontal lines illustrate the large sample limit of the respective method, showing the mean parameter shift evaluated on a large sample of $N_{\text{bin}} = 13 \cdot 10^6$ galaxies per tomographic bin. The parameter shift scatter decreases with increasing sample size for all methods. The relative sensitivity to errors in the redshift distribution strongly depends on the cosmological parameter. While the galaxy-dark matter bias parameter b_g is least sensitive to errors in the redshift distribution, the dark energy equation of state parameter w_0 shows large parameter shifts. The performance of the four bin width selection algorithms differs especially in terms of their mean parameter shift values. Algorithms that adapt the bin width with the shape of the distribution and the number of objects are consistent, i. e. the mean parameter shift vanishes in the large sample limit. In contrast the

histogram with $\Delta z = 0.05$ always produces a large mean parameter shift even in the large sample limit, where the estimator is very stable and the parameter shift scatter vanishes. The histogram with $\Delta z = 0.05$ therefore oversmooths the underlying distribution. The Knuth rule, being the most sophisticated bin width selection method considered in this work, tightly adapts the histogram to the underlying density and produces very small mean parameter shifts almost independent of the sample size. The parameter shift scatter is however significant for all considered methods even for moderate sample sizes of $N_{\text{bin}} = 30000$ per tomographic bin. We reiterate that the parameter shift obtained on a single catalogue will be a single sample from the distribution of parameter shifts, where the parameter shift scatter is its standard deviation. The simple bin width selection algorithms like the Scotts rule for the Kernel density estimate and the adaptive bin width selection rule for the histogram produce larger mean parameter shifts compared with the more elaborate Knuth rule. As the implementations of the Knuth rule currently do not support the application to weighted data, its practical applicability is limited for photometric redshift estimation.

So far we have considered the performance of density estimators applied to unweighted data. In practice, photometric galaxy samples are typically weighted for the cosmological analysis. These weights can parameterize the quality of a particular measurement like the error on the measured galaxy shape in cosmic shear. Furthermore empirical methods for photometric redshift estimation weight a spectroscopic training sample to resemble a photometric sample. In the next section we study how the introduction of weights can affect the distribution of parameter shifts.

4.2 Weighting Errors

The introduction of weights to a sample introduces an artificial correlation between previously independent samples, which increases the variance of density estimates constructed on the weighted sample. We study the resulting shifts in the cosmological parameters by slightly modifying the experimental setup described in the previous sections.

Instead of considering an unweighted sample, we generate a weighted sample choosing $\alpha = 0.84$ in Eq. 9, following the methodology described in §3.1.4. The resulting weighted sample then resembles the original distribution of the last tomographic bin. The MC experiment can then be performed as explained in the previous sections, with the only modification that we construct the respective density estimate for the last tomographic bin using a weighted sample.

We want to study the effect of introducing weights to a sample independently of possible modifications to the error of the density estimate that occurs from changing the bin width. Thus we concentrate on the histogram with a fixed bin width of $\Delta z = 0.05$ instead of adapting the bin width with the shape of the distribution and the effective sample size. The result of this experiment is shown in Fig. 5, where we compare the distribution of parameter shifts for the case of weighted data, with the result for unweighted data. To make the visual comparison between the unweighted and weighted case easier, we subtract the mean parameter shift obtained on the unweighted catalogues. In close analogy to the previous section, we show the distribution of parameter

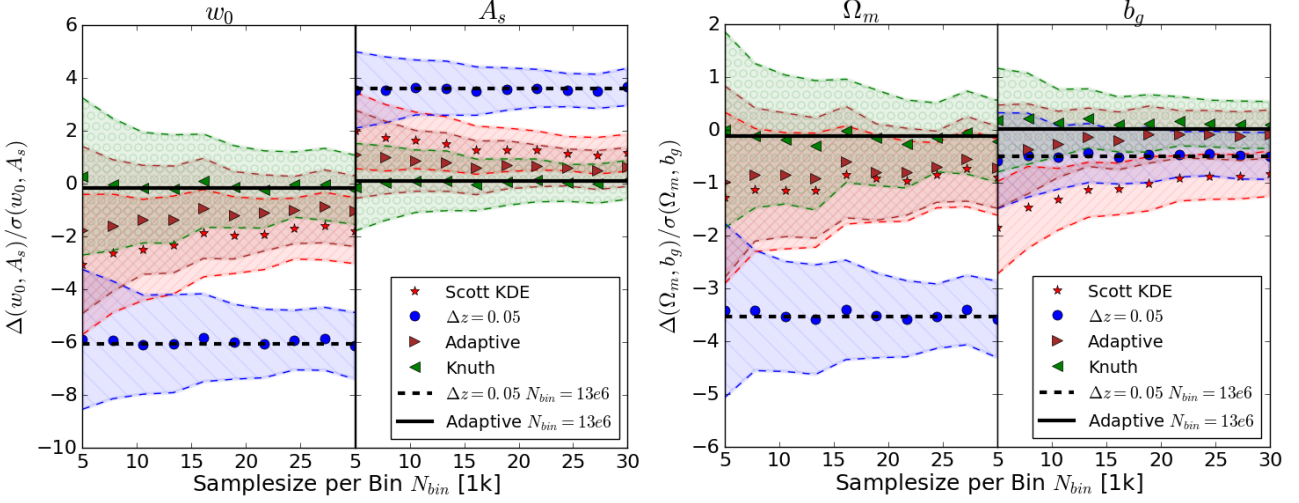


Figure 4. We show the distribution of cosmological parameter shifts caused by the statistical errors in the estimated redshift distributions as a function of the number N_{bin} of galaxy redshifts available to construct the estimate of the tomographic bin; given in units of 1000 galaxy redshifts. All parameter shifts are normalized by the respective fiducial constraints quoted in §3.3. The left plot considers shifts in the dark energy equation of state parameter w_0 and the primordial power spectrum amplitude A_s , the right plot in the matter density Ω_m and the galaxy-dark matter bias parameter b_g . We compare the performance of the Scotts rule applied to a kernel density estimate ‘Scott KDE’, a simple histogram with a bin width of $\Delta z = 0.05$, the adaptive bin width selection rule ‘Adaptive’ and the Knuth rule for histograms ‘Knuth’. The respective distributions of parameter shifts are constructed on 100 simulated catalogues drawn from Fig. 2 as described in the text. The points show the mean and the dashed curves enclose the $\pm 1\sigma$ error regions of the respective distribution. The mean of the distribution of relative parameter shifts evaluated on a large sample ($N_{bin} = 13 \cdot 10^6$) is illustrated by the horizontal lines. We show this large sample limit for the simple histogram with $\Delta z = 0.05$ (dashed black line) and the adaptive bin width selection rule (solid black line).

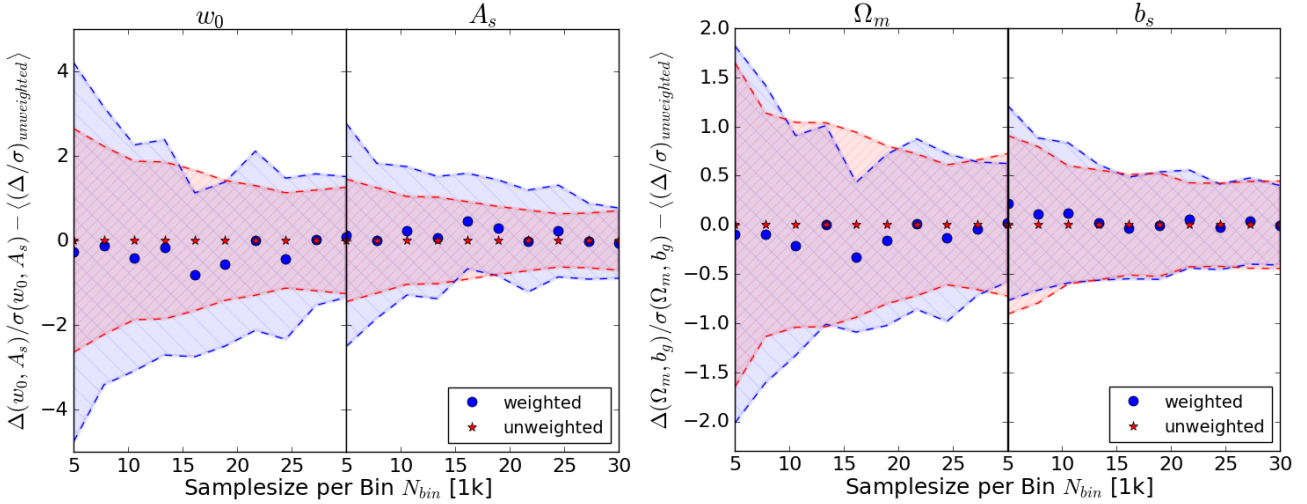


Figure 5. In analogy with Fig. 4 we compare the distribution of cosmological parameter shifts caused by the statistical errors in the estimated redshift distributions used to reconstruct each tomographic redshift bin for weighted and unweighted data, as a function of the sample size N_{bin} . The parameter shifts are normalized by the respective fiducial constraints quoted in §3.3, the sample size N_{bin} is given in units of 1000 galaxies. The dots show the mean, the dashed regions the standard deviation of the distribution of parameter shifts. For both the weighted and the unweighted case, we subtract the mean of the distribution of parameter shifts of the unweighted case. Therefore note that the red stars are centered at zero. The left plot considers the dark energy equation of state parameter w_0 and the primordial power spectrum amplitude A_s , the right plot the matter density Ω_m and the galaxy-dark matter bias parameter b_g . We use the histogram with $\Delta z = 0.05$ as the density estimate.

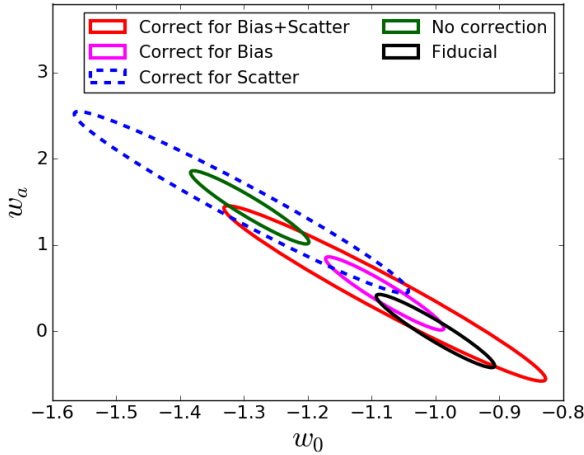


Figure 6. Typical example of the effect of the cosmological parameter shift correction. The black contour shows the fiducial dark energy parameter constraint. The dark green contour is shifted due to the typical photometric redshift distribution error expected from a KDE with a bandwidth selected by the Scotts rule and applied to a sample of galaxies with $N_{bin} = 5000$. The magenta contour shows the corrected bias using the smoothed bootstrap technique. The red ellipse uses the smoothed bootstrap to marginalize the corrected magenta contour over the remaining statistical uncertainty in the redshift distribution. The dashed blue contour shows the result of marginalizing over the statistical uncertainty using the normal bootstrap technique without the bias correction from the smoothed bootstrap. All contours are 1σ constraints.

shifts for the dark energy equation of state parameter w_0 , the primordial power spectrum amplitude A_s , the matter density Ω_m and the galaxy-dark matter bias parameter b_g as a function of the sample size per tomographic redshift bin N_{bin} . The parameter shift scatter for the weighted case is in general larger than for the unweighted case. The magnitude of this increase in parameter shift scatter is especially large for the dark energy of state parameter w_0 and small for the galaxy-dark matter bias parameter b_g . We further note that the mean parameter shifts are only weakly affected compared with the increase in parameter shift scatter. This is to be expected as the introduction of weights to a sample primarily decreases the effective sample size and in turn increases the variance of the density estimate.

We have seen in the last sections that cosmological parameter constraints can be significantly shifted by errors in the tomographic redshift distributions. The two main sources of error are the effect of oversmoothing and the introduction of weights. Efficient algorithms can closely adapt the bin width to the shape of the distribution and the available sample size to reduce the effect of oversmoothing. However there still remains a statistical error, especially in the presence of weighted samples. In the next section we investigate resampling techniques, that can be used to incorporate both sources of error into the parameter likelihood.

5 CORRECTING COSMOLOGICAL PARAMETER SHIFTS

In the last section we assumed perfect knowledge of the underlying parent redshift distribution to investigate how the systematic and statistical errors in the redshift histograms lead to shifts in the cosmological parameters. We have seen that the selection of a too large bin width oversmooths small scale structure in the density estimate. This systematic bias in the redshift distribution propagates into a global shift in the cosmological parameters; the mean parameter shift. This systematic shift is persistent in the large sample limit where the statistical noise in the density estimate vanishes. In addition to this systematic error in the density estimate, we also need to correct for the statistical uncertainty given by the noise in the density estimate. We can incorporate this error into the final parameter constraint by adding its covariance, i.e. the parameter shift scatter, to the fiducial covariance. In practice, the true parent distribution of the tomographic redshift bins is unknown and both sources of error need to be estimated on a single sample.

This can be done in two steps using the smoothed bootstrap technique as illustrated in Fig. 6. In dark green we show a parameter ellipse shifted by the typical error in the redshift distribution obtained from a KDE with a bandwidth selected by the Scotts rule and $N_{bin} = 5000$. This total sample size of 25k representative spectroscopic calibration objects amounts to approximately the number of spectra used by KIDS (Hildebrandt et al. 2016). Using an estimate of the systematic error, i. e. the mean parameter shift, we can correct this biased constraint by shifting it to the magenta contour. Marginalizing over the remaining statistical uncertainty, i. e. the parameter shift scatter⁴, we can produce the red contour which then almost completely overlaps with the unbiased fiducial contour (black). This has to be compared with the result from the normal bootstrap (dashed blue) that is, like the smoothed bootstrap, able to marginalize over the statistical uncertainty. However in contrast to the smoothed bootstrap, the normal bootstrap is not able to correct for the mean parameter shift. As a result the parameter contour produced by the normal bootstrap is still significantly biased in contrast to the result from the smoothed bootstrap.

In the following we compare the smoothed bootstrap technique with the normal bootstrap in a Monte Carlo (MC) experiment. We reiterate, that a simulation is necessary as the true underlying redshift distribution of real samples is unknown. Thus, in order to investigate the statistical performance of the bootstrap techniques, we need to define this true underlying distribution. This experiment is carried out by drawing 50 samples from the theoretical distributions in Fig. 2 and applying the kernel density estimate with a bandwidth selected by the Scotts rule. We choose the kernel density estimate because it is well suited for the generation of new samples which is an important step in the smoothed bootstrap method. To illustrate how we can correct parameter shifts even on a small data set, we choose a sample size of $N_{bin} = 5000$ objects per tomographic bin. For each

⁴ For simplicity we will refer to estimates of the covariance of the distribution of parameter shifts in two dimensions, e.g. w_0 and w_a , as the parameter shift scatter, too.

MC experiment, parameter shifts need to be determined for each of the 100 bootstrap samples. As this is computationally expensive, we choose to perform 50 MC experiments which gives us sufficient statistical accuracy while still being computationally manageable. In the following discussion we will refer to the 50 catalogues as the original catalogues and to the corresponding 50 density estimates as the original density estimates. For each of the original catalogues we estimate the mean parameter shift and the parameter shift scatter using the smoothed bootstrap and the normal bootstrap. We apply the Fisher forecast method to propagate the uncertainties in the redshift distribution into shifts in cosmological parameters in analogy to the previous sections. This gives us 50 estimates for the mean parameter shift and the parameter shift scatter from both resampling methods. The distribution of these estimates is then compared with the true mean parameter shift and scatter obtained on the original density estimates in Fig. 7. The left panel of this figure shows the quality of the estimation of the mean parameter shift using the smoothed bootstrap and the normal bootstrap for the example of the dark energy equation of state parameter. As the mean parameter shift is a constant offset, each estimate should give the same value, independent on the particular sample. In practise this is of course not the case, which we quantify by calculating the mean and the standard deviation of this distribution. For illustration we normalize the respective mean parameter shift estimates by the true mean parameter shift. On this x-axis scale, 0 corresponds to no mean parameter shift correction and 1 corresponds to a perfect correction.

We find that the normal bootstrap is not able to estimate the mean parameter shift, while the smoothed bootstrap is able to recover the majority of the mean parameter shift. The right panel of Fig. 7 shows the average quality of the parameter shift scatter estimation for the example of the covariance between w_0 and w_a . The smoothed bootstrap and the normal bootstrap produce the same estimation quality and are both able to accurately estimate the true parameter shift scatter. The dashed ellipses show the individual parameter shift scatter estimates from the 50 experiments. As can be seen, the scatter around the mean value is in general quite small for both methods.

We have shown that the smoothed bootstrap is able to correct shifts in cosmological parameter constraints produced by errors in the redshift distribution. This includes errors suffered from oversmoothing small scale features in the redshift distribution, as well as statistical errors produced by noisy density estimates.

The mean parameter shift correction using the smoothed bootstrap technique implicitly assumes that the bias between the true unknown distribution $p(z)$ and the original density estimate $\hat{p}(z)$ is approximately equal to the bias between the original density estimate $\hat{p}(z)$ and the density estimates constructed on its smoothed bootstrap samples $\hat{p}^*(z)$. In the context of this section, this implicitly assumes that the parameter shift bias is linear with respect to the smoothing parameter. This will likely not be the case in practise. Smoother density estimates are less sensitive to the smoothing parameter, than noisy density estimates, especially in the variance component (Eq. 2). The density estimates constructed on the smoothed bootstrap samples are smoother than the original estimate. As can be seen in Fig.

4, the original density estimate constructed using the Scott rule produces a small mean parameter shift. It is therefore a bit too smooth compared to the true unknown density estimate. As discussed, smoother density estimates will be more stable with respect to changes in the bin width compared with more noisy density estimates. Thus the mean parameter shift between the original density estimate constructed using the Scott rule and the unknown true density estimate will be higher than between the original density estimate and the smoothed bootstrap samples. We identify this non linear dependency of the mean parameter shift on the bin width as the reason why the smoothed bootstrap underestimates the mean parameter shift.

In contrast the parameter shift scatter is relatively insensitive to changes in the bin width. This can be seen in the Fig. 4 where the parameter shift scatter values for different bin width selection algorithms have been shown to be quite similar. Thus making the density estimate smoother in the smoothed bootstrap method has a negligible effect on the quality of the estimated parameter shift scatter as shown in Fig. 7. The likely reason for this is that the sample size often dominates the error of a statistical estimator over changes in the shape of the distribution. For instance the error on the sample mean scales with $\propto \hat{\sigma}/\sqrt{n}$ where n is the sample size and $\hat{\sigma}$ is the sample estimator for the standard deviation. A small increase in $\hat{\sigma}$ produced by e. g. the smoothed bootstrap is therefore strongly suppressed ($1/\sqrt{N_{\text{bin}}} = 0.01$) even for a relatively small sample size of $N_{\text{bin}} = 5000$ galaxies per tomographic redshift bin as considered here.

6 SUMMARY AND CONCLUSIONS

Current and next generation large area photometric surveys like DES or Euclid are expected to measure cosmological parameters with unprecedented accuracy. To enter this era of precision cosmology, our understanding of systematic errors needs to increase faster than the statistical power of these measurements. Errors in the distribution of distance, or redshift are already challenging for current multiband photometric surveys like CFHTLenS, DES or KIDS and are likely to become an even greater burden for next generation surveys like Euclid.

To prepare for these upcoming challenges, this work studies how photometric redshift distributions can be estimated without causing systematic errors in the cosmological parameters. We start the discussion in §4.1 by considering the statistical properties of a simple histogram estimate of the redshift distribution. We have seen that the selection of a bin width that is too large can bias redshift distributions in each tomographic bin. This ‘oversmoothing effect’ destroys information about small scale features in the density like multimodal or sharp peaks. As a result, the estimated density then no longer coincides with the true underlying distribution. This effect can even be significant for bin widths used in current analysis like the constant redshift binning of $\Delta z = 0.05$ used in Benjamin et al. (2013). We note that their continued usage can become a significant error source when parameter constraints become tighter in future surveys.

To reduce these errors in the redshift distribution, we studied methods to adaptively select the bin width as a function of the number of objects and the shape of the distri-

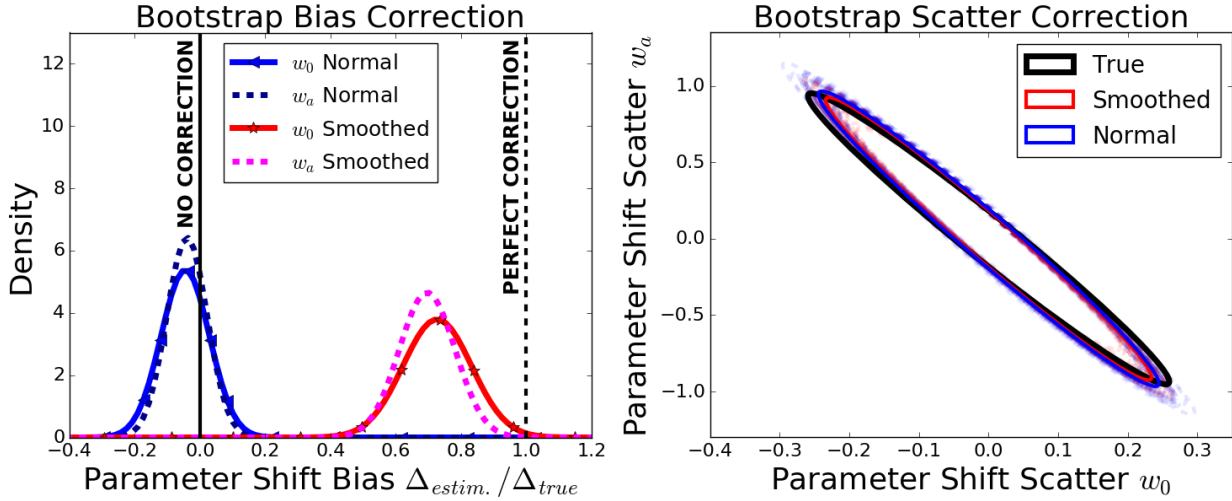


Figure 7. Left panel: Quality of the cosmological mean parameter shift correction using the smoothed bootstrap compared with the normal bootstrap. The x-axis scale shows the estimated over the true mean parameter shift. The vertical line at the origin corresponds to no mean parameter shift estimation and therefore no possible correction. The vertical line at 1.0 corresponds to a perfect mean parameter shift estimation and therefore a perfect correction. Right panel: Quality of the variance estimation using bootstrap methods. The black ellipse shows the true 1σ parameter shift scatter. The red and the meshed red contours show the respective parameter shift scatter estimates from the smoothed bootstrap. The blue and the meshed blue contours show the corresponding result for the normal bootstrap. As explained in the text, the results in both panels are obtained using 50 Monte Carlo experiments, where each time 100 bootstrap samples were drawn.

bution. We specifically investigate small sample sizes of 25k spectroscopic validation galaxies which are currently available in state-of-the-art photometric surveys, where KIDS and DES use approximately 25k and 50k spectroscopic validation galaxies respectively (Hildebrandt et al. 2016, Tab. 2 and Bonnett et al. 2016). In §4.1 we demonstrated for the case of a DES like galaxy clustering forecast, that these methods can reduce the relative shift in cosmological parameters by a factor of up to 6 for the dark energy equation of state parameters w_0 and w_a , as compared with the aforementioned constant redshift binning. We obtained the most accurate redshift distributions using the Knuth rule. Using this method we were able to produce cosmological parameter constraints with an especially low systematic error, even for small sample sizes of 5000 galaxies per tomographic redshift bin. However current implementations of the Knuth rule do not support the application to weighted data. This severely limits its practical applicability, as some form of weighting scheme is usually applied to redshift samples. The generalization of the algorithm to weighted data should be a straightforward modification of the multinomial data likelihood and the prior on the bin probabilities (see §3.1.3). We leave this for future work. Irrespective of the chosen method, there still remains a statistical uncertainty in the redshift distribution that cannot be removed even if the bin width is carefully selected as discussed in §4.

In §4.2 we demonstrated that the introduction of sample weights drastically deteriorates the quality of the redshift histograms. The size of this effect naturally depends on the weighting scheme and the shape of the considered distributions. For our choices the statistical error in the measured dark energy equation of state parameters w_0 and w_a increased by a factor of up to two while other parameters like

the matter density Ω_m or the galaxy-dark matter bias b_g were shown to be much more robust.

The magnitude of the aforementioned types of error scale with the size of the spectroscopic catalogue. In practise the amount of calibration data available for photometric redshift estimation is limited. Especially broad band photometric surveys require accurate calibration as their photometry often does not allow a unique and accurate evaluation of distance, independent of the chosen photometric redshift estimation algorithm. As spectroscopic surveys use different strategies to select their targets than their photometric counterparts, their selection functions in color-magnitude space are typically different. In particular the estimation of spectra for fainter objects requires long exposure times. Therefore faint regions of color-magnitude space are typically incompletely covered by spectroscopy.

In order to validate photometric redshifts using spectroscopic surveys, weights need to be introduced such that these incompatible selection functions are corrected (e.g. Cunha et al. 2014; Bonnett et al. 2016). In regions of color-magnitude space where no spectroscopic calibration is available, we even need to exclude subsets of the photometric sample to guarantee unbiased results. While the weighted spectroscopic calibration data will mimic the photometric science sample, the resulting density estimates will be more noisy, as the sparsely populated high redshift tail will be strongly upweighted (see §3.1.4). This uncertainty in the redshift distribution has to be incorporated into the final parameter likelihood, as it cannot be avoided even by the most accurate bin width selection methods. In §5 we compared two resampling methods to accomplish this. The first is the commonly used bootstrap method, that failed to correct for the effect of oversmoothing. Instead we showcase a modified version of the bootstrap. The ‘smoothed boot-

strap’ smears out the individual bootstrap samples in the same smoothing scale as used in the original density estimate. We demonstrate that this method is able to correct for the effect of oversmoothing to good accuracy. At the same time the smoothed bootstrap shows the same quality in estimating the statistical noise in the redshift distribution as the normal bootstrap. This means that we can accurately marginalize over this remaining statistical uncertainty after the systematic bias from oversmoothing is accounted for. In this way we can correct the final parameter likelihood from both sources of error.

While this work mainly addresses redshift distributions in photometric surveys, the results presented here are readily applicable to all problem settings where a distribution needs to be estimated. This includes for example the spectroscopic validation of photometric redshift algorithms that otherwise do not require a representative spectroscopic calibration dataset like template fitting or redshift estimation using cross correlations. In order to accurately validate photometric redshift distributions, we require a sample of representative spectroscopic galaxies. The photometric redshift distribution that has been estimated by any photometric redshift method can then be compared against an estimate of this reference distribution. This distribution of spectroscopic calibration redshifts is however subject to the sources of error discussed in this work. This limits our ability to calibrate photometric redshift distributions which indirectly contributes to the total photometric redshift error of the respective method. We want to highlight that this applies to photometric redshift methods that reconstruct redshift distributions for individual galaxies as well as special photometric redshift point estimates (see e.g. Rau et al. 2015) that estimate, in analogy to this paper, redshift distributions of samples of galaxies. It is however especially important for empirical photometric redshift methods based on machine learning. These methods estimate individual object redshift distributions and point predictions by reweighting accurate calibration data in color space. Estimates of the individual object redshift distributions are constructed as weighted density estimates of spectroscopic calibration data. Photometric redshift point predictions can be seen as the mean or median estimated on the weighted calibration dataset. We refer the interested reader to (Rau et al. 2015) for a more detailed explanation. As photometric redshift distributions estimated using machine learning are in essence density estimates constructed on a weighted spectroscopic calibration dataset, the methods discussed in this work readily apply to them. The estimation of these weights requires a density estimate in color space that can be the source of additional errors that haven’t been explicitly discussed in this work. However the same resampling techniques should also apply here, which we highlight as a direction for future research. While we focussed on the application to the modelling of angular correlation power spectra, we note that the methods developed in this work will also be potentially relevant for other two point statistics like cosmic shear spectra.

In summary, uncertainties in the photometric redshift distribution are a limiting source of systematic error for ongoing and future photometric surveys. Their quality can only be guaranteed by validating against highly accurate spectroscopic redshift measurements. Weighting methods are able to correct for the mismatch between the spectroscopic and

photometric selection functions and the efficient bin width selection algorithms investigated in this work are able to avoid being systematically biased by oversmoothing the resulting density estimates. We finally demonstrated, that the smoothed bootstrap can correct the remaining cosmological parameter biases without assuming a particular model for the redshift uncertainty. In this way future photometric surveys will be able to obtain unbiased cosmological parameter estimates using a minimum amount of spectroscopic calibration data.

7 ACKNOWLEDGEMENTS

MMRau thanks the anonymous referee, Gary Bernstein, Michael Troxel and Sarah Bridle for invaluable comments on the manuscript. This work was supported by SFB-Transregio 33 ‘The Dark Universe’ by the Deutsche Forschungsgemeinschaft (DFG) and the DFG cluster of excellence ‘Origin and Structure of the Universe’.

REFERENCES

- Akima H., 1970, *J. ACM*, 17, 589
- Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, *MNRAS*, 460, 1371
- Becker M. R. et al., 2015, *ArXiv e-prints*
- Bender R. et al., 2001, in *Deep Fields*, Cristiani S., Renzini A., Williams R. E., eds., p. 96
- Benítez N., 2000, *The Astrophysical Journal*, 536, 571
- Benjamin J. et al., 2013, *MNRAS*, 431, 1547
- Bonnett C., 2015, *MNRAS*, 449, 1043
- Bonnett C. et al., 2016, *PRD*, 94, 042005
- Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
- Choi A. et al., 2016, *MNRAS*, 463, 3737
- Cunha C. E., Huterer D., Lin H., Busha M. T., Wechsler R. H., 2014, *MNRAS*, 444, 129
- Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, *MNRAS*, 396, 2379
- de Jong J. T. A. et al., 2013, *The Messenger*, 154, 44
- Flaugher B., 2005, *International Journal of Modern Physics A*, 20, 3121
- Fry J. N., 1996, *ApJL*, 461, L65
- Gray P. G., 1969, *Journal of the Royal Statistical Society. Series A (General)*, 132, 272
- Hearin A. P., Zentner A. R., Ma Z., Huterer D., 2010, *ApJ*, 720, 1351
- Hildebrandt H. et al., 2016, *ArXiv e-prints*
- Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015, *MNRAS*, 450, 305
- Huterer D., 2002, *PRD*, 65, 063001
- Huterer D., Cunha C. E., Fang W., 2013, *MNRAS*, 432, 2945
- Jeffreys H., 1961, *Theory of Probability*, 3rd edn. Oxford, Oxford, England
- Joachimi B., Schneider P., 2009, *Astron. & Astrophys.*, 507, 105
- Jouvel S. et al., 2009, *Astron. & Astrophys.*, 504, 359
- Kitching T. D., Verde L., Heavens A. F., Jimenez R., 2016, *MNRAS*, 459, 971
- Knox L., Scoccimarro R., Dodelson S., 1998, *Physical Review Letters*, 81, 2004
- Knuth K. H., 2006, *ArXiv Physics e-prints*
- Koo D. C., 1985, *AJ*, 90, 418
- Laureijs R. et al., 2011, *ArXiv e-prints*
- Leistedt B., Mortlock D. J., Peiris H. V., 2016, *MNRAS*, 460, 4258
- Liau A., Wiener M., 2002, *R News*, 2, 18
- Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, *ArXiv e-prints*
- Newman J. A., 2008, *ApJ*, 684, 88
- Rahman M., Mendez A. J., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., Budavári T., 2016, *MNRAS*, 460, 163
- Rau M. M., Seitz S., Brimiouille F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, *MNRAS*, 452, 3710
- Sánchez C. et al., 2014, *MNRAS*, 445, 1482
- Scargle J. D., Norris J. P., Jackson B., Chiang J., 2013, *ApJ*, 764, 167
- Scott D. W., 1992, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 1st edn. Wiley
- Shafer D. L., Huterer D., 2015, *MNRAS*, 447, 2961
- Speagle J. S., Capak P. L., Eisenstein D. J., Masters D. C., Steinhardt C. L., 2016, *MNRAS*, 461, 3432
- Speagle J. S., Eisenstein D. J., 2015a, *ArXiv e-prints*
- Speagle J. S., Eisenstein D. J., 2015b, *ArXiv e-prints*
- Vanderplas J., Connolly A., Ivezić Ž., Gray A., 2012, in *Conference on Intelligent Data Understanding (CIDU)*, pp. 47–54
- Zuntz J. et al., 2015, *Astronomy and Computing*, 12, 45

Accurate photometric redshift validation

The previous chapters provided the reader with all necessary tools to obtain redshift information for photometrically observed galaxies using Machine Learning. We started in §3 by proposing methods to more accurately estimate photometric redshift distributions for individual galaxies and samples of galaxies. These techniques allow us to obtain photo- z distributions without making any assumptions about their shape or functional form. However the accuracy of photometric redshift estimation and validation is still limited by the number of spectroscopic reference galaxies available to construct the estimate. As discussed in §4, a low number of spectroscopic reference galaxies introduces significant errors in redshift distributions. We demonstrated how these errors can be reduced and incorporated into cosmological parameter constraints by the application of more accurate density estimates and resampling techniques. Using the methods proposed in §4, we were able to obtain unbiased cosmological parameter constraints using a very moderate number of accurate calibration galaxies.

This discussion however ignored additional sources of error, such as sample selection biases, or field-to-field variations in photometric noise, that strongly depend on the considered dataset. If we expect these errors to play a significant role for the accuracy of the photo- z estimate, we need to perform careful photometric redshift validation, which is especially important for large area photometric surveys like the Dark Energy Survey (DES) that provides deep photometry. This chapter therefore complements the previous discussion by considering how sample selection biases and field-to-field variations in photometric noise can lead to biased estimates of redshift performance. These effects will be discussed in the context of the photometric redshift validation strategy of the Dark Energy survey.

Contributions and Acknowledgements The text of the chapter is based on my first author paper ‘Sources of Error in photometric redshift calibration’, that I currently prepare for the final review circle within the DES collaboration. I wrote the text to both the original paper and the following chapter. I designed and performed all analyses presented in §5.3, §5.4 and §5.5 and prepared the plots. In §5.3, I performed the described matching procedures, obtained BPZ photometric redshift distributions and wrote the analysis pipeline to process the photometric redshift distributions. In §5.4, I obtained the weights that match the photometric input space to the spectrophotometric space for different numbers of nearest neighbors. All subsequent data processing steps have also been performed by me. In §5.5, I developed the

spline compression algorithm. I also provide a public implementation of this algorithm¹. Furthermore I ran the sparse representation code and implemented the Monte Carlo sampling procedures. All subsequent steps of data processing have also been performed by me.

The data products and procedures described in §5.2 are the joint work from several members of the DES photometric redshift group, most notably Ben Hoyle and Daniel Gruen. I initially started investigating the problem of field-to-field variations in photometric noise using photometric data, that was observed several times, with varying number of exposures. Later Ben Hoyle developed the initial version of the resampling algorithm; Daniel Gruen developed the final version. The photometric reduction of the data I use is described in detail in Drlica-Wagner et al. (2017). Ben Hoyle contributed a script that runs the *stilts* program (Taylor, 2006) and a script that calculates the k-nearest neighbor density estimate using the *sklearn* package (Pedregosa et al., 2011). I use the *cosmosis* software (Zuntz et al., 2015) to obtain the convergence power spectra models and the BPZ code (Benítez, 2011) using the standard configuration of the DES photometric redshift group (Will Hartley, Boris Leisted, Carles Sánchez). I thank my internal referees Dragan Huterer and Carles Sánchez as well as Ben Hoyle, Michael Troxel and Gary Bernstein for valuable feedback, that improved the content and presentation of the text.

5.1 Introduction

Data based photometric redshift validation estimates the quality of potentially inaccurate redshift distributions of photometric galaxy samples by comparing them with more accurate redshift distributions extracted from overlapping spectroscopic or multiband, narrow filter photometric surveys (Hildebrandt et al., 2010; Sánchez et al., 2014; Bonnett et al., 2016; Hildebrandt et al., 2017; Gruen & Brimiouille, 2016). Typically only a small fraction of the full footprint of the photometric survey overlaps spatially with these validation fields and many spectroscopic datasets only cover a small part of the full color-magnitude space of the photometric science sample. As a result, photometric redshift validation is only possible for those subsamples of the full photometric data, for which validation samples of sufficient completeness are available. To validate the redshift distributions of the full sample, we therefore need to estimate the quality of photometric redshifts using a subsample where spectroscopic, or alternatively high-precision multiband, narrow filter photo-*z*, are available. The error estimates obtained on these subsamples, can then be used in the science analysis of the full sample. The DES photometric redshift validation strategy uses the COSMOS field that spatially overlaps with DES in an $\approx 1 \text{ deg}^2$ large footprint as its main source of photometric redshift validation. Since this validation field is quite small, we have to additionally correct for the effect of cosmic variance, if we want to extrapolate photometric redshift performance estimates to the much larger footprint of the science sample. This can be achieved by performing redshift validation on small random patches extracted from simulations and interpreting the variance between the obtained redshift errors as a cosmic variance induced error contribution. The discussion presented here instead focuses on possible biases within a validation field as listed in Fig. 5.1. A detailed discussion of the cosmic variance effect can be found in Hoyle et al. (2017).

The DES photometry in the COSMOS field is significantly deeper than the photometry of the full DES science sample. The higher quality of the photometry can be expected to naturally translate into more accurate photometric redshift estimates as demonstrated in

¹<https://github.com/MarkusMichaelRau/PDFZip>

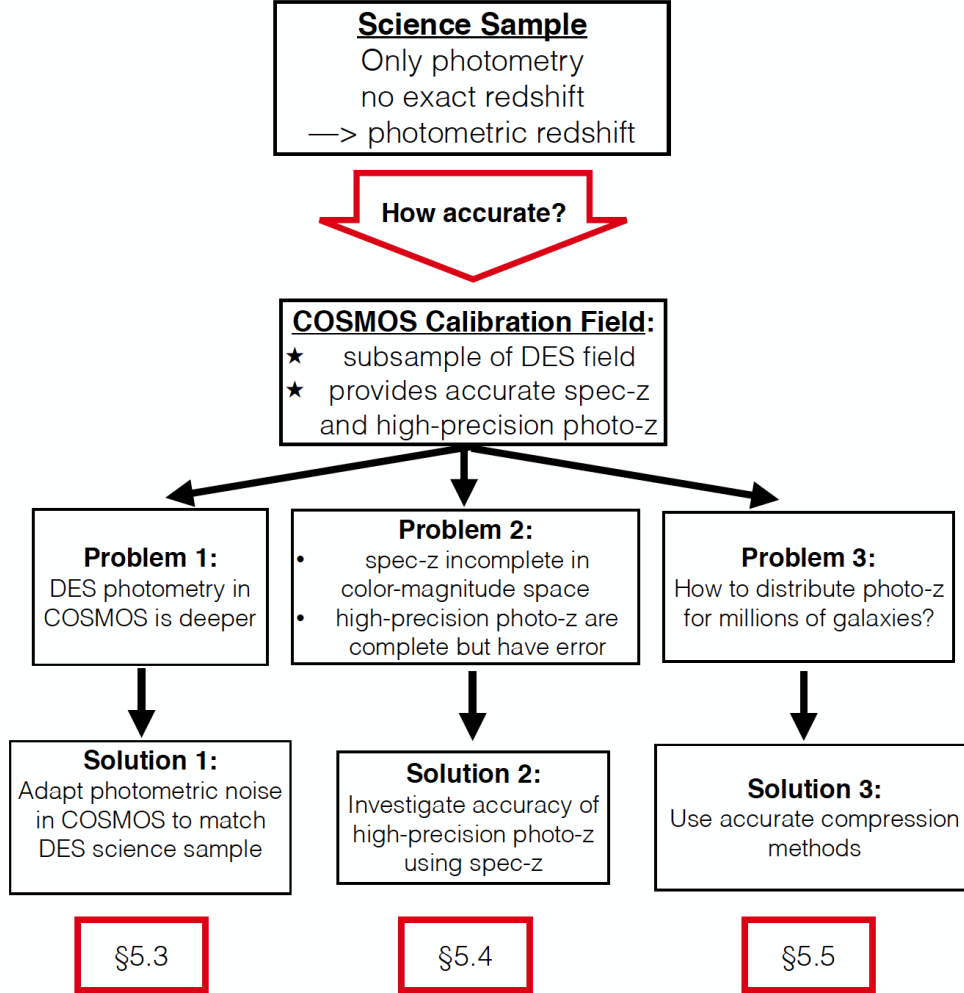


Figure 5.1: Schematic illustration of the different sources of error in photometric redshift (‘photo- z ’) validation investigated in this work. To estimate the error of the photometric redshifts obtained on the DES science sample, we compare them with high-precision photometric redshift estimates from the COSMOS field. To ensure that the resulting error estimate is unbiased, we need to match the photometric noise level of the DES photometry in the COSMOS field to the photometric depth of the DES science sample. Furthermore we investigate the intrinsic error in the high-precision photometric redshifts used for validation. In order to efficiently distribute and process the generated photo- z , we propose more accurate and efficient compression algorithms, compared with the current state-of-the art. The last row refers to the sections in this paper that address the respective problems.

Sánchez et al. (2014) on DES Science Verification data. A photometric redshift algorithm that is evaluated on this field would therefore perform significantly better compared with the expected performance in the remaining footprint, which would result in an underestimation of redshift error. The DES photometric redshift validation strategy therefore artificially degrades the photometry in the COSMOS field to match the noise level in the science fields. Since this method is based on data resampling to degrade the DES photometry in the COSMOS field, it increases the statistical error budget on the photometric redshift error estimates as discussed in §5.3.

As mentioned previously, unbiased photometric redshift validation requires a validation sample, that is representative of the color-magnitude space of the full photometric science sample and covers its full redshift range. This is typically not the case for spectroscopic samples, that are usually quite incomplete at the faint end of the color-magnitude space. In the DES photometric redshift validation strategy we therefore use high-precision photometric redshifts in the COSMOS field as a substitute. The decisive advantage of this approach is, that we find COSMOS multiband, narrow filter photometry for each DES galaxy, which guarantees a complete photometric redshift validation sample (see Hoyle et al., 2017). While the COSMOS high-precision photometric redshifts are of exceptional quality, they are not perfect and their intrinsic photometric redshift error needs to be incorporated into the full validation procedure. Incidentally this is done using spectroscopic redshifts (‘spec- z ’) which are, as noted earlier, often incomplete in color-magnitude space. The goal of §5.4 is therefore to identify those regions of color-magnitude space, where the COSMOS high-precision photometric redshifts can be calibrated well with spectra. This enables us to selectively set more stringent priors on the accuracy of the COSMOS high-precision photometric redshifts in regions where our spectroscopic validation is complete. On the other hand, we can remove those regions from the analysis, where we expect large photometric redshift errors.

In the era of large area photometric surveys that observe hundreds of millions of galaxies, it is important to efficiently process and store photometric redshift distributions (see e.g. Carrasco Kind & Brunner, 2014b). In order to reduce the storage requirements for DES, we compare several methods to efficiently store photometric redshift distributions using a minimum amount of floating point numbers per galaxy. While compression power is an important aspect of these algorithms, they can potentially bias redshift estimates by losing information about the accurate form of the redshift distribution during compression. One of the most popular techniques to compress photometric redshift distributions of individual galaxies (Carrasco Kind & Brunner, 2014b; Rau et al., 2015) is the Monte Carlo sampling approach. This method ‘stores’ photometric redshift distributions by drawing a sample from each individual galaxy redshift distribution. Having to store only a single floating point number per galaxy is extremely efficient in terms of storage requirements and computation time. However the reconstruction of the photometric redshift distribution of a galaxy sample using these Monte Carlo draws, requires the application of a density estimate, that can be subject to the types of error investigated in §4. In §5.5 we study these sources of error and discuss their impact on the modeling of the weak lensing convergence power spectrum introduced in Eq. 1.122.

Metrics

Efficient photometric redshift validation needs to judge photometric redshift accuracy with respect to a specific science goal, which requires the definition of a suitable performance metric. Traditionally these metrics are aimed towards the quality of point predictions like the bias

in the mean of the photometric redshift distribution of individual galaxies (e.g. Hildebrandt et al., 2008, 2010; Abdalla et al., 2011; Dahlen et al., 2013; Sánchez et al., 2014). However as discussed in §4, even if point predictions are perfect, unbiased samples from the underlying redshift distribution, the process of constructing histograms is imperfect and can lead to errors in the modeling of cosmological probes like angular correlation power spectra. These biases are caused by the intrinsic error in the density estimate due to the limited sample size and the inaccurate selection of the smoothing, i.e. the histogram bin width. As a result, it is in general advisable to quantify photometric redshift error by the redshift induced bias in the modeling of cosmological observables.

As weak gravitational lensing is one of the most important cosmological probes in DES, we choose to directly quantify the photometric redshift error as the relative bias in the convergence power spectrum (Eq. 1.122). We define this error metric in our analysis as the relative difference between the convergence power spectrum calculated using the error free, unbiased, distribution $C_\ell^{\kappa, \text{unbiased}}$ and the biased redshift distribution C_ℓ^{bias} :

$$\Delta(C_\ell^\kappa) = \frac{C_\ell^{\kappa, \text{bias}} - C_\ell^{\kappa, \text{unbiased}}}{C_\ell^{\kappa, \text{unbiased}}} . \quad (5.1)$$

The following discussion will refer to this metric as the ‘relative bias in the convergence power spectrum’. We also define its average over the modes $20 \leq \ell \leq 10,000$

$$\langle \Delta(C_\ell^\kappa) \rangle = \sum_{\ell=20}^{10,000} \Delta(C_\ell^\kappa) \quad (5.2)$$

as a summary statistic. The convergence power spectra used in this analysis were obtained using the cosmosis software (Zuntz et al., 2015).

Probability Density Estimation

In this work we will often estimate redshift distributions of samples of galaxies from discrete point values like spectroscopic redshifts. We will use the histogram as a density estimate and select the histogram bin width using the Scott’s rule (Scott, 1992), which is one of the standard bin width selection criteria implemented in many statistics packages (e.g. Jones et al., 2001–; Astropy Collaboration et al., 2013; R Core Team, 2015). The bin width h is then selected as:

$$h = 3.5 \frac{\sigma}{n^{1/3}} \quad (5.3)$$

where σ is the sample standard deviation of the redshift values and n is the number of galaxies in the sample. We note that there exists a plethora of different methods to select the bin width of a histogram (see e.g. Scott, 1992; Sheather, 2004; Rau et al., 2017). However the Scott’s rule in the form of Eq. 5.3 is a standard method that can be applied efficiently to large catalogs.

We note that, if the sample is weighted, the mean and the standard deviation are modified as

$$\langle z \rangle = \sum_{i=1}^n w_i z_i \quad (5.4)$$

and

$$\sigma^2(z) = \sum_{i=1}^n w_i (z_i - \langle z \rangle)^2 \quad (5.5)$$

respectively. Here the weights w_i are normalized to sum to unity and n denotes the sample size. In the case of a weighted sample, we modify the sample size used in Eq. 5.3, to be the effective sample size defined as

$$n_{\text{eff}} = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}. \quad (5.6)$$

The following discussion refers to the ‘normalized effective sample size’ N_{eff} as the effective sample size divided by number of galaxies n :

$$N_{\text{eff}} = \frac{n_{\text{eff}}}{n}. \quad (5.7)$$

We will now describe the datasets used in this analysis, concentrating on the COSMOS validation sample. For a further description of the data and the DES survey, we refer the reader to §8.2.

5.2 The Cosmos Validation Data

The DES Y1 analysis validates photometric redshifts in the 1 deg^2 large footprint that overlaps with the Cosmic Evolution Survey (COSMOS) (Scoville et al., 2007; Laigle et al., 2016). The COSMOS survey provides accurate imaging information in 16 narrow photometric bands. As a result, a complete sample of high-precision template fitting photo- z is available (Laigle et al., 2016) to compare against the broad band DES photo- z . In addition to multiband photo- z , spectroscopic surveys provide redshifts for 22,429 galaxies, where the zCosmos (Lilly et al., 2007a), PRIMUS (Coil et al., 2011) and 3D-HST (Momcheva et al., 2016) surveys provide 9841, 9620 and 2176 spectroscopic redshifts each. About 792 additional spec- z have been matched from other surveys like VVDS (Le Fèvre et al., 2005) and FMOS_COSMOS (Silverman et al., 2015).

Using this data we generate two catalogs by combining the spectroscopic redshifts (‘COSMOS spec- z ’) and the multiband photo- z (‘COSMOS photo- z ’) with the DES photometry in the COSMOS field. Fig. 5.2 shows, that the density of spectroscopic redshifts drops significantly for $z > 1.0$, while the COSMOS photo- z sample still contains a large number of high redshift galaxies. This is also reflected in the magnitude distributions plotted in Fig. 5.3, where the COSMOS spec- z sample becomes very sparse for $\text{MAG } i > 23$. In addition we overplot the 10% and 90% percentiles of the i band magnitude distribution of the WIDE field sample (‘DES WIDE’), that contains the SPT and Stripe 82 regions. We see, that the DES photometry in the COSMOS field is about 2 magnitudes deeper compared with the WIDE field. In addition to a larger magnitude coverage, the DES photometry also has a lower photometric noise in the COSMOS field compared with the WIDE field as shown in Fig. 5.4. We already discussed in §2.6.2, that differences in photometric noise need to be corrected to avoid biases in the derived regression functions. To use the COSMOS field for photo- z validation, we degrade the photometry in the COSMOS field to match the depth of the DES WIDE field (Hoyle et al., 2017). We also note that DES is currently not using Y band photometry.²

In a first step random Stellar Locus Regression (SLR) corrections (High et al., 2009) are drawn from the DES WIDE footprint and applied to all galaxies in COSMOS. A random

²Adding Y band is unlikely to improve the photometric redshift performance, due to its wavelength overlap with the z band. Furthermore Y band photometry doesn’t have the same depth and coverage as the remaining bands (see Hoyle et al., 2017).

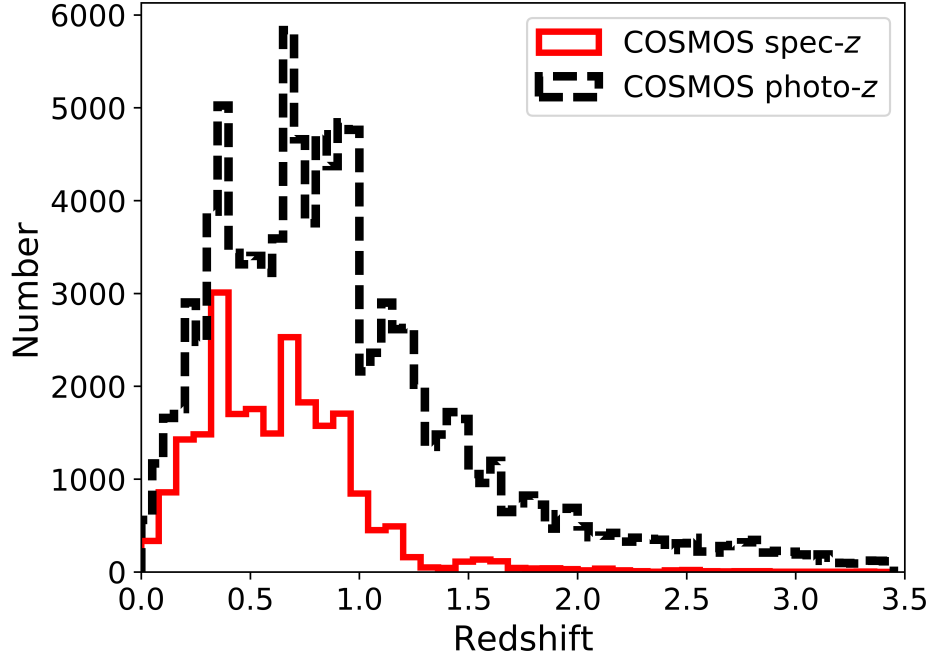


Figure 5.2: Redshift distributions for the spectroscopic dataset ('COSMOS spec-z') and the full COSMOS high-precision calibration sample ('COSMOS photo-z').

galaxy 'Rgal' is selected in the DES Y1 footprint outside the COSMOS field and all COSMOS galaxies that have a flux error in the DES i-band larger than that DES Y1 reference galaxy are removed. Subsequently we determine the additional error component required to equalize the magnitude of the flux error of Rgal and each remaining COSMOS galaxy using,

$$\sigma_{\text{add-error}} = \sqrt{\sigma(\text{FLUX_Rgal})^2 - \sigma(\text{FLUX_COSMOS})^2}. \quad (5.8)$$

Assuming a Gaussian noise model, we then perturb the mean value of each COSMOS flux by the corresponding value of $\sigma_{\text{add-error}}$.

The result of this procedure is a catalog with COSMOS galaxies that have the same signal-to-noise ratio as the DES Y1 reference galaxy. The galaxy that has the most similar magnitude values compared with the DES Y1 reference galaxy is then identified using a χ^2 fit.

In order to obtain a single catalog of $N=100,000$ galaxies, one repeats these steps by selecting random galaxies from DES Y1 and using the same SLR corrections N times. After generating a single catalog, the full procedure is iterated, including the random SLR corrections, to obtain 180 catalogs.

The result of this resampling procedure is shown in Fig. 5.4 where we plot the fluxerror distribution in the i-band for the datasets ('res DES COSMOS') that have been resampled by the aforementioned recipe. The 2σ areas of photometric error (red) obtained on the resampled datasets is in excellent agreement with the corresponding distribution of DES WIDE. The χ^2 fit used in the resampling procedure selects only those COSMOS galaxies that have similar photometry than the galaxies in the WIDE field. As a result, the photometric selection function of the DES WIDE sample is automatically implemented in our resampled files, i.e. the i band magnitude distribution of the resampled COSMOS validation files ('res DES

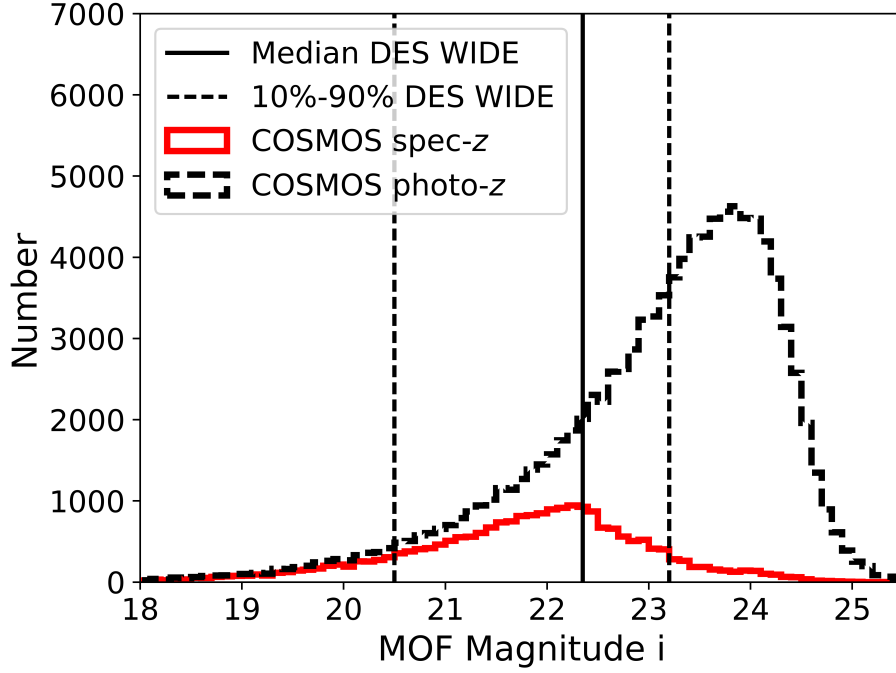


Figure 5.3: DES i band magnitude distributions for the spectroscopic dataset (‘COSMOS spec- z ’) and the COSMOS high-precision calibration sample (‘COSMOS photo- z ’). The solid vertical line shows the median, the dashed lines the 10% and 90% percentiles of the DES i band magnitude distribution of the DES Year 1 WIDE sample (‘DES WIDE’).

COSMOS’) matches the one of DES WIDE as shown in Fig. 5.5. We now match the available spectra to each of the 180 resampled COSMOS validation files to obtain the corresponding 180 spectrophotometric catalogs. The green contours in Fig. 5.5 compare the i band magnitude distributions of these spectrophotometric subsamples (‘res COSMOS spec- z ’) with the i -band magnitude distribution of the resampled photometry ‘res DES COSMOS’. We see that there persists a shortage of spectroscopic redshifts for faint galaxies between the resampled validation files and the spectroscopic subsample. However this discrepancy is much less pronounced than for the original COSMOS sample prior to the application of the matching algorithm.

To avoid confusion between the various datasets used in this chapter, we strictly adhere to the following naming convention. The spectroscopic and COSMOS high-precision photometric redshifts available for DES galaxies in the COSMOS footprint will be referred to as ‘COSMOS spec- z ’ and ‘COSMOS photo- z ’ respectively. The prefix ‘res’ indicates that the catalog has been resampled using the recipe described previously. If the respective catalog contains DES photometry, this method additionally degrades the flux measurements to match the signal-to-noise ratio in DES WIDE. If the respective catalog does not contain DES photometry it simply means that the catalog is matched against the resampled DES photo- z catalogs. Furthermore the suffix ‘sp’, abbreviating ‘spectrophotometric’, indicates that only those galaxies are selected for which spectroscopic redshifts are available.

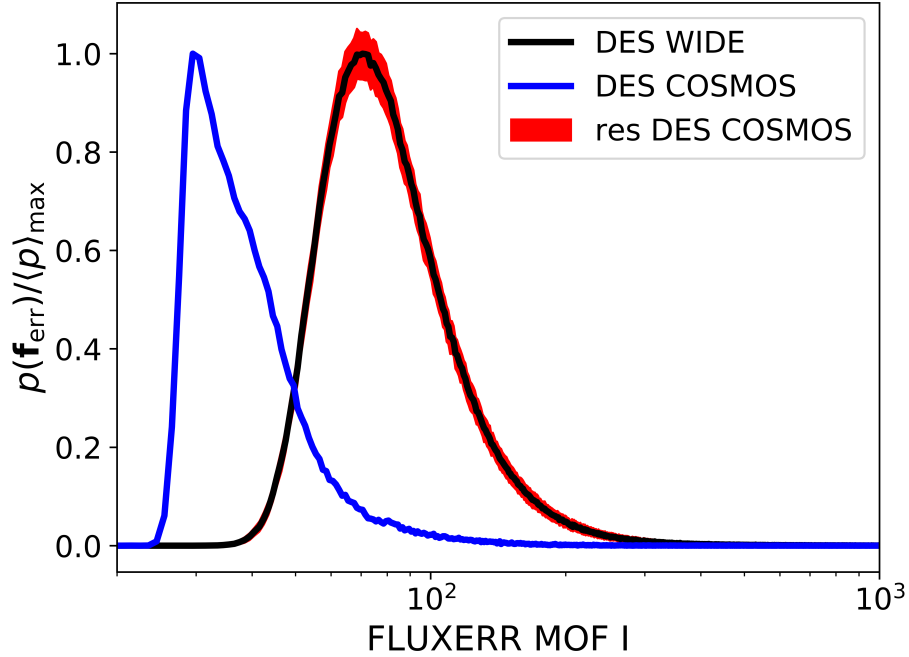


Figure 5.4: Distribution of i band flux error for the full DES science sample (‘DES WIDE’), the original DES photometry in the COSMOS field (‘DES COSMOS’) and the resampled DES photometry in the COSMOS field (‘res DES COSMOS’). The 2σ errorbars are obtained using the 180 resampled catalogs as described in the text. Note that we show the distribution of flux error in the i band on a logarithmic scale.

DES Photometric Redshifts

We estimate photometric redshift distributions using the template fitting code BPZ (Benítez, 2000) using the four band DES photometry in the resampled and degraded datafiles. Fig. 5.6 compares the sample redshift distribution of the DES photometric redshifts obtained on the degraded DES photometry in the COSMOS field (‘res DES photo- z sp’) with the COSMOS high-precision photometric redshifts (‘res COSMOS photo- z sp’) and the spectroscopic redshift distribution (‘res COSMOS spec- z ’). We reiterate, that these results have been obtained on the resampled catalogs with artificially degraded photometry as explained in the previous section, and only galaxies with spec- z are considered. We see that the COSMOS high-precision photometric redshift distributions are of higher quality, as measured by their agreement with the spec- z distribution, compared with the photometric redshift distributions obtained on the broad band DES photometry. Especially the peaked structure of the spectroscopic redshift distribution is washed out by the broad kernel of the DES photometric redshift estimates. The better performance of the COSMOS multiband photometric redshifts is a result of the higher number of photometric bands that are available in COSMOS. In contrast to DES, this enables the COSMOS multiband photometric redshifts to more accurately trace the peaked structure of the spectroscopic redshift distribution.

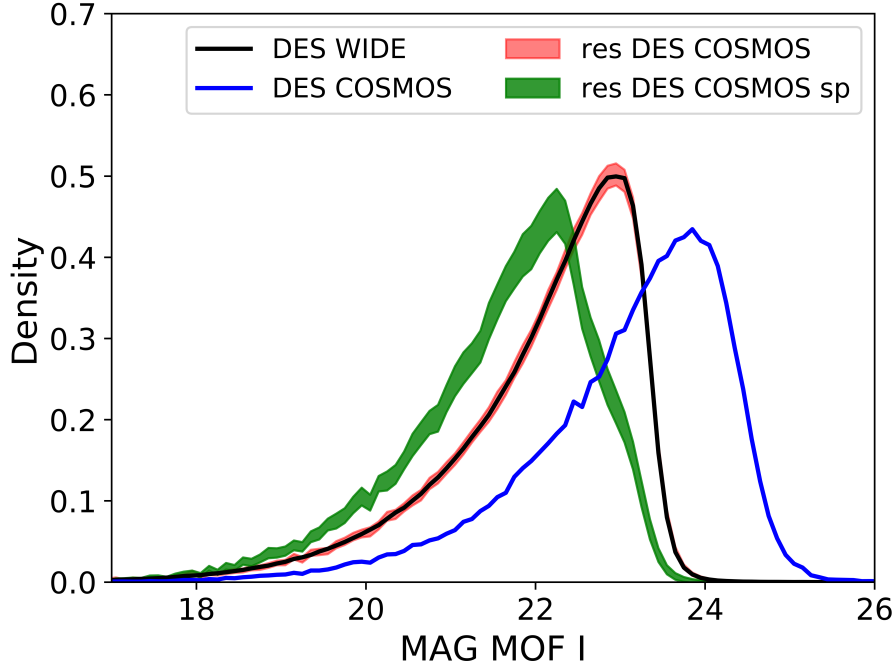


Figure 5.5: Distribution of i band magnitude for the full DES science sample (‘DES WIDE’), the original DES photometry in the COSMOS field (‘DES COSMOS’) and the resampled DES photometry in the COSMOS field (‘res DES COSMOS’). The 2σ errorbars are obtained using the 180 resampled catalogs as described in the text.

5.3 Variations in photometric noise

In this section we investigate how an increased photometric noise level degrades the photo- z quality in the context of the COSMOS validation sample.

For this study we will compare photometric redshift estimates obtained on the DES photometry with the corresponding spectroscopic redshifts and therefore exclusively consider the spectrophotometric subsample (‘DES photo- z sp’). To simplify the notation we will abbreviate ‘DES photo- z sp’ as ‘photo- z ’ within this section. In addition to the photo- z obtained on the degraded photometry (‘res photo- z ’), we obtain BPZ photometric redshift estimates on the nondegraded, original DES photometry (‘orig photo- z ’), which are then matched to the 180 resampled data files described in the last section.

Fig. 5.7 shows the residual distributions for res photo- z and orig photo- z , where we define the residuals as the difference between the mean z_{photo} of the individual galaxy photo- z distribution and the spectroscopic redshift z_{spec}

$$\Delta z = z_{\text{photo}} - z_{\text{spec}}. \quad (5.9)$$

We see that the error in Δz is much larger for the photo- z estimates obtained on the resampled photometry compared with the results obtained on the original photometry. Furthermore the mean redshift bias for ‘res photo- z ’ is larger than for ‘orig photo- z ’. This suggests that the effect of increasing the photometric noise level not only induces a bias, but also increases the residual variance of the photo- z estimate, as to be expected from the simplified discussion

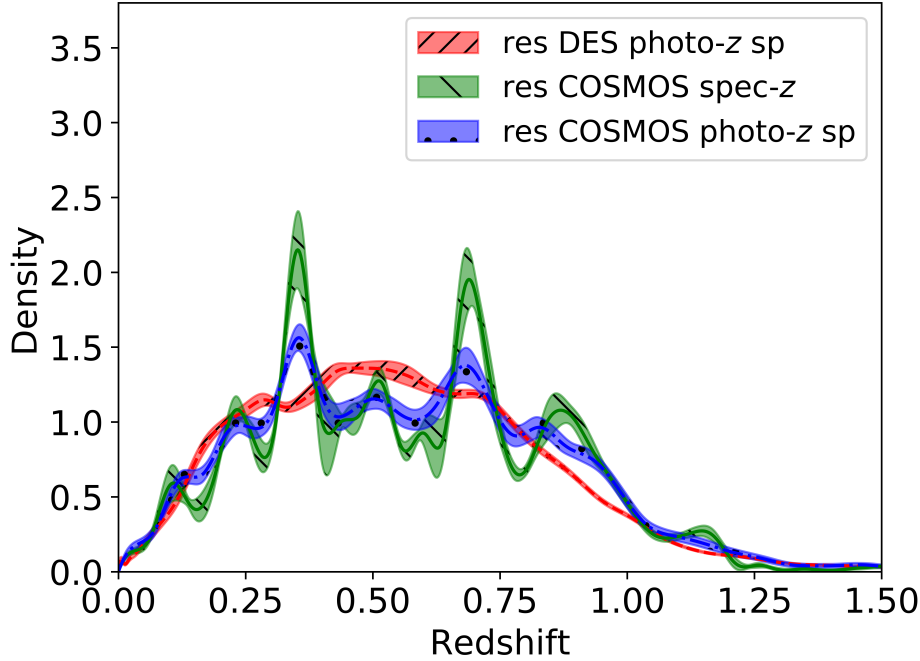


Figure 5.6: Photometric redshift distributions obtained with the BPZ template fitting code on the resampled and degraded photometry (`res DES photo- z sp`), the corresponding COSMOS high-precision photometric redshifts (`res COSMOS photo- z sp`) and the spectroscopic redshift distributions (`res DES photo- z sp`). Note that we consider only galaxies which have spectroscopic redshift measurements. The 2σ errorbars are obtained using the resampled datafiles as explained in the text.

in §2.6.2.

To investigate how this would affect a weak gravitational lensing analysis, we estimate the distribution of convergence power spectra using the stacked redshift distributions, i.e. the sum over the individual galaxy photometric redshift distributions, obtained on the resampled and the original photometry. In order to compare with the spectroscopic ‘truth’, we obtain the corresponding spectroscopic redshift distributions using histograms as described in §5.1.

Fig. 5.8 shows that the distribution of relative biases in the convergence power spectrum (Eq. 5.1) is significantly wider and shows a larger bias for the BPZ predictions obtained on the degraded photometry (`res photo- z`) compared with the corresponding result using the original photometry (`orig photo- z`). Furthermore, the relative bias in the lensing convergence power spectrum is quite high in both cases, which can be explained by the very peaked spectroscopic redshift distribution of the spectrophotometric COSMOS sample. As shown in Fig. 5.6 the photometric redshifts obtained on the DES photometry are not able to resolve these large scale structure peaks. The poor performance of the DES photometric redshifts on the spectrophotometric COSMOS sample then propagates into large biases in the convergence power spectrum. We further note, that the relative biases in the convergence power spectrum as a function of ℓ obtained for the different resampled catalogs, to good accuracy, only differ by an offset and are highly correlated. This justifies using Eq. 5.2 as a summary statistic for the relative bias in the convergence power spectrum.

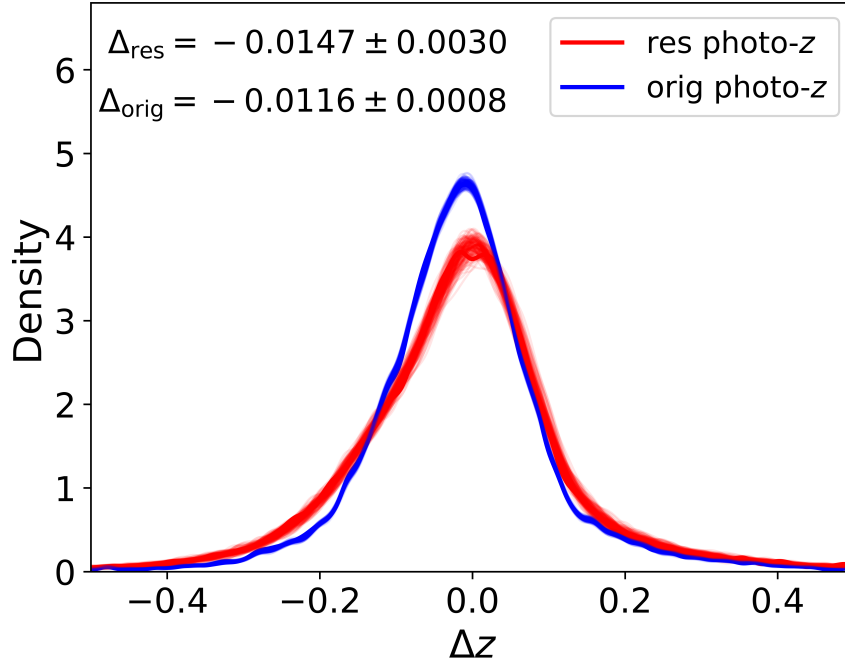


Figure 5.7: Distribution of photometric redshift residuals $\Delta z = z_{\text{phot}} - z_{\text{spec}}$, where z_{phot} denotes the mean of the individual galaxy photo- z distribution and z_{spec} the spectroscopic reference redshifts. We show the residual distributions for the BPZ template fitting method applied to the degraded DES photometry (‘res photo- z ’) and the original DES photometry (‘orig photo- z ’) in the COSMOS field. The blue and red regions are obtained by plotting the corresponding results for the 180 resampled datasets as described in the text.

Current weak lensing analyses (e.g. Bonnett et al., 2016; Troxel et al., 2017; DES Collaboration et al., 2017) assume that uncertainties in the mean of the photometric redshift distribution are the main contributor to biases in weak gravitational lensing analyses. We repeat the previous analysis, but center all photometric redshift distributions to the mean of the respective spectroscopic redshift distribution. The result is shown in Fig. 5.8 for both the original (‘orig photo- z corr’) and the resampled photometry (‘res photo- z corr’). We see, that indeed both the bias and the variance in $\Delta(C_\ell^\kappa)$ are much smaller for the corrected photometric redshift distributions. However there remains a significant residual error of $\approx 1\% - 2\%$ that suggests, that errors in other statistics of the photometric redshift distribution also contribute in a significant way to the total error budget.

5.4 Spectroscopic Incompleteness

As described in the previous sections, photometric redshifts obtained on the DES photometry are not accurate enough to trace the narrow large scale structure peaks in the COSMOS spec- z distribution and therefore show substantial photo- z errors. As we can reasonably expect the redshift distribution of the full DES sample to be much smoother than this spectroscopic reference sample, the error estimates obtained in the previous section are likely to be quite conservative for the photo- z accuracy of DES WIDE. Furthermore, the incompleteness of the

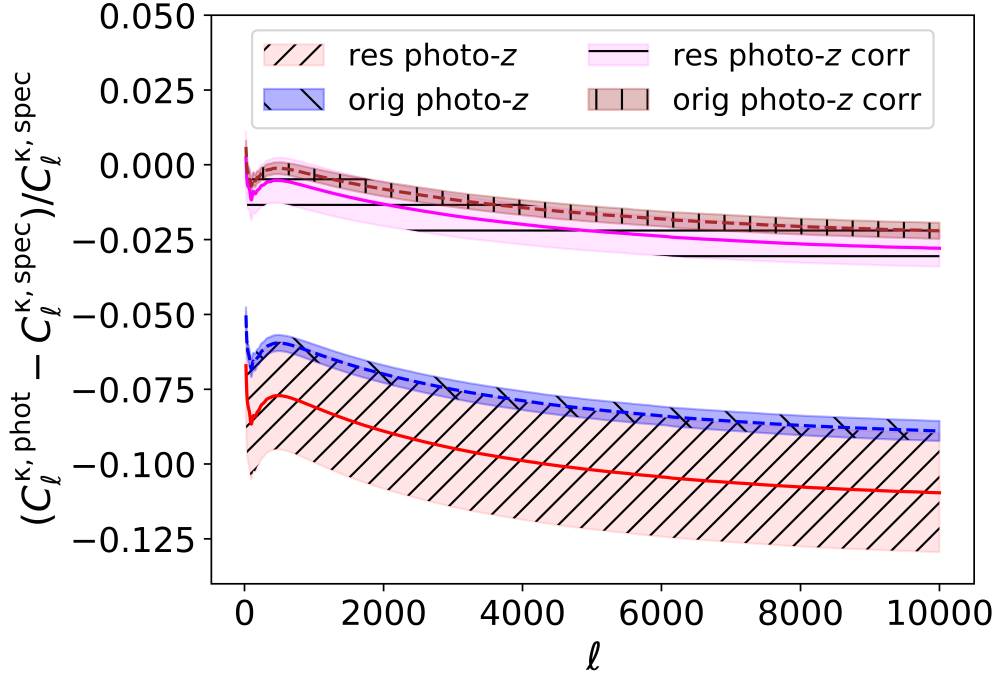


Figure 5.8: Bias in the convergence power spectrum for the BPZ photometric redshift predictions using the degraded (‘res photo- z ’) and original (‘orig photo- z ’) DES photometry in the COSMOS field. We also show the respective results if the photometric redshift distributions are shifted such, that their mean coincides with the true, spectroscopic result (‘res photo- z corr’ and ‘orig photo- z corr’). The 1σ error contours have been obtained using the resampled datasets as explained in the text.

spectroscopic validation sample at the faint end of the color-magnitude space would require removing a substantial fraction of faint, high redshift galaxies from DES WIDE. The DES photometric redshift validation strategy therefore uses the multiband photometric redshifts in COSMOS, as a substitute for spectroscopic redshift validation. Since the COSMOS data extends to much greater photometric depth, we are able to maintain the high statistical accuracy of the DES analysis without the need to remove data from our science samples.

However while the COSMOS high-precision photometric redshifts are obtained on narrow band photometry and are therefore substantially more accurate than the DES photo- z , they can still be subject to photo- z errors. In the following section we will quantify this error by comparing the COSMOS multiband photo- z with spectroscopic redshifts in those regions of color-magnitude space where spectroscopic redshifts are available.

Methodology

The general ansatz for the selection function of the spectrophotometric dataset, i.e. a dataset with both photometry and spectroscopic redshift information, $S(\mathbf{f}, z)$ would depend on both the photometric color space \mathbf{f} as well as the spectroscopic redshift z . A selection function in redshift dimension cannot be derived without either accurately modeling the data generation process or using detailed simulations, as the redshift is a ‘hidden variable’ in the photomet-

ric dataset. The simplifying assumption made in this chapter therefore assumes a selection function that depends on the photometry alone.

We define the selection function $S(\mathbf{f})$ for the joint probability distribution of the photometric features \mathbf{f} , i.e. the set of magnitudes, colors and fluxes, and the redshifts z of the spectrophotometric sample $p_{\text{spec}}(\mathbf{f}, z)$ as

$$p_{\text{spec}}(\mathbf{f}, z) = S(\mathbf{f}) p_{\text{photo}}(\mathbf{f}, z). \quad (5.10)$$

Here $p_{\text{photo}}(\mathbf{f}, z)$ denotes the joint probability distribution of the color-redshift space of the full photometric sample, where the redshift z is unknown. For simplicity, we assume in Eq. 5.10 that the spectroscopic redshifts have negligible error compared with the true redshift. The selection function $S(\mathbf{f}) = p_{\text{spec}}(\mathbf{f})/p_{\text{photo}}(\mathbf{f})$ is therefore given as the density ratio between the color space populated by the spectrophotometric subsample $p_{\text{spec}}(\mathbf{f})$ and the full photometric color-magnitude space $p_{\text{photo}}(\mathbf{f})$. Given an estimate for $S(\mathbf{f})$, we can obtain weights for each galaxy in the photometric dataset. The distribution of these weights in color-magnitude space quantifies relative under and overdensities of spectroscopic redshifts and is therefore a metric for the spectroscopic ‘representativeness’.

In this work we use the method of Lima et al. (2008); Cunha et al. (2009) which is based on the nearest neighbor density estimate, that is constructed by selecting n_{photo} nearest neighbors for each galaxy in the photometric dataset. This effectively constructs a hypersphere around each photometric galaxy with radius given as the distance to its farthest nearest neighbor. Afterwards we count the number of spectrophotometric galaxies n_{spec} in this hyper volume. The weights for each galaxy in the photometric dataset are then given as

$$w_i = \frac{1}{n_{\text{spec,tot}}} \left(\frac{n_{\text{spec}}(\mathbf{f}_i)}{n_{\text{photo}}(\mathbf{f}_i)} \right) \quad (5.11)$$

where $n_{\text{spec,tot}}$ is the total number of galaxies in the spectrophotometric catalog. The result is a weighted photometric dataset, that has the same (weighted) color-magnitude space as the corresponding spectrophotometric dataset, where we use the four magnitudes in g, r, i, z as input features. The number of nearest neighbors n_{photo} is a tuning coefficient that governs the ‘smoothing’ of the density ratio estimate. It is important to note that this parameter requires careful tuning. Choosing a small value leads to a very ‘local’ density ratio estimate, where the weighted photometric sample will then have a very similar color-redshift distribution compared with the spectrophotometric reference sample. On the other hand, this can also lead to more galaxies being downweighted. The correct number of nearest neighbors therefore needs to be chosen such, that there exists a good tradeoff between the effective sample size and the agreement between the weighted redshift distribution and the reference distribution. The metric that quantifies the similarity between two redshift distributions ultimately depends on the particular science case, where we choose the relative bias in the convergence power spectrum (Eq. 5.2).

To quantify over- and underdensities in the weighting, we define the relative weight of the photometric galaxy as

$$w_{\text{rel},i} = \frac{w_i}{1/n_{\text{photo}}}, \quad (5.12)$$

which is the weight excess over a uniformly weighted dataset. We assume here that the set of weights w_i has already been normalized to sum to unity. We finally note, that this approach is the inverse from the traditional methodology of weighting a spectrophotometric dataset to

the full photometric sample. The objective of this more traditional approach is to validate a photometric sample given a representative spectrophotometric sample. If the spectrophotometric sample does not cover the full color-magnitude space of the photometric sample, the obtained weighting scheme is not well behaved, as very few spectrophotometric galaxies are attributed a very large weight. In contrast, our methodology would weight down the galaxies in the full photometric sample in those regions of color-magnitude space, that are insufficiently populated by spectroscopic calibration redshifts. This will give us both information about the spatial extend of the incompleteness and makes these weights usable in following analyses by defining a photometric sample of galaxies with well calibrated photometric redshifts.

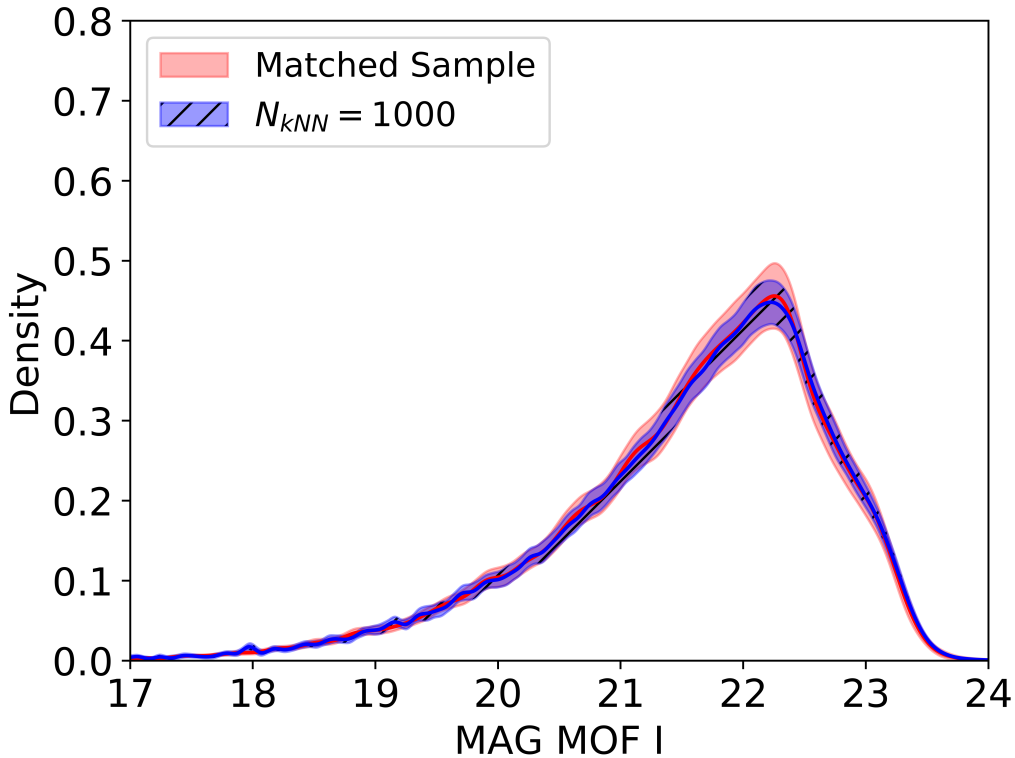


Figure 5.9: Distribution of i band magnitude for the matched sample and the weighted COSMOS multiband sample ($N_{kNN} = 1,000$). The 2σ error contours are obtained using the resampled datafiles.

5.4.1 Analysis and Results

The following analysis uses the high-precision multiband photo- z from COSMOS (‘res COSMOS photo- z ’), the respective spectrophotometric subsample (‘res COSMOS photo- z sp’) as well as the spectroscopic redshifts (‘res COSMOS spec- z ’). We will use the weighting scheme described in the previous section to weight ‘res COSMOS photo- z ’ onto ‘res COSMOS photo- z sp’. This means instead of matching the multiband photo- z from COSMOS onto the spectroscopic dataset (‘res COSMOS photo- z sp’), we attach weights to the full COSMOS photo- z sample (‘res COSMOS photo- z ’), such that it matches the color-magnitude distribution of the

spectroscopic subsample. Ideally, the accuracy of the weighted ‘res COSMOS photo- z ’, as measured by $\Delta(C_\ell^k)$, has then to coincide with the accuracy of ‘res COSMOS photo- z sp’.

To unclutter the notation, this section refers to the weighted ‘res COSMOS photo- z ’ sample, simply with the number of nearest neighbors selected during weighting, e.g. $N_{kNN} = 1000$. The corresponding spectrophotometric sample, i.e. ‘res COSMOS photo- z sp’, will be simply called the ‘matched sample’. All performance estimates are obtained with respect to the respective spectroscopic distributions. The statistical error estimates are obtained from the 180 resampled datafiles.

As noted earlier, we obtain a weight for each photometric galaxy, that quantifies the relative density of surrounding spectroscopic galaxies in color-magnitude space. In order to calculate the weights using Eq. 5.11, we need to tune the number of nearest neighbors N_{kNN} to consider in the density ratio estimate. Typically the accuracy of these weighting schemes is quantified as the agreement between weighted and target magnitude distribution (see e.g. Sánchez et al., 2014; Bonnett et al., 2016). This is shown in Fig. 5.10, where we compare the distribution in the DES i band magnitude for the matched sample and $N_{kNN} = 1,000$. Both distributions are in excellent agreement, which suggests that the weighting scheme is sufficient to mimic the color-magnitude space of the spectrophotometric sample.

Fig. 5.10 shows the corresponding mean relative bias in the convergence power spectrum for various numbers of nearest neighbors. Despite the consistency in terms of the magnitude distribution, the weighted photometric redshift distribution using 1,000 nearest neighbors ($N_{kNN} = 1000$), shows a noticeably larger bias, compared with the results that use a lower number of nearest neighbors and therefore a more local culling. The improved accuracy gained by reducing the number of nearest neighbors comes at the price of a reduced normalized effective sample size (Eq. 5.7). Comparing with the error obtained using the matched sample, we see that the weighted COSMOS multiband photo- z still perform considerably worse even if we cull away almost half of the data, as shown by the extreme point $N_{kNN} = 10$.

We can therefore conclude that while the matched sample of COSMOS has almost an order of magnitude lower error than the DES photometric redshifts, it is difficult to extrapolate this result in color space. Even a very stringent weighting scheme using $N_{kNN} = 10$ will increase the error budget by a couple of percent. Furthermore, judging the number of nearest neighbors purely in terms of the agreement in the magnitude distributions is not accurate enough, if the selection function is very local, as is the case for the spectrophotometric COSMOS sample.

We further illustrate the sample selection bias in Fig. 5.11 using a scatter plot of the $g - r$ color against the $r - i$ color of the DES COSMOS sample in several bins of the COSMOS high-precision photometric redshift using $N_{kNN} = 100$, which is a modest weighting scheme that still maintains a decent sample size ($N_{eff} = 0.75$). The color coding illustrates the relative weight (Eq. 5.12), which is a measure of how representative the spectrophotometric sample is for the full DES photometric color space in the COSMOS footprint. Each tomographic bin contains the same number of galaxies, where we plot 3,000 galaxies in each bin, randomly drawn from all resampled datasets of the DES photometry in the COSMOS field. Dark areas correspond to a low relative weight, i.e. a weight decrement over an equally weighted sample, light regions indicate overdensities of spectroscopic calibration redshifts. We see that underpopulated regions can be especially found in the highest photometric redshift bin $z_{phot} = [0.98, 5.87]$ and on the edges of the color-magnitude space in lower redshift bins. This shows, that the spectroscopic surveys cover the color-magnitude space very irregularly and especially sparse at the faint end.

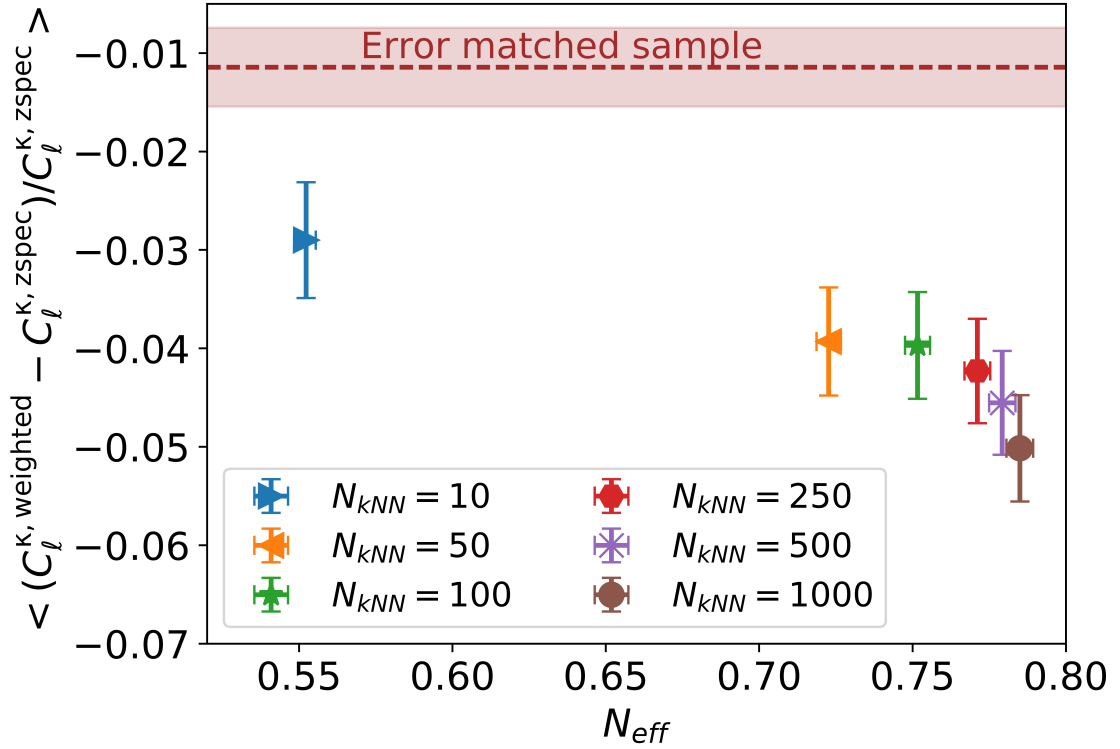


Figure 5.10: Average relative bias in the convergence power spectrum as a function of the normalized effective sample size N_{eff} (Eq. 5.7). A low value indicates that many galaxies in the photometric dataset (res COSMOS photo- z sp) are downweighted due to the incomplete coverage with spectroscopic galaxies in these regions. The points correspond to a different number of nearest neighbors N_{kNN} used during weighting. The brown 1σ contour corresponds to the relative bias obtained using the matched sample. All 1σ errorbars are estimated using the resampled datasets as described in the text.

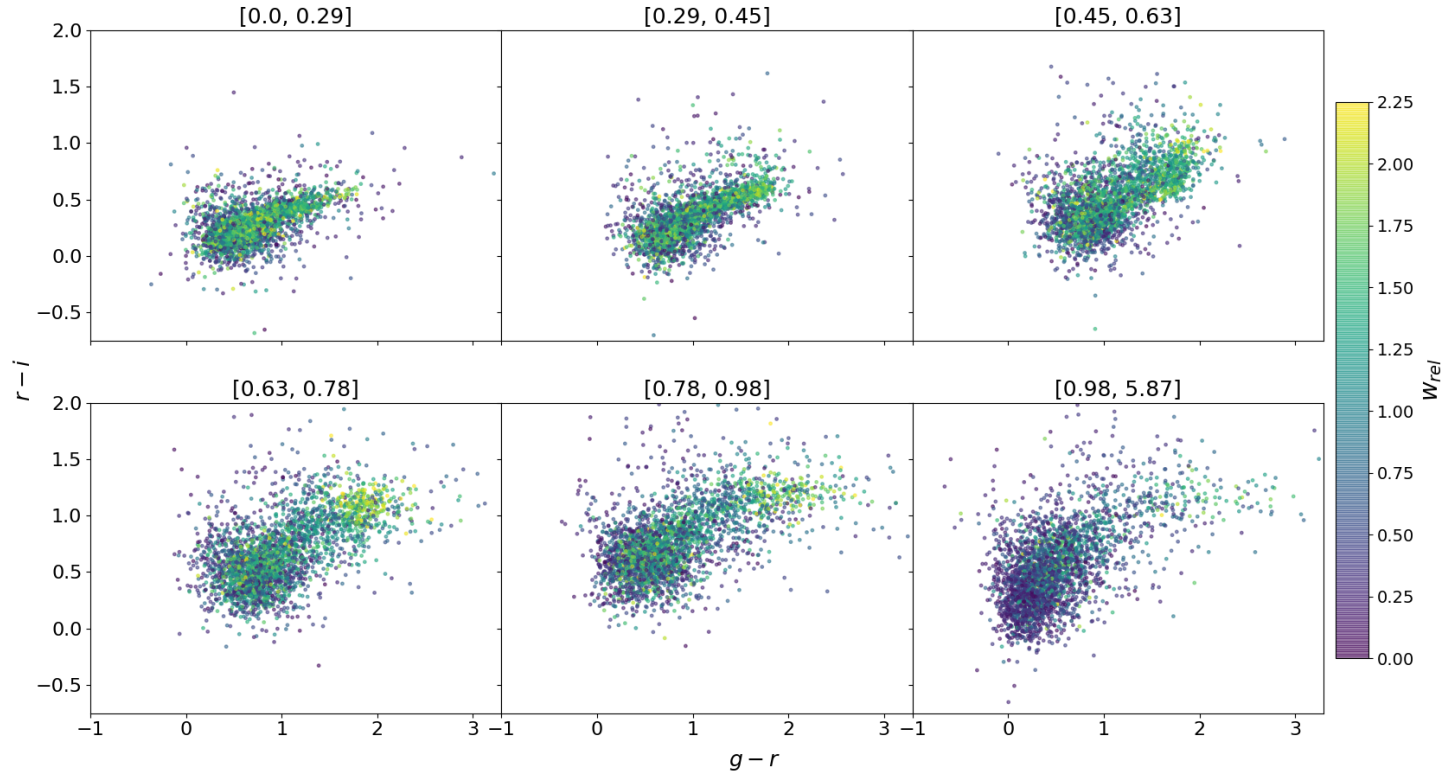


Figure 5.11: Scatter plot of the $g-r$ color against the $r-i$ color of the full DES photometric color space in the COSMOS field. The subplots show the color space $g-r$ vs $r-i$ equal frequency binned in terms of COSMOS photo- z , where the interval limits are quoted in the subplot titles. The color coding is a measure of how representative the spectrophotometric subsample is for this full photometric color space. Light yellow areas correspond to a large relative weight (Eq. 5.12) and dark blue areas to a low relative weight.

5.5 Compressing photometric redshift distributions

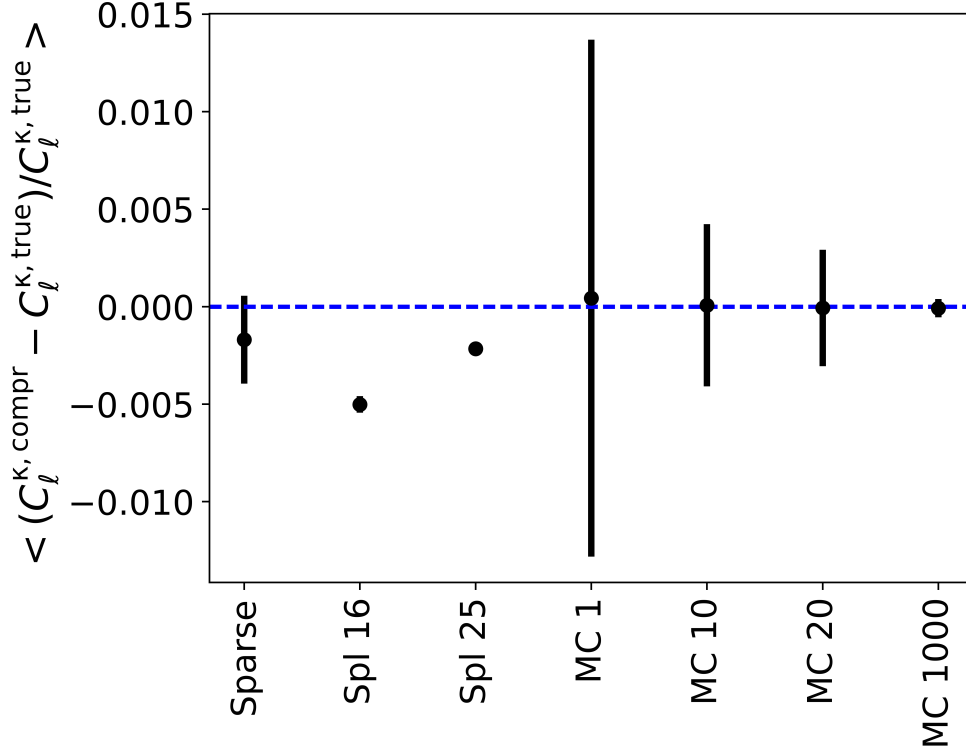


Figure 5.12: Average relative bias in the convergence power spectrum (Eq. 5.2) obtained for different compression methods, where we compare the accuracy of the compressed distribution with the original uncompressed distribution. We show the performance of the Sparse representation method (‘Sparse’), the Spline compression method that uses 16/25 floating point numbers per galaxy (‘Spl 16/25’), and the Monte Carlo sampling method that draws 10 (‘MC 10’), 20 (‘MC 20’) and 1,000 (‘MC 1000’) samples per individual galaxy redshift distribution. The case MC 1000 used the same bin width ($\Delta z = 0.01$) as the Y1 science analysis. In all other applications of the MC method, we used the Scotts rule (Eq. 5.3) to select the bin width. The dashed blue line highlights the origin, i.e. the best possible performance. The 2σ errorbars are obtained using the resampled datasets as described in the text.

The accurate and efficient representation of photometric redshift distributions becomes increasingly important, especially if hundreds of millions of galaxies need to be processed in current and future large area photometric surveys. It is therefore timely to devise efficient and accurate methods to compress and represent photometric redshift distributions.

This section investigates, how inaccurate compression and representation of redshift distributions can lead to biases in the convergence power spectrum. We evaluate several compression algorithms using the BPZ photo- z distributions obtained on the 180 resampled datafiles of DES photometry within the spectrophotometric subsample, that each contain about 15,500 galaxies. The following section will introduce the methods and the evaluation strategy used in this section.

5.5.1 Compression Algorithms

The current DES photometric redshift strategy uses the Monte Carlo Sampling algorithm as their primary technique to distribute photometric redshift distributions. The following sections will quantify the error in this technique in comparison with contemporary methods.

Monte Carlo Sampling

The Monte Carlo (MC) sampling approach draws a single redshift from each individual galaxy redshift distribution in the galaxy catalog. These Monte Carlo samples are then used to reconstruct the redshift distribution of the galaxy sample. The MC approach is extremely storage efficient, as it allows the estimation of sample redshift distributions using a single floating point number per galaxy. On the other hand, this recipe loses information about the exact shape of the individual galaxy redshift distributions in the sampling process. The application of density estimates, e.g. histograms, can then lead to an intrinsic error in the reconstruction as discussed in §4.

Spline Compression

The Spline compression algorithm is based on the simple idea to interpolate a smooth photometric redshift distribution using Spline interpolation and evaluate this density on a coarser grid. As photometric redshift distributions are typically peaked closely around a best fitting redshift, we can significantly reduce the total number of density values to store, by cutting out regions of the photometric redshift distribution that do not contribute much to the overall density. The total number of data points that need to be stored are the density values at the grid points, that are above a predefined threshold, as well as the start and end indices on the grid. In this way we are cutting away the leftmost and rightmost part of the distribution until we reach density values above the predefined threshold. The accuracy and efficiency of this discretization will therefore depend on the number of grid points used in the interpolation, the threshold and the redshift range of the photometric redshift distribution. The number of floating point numbers that need to be stored using this method is typically around 20 numbers per galaxy, even for a large redshift range of $0.0 < z < 3.0$. The accuracy of the Spline compression method is quite insensitive to the floating point accuracy of the stored density values. We can therefore store them as 16 bit floats and the two indexes as 8 bit unsigned integers. In contrast to using Monte Carlo samples from the distribution, the Spline compression method directly compresses the shape of the distribution. As noted earlier, the method interpolates each individual galaxy redshift distribution on a fixed grid. Redshift distributions of samples of galaxies are therefore efficiently constructed by summing up the density values evaluated at these grid points. It has to be emphasized that this method is computationally extremely efficient and easy to implement and use.

Sparse Representation

The Sparse representation method (Carrasco Kind & Brunner, 2014b) is based on the concept of ‘dictionary learning’ (Mairal et al., 2009) and decomposes a photometric redshift distribution $\mathbf{p}_k(z)$ of the galaxy k as a linear combination of basis functions at each grid point

$$\mathbf{p}_k(z) = \mathbf{D} \delta_k, \quad (5.13)$$

where \mathbf{D} is the ‘dictionary matrix’ and δ_k is the solution vector. The dictionary matrix has $n \times m$ dimensions where n is the number of points in the original photometric redshift distribution and m is the number of basis functions. In general the number of basis functions is much larger than the number of points in the original redshift distribution $m > n$. The Sparse representation technique optimizes this overdetermined system of equations, to optimally represent the original shape of the redshift distribution. The basis functions are selected to be Gaussian and additionally include Voigt profiles to mimic the extended wings that are present in some photometric redshift distributions. We refer the interested reader to Carrasco Kind & Brunner (2014b) for a more in depth description of the method.³ For the following analysis we use a set of 20 basis functions.

5.5.2 Analysis and Results

To evaluate the quality of the compression, we will first compress and subsequently decompress the individual galaxy redshift distributions. If the compression scheme is inaccurate, the decompressed sample redshift distributions will no longer coincide with the original distributions. We then quantify this error in terms of the average bias in the convergence power spectrum (Eq. 5.2).

In analogy to the previous sections, Fig. 5.12 shows the photometric redshift accuracy in terms of the relative bias in the convergence power spectrum (Eq. 5.2), where we plot 2σ error bars. We see that the Sparse representation and Spline compression algorithm that use 25 floating point numbers per galaxy perform quite well, while the Sparse representation algorithm shows a much greater variance compared with Spl 25⁴. The compression power of both algorithms is comparable, where the Sparse representation algorithm uses 20 bases per galaxy and the Spline compression about 25 step points. Decreasing the number of step points to be stored in the Spline compression method leads to a slight increase in error as shown for the case of Spl 16. However even this accuracy is still well below 0.5 % and therefore negligible compared with the typical biases to be expected for the high precision photo- z in COSMOS. While the Sparse representation method produces quite accurate results, we note that it requires considerable computation time. We therefore evaluated its performance not on the full set of redshift distributions but on a smaller random subset containing about 1,500 individual galaxy redshift distributions. We also compared the performance of the Spline compression algorithm with a Wavelet compression algorithm (Julia Gschwend personal communication) of similar compression power, finding that both techniques produce a comparable error.

In contrast, the Monte Carlo sampling methods can lead to significant errors, especially if only a single Monte Carlo draw (‘MC 1’) is used. However this error is quickly reduced, if we increase the number of MC samples. For 20 Monte Carlo samples we even reach comparable accuracy with the more sophisticated compression strategies, which suggests that for the purpose of storing photometric redshift distributions to reconstruct redshift distributions of samples of galaxies, this method suffices. We further simulated a ‘DES-like’ sample size by drawing 1000 Monte Carlo samples. Since the spectrophotometric datasets we consider contain about 15,500 galaxies, the ‘MC 1000’ result corresponds approximately to the expected performance of the MC method on a weak lensing photometric redshift distribution in DES. For this case we made sure to use the same binning scheme ($h = 0.01$) that is used in the DES cosmic shear analysis. As can be seen ‘MC 1000’ performs very well and we do not expect that

³We used the implementation of this method as available at <https://github.com/mgckind/SparsePz>.

⁴If the errorbars are not visible, they are smaller than the pointsize.

the Monte Carlo method contributes in any significant way to the error budget at these large sample sizes, at least if we consider smooth redshift distributions in a weak lensing analysis.

5.6 Summary

The accurate estimation of photometric redshift distributions for large area photometric surveys like DES is an important prerequisite to use these surveys for precision cosmology. The broad band filters used in these surveys however provide only limited information about the spectral energy distribution of the galaxies. Measurements of their redshift are therefore typically subject to errors, that need to be accurately controlled, to ensure unbiased cosmological parameter constraints.

Photometric redshift calibration compares the photometric redshift estimates obtained on this broad band photometry with more accurate calibration data typically taken from overlapping spectroscopic surveys. However previous studies in DES Science Verification (Bonnett et al., 2016) suggest that the available spectroscopic samples for DES are highly incomplete, especially at the faint end of the color-magnitude space. In the DES Y1 photometric redshift validation pipeline we therefore use complete and accurate multiband, narrow filter photometric redshifts in the COSMOS field as the main source of redshift calibration. To ensure that the redshift performance estimates obtained on this COSMOS validation data agree with the performance on the full Dark Energy Survey Year 1 (DES WIDE) sample, we need to accurately control several sources of systematic error. This paper focuses on three of the most dominant effects, that need to be controlled to ensure accurate photometric redshift validation. The first effect we addressed, is the larger depth of the DES photometry in the COSMOS field compared with the DES WIDE sample. As higher quality photometry naturally translates into more accurate photometric redshift estimates, we would underestimate the photometric redshift error budget without adapting the DES photometry in the COSMOS field to be representative of the signal-to-noise in the DES WIDE sample. The DES photometric redshift validation pipeline corrects this, by degrading the DES photometry in the COSMOS field to ensure a consistent signal-to-noise ratio between the COSMOS validation field and the DES Y1 science sample. This is done using a resampling technique, that generates a set of validation files that can then be used to validate the DES photometric redshifts. In §5.3 we demonstrate that the differences in photometric noise significantly degrade the quality of the photometric redshift estimate both in terms of a systematic bias, as well as in terms of an increased statistical variance. We also discuss that the large biases in the lensing convergence power spectrum we obtain using the DES photo- z as compared with the results obtained using spec- z is a result of the large-scale-structure peaks in the spectroscopic dataset, that are largely oversmoothed by the broad resolution of the DES photometric redshifts. The distributions in larger fields will very likely be much smoother and the DES photo- z will perform significantly better.

The ill-controlled selection functions in spectroscopic data make the validation of DES photo- z using spectroscopic redshifts difficult. Furthermore, the spectroscopic validation data available for DES is very sparse at the faint end of the color-magnitude space. The DES Y1 photometric redshift validation strategy is therefore based on using high-precision photometric redshifts in the COSMOS field as the main source of photometric redshift validation (Hoyle et al., 2017). While this strategy largely provides a complete source of validation, it is still important to validate the high-precision photo- z in COSMOS, to control their intrinsic photo- z error. In §5.4 we analyze which areas of color-magnitude space are populated by accurate

spectroscopic validation data, by weighting the full DES photometric sample in the COSMOS footprint onto the spectrophotometric subsample. This method gives very low weight to all galaxies in the photometric sample that occupy regions with few spectroscopic redshifts and therefore effectively removes all regions of color-magnitude space with incomplete spectroscopic coverage. The photo- z performance of this weighted photometric sample can then be compared with the matched spectrophotometric sample, where an exact spectroscopic redshift estimate is available for each photometric galaxy. By carefully tuning this weighting procedure, we find that we can validate the full DES photometric sample in the COSMOS footprint up to an normalized effective sample size (Eq. 5.7) of 75% of its full, unweighted sample size. However we still obtain a relatively large relative bias of $\approx 4\%$ in the lensing convergence power spectrum compared with the results obtained using the spectroscopic distribution. This is larger than the corresponding 1% relative bias for the spectrophotometric subsample. The inconsistency between the performance obtained on the spectrophotometric dataset and the weighted photometric dataset makes it problematic to extrapolate the 1% error to other regions of color magnitude space. We therefore suggest the usage of conservative error estimates for the accuracy of the COSMOS multiband photo- z , as their accuracy cannot be quantified in a reliable way using the available spectroscopic data.

Ongoing and future large area photometric surveys will obtain photometric information for hundreds of millions of galaxies. The rapid increase in the size of the photometric data sets, raises the demand for the accurate and efficient storage of photo- z distributions of individual galaxies. A commonly used technique to efficiently compress photometric redshift distributions is the Monte Carlo sampling method, that draws a single sample from each individual galaxy redshift distribution. The redshift distribution of samples of galaxies is then reconstructed using these Monte Carlo draws. As shown in §5.5.2 this methodology can lead to a significant error of 1% within 2σ when applied to a relatively small sample size of about 15,500 galaxies. In these cases, where the sample size is small, we suggest using the more accurate compression algorithms described in this work e.g. the Spline compression method. However for a large, ‘DES Y1-like’ sample size of $\approx 15,000,000$ galaxies, this error component is negligible. This is also in agreement with similar work from Carrasco Kind & Brunner (2014b); Rau et al. (2015) and the analyses presented in §4. We however note that the accuracy of density estimates, and therefore of the MC technique, is strongly dependent on the shape of the photometric redshift distribution and the cosmological probe under consideration. In this regard, the smooth DES photo- z distribution did favour this method.

Future large area photometric surveys are expected to observe hundreds of millions of galaxies to an unprecedented depth and accuracy. While these datasets provide exciting chances to improve our understanding of the Universe and the growth of structure, it also sets high requirements on the control of systematic sources of error. Photometric redshift uncertainty is a major component of the overall error budget in these surveys. It is therefore paramount to incorporate all sources of error into the photometric redshift validation pipeline. This work contributes to this effort by studying the effect of sources of error that bias photo- z distributions: variations in the photometric noise level, incomplete spectroscopic calibration samples and the intrinsic error from misrepresenting photometric redshift distributions. The combined error budget from these sources of error can then be incorporated into the cosmological analysis.

Conclusions and Future Work

Photometric redshift estimation is one of the most exciting technical challenges for precision cosmology. This thesis developed a complete photometric redshift methodology based on Machine Learning, that not only enables the accurate estimation of photometric redshifts, but also incorporates modeling errors in photometric redshift distributions into the subsequent cosmological analysis. In addition to the development of algorithms to estimate photometric redshifts, I also investigated several different sources of error that can bias the validation of photometric redshifts in the context of the Dark Energy Survey.

I developed accurate methods to estimate photometric redshift distributions for individual galaxies and samples of galaxies using Machine Learning. As demonstrated in §3 using public CFHTLenS data, these methods significantly improve the modeling of cosmological probes, such as angular correlation functions and cosmic shear correlation functions, as well as cluster mass estimates, over results obtained using more traditional methodologies like Neural Networks. The algorithms developed in §3 not only enable the accurate modeling of photometric redshift distributions, but make their storage and processing more efficient. The Highest Weight Element method enables the accurate estimation of photometric redshift distributions for samples of galaxies using a single floating point number per galaxy, and a method based on Gaussian Mixture models expresses individual galaxy redshift distributions using on average between 2 to 5 floating point numbers.

However the accuracy of the constructed redshift distributions depends on the number of available calibration galaxies and the selected model parameters. Especially their estimation using histograms or kernel density estimates can lead to biased distributions, if the sample size is low or the bin width improperly chosen. It is worthy to note that this issue affects all cases where a distribution needs to be estimated on a finite sample, like photometric redshift validation using spectroscopic redshifts. To avoid cosmological parameter biases, I investigated in §4 how this statistical error can be incorporated into the cosmological analysis. This will be especially important for cosmological probes that are particularly sensitive to photometric redshift errors like galaxy clustering. I demonstrated that the selection of a too large bin width oversmooths the small scale structure of the cosmic density field, which can lead to systematic biases in the derived cosmological parameters. To reduce these errors, I discussed more accurate bin width selection algorithms that adapt the size of the bins with the shape of the distribution and the available sample size. However even if more advanced bin width selection algorithms are used, errors can still persist. I successfully incorporated these sources of

error into the cosmological analysis using a resampling technique, that quantifies the bias and the variance in the redshift distribution estimate and then corrects the parameter constraints accordingly. I demonstrate the success of this method using a DES-like galaxy clustering forecast.

In §5, I investigated sources of error that bias photometric redshift distributions in the context of the Dark Energy Survey photometric redshift validation strategy, that uses the COSMOS field as their primary source of redshift validation. I demonstrated, that field-to-field variations in photometric noise severely degrade the accuracy of photometric redshift estimates and can therefore severely bias the photometric redshift validation. In a subsequent analysis, I investigated which areas of the COSMOS color-magnitude space are sufficiently populated with spectroscopic reference galaxies. I find that the spectroscopic subsample covers the color-magnitude space of COSMOS only insufficiently. As a result, spectroscopic validation samples in the Dark Energy Survey are too incomplete to provide accurate estimates of photometric redshift performance. The alternative option is a validation using the high-precision photometric redshifts provided by the COSMOS survey. These redshifts are of much higher quality than the broad band photometric redshifts obtained on the DES photometry. While these photometric redshifts still have an error, I find that by adding a relatively modest error budget of 3-4 times the expected photometric redshift accuracy to the analysis, one can extend the validation to about 75% of the full photometric sample. As large area photometric surveys require the estimation of photometric redshifts for hundreds of millions of galaxies, compression algorithms for photometric redshift distributions are an important aspect of the photometric redshift strategy. I therefore investigated the compression accuracy of several compression methods in the context of the accurate modeling of convergence power spectra. I found that for this science goal, the simple Monte Carlo sampling method provides sufficient accuracy given the large sample sizes in the DES catalogs. However especially for smaller sample sizes, alternative methods that store the full shape of the individual galaxy redshift distributions, like the Spline compression method I developed, can be preferable. My work in the DES photometric redshift group significantly contributed to the development of the photometric redshift validation procedure and therefore to the accurate validation of the DES Y1 photometric redshift distributions. These derived errors are an important part of the DES Y1 cosmological analysis (DES Collaboration et al., 2017).

Looking ahead, photometric redshift uncertainty will likely remain an important systematic in large area photometric surveys for years to come. Especially the lack of complete and accurate spectroscopic validation data at the faint end of the color-magnitude space is a significant problem, as it prevents a direct estimation of redshift quality. Extended follow-up programs using multiband photometric and spectroscopic surveys, will definitely contribute to a solution. However performing representative spectroscopic follow-up to magnitudes of $i < 24$, in a large enough footprint to avoid cosmic variance, is certainly ambitious and perhaps even impractical. A more realistic alternative might be follow-up by narrow filter, multiband photometric surveys, which would certainly reduce the danger of incomplete calibration samples. However we might also run into the risk of misestimating the error in the derived redshifts that have been, after all, obtained using template fitting or Machine Learning techniques. These considerations should however not be seen as a statement against these programs, but against the overreliance on them. The problem of photometric redshift estimation will not be solved by a single method or strategy, but requires the combination of various approaches and follow-up programs to complement each other.

Alternative methods that cross correlate spectroscopic with photometric samples to derive

photometric redshift information (e.g. Newman, 2008; Ménard et al., 2013; Morrison et al., 2017) are already an integral part of modern photometric redshift strategies (Hildebrandt et al., 2017; Hoyle et al., 2017). As these methods do not require the spectroscopic reference sample to be representative of the photometric sample, they provide complementary information to template fitting and Machine Learning approaches. Especially if photometric redshift strategies begin to rely heavily on cross correlation estimates, we have to account for additional systematics and might even reconcile future survey designs. The shape of the photometric redshift distribution, as estimated by cross correlations, is degenerate with a redshift dependent galaxy-dark matter bias of both the reference sample and the photometric sample. As a result, it is advantageous for these methods, if photometric samples can be selected within thin redshift slices. This can be done, for instance, by traditional photometric redshift techniques using photometry of high enough quality in a sufficient number of bands. Besides relying on spatial overlap, cross correlation methods have to cover the full redshift range of the unknown sample. This is by no means a trivial assumption and the clustering redshift analysis in DES Y1 is currently unable to calibrate redshifts $z > 0.9$, because of the redshift limit in their reference sample (Hoyle et al., 2017). It is therefore vital to ensure that samples of suitable, high redshift sources, e.g. emission-line galaxies, quasars or luminous red galaxies, are available in sufficient density and spatial coverage, to optimize the effectiveness of these methods. Future spectroscopic follow-up programs dedicated to improve redshift validation for large area photometric surveys, therefore have to optimize their target selection not only to meet the demands of data based photometric redshift methodologies, but also to meet those of cross correlation methods.

The possible degeneracy of cross correlation redshifts with cosmological parameters might be avoided by including the respective cross correlation measurements into the joint data vector. This will also enable the consistent parametrization of various sources of systematic error, to facilitate self calibration of redshift systematics by cross correlations between neighboring photometric redshift bins and cosmological probes (e.g. Zhang et al., 2010; Benjamin et al., 2010; Zhang et al., 2016). In addition, redshift calibration based on shear ratios (Jain & Taylor, 2003; Taylor et al., 2007; Kitching et al., 2015; Schneider, 2016) can provide not only a diagnostic tool for shear systematics but also for photometric redshift error.

The field of photometric redshift estimation is clearly not lacking ideas and techniques that can complement the well established methods based on template fitting and Machine Learning. However combining all these approaches into a self consistent framework, will be one of the main challenges in the following years. State-of-the-art weak gravitational lensing and galaxy clustering analyses already parametrize the mean of the tomographic redshift distributions as free parameters (e.g. Bonnett et al., 2016; Troxel et al., 2017; DES Collaboration et al., 2017) and select the priors based on the result of photometric redshift validation. However extensive studies that investigate more flexible parametrizations, as well as the synergy between different photometric redshift methods are still somewhat lacking. Future work in this direction would develop a better understanding of the complementarity of different photometric redshift techniques and therefore prevent systematic biases in the cosmological analysis.

Additional Projects

During my PhD I contributed to several projects that have not been explicitly mentioned in the previous chapters. In the following, I would like to present a selection of these projects, where my complete publication list is attached at the end of this thesis.

7.1 Dark Energy Survey Year 1 Results: Redshift distributions of the weak lensing source galaxies

Motivation and Goal The Dark Energy Survey (DES) is a large area photometric survey in the southern sky that covers a total area of 5000 deg^2 . Its main science goal is to study dark energy and cosmic acceleration to high statistical precision, where we give a more detailed description of the survey and its science goals in §8. DES is one of the most powerful experiments for weak gravitational lensing today, due to its large area and deep photometry. The modeling of lensing observables however requires accurate photometric redshift estimates for the weak lensing source galaxies. In the paper Hoyle et al. (2017), we present the photometric redshift distributions and the validation results for the Dark Energy Survey Year 1 (DES Y1) weak lensing source galaxies, that are used in the DES Y1 science analyses (Troxel et al., 2017; Prat et al., 2017; DES Collaboration et al., 2017).

Analysis We use the template fitting photometric redshift code BPZ (Benítez, 2000) to derive the tomographic redshift distributions¹ within the redshift ranges (z -range) quoted in Tab. 7.1. However, as discussed throughout this thesis, these photometric redshift distributions are likely to be inaccurate. The DES Y1 analysis therefore parametrizes the redshift error in a tomographic photometric redshift bin $n^i(z)$ as a shift δz in its mean as $n^i(z) = n_{\text{photo}-z}^i(z - \Delta z^i)$ (Hoyle et al., 2017). We demonstrated in the previous section §5.3, that biases in the mean of a redshift distribution are the dominant source of error for the modelling of the lensing convergence power spectrum. In this paper (Hoyle et al., 2017) we further test this assumption and verify, that cosmological parameter shifts induced by differences in the shape of the redshift distribution, as opposed to shifts in the mean, are negligible for the DES Y1 analysis. Photometric redshift error estimates are therefore quoted as shifts in the mean of the tomographic

¹Tomographic photometric redshift distributions are redshift distributions for samples of galaxies that are binned using a photometric redshift point estimate into photometric redshift bins.

Table 7.1: Photometric redshift bias parameters Δz for the tomographic photometric redshift distributions estimated using BPZ. Adopted from Hoyle et al. (2017).

Value	Bin 1	Bin 2	Bin 3	Bin 4
z -range	0.20–0.43	0.43–0.63	0.63–0.90	0.90–1.30
COSMOS Δz^i	± 0.013	± 0.013	± 0.011	± 0.014
METACALIBRATION				
COSMOS final Δz^i	-0.006 ± 0.020	-0.014 ± 0.021	$+0.018 \pm 0.018$	-0.018 ± 0.022
WZ final Δz^i	$+0.007 \pm 0.026$	-0.023 ± 0.017	$+0.003 \pm 0.014$	—
Total final Δz^i	-0.001 ± 0.016	-0.019 ± 0.013	$+0.009 \pm 0.011$	-0.018 ± 0.022
IM3SHAPE				
COSMOS final Δz^i	$+0.001 \pm 0.020$	-0.014 ± 0.021	$+0.008 \pm 0.018$	-0.057 ± 0.022
WZ final Δz^i	$+0.008 \pm 0.026$	-0.031 ± 0.017	-0.010 ± 0.014	—
Total Δz^i	$+0.004 \pm 0.016$	-0.024 ± 0.013	-0.003 ± 0.011	-0.057 ± 0.022

photometric redshift distributions

$$\Delta z^i = \Delta_{\text{sys}}^i \pm \Delta_{\text{stat}}^i, \quad (7.1)$$

where Δ_{sys}^i is the systematic error contribution and Δ_{stat}^i is the statistical error of the i th tomographic photometric redshift bin. The subsequent cosmological analysis then marginalizes over these shifts.

We derive photo- z error estimates using both clustering redshift methods and data based photometric redshift validation using the COSMOS field. Both methods are complementary in their sources of systematic error and their data dependence. While the data based validation strategy uses the multiband photometric redshifts from COSMOS as a source of redshift validation, the cross-correlation methods employ high-precision photometric redshifts from red-sequence galaxies, obtained using the redMaGiC algorithm (Roza et al., 2016), as their reference sample². The complementarity of both the calibration data and the validation methodology, is a vital consistency check to derive meaningful photometric redshift errors.

In §5 we discussed three of the most important sources of photometric redshift error that are incorporated into the data based DES photometric redshift validation strategy: field-to-field variations in photometric noise, nonrepresentative calibration redshifts and inaccurate compression methods for individual galaxy photometric redshift distributions. An additional, important adverse effect, that was neglected in these previous discussions is cosmic variance. Since the accuracy of photometric redshifts varies spatially due to local variations in the galaxy density field, error estimates obtained in a small field, are not representative for the photo- z performance of the full survey. With an area of $\approx 1 \text{ deg}^2$, the COSMOS field is a relatively small validation field compared with the 1800 deg^2 large area of DES Y1. Cosmic variance therefore contributes significantly to the overall photometric redshift error budget and needs to be incorporated into the photo- z validation. This is done by extracting small COSMOS-like patches from the Buzzard simulations (DeRose et al. in prep.; Wechsler et

²The redMaGiC photo- z are used as a substitute for spectroscopic data, because there are currently no suitable spectroscopic surveys available that have sufficient spatial overlap with the Y1 footprint.

al. in prep.; MacCrann et al. in prep.) to measure the corresponding photometric redshift error on these subsamples. The distribution of photometric redshift error obtained on these simulation samples then quantifies the cosmic variance error budget. For a detailed discussion of the various error components, we refer the reader to Hoyle et al. (2017).

The DES photometric redshift validation pipeline evaluates the performance of a particular photometric redshift code for a given set of tomographic photometric redshift bins. The error estimates are quantified as offsets in the mean of the tomographic redshift distributions and incorporate all of the aforementioned sources of photometric redshift error, that are to be expected in the DES Y1 analyses. The algorithm is then applied to a set of resampled datafiles that resemble the photometric noise properties of the DES Y1 science sample. The dispersion of redshift performance evaluated on these datafiles is then added in quadrature to the other sources of error from e.g. cosmic variance or the intrinsic uncertainty of the COSMOS photometric redshifts. The final result of the pipeline is a set of offsets in the mean of the tomographic photometric redshift distributions, that is then combined with the clustering based photometric redshift validation performance estimates.

Clustering based redshift estimation (‘WZ’) cross-correlates a reference catalog with accurate redshifts with a spatially overlapping photometric redshift sample to derive photometric redshift information. These clustering redshift estimates can then be used as a complementary photometric redshift validation technique, alongside the data based approaches. However, clustering based photometric redshift estimation can also be subject to a variety of systematic biases. The clustering estimate depends on accurately accounting for a redshift dependent galaxy-dark matter bias in both the reference sample, as well as in the photometric sample. Furthermore, the intrinsic redshift error in the redMaGiC sample, which is $|\Delta z| < 0.01$ (Gatti et al., 2017), contributes as an additional source of error, that needs to be quantified, to avoid systematic biases in the derived cross correlation photometric redshift estimates. We refer to our paper (Gatti et al., 2017) for an in-depth study of these sources of error. Most importantly however we note, that the redMaGiC galaxies only cover a redshift range of $0.15 < z < 0.9$ (Rozo et al., 2016). Accordingly the last tomographic bin can only be validated using the data based methodology on the COSMOS field.

Results Fig. 7.1 shows the photometric redshift distributions of the 4 tomographic photometric redshift bins of the weak lensing source galaxy sample, with error estimates obtained from both photometric redshift validation strategies. We see, that the results obtained using the two redshift validation approaches differ significantly in shape for the first three tomographic photometric redshift bins. The large error contours around the ‘COSMOS’ estimate, that corresponds to the data based strategy in the COSMOS field, is a result of the aforementioned error contributions, i.e. cosmic variance, field-to-field variations in photometric noise and the intrinsic uncertainty of the COSMOS photo- z . The error bars on the clustering redshift estimates (WZ) account for its statistical shot noise. As noted earlier, we verified that differences in the shape of the tomographic photometric redshift distributions don’t contribute significantly to the overall error budget, at least at the level of accuracy required in the DES Y1 analysis. Tab. 7.1 summarizes the final error estimates in terms of shifts in the mean of the tomographic photometric redshift distributions Δz . The combined error from effects like e.g. cosmic variance or photometric calibration uncertainties, shown in the ‘COSMOS Δz^i ’ row, are corrected for correlations between tomographic redshift bins to yield the statistical error estimate Δ_{stat} (see Eq. 7.1), in the ‘COSMOS final Δz^i ’ row. The systematic error Δ_{sys}

(see Eq. 7.1) is obtained by validating the BPZ photometric redshift distributions against the COSMOS validation data. We then combine this result with the calibration from the clustering redshift estimate, quoted in the ‘WZ final Δz^i ’ row. Both of these results directly depend on the BPZ photometric redshift estimates and therefore on the photometry. We therefore consider the two photometric reductions used in the DES Y1 analysis, ‘METACALIBRATION’ and ‘IM3SHAPE’ separately, as explained in detail in Rozo et al. (2016). For each separate photometry, we compute the total error, quoted in the ‘Total Δz^i ’ row, by combining the error from the clustering redshift method and the data based COSMOS photometric redshift validation pipeline, as independent measurements³.

Contributions and Acknowledgements During my work in the DES photo- z group, I contributed to the development of the various parts of the photo- z validation pipeline. Most notably I was involved in the early study of field-to-field variations in photometric noise and investigated sample selection biases using simulation data. The goal was to estimate how many galaxies with inaccurately validated photometric redshifts need to be removed to ensure a good enough overall redshift accuracy. This work was one of the main motivations to rely on multiband photometric redshifts as the main source of photometric redshift calibration. I further provided photometric redshift estimates to the DES collaboration and significantly contributed to the cross correlation redshift estimation efforts, that contribute to this paper. Specifically I prepared a blind-challenge to investigate the accuracy of cross correlation methods for photometric redshift estimation. The preparation of this dataset involved estimating photometric redshifts on the simulation data, performing selection cuts, applying systematics masks, as well as performing consistency checks by measuring correlation functions. These efforts contributed to another paper Gatti et al. (2017), on which I am one of the leading coauthors.

7.2 Anomaly detection for machine learning redshifts applied to SDSS galaxies

Motivation and Goal In this thesis we generally assumed that the spectroscopic redshift (spec- z) measurements have a negligible error. This is in general a sound assumption in comparison with the much larger photometric redshift errors to be expected from a broad band photometric survey like DES. However, especially if we consider low resolution spectroscopy from surveys like PRIMUS (Coil et al., 2011; Cool et al., 2013), the spectroscopic redshift error can be significant. Cool et al. (2013) compared the highest quality PRIMUS redshifts (quality flag $Q = 4$) with more accurate data from higher resolution experiments like COSMOS (Lilly et al., 2007b), DEEP2 (Newman et al., 2013), and VVDS (Le Fèvre et al., 2005), finding that 8% deviated by more than 5σ from these high precision spec- z . Since the low resolution PRIMUS spectroscopy cannot resolve individual spectral lines, redshift estimation is based on fitting spectral template models to the observed spectrum, similar to the template fitting photo- z approach. As a result, PRIMUS redshift estimates can be subject to degeneracies between spectral templates, which contributes to the redshift error. However even spectra from high resolution instruments that are redshifted by humans can be subject to errors, and can

³We obtain the total error σ_{tot} from the independent measurements σ_i as $\sigma_{\text{tot}} = \sqrt{1/\sum_i \sigma_i^{-2}}$ (see e.g. Ku, 1966).

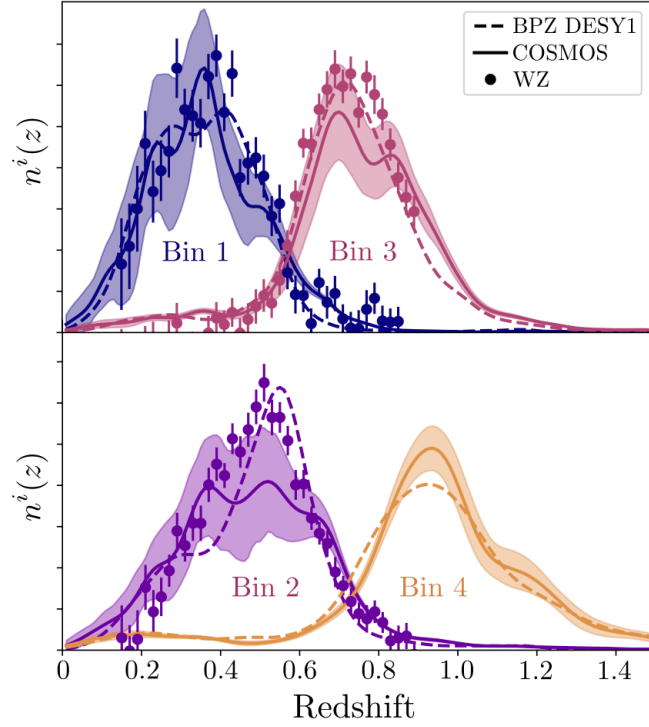


Figure 7.1: Tomographic photometric redshift distributions used in the Y1 analysis. The dashed lines show the photometric redshift distributions of the WL source bins obtained using the BPZ template fitting code. The error contours show the results from the photometric redshift validation pipeline, that incorporates various sources of error as explained in the text. The points with corresponding errorbars denote the result of the clustering redshift method within a redshift range of $0.15 < z < 0.9$, where the errorbars correspond to the statistical noise in the clustering redshift estimate. Outside the restricted redshift range we chose an arbitrary normalization for the clustering redshift distributions. Image credit: (Hoyle et al., 2017)

have a significant failure rate, especially at the faint end of the color magnitude space (Hartley, et al. in prep.). This can lead to highly incomplete spectroscopic samples, thereby introducing punishing selection effects in photometric redshift estimation and validation (Hartley, et al. in prep.). Particularly, Machine Learning derived photometric redshift estimates that use these samples as training sets, can be severely affected by spectroscopic redshift errors. In Hoyle et al. (2015b) we therefore investigate the usage of outlier detection methods to remove faulty spec- z from the training and validation samples. These methods identify galaxies with spectroscopic redshift values that deviate strongly from the typical spec- z values of the bulk of galaxies with similar photometry, and removes them from the training sample.

The goal of this work is to study the effectiveness of outlier detection methods to remove these faulty spectroscopic redshift measurements. We further investigate the photometric redshift accuracy of Machine Learning methods, so called robust regression methods, that are less biased by contaminated training samples.

Analysis We demonstrate the effectiveness of our approach using spectroscopic data from the Sloan Digital Sky survey (SDSS) III data release 12 (Alam et al., 2015), that contains photometric and spectroscopic observations for 2.5M galaxies. A subsample of 9115 galaxies has been spectroscopically observed multiple times, as the initial observations produced faulty redshift measurements. These galaxies therefore have both SDSS photometry, as well as inaccurate z_{error} and accurate spectroscopic redshift measurements z_{true} from the initial and follow-up observations. We select all unique galaxies with a difference between accurate z_{true} and faulty z_{error} spectroscopic redshifts of $|z_{\text{true}} - z_{\text{error}}| > 0.01$. The selected galaxies then provide a suitable dataset to study the effect of inaccurate spectroscopic measurements on photometric redshift accuracy.

We contaminate a clean training sample, that contains galaxies with a small spectroscopic redshift error, with inaccurate spectroscopic redshift measurements, which are subsequently identified and removed with an outlier detection algorithm. In our analysis we use the elliptical envelope method (Rousseeuw & Driessen, 1998), as implemented in the scikit-learn package (Pedregosa et al., 2011), and vary a hyperparameter, defined as contamination fraction n_c , that introduces the expected fraction of galaxies with a large spectroscopic redshift error. The elliptical envelope is then constructed such, that approximately this fraction of objects is removed from the dataset. In this process, the algorithm prefers likely outliers, i.e. galaxies that have very untypical spec- z values given their photometry.

Since we know, which galaxies have inaccurate spectroscopic redshift measurements, we can investigate how well the outlier detection algorithm performs. For this, we train a Machine Learning photometric redshift code on a dataset that is augmented with a fraction of inaccurate spectroscopic redshifts and a dataset that was cleaned by the algorithm. The relative performance improvement due to the cleaning process then quantifies the accuracy of our approach. We note, that the model performance is always evaluated on an independent test set, with well known spectroscopic redshifts. If the training set however contains a significant fraction of inaccurate redshift measurements, the model will ‘learn’ these errors and produce suboptimal photometric redshift predictions. To evaluate the performance of our model, we define the redshift scaled residuals as $\Delta_{z'} = (z_{\text{phot}} - z_{\text{spec}}) / (1 + z_{\text{spec}})$, where z_{phot} denotes the photometric redshift estimate and z_{spec} the spectroscopic redshift. The photometric redshift performance metrics used in this work are the median of the scaled residual distribution $|\mu|$ and σ_{68} , σ_{95} , that denote the 68%, 95% spread around $|\mu|$. The outlier rate is defined as the fraction of objects at the tails of the scaled residual distribution $|\Delta_{z'}| > 0.15$. We compare the performance of two Machine Learning algorithms: median regression and mean regression. Mean regression is the classical regressor, that predicts the mean of the predictive distribution, whereas median regression predicts its median. The median however can be expected to be much more robust against outlier in the data. For both schemes we use Gradient boosted trees (Friedman, 1999, 2001) implemented in the Scikit-learn software package (Pedregosa et al., 2011), a Machine Learning method, that is also based on the concept of the decision tree, that was introduced in §2.5.

Results We generate 250 datasets with accurate spectroscopic redshifts and add a random fraction of inaccurate spectroscopic redshift measurements. We then apply the elliptical envelope technique with various contamination fraction hyperparameters n_c and apply both the median and the mean regression techniques to this training dataset. We evaluate the performance of these methods relative to the original estimates obtained without outlier removal in

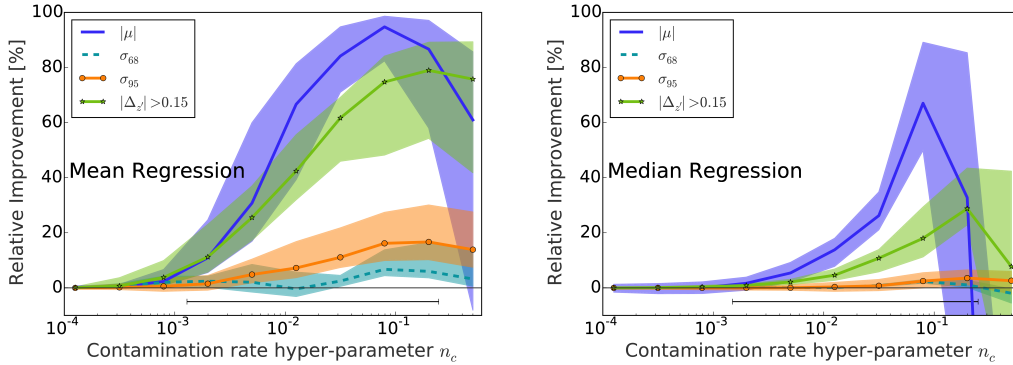


Figure 7.2: Relative improvement in photometric redshift performance after outlier removal as a function of the contamination rate hyper-parameter n_c of the Elliptical Envelope method. The contours correspond to the results obtained on 250 datasets, where each dataset has a different fraction of contaminated galaxies as drawn from the range of values shown by the black horizontal line. For a small n_c we therefore keep a large number of contaminated galaxies in the sample, whereas a large n_c culls a significant amount of data. The left/right panel shows the performance for the mean/median regression method as described in the text. Image credit: (Hoyle et al., 2015b)

Fig. 7.1. We note that a higher contamination rate parameter forces the elliptical envelope to remove more data and the performance therefore increases for all metrics. If too much data is removed, the performance of the model decreases again, due to the higher statistical noise. We further note, that median regression is much more robust against the contamination fraction and the corresponding performance curve begins to rise much later. This implies, that one has to remove a very significant fraction of the data, before the median regression algorithm shows better performance. Accordingly, median regression is much more robust against outliers in the data as compared with mean regression. It can therefore be advantageous to use median regression instead of the more traditional mean regression for photometric redshift estimation, if one expects a large fraction of faulty spectroscopic redshift measurements in the training data.

Contributions and Acknowledgements In this project I contributed the idea of using robust regression methods, more specifically median regression, for photometric redshift estimation, if the spectroscopic training sample contains a significant fraction of spectroscopic failures. I demonstrated the effectiveness of this approach and helped the first author (Ben Hoyle) to include this analysis into the paper. I also contributed by commenting and proof reading the text.

7.3 Utility Optimization in Regression Tasks

Motivation and Goal Point predictions like the conditional mean are not always the best photometric redshift estimates in all settings. Specifically for the modelling of angular correlation power spectra and cluster mass estimates, we saw that the Highest Weight Element

point prediction produces more accurate models than more traditional estimates like the conditional mean. This is not a fundamental weakness of the traditional regression approach, but rather a result of the different optimization directive that needs to be fulfilled in these science cases. More specifically we saw in §1.3, that the goal of photometric redshift estimation is to facilitate the accurate modeling of cosmological observables, and there is no a priori reason to expect this to perfectly coincide with the optimization of the sum-of-squared error in regression. Similar problems appear quite often in regression applications, especially if data is sparse, or there exists a large uncertainty in the input variables. As a fictive example consider predicting the impact time of a hurricane or a water wave: it will be much better to predict an earlier arrival time than one that is too late.⁴

If the optimization directive is more complex, we need to take into account the ‘utility’ and the ‘cost’ of a prediction, rather than just minimizing the sum-of-squared error. Utility based regression is a novel Machine Learning technique that incorporates this problem dependent optimization directive into the prediction task by convolving the predictive distribution with a utility function, that describes the utility-cost tradeoff in regression. It can be seen as the generalization of the well known concept of cost-sensitive learning in classification (Domingos, 1999; Elkan, 2001; O’Brien et al., 2008) to regression. However besides the early works by Torgo & Ribeiro (2007); Ribeiro (2011), very little has been done on refining utility based regression. In our paper Branco et al. (2017) we study the usage of conditional density estimation as a powerful tool for utility based learning.

Analysis The basis of utility based learning is the utility function $U(t, \hat{t})$ that attributes to each target value t a number between $[-1, 1]$. Negative values represent a cost, positive values a benefit, or utility. This means that the algorithm will give more weight to training samples, that have a higher utility than to those that have a higher cost. The optimal prediction \hat{t}_{\max} then maximizes the predictive distribution $p(t|\mathbf{x})$ convolved with the utility function as

$$\hat{t}_{\max} = \operatorname{argmax}_{\hat{t}} \int p(t|\mathbf{x}) U(t, \hat{t}) dt. \quad (7.2)$$

While there exists an infinite number of utility functions, not all functional forms can be considered reasonable. Specifically we require that every utility function (Torgo & Ribeiro, 2007; Ribeiro, 2011)

$$U(t, \hat{t}) = \mathcal{B}(\hat{t}, t) - \mathcal{C}(\hat{t}, t) \quad (7.3)$$

should consist of a term $\mathcal{B}(\hat{t}, t)$ that parametrizes the benefit of a certain prediction, and a cost function $\mathcal{C}_p(\hat{t}, t)$ that will depend on a tuning parameter $p \in [0, 1]$, indicating how much certain target values are penalized. A detailed description and an algorithm to obtain meaningful utility functions is presented in Ribeiro (2011). As a conditional density estimator we use the classification based method described in §3, where we evaluate the quality of the utility based regression estimate using the Normalized Mean Utility (NMU) metric (Ribeiro, 2011)

$$\text{NMU} = \frac{\sum_{i=1}^N U(t_i, \hat{t}_i) + N}{2N} \quad (7.4)$$

⁴This is very similar to the problem of imbalanced datasets in the context of the evaluation of classifiers (§2.3). The under representation of a certain class in the training set, can introduce a bias towards preferentially predicting the majority class.

where \hat{t} is the model prediction and t is the true response value in the test set that contains N objects.

Results We apply the utility based regression framework to a large variety of datasets and utility surfaces using the Random Forest (§2.5.3) and the Support Vector Machine algorithms⁵ as a classifier. We could show that our methodology significantly improves over the traditional regression estimate for a variety of datasets and utility functions. We refer to Branco et al. (2017) for further details. However our analyses also showed that the benefit in applying utility based regression strongly depends on the dataset and the choice of utility surface. While these studies have been performed in a general Machine Learning context, utility based regression can be readily applied to photometric redshift estimation. As mentioned previously, different cosmological probes are not necessarily sensitive to the same types of photometric redshift error. Utility based regression could therefore provide a way to optimize the prediction towards a certain science goal or provide a framework to compensate for potential selection effects. We leave these problems for future work.

Contributions and Acknowledgements An important aspect of the proposed method of utility based regression is the estimation of conditional distributions $p(t|\mathbf{x})$ of the response t given the input variables \mathbf{x} . I contributed to this project by providing source code and explanatory material for the conditional density estimation methods described in §3. I also contributed to the text. I want to highlight that the conditional density estimation routine implemented in the UBL package⁶ (Branco et al., 2016), is based on my original implementation and was modified and adapted by Paula Branco.

⁵We refer to the literature (Suykens & Vandewalle, 1999; Smola & Schölkopf, 2004; Bishop, 2006; Hastie et al., 2009) for a detailed description of the Support Vector Machine.

⁶<https://cran.r-project.org/web/packages/UBL/UBL.pdf>

Appendix

The following sections give a brief overview over two large area photometric surveys: the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS) survey and the Dark Energy survey (DES). This thesis uses data from CFHTLenS in §3 and DES data in §5. A general description of the CFHTLenS survey, on which the following discussion is based, can be found in Erben et al. (2013); Hildebrandt & CFHTLenS collaboration (2014); CFHTLenS-collaboration (2017); CFHTLS-collaboration (2017a). By the time of this writing, the DES collaboration completed the analysis of the first year (Y1) DES observations, that will be described in §8.2. The description of the DES survey, its current status and future prospects is based on The Dark Energy Survey Collaboration (2005); Dark Energy Survey Collaboration et al. (2016); Drlica-Wagner et al. (2017).

8.1 The CFHTLenS survey

Survey characteristics and systematics CFHTLenS is a 154 deg^2 large photometric survey (see e.g. Erben et al., 2013), that provides deep imaging information up to a magnitude of $i' \sim 24.5$, in the 5 photometric filters shown in Fig. 8.2. The survey uses the MegaCam instrument (CFHTLS-collaboration, 2017b), that has a $1 \times 1 \text{ deg}^2$ field-of-view and with a resolution of $0.187 \text{ arcsec}^2/\text{pixel}$. MegaCam is mounted at Mount Kea and has a median seeing of 0.7 arcsec , where the i' band has the best seeing conditions and is therefore used as the main detection band (Brimioulle, 2013; Hildebrandt & CFHTLenS collaboration, 2014). Tab. 8.1 summarizes important properties of the CFHTLenS photometric data (see also Erben et al., 2013) including the $5\text{-}\sigma$ depth in the respective photometric filter bands. We note that the i' filter broke during the observations and was replaced by the y' . While a total of 33 fields only have imaging information in y' , this description only considers the i' band, since we do not use y' band photometry in this thesis. The large area of the survey in combination with its deep photometry and excellent seeing conditions make CFHTLenS perfectly suited for weak gravitational lensing analyses. One of the most important systematics for the CFHTLenS, as for all imaging surveys, is the estimation of galaxy shapes and photometric redshifts, which requires accurate photometry. An important systematic in this context are inhomogeneities in the point spread function (PSF).

When the light from an e.g. point like source is processed by an optical instrument, its shape is ‘blurred out’ and distorted by the convolution with the PSF of the imaging system. A

Table 8.1: Basic properties of the CFHTLenS science data. Taken from Erben et al. (2013).

Filter	expos. time [s] in a $2''.0$ aperture	m_{lim} [AB mag] 5- σ lim. mag.	seeing [$''$]
u^*	5×600 (3000)	25.24 ± 0.17	0.88 ± 0.11
g'	5×500 (2500)	25.58 ± 0.15	0.82 ± 0.10
r'	4×500 (2000)	24.88 ± 0.16	0.72 ± 0.09
i'	7×615 (4305)	24.54 ± 0.19	0.68 ± 0.11
y'	7×615 (4305)	24.71 ± 0.13	0.62 ± 0.09
z'	6×600 (3600)	23.46 ± 0.20	0.70 ± 0.12

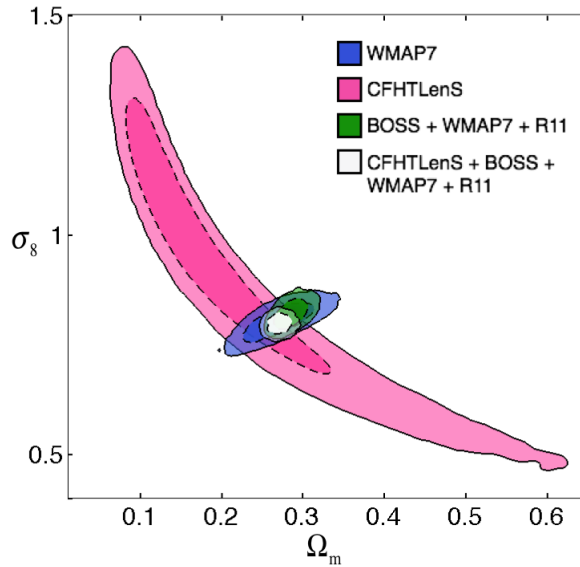


Figure 8.1: Constraint on the power spectrum amplitude σ_8 and Ω_m for a flat Λ -CDM model, obtained from the tomographic cosmic shear analysis presented in Heymans et al. (2013). The blue, green and white contours show the constraints incorporating external datasets as explained in the text. Image credit: Heymans et al. (2013)

varying PSF between photometric bands and between different seeing conditions can therefore lead to inaccurate galaxy colors (cf. Hildebrandt & CFHTLenS collaboration, 2014; Brimiouille, 2013). Similarly, if the PSF is elliptical, weak gravitational lensing measurements will be biased by this systematic distortion of galaxy shapes¹. It is therefore essential to correct the photometry and shape measurements for this important systematic. The CFHTLenS collaboration uses the method of Kuijken (2008) to correct the PSF in the different bands for a single pointing, such that it is gaussian and of equal size (cf. Hildebrandt & CFHTLenS

¹The ellipticity of a galaxy is quantified by the components e_1 and e_2 that quantify the shape of an ellipse parametrized by the axes a and b , where $b < a$. The total ellipticity is then given as $|e| = \sqrt{e_1^2 + e_2^2} = \frac{a-b}{a+b}$ (cf. Bartelmann & Schneider, 2001).

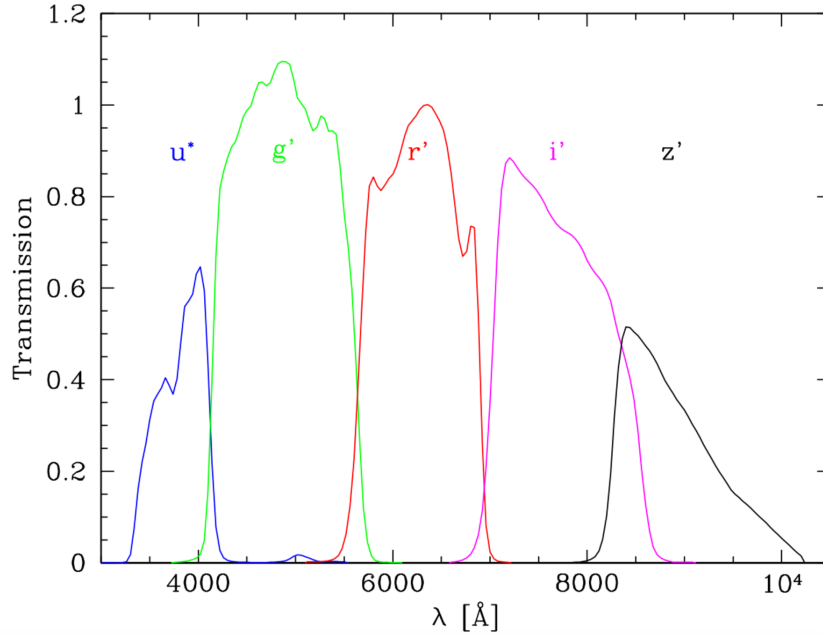


Figure 8.2: Filter bands of the CFHTLenS survey. Image credit: Brimiouille (2013)

collaboration, 2014; Brimiouille, 2013). To correct for elliptical PSFs one uses stellar objects to estimate the PSF ellipticity. The galaxy shapes can then be corrected by this PSF pattern (Brimiouille, 2013). The upper left panel of Fig. 8.3 shows the PSF anisotropy pattern for a particular pointing (m1m2 in the W1 field). The upper left panel shows the orientation of the PSF anisotropy pattern before the shape correction, where the corresponding values for e_1 and e_2 are quoted in the corresponding lower left panel. We clearly see a correlation of the ellipticity coefficients that will bias any shape measurements, if left uncorrected. The right panels show the corresponding pattern after a PSF correction has been applied. We see that the anisotropy has been significantly reduced and the correlation between the ellipticity parameters is greatly improved (Brimiouille, 2013).

Correcting for the biasing effect of the PSF is not always possible and additional consistency checks have to be performed to detect fields with ill controlled PSF. The CFHTLenS collaboration therefore rejected 25% of the fields with a bad PSF, where we refer to Hildebrandt & CFHTLenS collaboration (2014) for a more indepth discussion on the shape measurement procedure used for the CFHTLenS analyses.

Science Goals The CFHTLenS survey is primarily focussed on weak gravitational lensing as its main cosmological probe. As discussed in §1.3.3 weak gravitational lensing is a powerful probe of the expansion of the universe and the growth of structure. Especially the tomographic cosmic shear analyses of Heymans et al. (2013) provided accurate constraints of important cosmological parameters like the amplitude of the linear matter power spectrum² at $8h^{-1}\text{Mpc}$. Specifically we see that the constraints in the $\Omega_m - \sigma_8$ plane for weak gravitational lensing are complementary, i.e. orthogonal, to those of other surveys that measure the CMB (WMAP7,

²Like A_s which we introduced in Eq. 1.65, σ_8 is a parameter that describes structure growth. In a simplified interpretation it can be seen as the analog to A_s at late times. (see e.g. Peacock, 1999)

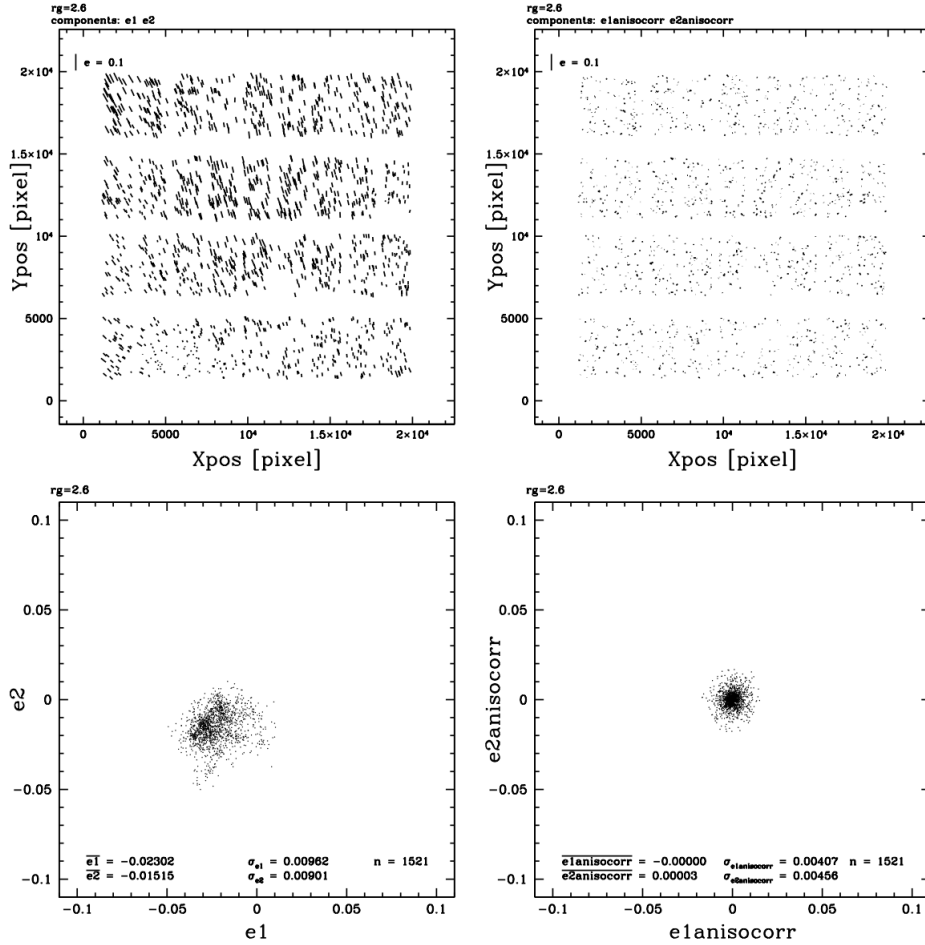


Figure 8.3: Illustration of PSF anisotropy for the case of the m1m2 pointing in the W1 field. The respective upper left and lower left panel show the PSF anisotropy pattern and its corresponding ellipticity values before applying a correction. The upper and lower right panels show the corrected PSF anisotropy. Image credit: Brimioulle (2013)

Larson et al. 2011), the clustering of galaxies (BOSS, Anderson et al. 2012) and supernovae (R11, Riess et al. 2011). This showcases the great importance of including constraints from weak gravitational lensing into a multiprobe analysis. Assuming a flat cosmology, Heymans et al. (2013) was able to constrain the dark energy equation of state parameter $w_0 = -1.02 \pm 0.09$, by combining CFHTLenS constraints with the aforementioned external probes.

Cosmic shear measurements are subject to several sources of systematic uncertainty, most importantly uncertainties in the photometric redshift distributions, as discussed in great detail in this thesis, and systematics in the shear measurement from e.g. PSF anisotropies. An important science goal of CFHTLenS was therefore the development of methods to control these sources of systematic error precisely enough to avoid major systematics in the cosmological measurements.

For example, Heymans et al. (2012) developed an important technique to control system-

atics in the shape measurement pipeline: the method cross correlates corrected galaxy shapes with the uncorrected stellar shapes. In the absence of residual shape systematics, this correlation should be zero, where the significance of a nonzero signal requires calibration using simulations (see e.g. Hildebrandt & CFHTLenS collaboration, 2014). We refer to Heymans et al. (2012) for further details.

Besides errors in the measured shear, photometric redshift uncertainty is another important systematic in the CFHTLenS survey. To ensure high quality photometric redshift distributions, the CFHTLenS collaboration worked extensively on improving both photometry and photometric redshift methodologies (Hildebrandt et al., 2012). Most notably CFHTLenS was one of the main motivations to develop methods that self calibrate photometric redshift distributions using angular cross correlations between neighboring tomographic photometric redshift bins (Benjamin et al., 2013).

The statistical power of CFHTLenS even allowed the study of alternatives to General Relativity in the form of modified gravity models, showcasing the great potential of cosmic shear as a tool for cosmology. In combination with redshift space distortion measurements³ using spectroscopic data from WiggleZ (Drinkwater et al., 2010) and 6dFGS (Jones et al., 2009), CFHTLenS could exclude a variety of alternative models to GR (Simpson et al., 2013). One of the most exciting properties of gravitational lensing is its sensitivity to the full dark matter structure of the universe, which makes it a potent tool to generate maps of the dark matter distribution (Van Waerbeke et al., 2013). These dark matter maps provide exciting opportunities to study the relation between luminous and dark matter and facilitate interesting analyses from their cross correlation with other probes (Van Waerbeke et al., 2014).

The scientific applications described in this paragraph only cover a fraction of the potential scientific usecases for this dataset. In fact, the CFHTLenS survey is itself part of the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS), that additionally consists of the Super Novae Legacy Survey (SNLS) and the CFHTL Deep programs, that obtain deep photometry up to a magnitude of $r' = 28$ (CFHTLS-collaboration, 2017a) in two selected fields. Furthermore CFHTLS also has a specialized ‘very wide’ program (CFHTLS-VW), that covers 410 deg^2 in 3 photometric bands (g' , r' , i') (CFHTLS-collaboration, 2017a). SNLS is mainly targeted towards supernovae and transient science, whereas CFHTLS-VW facilitates solar system science.

8.2 The Dark Energy Survey

Survey characteristics and systematics The Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005) is an ongoing large area imaging survey that covers a total area of 5000 deg^2 in the southern sky. DES observations are performed in five years and are expected to produce accurate photometry for 300 million galaxies, as well as accurate shape measurements for 200 million. The galaxies are observed in the 5 photometric filter bands g , r , i , z and Y shown in Fig. 8.6 with the Dark Energy Camera (DECAM; Diehl & Dark Energy Survey Collaboration 2012; Flaugher et al. 2015) that is installed on the Blanco-4m telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile. In addition to its Wide field program, DES facilitates supernovae science with deep observations in g , r , i , z in

³The peculiar velocity of galaxies causes a doppler shift that ‘alongates’ galaxies in redshift space. These redshift space distortions can be used as a cosmological probe. We refer to Percival et al. (2011) for a detailed explanation.

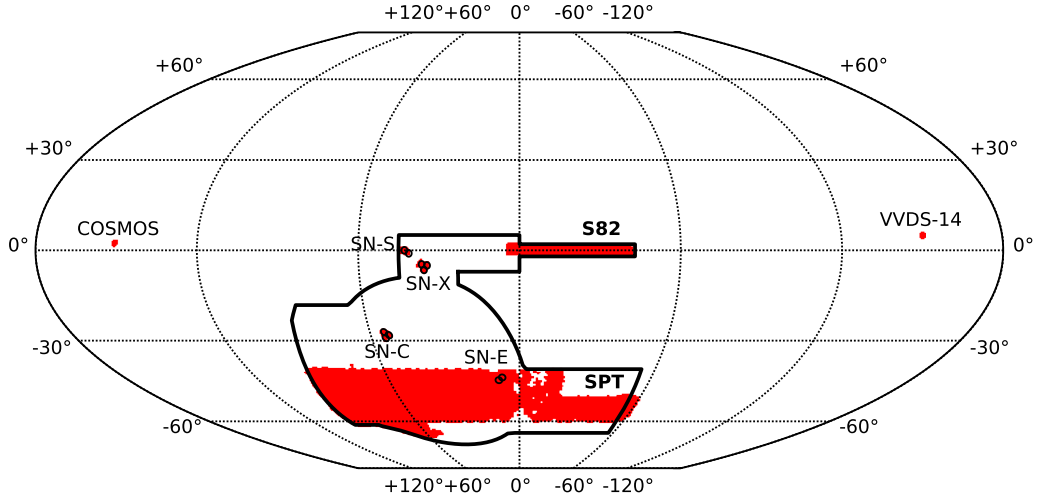


Figure 8.4: Footprint of the DES Y1 sample. We show the current science samples in the Stripe 82 (S82) and SPT regions, as well as the supernovae (SN) fields and the two auxiliary calibration fields VVDS-14h and COSMOS. The footprint of the final Y5 observations is indicated by the solid black contour. Image credit: Drlica-Wagner et al. (2017)

Table 8.2: Properties of the Y1A1 GOLD dataset. We note that the magnitude limits m_{lim} and their respective error intervals are obtained using the mode and the 10% and 84% percentile of the magnitude limit distribution, where using the median instead reduces this value by ~ 0.05 mag (see Drlica-Wagner et al., 2017). Taken from Drlica-Wagner et al. (2017).

Parameter	g	r	i	z	Y
Median PSF FWHM	1.25''	1.07''	0.97''	0.89''	1.07''
All Band Sky Coverage	1786 deg ²				1773 deg ²
Absolute Photometric Error (mmag)	14	4	2	15	32
Completeness Limit (95%)	23.6	23.4	22.9	22.4	
Coadd Galaxy m_{lim} (10σ)	$23.4^{+0.14}_{-0.40}$	$23.2^{+0.13}_{-0.37}$	$22.5^{+0.14}_{-0.34}$	$21.8^{+0.12}_{-0.37}$	$20.1^{+0.18}_{-0.33}$
Multi-Epoch Galaxy m_{lim} (10σ)	$23.7^{+0.07}_{-0.40}$	$23.5^{+0.16}_{-0.29}$	$22.9^{+0.14}_{-0.30}$	$22.2^{+0.14}_{-0.32}$	
Galaxy Selection ($i \leq 22$)	Efficiency > 98%; Contamination < 3%				
Stellar Selection ($i \leq 22$)	Efficiency > 86%; Contamination < 6%				

10 smaller fields.

The survey saw first light in September 2012 and started its first observation circle from Nov 2012 - Feb. 2013, to obtain a representative dataset for science verification (Dark Energy Survey Collaboration et al., 2016). For this purpose, observations were concentrated on a smaller field, with comparable photometric depth to the completed survey. In addition, two smaller fields, that overlap with the ‘COSMOS’ (Scoville et al., 2007) and ‘VVDS-14h’ (Le Fèvre et al., 2005) surveys were observed, to provide deep spectroscopic calibration data for photometric redshift estimation and star-galaxy classification algorithms. At the time of this

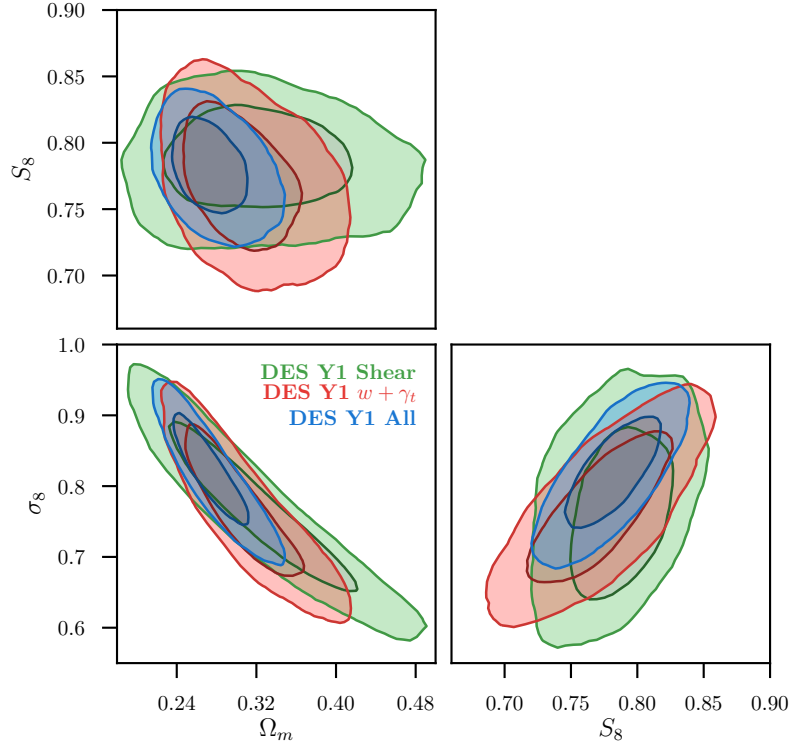


Figure 8.5: Cosmological parameter constraints of the DES Y1 combined analyses of weak gravitational lensing and galaxy clustering. The green ellipse shows the result obtained using shear alone, the red contours show the constraint from galaxy-galaxy lensing and galaxy clustering and the blue ellipse their combination. Image credit: DES Collaboration et al. (2017)

writing, the DES collaboration finished the main analyses of the first year data (DES Y1), that was taken from August 2013 - Feb. 2014. In the following paragraph, I will give a brief overview over this dataset and summarize the main science interests of DES Y1.

Fig. 8.4 shows the footprint of the DES Y1 observations, that covers a total area of 1839 deg^2 . The main science fields for galaxy clustering and weak gravitational lensing are the Stripe 82 (S82) field at the celestial equator and a larger region in the south, that overlaps with the footprint of the South Pole telescope (SPT)⁴ (Ruhl et al., 2004; Carlstrom et al., 2011; Austermann et al., 2012). Tab. 8.2 summarizes the basic properties of the DES Y1 science sample (Y1A1 GOLD). We note that the observations in the Y band have a lower depth than observations in the other filters. Furthermore Y band observations have an incomplete coverage of the full Y1 area. We therefore currently do not include them into the analysis, especially due to the high overlap of the Y band with the z band (cf. Hoyle et al., 2017). We note that DES provides both coadded photometry as well as multi-epoch photometry (see e.g. Drlica-Wagner et al., 2017), where we use the latter in the analyses described in §5.

Comparing with the CFHTLenS dataset described earlier, DES Y1 already has an area that is larger by a factor of 10. DES Y1 is therefore, alongside KiDS (de Jong et al., 2013),

⁴The South Pole telescope is a 10 meter radio telescope operating in the millimeter-submillimeter spectral range to facilitate galaxy cluster science and the study of CMB anisotropies.

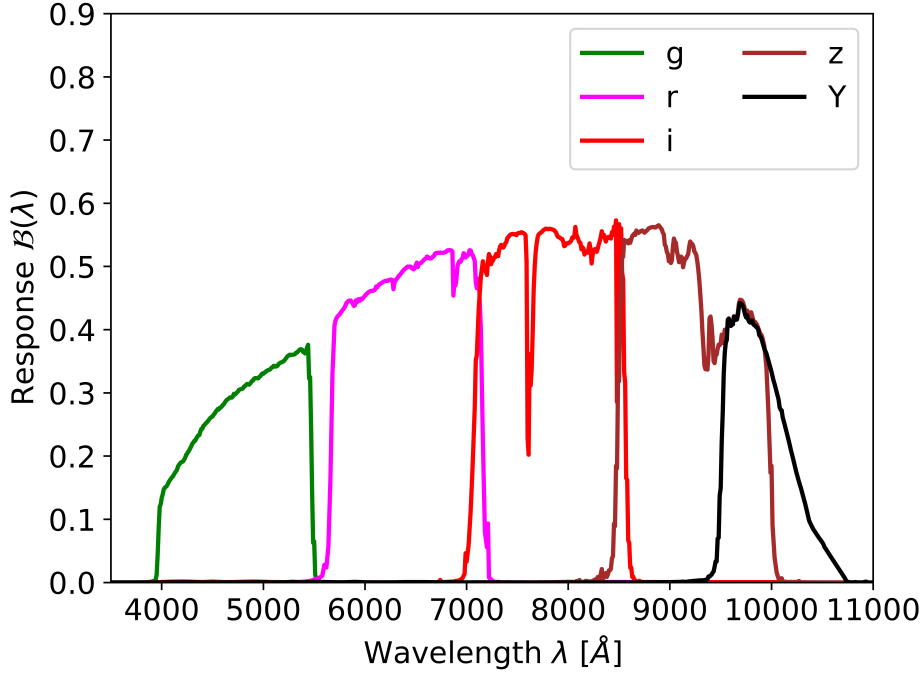


Figure 8.6: Filter functions of the Dark Energy Survey.

one of the most powerful large area photometric surveys to date. In the coming years, DES will increase its survey area to 5000 deg^2 and its photometric depth by $\sim 2 \text{ mag}$ (The Dark Energy Survey Collaboration, 2005).

Science Goals The main science goal of DES is the study of the dark energy equation of state using galaxy clusters, weak gravitational lensing, galaxy clustering and supernovae measurements. It is therefore quite similar to the program pursued by the CFHTLenS survey, however using a dataset that facilitates much better cosmological parameter constraints.

While weak gravitational lensing was the main cosmological probe used in the context of the CFHTLenS survey, DES performs joint measurements of cosmic shear and angular clustering, constraining the accelerated expansion of the universe and the growth of structure to unprecedented precision (DES Collaboration et al., 2017). We illustrate this in Fig. 8.5 that shows the cosmological parameter constraints obtained using a joint analyses of cosmic shear, galaxy-galaxy lensing⁵ and angular clustering for a flat Λ -CDM model. Comparing with Fig. 8.1, we note that the statistical power of DES Y1 exceeds the results of CFHTLenS by a factor of 4 in σ_8 and a factor of 3 in Ω_m (DES Collaboration et al., 2017). The DES Y1 analyses alone, i.e. without using external data, is already able to constrain $w = -0.8^{+0.2}_{-0.22}$ (DES Collaboration et al., 2017). In the ongoing Y3 analysis DES will increase its area by a factor of 3 to greater photometric depth and include additional cosmological probes like supernovae and galaxy cluster constraints into the cosmological data vector, which will lead to significant improvements in the constraining power (DES Collaboration et al., 2017).

⁵Galaxy-galaxy lensing uses the lensing effect from background galaxies on individual foreground galaxies. We refer to Brimiouille (2013) for an in depth introduction.

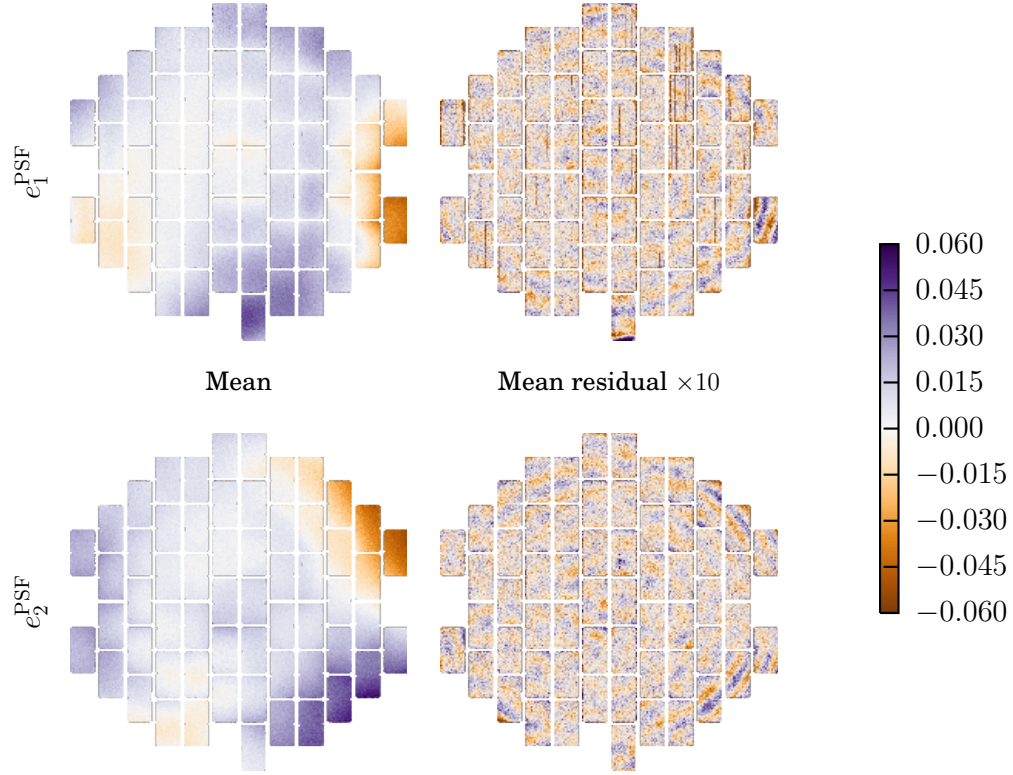


Figure 8.7: Left: Mean ellipticity of the PSF binned by focal plane position. Right: Residual ellipticity after correcting using PSFEx (Bertin, 2011). The residuals are inflated by a factor of 10 for visualization purposes. Image credit: Zuntz et al. (2017)

The significant increase in statistical precision also sets high requirements on the control of systematic uncertainties in the analysis. As for the CFHTLenS survey, DES also needs to accurately control shape and photometric redshift measurements to high accuracy. The left panel of Fig. 8.7 shows the pattern of the mean PSF ellipticity at different focal positions. The corresponding right side shows the residual pattern after the PSF correction. While these residuals are in general quite small, they still have a significant pattern that needs to be corrected in the analysis (see e.g. Zuntz et al., 2017). Another important systematic in the Dark Energy Survey is photometric redshift uncertainty, where I refer to the detailed discussions in §5 and §7.1.

Besides providing better constraints on cosmological parameters like the dark energy equation of state, DES can be expected to provide unprecedented constraints on modified gravity models, especially in its final year (Kirk et al., 2015). Furthermore DES provides the largest weak lensing mass maps to date (Chang et al., 2017) and also has a dedicated supernovae survey that will significantly contribute to the cosmological analyses in DES Y3 (Dark Energy Survey Collaboration et al., 2016).

Bibliography

- Abbott T., Abdalla F. B., Allam S., et al., 2016, Cosmology from cosmic shear with Dark Energy Survey Science Verification data, PRD, 94, 2, 022001
- Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies, MNRAS, 417, 1891
- Abdelsalam H. M., Saha P., Williams L. L. R., 1998, Nonparametric Reconstruction of Abell 2218 from Combined Weak and Strong Lensing, AJ, 116, 1541
- Alam S., Albareti F. D., Allende Prieto C., et al., 2015, The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III, ApJS, 219, 12
- Alam S., Ata M., Bailey S., et al., 2017, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample, Monthly Notices of the Royal Astronomical Society, 470, 3, 2617
- Alborno S. S., 2016, A tomographic approach to the statistical analysis of the large-scale structure of the universe
- Albrecht A., Bernstein G., Cahn R., et al., 2006, Report of the Dark Energy Task Force, ArXiv Astrophysics e-prints
- Albrecht A., Steinhardt P. J., 1982, Cosmology for grand unified theories with radiatively induced symmetry breaking, Physical Review Letters, 48, 1220
- Anderson L., Aubourg E., Bailey S., et al., 2012, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample, MNRAS, 427, 3435
- Arnborg S., Sjödin G., 2001, On the foundations of Bayesianism, AIP Conference Proceedings, 568, 1, 61
- Arnouts S., Cristiani S., Moscardini L., et al., 1999, Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North, MNRAS, 310, 540
- Asorey J., Carrasco Kind M., Sevilla-Noarbe I., Brunner R. J., Thaler J., 2016, Galaxy clustering with photometric surveys using PDF redshift information, MNRAS, 459, 1293

- Asorey J., Crocce M., Gaztañaga E., 2014, Redshift-space distortions from the cross-correlation of photometric populations, *MNRAS*, 445, 2825
- Asorey J., Crocce M., Gaztañaga E., Lewis A., 2012, Recovering 3D clustering information with angular correlations, *MNRAS*, 427, 1891
- Astashenok A. V., Nojiri S., Odintsov S. D., Yurov A. V., 2012, Phantom cosmology without Big Rip singularity, *Physics Letters B*, 709, 396
- Astropy Collaboration, Robitaille T. P., Tollerud E. J., et al., 2013, Astropy: A community Python package for astronomy, *Astron. & Astrophys.*, 558, A33
- Auder B., Lebrete R., Iovleff S., Langrognet F., 2014, Rmixmod: An interface for MIXMOD, R package version 2.0.2
- Austermann J. E., Aird K. A., Beall J. A., et al., 2012, in *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VI*, vol. 8452 of , 84521E
- Ball N. M., Brunner R. J., 2010, Data Mining and Machine Learning in Astronomy, *International Journal of Modern Physics D*, 19, 1049
- Banday A. J., Górski K. M., Bennett C. L., et al., 1997, Root Mean Square Anisotropy in the COBE DMR Four-Year Sky Maps, *ApJ*, 475, 393
- Bartelmann M., Schneider P., 2001, Weak gravitational lensing, *Physics Reports*, 340, 291
- Bassett B., Hlozek R., 2010, Baryon acoustic oscillations, 246
- Baugh C., 2000, Correlation Function and Power Spectra in Cosmology, *Encyclopedia of Astronomy and Astrophysics*
- Bender R., Appenzeller I., Böhm A., et al., 2001, in *Deep Fields*, edited by S. Cristiani, A. Renzini, R. E. Williams, 96
- Benítez N., 2000, Bayesian Photometric Redshift Estimation, *ApJ*, 536, 571
- Benítez N., 2000, Bayesian Photometric Redshift Estimation, *The Astrophysical Journal*, 536, 2, 571
- Benítez N., 2011, BPZ: Bayesian Photometric Redshift Code, *Astrophysics Source Code Library*
- Benítez N., Ford H., Bouwens R., et al., 2004, Faint Galaxies in Deep Advanced Camera for Surveys Observations, *The Astrophysical Journal Supplement Series*, 150, 1, 1
- Benjamin J., Van Waerbeke L., Heymans C., et al., 2013, CFHTLenS tomographic weak lensing: quantifying accurate redshift distributions, *MNRAS*, 431, 1547
- Benjamin J., van Waerbeke L., Ménard B., Kilbinger M., 2010, Photometric redshifts: estimating their contamination and distribution using clustering information, *MNRAS*, 408, 1168

- Bertin E., 2011, in *Astronomical Data Analysis Software and Systems XX*, edited by I. N. Evans, A. Accomazzi, D. J. Mink, A. H. Rots, vol. 442 of *Astronomical Society of the Pacific Conference Series*, 435
- Biernacki C., Celeux G., Govaert G., 1999, An improvement of the NEC criterion for assessing the number of clusters in a mixture model., *Pattern Recognition Letters*, 20, 3, 267
- Bishop C., 1995, *Neural Networks for Pattern Recognition*, Oxford University Press
- Bishop C. M., 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA
- Blandford R. D., Saust A. B., Brainerd T. G., Villumsen J. V., 1991, The distortion of distant galaxy images by large-scale structure, *MNRAS*, 251, 600
- Blanton M. R., Bershadsky M. A., Abolfathi B., et al., 2017, Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe, *AJ*, 154, 28
- Blanton M. R., Brinkmann J., Csabai I., et al., 2003, Estimating Fixed-Frame Galaxy Magnitudes in the Sloan Digital Sky Survey, *AJ*, 125, 2348
- Blas D., Lesgourgues J., Tram T., 2011, The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes, *JCAP*, 7, 034
- Bolzonella M., Miralles J.-M., Pelló R., 2000, Photometric redshifts based on standard SED fitting procedures, *Astron. & Astrophys.*, 363, 476
- Bonnett C., 2015, Using neural networks to estimate redshift distributions. An application to CFHTLenS, *MNRAS*, 449, 1043
- Bonnett C., Troxel M. A., Hartley W., et al., 2016, Redshift distributions of galaxies in the Dark Energy Survey Science Verification shear catalogue and implications for weak lensing, *PRD*, 94, 4, 042005
- Bosma A., 1978, The distribution and kinematics of neutral hydrogen in spiral galaxies of various morphological types, Ph.D. thesis, PhD Thesis, Groningen Univ., (1978)
- Bosma A., 1981a, 21-cm line studies of spiral galaxies. I - Observations of the galaxies NGC 5033, 3198, 5055, 2841, and 7331, *AJ*, 86, 1791
- Bosma A., 1981b, 21-cm line studies of spiral galaxies. II. The distribution and kinematics of neutral hydrogen in spiral galaxies of various morphological types., *AJ*, 86, 1825
- Brammer G. B., van Dokkum P. G., Coppi P., 2008, EAZY: A Fast, Public Photometric Redshift Code, *ApJ*, 686, 1503
- Branco P., Ribeiro R. P., Torgo L., 2016, UBL: an R package for Utility-based Learning, *ArXiv e-prints*
- Branco P., Torgo L., Ribeiro R. P., Frank E., Pfahringer B., Rau M. M., 2017, Learning Through Utility Optimization in Regression Tasks, accepted to 4th IEEE International Conference on Data Science and Advanced Analytics - DSAA2017 proceedings

- Brimioulle F., 2013, Dark matter halo properties from galaxy-galaxy lensing
- Bulmer M., 1967, Principles of statistics, M.I.T. Press, Cambridge, Mass., 2nd edn.
- Cai R.-G., Tuo Z.-L., 2011, Detecting the cosmic acceleration with current data, Physics Letters B, 706, 116
- Caldwell R. R., Kamionkowski M., Weinberg N. N., 2003, Phantom Energy: Dark Energy with $w = -1$ Causes a Cosmic Doomsday, Physical Review Letters, 91, 7, 071301
- Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, Random Forests for Photometric Redshifts, ApJ, 712, 511
- Carlstrom J. E., Ade P. A. R., Aird K. A., et al., 2011, The 10 Meter South Pole Telescope, PASP, 123, 568
- Carrasco Kind M., Brunner R. J., 2013, TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests, MNRAS, 432, 1483
- Carrasco Kind M., Brunner R. J., 2014a, SOMz: photometric redshift PDFs with self-organizing maps and random atlas, MNRAS, 438, 3409
- Carrasco Kind M., Brunner R. J., 2014b, Sparse representation of photometric redshift probability density functions: preparing for petascale astronomy, MNRAS, 441, 3550
- Carroll B. W., Ostlie D. A., 2007, An Introduction to Modern Astrophysics, 2nd edn.
- Carroll R. J., Ruppert D., Stefanski L. A., Crainiceanu C. M., 2006, Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition, Chapman and Hall/CRC, 2nd edn.
- Carroll S. M., 1997, Lecture Notes on General Relativity, ArXiv General Relativity and Quantum Cosmology e-prints
- Carroll S. M., Press W. H., Turner E. L., 1992, The cosmological constant, Annu. Rev. Astron. Astrophys., 30, 499
- Cattoën C., Visser M., 2008, Cosmographic Hubble fits to the supernova data, PRD, 78, 6, 063501
- CFHTLenS-collaboration, 2017, The CFHT Lensing Survey, <http://www.cfhtlens.org/astromers/content-suitable-astromers>, Accessed: 31 Aug 2017
- CFHTLS-collaboration, 2017a, Canada-France-Hawaii Telescope Legacy Survey, <http://www.cfht.hawaii.edu/Science/CFHTLS/>, Accessed: 31 Aug 2017
- CFHTLS-collaboration, 2017b, MegaCam - Instrument Description: the various components, <http://www.cfht.hawaii.edu/Instruments/Imaging/Megacam/megaprimecomponents.html>, Accessed: 31 Aug 2017
- Chan B. M. Y., Broadhurst T., Lim J., et al., 2017, Geometric Corroboration of the Earliest Lensed Galaxy at $z = 10.8$ from Robust Free-form Modelling, ApJ, 835, 44

- Chang C., Pujol A., Mawdsley B., et al., 2017, Dark Energy Survey Year 1 Results: Curved-Sky Weak Lensing Mass Map, ArXiv e-prints
- Charnock T., Moss A., 2017, Deep Recurrent Neural Networks for Supernovae Classification, *ApJL*, 837, L28
- Clerkin L., Kirk D., Lahav O., Abdalla F. B., Gaztañaga E., 2015, A prescription for galaxy biasing evolution as a nuisance parameter, *MNRAS*, 448, 1389
- Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, Galaxies in the Hubble Ultra Deep Field. I. Detection, Multiband Photometry, Photometric Redshifts, and Morphology, *AJ*, 132, 926
- Coe D., Zitrin A., Carrasco M., et al., 2013, CLASH: Three Strongly Lensed Images of a Candidate $z \sim 11$ Galaxy, *ApJ*, 762, 32
- Coil A. L., Blanton M. R., Burles S. M., et al., 2011, The PRISM Multi-object Survey (PRIMUS). I. Survey Overview and Characteristics, *The Astrophysical Journal*, 741, 1, 8
- Coil A. L., Blanton M. R., Burles S. M., et al., 2011, The PRISM Multi-object Survey (PRIMUS). I. Survey Overview and Characteristics, *ApJ*, 741, 8
- Coles P., 2001, in *Phase Transitions in the Early Universe: Theory and Observations*, edited by H. J. de Vega, I. M. Khalatnikov, N. G. Sanchez, 217
- Collister A. A., Lahav O., 2004, ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks, *PASP*, 116, 345
- Cool R. J., Moustakas J., Blanton M. R., et al., 2013, The PRISM Multi-object Survey (PRIMUS). II. Data Reduction and Redshift Fitting, *ApJ*, 767, 118
- Cox R. T., 1946, Probability, Frequency and Reasonable Expectation, *American Journal of Physics*, 14, 1
- Crocce M., Cabré A., Gaztañaga E., 2011, Modelling the angular correlation function and its full covariance in photometric galaxy surveys, *MNRAS*, 414, 329
- Cun Y. L., Boser B., Denker J. S., et al., 1990, chap. Handwritten Digit Recognition with a Back-propagation Network, 396–404, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
- Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, Estimating the redshift distribution of photometric galaxy samples - II. Applications and tests of a new method, *MNRAS*, 396, 2379
- Dahlen T., Mobasher B., Faber S. M., et al., 2013, A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation, *ApJ*, 775, 93
- Dark Energy Survey Collaboration, Abbott T., Abdalla F. B., et al., 2016, The Dark Energy Survey: more than dark energy - an overview, *MNRAS*, 460, 1270

- Davis M., Huchra J., Latham D. W., Tonry J., 1982, A survey of galaxy redshifts. II - The large scale space distribution, *ApJ*, 253, 423
- Davis M., Peebles P. J. E., 1983, A survey of galaxy redshifts. V - The two-point position and velocity correlations, *ApJ*, 267, 465
- Dawson K. S., Schlegel D. J., Ahn C. P., et al., 2013, The Baryon Oscillation Spectroscopic Survey of SDSS-III, *AJ*, 145, 10
- De Gooijer J. G., Zerom D., 2003, On Conditional Density Estimation, *Statistica Neerlandica*, 57, 2, 159
- de Jong J. T. A., Kuijken K., Applegate D., et al., 2013, The Kilo-Degree Survey, *The Messenger*, 154, 44
- DES Collaboration, Abbott T. M. C., Abdalla F. B., et al., 2017, Dark Energy Survey Year 1 Results: Cosmological Constraints from Galaxy Clustering and Weak Lensing, *ArXiv e-prints*
- Desjacques V., Jeong D., Schmidt F., 2016, Large-Scale Galaxy Bias, *ArXiv e-prints*
- Diehl T., Dark Energy Survey Collaboration, 2012, The Dark Energy Survey Camera (DECam), *Physics Procedia*, 37, 1332
- Dieleman S., Willett K. W., Dambre J., 2015, Rotation-invariant convolutional neural networks for galaxy morphology prediction, *MNRAS*, 450, 1441
- Dodelson S., 2003, *Modern cosmology*, Academic Press, San Diego, CA
- Domingos P., 1999, in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 155–164, ACM
- Drinkwater M. J., Jurek R. J., Blake C., et al., 2010, The WiggleZ Dark Energy Survey: survey design and first data release, *MNRAS*, 401, 1429
- Drlica-Wagner A., Sevilla-Noarbe I., Rykoff E. S., et al., 2017, Dark Energy Survey Year 1 Results: Photometric Data Set for Cosmology, *ArXiv e-prints*
- Dyson F. W., Eddington A. S., Davidson C., 1920, A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919, *Philosophical Transactions of the Royal Society of London Series A*, 220, 291
- Einstein A., 1915, Die Feldgleichungen der Gravitation, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*, Seite 844-847.
- Einstein A., 1916, Die Grundlage der allgemeinen Relativitätstheorie, *Annalen der Physik*, 354, 769
- Einstein A., 1917, Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)*, Seite 142-152.
- Eisenstein D. J., 1997, An Analytic Expression for the Growth Function in a Flat Universe with a Cosmological Constant, *ArXiv Astrophysics e-prints*

- Eisenstein D. J., Hu W., 1998, Baryonic Features in the Matter Transfer Function, *ApJ*, 496, 605
- Elkan C., 2001, in *IJCAI'01: Proc. of 17th Int. Joint Conf. of Artificial Intelligence*, vol. 1, 973–978, Morgan Kaufmann Publishers
- Erben T., Hildebrandt H., Miller L., et al., 2013, CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey - imaging data and catalogue products, *MNRAS*, 433, 2545
- Escoffier S., Comparat J., Ealet A., Kneib J.-P., Zoubian J., Lamareille F., 2013, The ELG target selection with the BOSS survey, *ArXiv e-prints*
- Feldmann R., Carollo C. M., Porciani C., et al., 2006a, The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS, *MNRAS*, 372, 565
- Feldmann R., Carollo C. M., Porciani C., et al., 2006b, The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS, *MNRAS*, 372, 565
- Firth A. E., Lahav O., Somerville R. S., 2003, Estimating photometric redshifts with artificial neural networks, *MNRAS*, 339, 1195
- Fixsen D. J., 2009, The Temperature of the Cosmic Microwave Background, *ApJ*, 707, 916
- Flaugher B., Diehl H. T., Honscheid K., et al., 2015, The Dark Energy Camera, *AJ*, 150, 150
- Fließbach T., 1990, *Allgemeine Relativitätstheorie*, BI-Wiss.-Verlag
- Frank E., Hall M., 2001, A Simple Approach to Ordinal Classification, Tech. Rep. 01/05, Department of Computer Science, University of Waikato
- Freeman K. C., 1970, On the Disks of Spiral and S0 Galaxies, *ApJ*, 160, 811
- Freeman P. E., Newman J. A., Lee A. B., Richards J. W., Schafer C. M., 2009, Photometric redshift estimation using spectral connectivity analysis, *MNRAS*, 398, 2012
- Friedman J. H., 1999, Stochastic Gradient Boosting, *Computational Statistics and Data Analysis*, 38, 367
- Friedman J. H., 2001, Greedy Function Approximation: A gradient boosting machine, *The Annals of Statistics*, 29, 5, 1189
- Frieman J. A., Turner M. S., Huterer D., 2008, Dark Energy and the Accelerating Universe, *Annu. Rev. Astron. Astrophys.*, 46, 385
- Fry J. N., 1996, The Evolution of Bias, *ApJL*, 461, L65
- Gatti M., Vielzeuf P., Davis C., et al., 2017, Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts - Methods and Systematics Characterization, *ArXiv e-prints*
- Gerdes D. W., Sypniewski A. J., McKay T. A., et al., 2010, ArborZ: Photometric Redshifts Using Boosted Decision Trees, *ApJ*, 715, 823
- Giannantonio T., Porciani C., Carron J., Amara A., Pillepich A., 2012, Constraining primordial non-Gaussianity with future galaxy surveys, *MNRAS*, 422, 2854

- Goobar A., Paech K., Stanishev V., et al., 2009, Near-IR search for lensed supernovae behind galaxy clusters. II. First detection and future prospects, *Astron. & Astrophys.*, 507, 71
- Greisel N., 2015, Photometric redshifts and properties of galaxies from the sloan digital sky survey
- Gruen D., Brimiouille F., 2016, Selection biases in empirical $p(z)$ methods for weak lensing, *ArXiv e-prints*
- Gu J., Wang Z., Kuen J., et al., 2015, Recent Advances in Convolutional Neural Networks, *ArXiv e-prints*
- Guth A. H., 1981, Inflationary universe: A possible solution to the horizon and flatness problems, *PRD*, 23, 347
- Guth A. H., Pi S.-Y., 1982, Fluctuations in the new inflationary universe, *Physical Review Letters*, 49, 1110
- Guzzo L., Pierleoni M., Meneux B., et al., 2008, A test of the nature of cosmic acceleration using galaxy redshift distortions, *Nature*, 451, 541
- Hald A., 2008, A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935, *Sources and Studies in the History of Mathematics and Physical Sciences*, Springer New York
- Hamilton A. J. S., 2001, Formulae for growth factors in expanding universes containing matter and a cosmological constant, *MNRAS*, 322, 419
- Hamilton A. J. S., Kumar P., Lu E., Matthews A., 1991, Reconstructing the primordial spectrum of fluctuations of the universe from the observed nonlinear clustering of galaxies, *ApJL*, 374, L1
- Harrison E. R., 1970, Fluctuations at the Threshold of Classical Cosmology, *PRD*, 1, 2726
- Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics, Springer New York
- Haykin S., 2009, *Neural networks and learning machines*, Pearson Education, 3rd edn.
- Hearst M. A., Dumais S. T., Osuna E., Platt J., Scholkopf B., 1998, Support vector machines, *IEEE Intelligent Systems and their Applications*, 13, 4, 18
- Heath D. J., 1977, The growth of density perturbations in zero pressure Friedmann-Lemaitre universes, *MNRAS*, 179, 351
- Heymans C., Grocutt E., Heavens A., et al., 2013, CFHTLenS tomographic weak lensing cosmological parameter constraints: Mitigating the impact of intrinsic galaxy alignments, *MNRAS*, 432, 2433
- Heymans C., Van Waerbeke L., Miller L., et al., 2012, CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey, *MNRAS*, 427, 146

- High F. W., Stubbs C. W., Rest A., Stalder B., Challis P., 2009, Stellar Locus Regression: Accurate Color Calibration and the Real-Time Determination of Galaxy Cluster Photometric Redshifts, *AJ*, 138, 110
- Hildebrandt H., Arnouts S., Capak P., et al., 2010, PHAT: PHoto-z Accuracy Testing, *Astron. & Astrophys.*, 523, A31
- Hildebrandt H., CFHTLenS collaboration, 2014, An overview of the completed Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS), *ArXiv e-prints*
- Hildebrandt H., Erben T., Kuijken K., et al., 2012, CFHTLenS: improving the quality of photometric redshifts with precision photometry, *MNRAS*, 421, 2355
- Hildebrandt H., Viola M., Heymans C., et al., 2017, KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing, *MNRAS*, 465, 1454
- Hildebrandt H., Wolf C., Benítez N., 2008, A blind test of photometric redshifts on ground-based data, *Astron. & Astrophys.*, 480, 703
- Hinton G., Deng L., Yu D., et al., 2012, Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Processing Magazine*, 29, 6, 82
- Hinton S. R., Kazin E., Davis T. M., et al., 2017, Measuring the 2D baryon acoustic oscillation signal of galaxies in WiggleZ: cosmological constraints, *MNRAS*, 464, 4807
- Hoekstra H., Bartelmann M., Dahle H., Israel H., Limousin M., Meneghetti M., 2013, Masses of Galaxy Clusters from Gravitational Lensing, , 177, 75
- Hogg, 1999, Distance measures in cosmology
- Hogg D. W., Baldry I. K., Blanton M. R., Eisenstein D. J., 2002, The K correction, *ArXiv Astrophysics e-prints*
- Hojjati A., Linder E. V., 2014, Next generation strong lensing time delay estimation with Gaussian processes, *Phys. Rev. D*, 90, 123501
- Holmes M. P., Gray A. G., Isbell C. L., 2012, Fast Nonparametric Conditional Density Estimation, *ArXiv e-prints*
- Horn K. S. V., 2003, Constructing a logic of plausible inference: a guide to Cox's theorem, *International Journal of Approximate Reasoning*, 34, 1, 3
- Hoyle B., Gruen D., Bernstein G. M., et al., 2017, Dark Energy Survey Year 1 Results: Redshift distributions of the weak lensing source galaxies, *ArXiv e-prints*
- Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015a, Data augmentation for machine learning redshifts applied to Sloan Digital Sky Survey galaxies, *MNRAS*, 450, 305
- Hoyle B., Rau M. M., Paech K., Bonnett C., Seitz S., Weller J., 2015b, Anomaly detection for machine learning redshifts applied to SDSS galaxies, *MNRAS*, 452, 4183

- Hu W., Dodelson S., 2002, Cosmic Microwave Background Anisotropies, *Annu. Rev. Astron. Astrophys.*, 40, 171
- Hubble E., 1929, A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae, *Proceedings of the National Academy of Science*, 15, 168
- Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification, *Astron. & Astrophys.*, 525, A157
- Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment, *Computing in Science Engineering*, 9, 3, 90
- Huterer D., Turner M. S., 2001, Probing dark energy: Methods and strategies, *PRD*, 64, 12, 123527
- Ilbert, O., Arnouts, S., McCracken, H. J., et al., 2006, Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey, *AA*, 457, 3, 841
- Jain B., Taylor A., 2003, Cross-Correlation Tomography: Measuring Dark Energy Evolution with Weak Lensing, *Physical Review Letters*, 91, 14, 141302
- Jaynes E. T., 1989, *Clearing up Mysteries — The Original Goal*, 1–27, Springer Netherlands, Dordrecht
- Joachimi B., Bridle S. L., 2010, Simultaneous measurement of cosmology and intrinsic alignments using joint cosmic shear and galaxy number density correlations, *Astron. & Astrophys.*, 523, A1
- Jones D. H., Read M. A., Saunders W., et al., 2009, The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures, *MNRAS*, 399, 683
- Jones E., Oliphant T., Peterson P., et al., 2001–, SciPy: Open source scientific tools for Python
- Jones M., Marron J., Sheather S., 1996, A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association*, 91, 433, 401
- Jouvel S., Kneib J.-P., Ilbert O., et al., 2009, Designing future dark energy space missions. I. Building realistic galaxy spectro-photometric catalogs and their first applications, *Astron. & Astrophys.*, 504, 359
- Kaiser N., 1992, Weak gravitational lensing of distant galaxies, *ApJ*, 388, 272
- Kaiser N., 1995, Nonlinear cluster lens reconstruction, *ApJL*, 439, L1
- Kessler R., Bassett B., Belov P., et al., 2010, Results from the Supernova Photometric Classification Challenge, *PASP*, 122, 1415
- Kilbinger M., 2015, Cosmology with cosmic shear observations: a review, *Reports on Progress in Physics*, 78, 8, 086901
- Kirk D., 2011, *The Practical use of Cosmic Shear as a Probe of Gravity*, Ph.D. thesis, Department of Physics & Astronomy, University College London

- Kirk D., Lahav O., Bridle S., Jouvel S., Abdalla F. B., Frieman J. A., 2015, Optimizing spectroscopic and photometric galaxy surveys: same-sky benefits for dark energy and modified gravity, *MNRAS*, 451, 4424
- Kitching T. D., Viola M., Hildebrandt H., et al., 2015, RCSLenS: Cosmic Distances from Weak Lensing, *ArXiv e-prints*
- Ku H. H., 1966, Notes on the use of propagation of error formulas, *Journal of Research of the National Bureau of Standards. Section C: Engineering and Instrumentation*, 70C, 4, 263
- Kuijken K., 2008, GaaP: PSF- and aperture-matched photometry using shapelets, *Astron. & Astrophys.*, 482, 1053
- Laigle C., McCracken H. J., Ilbert O., et al., 2016, The COSMOS2015 Catalog: Exploring the $1 < z < 6$ Universe with Half a Million Galaxies, *ApJS*, 224, 24
- Larson D., Dunkley J., Hinshaw G., et al., 2011, Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Power Spectra and WMAP-derived Parameters, *ApJS*, 192, 16
- Le Fèvre O., Vettolani G., Garilli B., et al., 2005, The VIMOS VLT deep survey. First epoch VVDS-deep survey: 11 564 spectra with $17.5 \leq IAB \leq 24$, and the redshift distribution over $0 \leq z \leq 5$, *Astron. & Astrophys.*, 439, 845
- Leistedt B., Mortlock D. J., Peiris H. V., 2016, Hierarchical Bayesian inference of galaxy redshift distributions from photometric surveys, *MNRAS*, 460, 4258
- Lewis A., Challinor A., Lasenby A., 2000, Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models, *ApJ*, 538, 473
- Liaw A., Wiener M., 2002, Classification and Regression by randomForest, *R News*, 2, 3, 18
- Lilly S. J., Le Fèvre O., Renzini A., et al., 2007a, zCOSMOS: A Large VLT/VIMOS Redshift Survey Covering $0 < z < 3$ in the COSMOS Field, *ApJS*, 172, 70
- Lilly S. J., Le Fèvre O., Renzini A., et al., 2007b, zCOSMOS: A Large VLT/VIMOS Redshift Survey Covering $0 < z < 3$ in the COSMOS Field, *ApJS*, 172, 70
- Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, Estimating the redshift distribution of photometric galaxy samples, *MNRAS*, 390, 118
- Limber D. N., 1953, The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field., *ApJ*, 117, 134
- Linde A. D., 1982, A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems, *Physics Letters B*, 108, 389
- Loeb A., 2006, The Dark Ages of the Universe, *University Lecture*
- Maartens R., 2011, Is the Universe homogeneous?, *Philosophical Transactions of the Royal Society of London Series A*, 369, 5115

- Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990, Galaxy correlations on large scales, *MNRAS*, 242, 43P
- Mairal J., Bach F., Ponce J., Sapiro G., 2009, in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, 689–696, ACM, New York, NY, USA
- Mandelbaum R., Seljak U., Hirata C. M., et al., 2008, Precision photometric redshift calibration for galaxy-galaxy weak lensing, *MNRAS*, 386, 781
- McKinney W., 2010, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt, J. Millman, 51 – 56
- Meinshausen N., 2006, Quantile Regression Forests, *J. Mach. Learn. Res.*, 7, 983
- Ménard B., Scranton R., Schmidt S., et al., 2013, Clustering-based redshift estimation: method and application to data, *ArXiv e-prints*
- Miralda-Escude J., 1991, Gravitational lensing by clusters of galaxies - Constraining the mass distribution, *ApJ*, 370, 1
- Momcheva I. G., Brammer G. B., van Dokkum P. G., et al., 2016, The 3D-HST Survey: Hubble Space Telescope WFC3/G141 Grism Spectra, Redshifts, and Emission Line Measurements for $\sim 100,000$ Galaxies, *ApJS*, 225, 27
- Monna A., Covone G., 2012, Searching primeval galaxies through gravitational telescopes., *Memorie della Societa Astronomica Italiana Supplementi*, 19, 258
- Monna A., Seitz S., Balestra I., et al., 2017, Precise strong lensing mass profile of the CLASH galaxy cluster MACS 2129, *MNRAS*, 466, 4094
- Morrison C. B., Hildebrandt H., Schmidt S. J., et al., 2017, the-wizz: clustering redshift estimation for everyone, *MNRAS*, 467, 3576
- Mukhanov V., 2005, *Physical Foundations of Cosmology*, Cambridge University Press
- Narayan R., Bartelmann M., 1996, *Lectures on Gravitational Lensing*, *ArXiv Astrophysics e-prints*
- Newman J. A., 2008, Calibrating Redshift Distributions beyond Spectroscopic Limits with Cross-Correlations, *ApJ*, 684, 88
- Newman J. A., Cooper M. C., Davis M., et al., 2013, The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts, *ApJS*, 208, 5
- O’Brien D. B., Gupta M. R., Gray R. M., 2008, in *Proceedings of the 25th international conference on Machine learning*, 712–719, ACM
- Oke J. B., Gunn J. E., 1983, Secondary standard stars for absolute spectrophotometry, *ApJ*, 266, 713
- Oke J. B., Sandage A., 1968, Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies, *ApJ*, 154, 21

- Olive K. A., et al., 2014, Review of Particle Physics, *Chin. Phys.*, C38, 090001
- Peacock J. A., 1999, *Cosmological Physics*, Cambridge University Press
- Peacock J. A., Cole S., Norberg P., et al., 2001, A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey, *Nature*, 410, 169
- Pedregosa F., Varoquaux G., Gramfort A., et al., 2011, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825
- Peebles P. J. E., 1973, Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory, *ApJ*, 185, 413
- Peebles P. J. E., 1980, The large-scale structure of the universe
- Percival W. J., Samushia L., Ross A. J., Shapiro C., Raccanelli A., 2011, Redshift-space distortions, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369, 1957, 5058
- Perez F., Granger B. E., 2007, IPython: A System for Interactive Scientific Computing, *Computing in Science Engineering*, 9, 3, 21
- Perlmutter S., Aldering G., Goldhaber G., et al., 1999, Measurements of Ω and Λ from 42 High-Redshift Supernovae, *ApJ*, 517, 565
- Pischke S., 2007, Lecture Notes on Measurement Error, http://econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf, Accessed: Mo 28 Aug 2017 15:31:14 CEST
- Planck Collaboration, Ade P. A. R., Aghanim N., et al., 2014, Planck 2013 results. XVI. Cosmological parameters, *Astron. & Astrophys.*, 571, A16
- Planck Collaboration, Ade P. A. R., Aghanim N., et al., 2016, Planck 2015 results. XIII. Cosmological parameters, *Astron. & Astrophys.*, 594, A13
- Polletta M., Tajer M., Maraschi L., et al., 2007, Spectral Energy Distributions of Hard X-Ray Selected Active Galactic Nuclei in the XMM-Newton Medium Deep Survey, *ApJ*, 663, 81
- Prat J., Sánchez C., Fang Y., et al., 2017, Dark Energy Survey Year 1 Results: Galaxy-Galaxy Lensing, *ArXiv e-prints*
- Primack J. R., Gross M. A. K., 2001, Hot dark matter in cosmology, 287–308
- Quionero-Candela J., Sugiyama M., Schwaighofer A., Lawrence N. D., 2009, *Dataset Shift in Machine Learning*, The MIT Press
- R Core Team, 2015, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
- Rau M. M., Hoyle B., Paech K., Seitz S., 2017, Correcting cosmological parameter biases for all redshift surveys induced by estimating and reweighting redshift distributions, *MNRAS*, 466, 2927
- Rau M. M., Seitz S., Brimiouille F., et al., 2015, Accurate photometric redshift probability density estimation - method comparison and application, *MNRAS*, 452, 3710

- Rauch M., Miralda-Escudé J., Sargent W. L. W., et al., 1997, The Opacity of the Ly α Forest and Implications for Ω_b and the Ionizing Background, *ApJ*, 489, 7
- Rawat W., Wang Z., 2017, Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, 1–98
- Ribeiro R. P. A., 2011, Utility-based Regression, Ph.D. thesis, Department of Computer Science, Faculty of Sciences University of Porto
- Richards J. W., Freeman P. E., Lee A. B., Schafer C. M., 2009, Exploiting Low-Dimensional Structure in Astronomical Spectra, *ApJ*, 691, 32
- Richards J. W., Starr D. L., Butler N. R., et al., 2011, On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data, *ApJ*, 733, 10
- Riess A. G., Filippenko A. V., Challis P., et al., 1998, Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant, *AJ*, 116, 1009
- Riess A. G., Macri L., Casertano S., et al., 2011, A 3% Solution: Determination of the Hubble Constant with the Hubble Space Telescope and Wide Field Camera 3, *ApJ*, 730, 119
- Rosenblatt M., 1969, Multivariate analysis - II: proceedings, Aerospace Research Laboratories (U.S.) and Wright State University, Academic Press
- Rossum G., 1995, Python Reference Manual, Tech. rep., Amsterdam, The Netherlands, The Netherlands
- Rousseeuw P. J., Driessen K. V., 1998, A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, 41, 212
- Rozo E., Rykoff E. S., Abate A., et al., 2016, redMaGiC: selecting luminous red galaxies from the DES Science Verification data, *MNRAS*, 461, 1431
- Rubakov V. A., Vlasov A. D., 2012, What do we learn from the CMB observations?, *Physics of Atomic Nuclei*, 75, 1123
- Rubin V. C., Ford Jr. W. K., Thonnard N., 1980, Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 / $R = 4\text{kpc}/$ to UGC 2885 / $R = 122\text{ kpc}/$, *ApJ*, 238, 471
- Ruhl J., Ade P. A. R., Carlstrom J. E., et al., 2004, in *Z-Spec: a broadband millimeter-wave grating spectrometer: design, construction, and first cryogenic measurements*, edited by C. M. Bradford, P. A. R. Ade, J. E. Aguirre, J. J. Bock, M. Dragovan, L. Duband, L. Earle, J. Glenn, H. Matsuhara, B. J. Naylor, H. T. Nguyen, M. Yun, J. Zmuidzinas, vol. 5498 of , 11–29
- Rydberg C.-E., Zackrisson E., Zitrin A., et al., 2015, A Search for Population III Galaxies in CLASH. I. Singly-imaged Candidates at High Redshift, *ApJ*, 804, 13
- Salazar-Albornoz S., Sánchez A. G., Grieb J. N., et al., 2017, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: angular clustering tomography and its cosmological implications, *MNRAS*, 468, 2938

- Sánchez A. G., 2017, The formation and evolution of cosmic structures, University Lecture
- Sánchez C., Carrasco Kind M., Lin H., et al., 2014, Photometric redshift analysis in the Dark Energy Survey Science Verification data, *MNRAS*, 445, 1482
- Schmidhuber J., 2014, Deep Learning in Neural Networks: An Overview, ArXiv e-prints
- Schneider P., 2006a, Einführung in die extragalaktische Astronomie und Kosmologie, Springer
- Schneider P., 2006b, Weak Gravitational Lensing, 269–451, Springer Berlin Heidelberg, Berlin, Heidelberg
- Schneider P., 2016, Generalized shear-ratio tests: A new relation between cosmological distances, and a diagnostic for a redshift-dependent multiplicative bias in shear measurements, *Astron. & Astrophys.*, 592, L6
- Scott D. W., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, 1st edn.
- Scott D. W., Sain S. R., 2005, Multidimensional Density Estimation, *Handbook of statistics*, 24, 229
- Scoville N., Aussel H., Brusa M., et al., 2007, The Cosmic Evolution Survey (COSMOS): Overview, *ApJS*, 172, 1
- Seabold S., Perktold J., 2010, in 9th Python in Science Conference
- Seldner M., Siebers B., Groth E. J., Peebles P. J. E., 1977, New reduction of the Lick catalog of galaxies, *AJ*, 82, 249
- Senn S., 2003, Dicing with Death: Chance, Risk and Health, Cambridge University Press
- Sesar B., Hernitschek N., Mitrović S., et al., 2017, Machine-learned Identification of RR Lyrae Stars from Sparse, Multi-band Data: The PS1 Sample, *AJ*, 153, 204
- Sheather S. J., 2004, Density Estimation, *Statist. Sci.*, 19, 4, 588
- Silva L., Granato G. L., Bressan A., Danese L., 1998, Modeling the Effects of Dust on Galactic Spectral Energy Distributions from the Ultraviolet to the Millimeter Band, *ApJ*, 509, 103
- Silverman J. D., Kashino D., Sanders D., et al., 2015, The FMOS-COSMOS Survey of Star-forming Galaxies at $z \sim 1.6$. III. Survey Design, Performance, and Sample Characteristics, *The Astrophysical Journal Supplement Series*, 220, 1, 12
- Simon P., 2007, How accurate is Limber’s equation?, *Astron. & Astrophys.*, 473, 711
- Simpson F., Heymans C., Parkinson D., et al., 2013, CFHTLenS: testing the laws of gravity with tomographic weak lensing and redshift-space distortions, *MNRAS*, 429, 2249
- Sivia D., Skilling J., 2006, Data Analysis: A Bayesian Tutorial, Oxford science publications, OUP Oxford

- Smee S. A., Gunn J. E., Uomoto A., et al., 2013, The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey, *AJ*, 146, 32
- Smith R. E., Peacock J. A., Jenkins A., et al., 2003, Stable clustering, the halo model and non-linear cosmological power spectra, *MNRAS*, 341, 1311
- Smith S., 1936, The Mass of the Virgo Cluster, *ApJ*, 83, 23
- Smola A. J., Schölkopf B., 2004, A tutorial on support vector regression, *Statistics and Computing*, 14, 3, 199
- Speagle J. S., Eisenstein D. J., 2015a, Deriving Photometric Redshifts using Fuzzy Archetypes and Self-Organizing Maps. I. Methodology, *ArXiv e-prints*
- Speagle J. S., Eisenstein D. J., 2015b, Deriving Photometric Redshifts using Fuzzy Archetypes and Self-Organizing Maps. II. Comparing Sampling Techniques Using Mock Data, *ArXiv e-prints*
- Suykens J., Vandewalle J., 1999, Least Squares Support Vector Machine Classifiers, *Neural Processing Letters*, 9, 3, 293
- Suyu S. H., Auger M. W., Hilbert S., et al., 2013, Two Accurate Time-delay Distances from Strong Lensing: Implications for Cosmology, *ApJ*, 766, 70
- Takada M., Jain B., 2004, Cosmological parameters from lensing power spectrum and bispectrum tomography, *MNRAS*, 348, 897
- Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, Revising the Halofit Model for the Nonlinear Matter Power Spectrum, *ApJ*, 761, 152
- Taylor A. N., Kitching T. D., Bacon D. J., Heavens A. F., 2007, Probing dark energy with the shear-ratio geometric test, *MNRAS*, 374, 1377
- Taylor M. B., 2006, in *Astronomical Data Analysis Software and Systems XV*, edited by C. Gabriel, C. Arviset, D. Ponz, S. Enrique, vol. 351 of *Astronomical Society of the Pacific Conference Series*, 666
- Tegmark M., 1994, Probes of the Early Universe, Ph.D. thesis, U.C. Berkeley Ph.D. thesis, May 1994
- The Dark Energy Survey Collaboration, 2005, The Dark Energy Survey, *ArXiv Astrophysics e-prints*
- Tonry J., Davis M., 1979, A survey of galaxy redshifts. I - Data reduction techniques, *AJ*, 84, 1511
- Torgo L., Ribeiro R., 2007, *Utility-Based Regression*, 597–604, Springer Berlin Heidelberg, Berlin, Heidelberg
- Treu T., Marshall P. J., 2016, Time delay cosmography, , 24, 11

- Trodden M., Carroll S. M., 2004, TASI Lectures: Introduction to Cosmology, ArXiv Astrophysics e-prints
- Troxel M. A., MacCrann N., Zuntz J., et al., 2017, Dark Energy Survey Year 1 Results: Cosmological Constraints from Cosmic Shear, ArXiv e-prints
- Umetsu K., 2010, Cluster Weak Gravitational Lensing, ArXiv e-prints
- Valls-Gabaud D., 2006, in Albert Einstein Century International Conference, edited by J.-M. Alimi, A. Füzfa, vol. 861 of American Institute of Physics Conference Series, 1163–1163
- van der Walt S., Colbert S. C., Varoquaux G., 2011, The NumPy Array: A Structure for Efficient Numerical Computation, *Computing in Science Engineering*, 13, 2, 22
- Van Waerbeke L., Benjamin J., Erben T., et al., 2013, CFHTLenS: mapping the large-scale structure with gravitational lensing, *MNRAS*, 433, 3373
- Van Waerbeke L., Hinshaw G., Murray N., 2014, Detection of warm and diffuse baryons in large scale structure from the cross correlation of gravitational lensing and the thermal Sunyaev-Zeldovich effect, *PRD*, 89, 2, 023508
- Vanderplas J., Connolly A., Ivezić Ž., Gray A., 2012, in Conference on Intelligent Data Understanding (CIDU), 47–54
- Wadadekar Y., 2005, Estimating Photometric Redshifts Using Support Vector Machines, *Publications of the Astronomical Society of the Pacific*, 117, 827, 79
- Walker T. P., Steigman G., Kang H.-S., Schramm D. M., Olive K. A., 1991, Primordial nucleosynthesis redux, *ApJ*, 376, 51
- Wang D., Zhang Y., Zhao Y., 2008, in Astronomical Data Analysis Software and Systems XVII, edited by R. W. Argyle, P. S. Bunclark, J. R. Lewis, vol. 394 of Astronomical Society of the Pacific Conference Series, 509
- Weller J., Albrecht A., 2002, Future supernovae observations as a probe of dark energy, *PRD*, 65, 10, 103512
- Witten I. H., Frank E., Hall M. A., 2011, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edn.
- Wu X., Kumar V., Ross Quinlan J., et al., 2007, Top 10 Algorithms in Data Mining, *Knowl. Inf. Syst.*, 14, 1, 1
- Yadav J. K., Bagla J. S., Khandai N., 2010, Fractal dimension as a measure of the scale of homogeneity, *MNRAS*, 405, 2009
- Zambom A. Z., Dias R., 2013, A Review of Kernel Density Estimation with Applications to Econometrics, *International Econometric Review (IER)*, 5, 1, 20
- Zel'Dovich Y. B., 1967, Cosmological Constant and Elementary Particles, *Soviet Journal of Experimental and Theoretical Physics Letters*, 6, 316

- Zeldovich Y. B., 1972, A hypothesis, unifying the structure and the entropy of the Universe, MNRAS, 160, 1P
- Zhang B. R., Childress M. J., Davis T. M., et al., 2017, A blinded determination of H_0 from low-redshift Type Ia supernovae, calibrated by Cepheid variables, ArXiv e-prints
- Zhang L., Yu Y., Zhang P., 2016, Non-negative matrix factorization for self-calibration of photometric redshift scatter in weak lensing surveys, ArXiv e-prints
- Zhang P., Pen U.-L., Bernstein G., 2010, Self-calibration of photometric redshift scatter in weak-lensing surveys, MNRAS, 405, 359
- Zuntz J., Paterno M., Jennings E., et al., 2015, CosmoSIS: Modular cosmological parameter estimation, Astronomy and Computing, 12, 45
- Zuntz J., Sheldon E., Samuroff S., et al., 2017, Dark Energy Survey Year 1 Results: Weak Lensing Shape Catalogues, ArXiv e-prints
- Zwicky F., 1933, Die Rotverschiebung von extragalaktischen Nebeln, Helvetica Physica Acta, 6, 110
- Zwicky F., 1937, On the Masses of Nebulae and of Clusters of Nebulae, ApJ, 86, 217

Acknowledgements

First of all, I would like to thank Stella Seitz and Ralf Bender for their support, that made this thesis possible. I am immensely grateful for the opportunities they gave me and their invaluable scientific advice.

Furthermore I would like to thank Ben Hoyle and Kerstin Paech for their help and support. I cannot thank you guys enough for your personal advice, scientific guidance and motivating words that helped me so much. I will never forget this.

I am immensely grateful to have the most awesome colleagues, that I can imagine and I thank all of you for the great time here in Munich. I would also like to especially thank Steffen Hagstotz and Romy Rehmann for valueable discussions.

I want to thank Kristina and my parents for all their love and tremendous patience with me. I love you with all my heart.

Finally I would like to thank all the people who believed in me and who influenced me in numerous ways. I would like to thank my friend Rolf who taught me so much about art, theatre and literature. I also thank my friends who fought alongside me during the years of exercise sheets, lectures, tutorials and lab courses. I also thank Kristinas parents for hosting me from time to time.

Lebenslauf

Markus Michael Rau
Pfeufferstrasse 47
81373 Munich, Germany
0049 176 78776617
markusmichael.rau@googlemail.com

CURRENT POSITION

PhD student, Cosmology and Astrophysics 2014 - Present
Supervisor: Ralf Bender
Ludwig-Maximilians-Universität, Munich, Germany
Max-Planck Institute for Extraterrestrial Physics, Munich, Germany

EDUCATION

Master of Physics, sehr gut (very good) 2012 - 2014
Specialized in Astrophysics
Ludwig-Maximilians-Universität, Munich, Germany

Bachelor of Physics 2009 - 2012
Ludwig-Maximilians-Universität, Munich, Germany

PUBLICATIONS

Markus Michael Rau, Ben Hoyle, Kerstin Paech, Stella Seitz
Correcting cosmological parameter biases for all redshift surveys induced by estimating and reweighting redshift distributions
MNRAS, 466, 2927-2938

Markus Michael Rau, Stella Seitz, Fabrice Brimiouille, Eibe Frank, Oliver Friedrich, Daniel Gruen, Ben Hoyle
Accurate photometric redshift probability density estimation - method comparison and application
MNRAS, 452, 3710-3725, October 2015

Paula Branco, Luis Torgo, Rita P. Ribeiro, Eibe Frank, Bernhard Pfahringer,
Markus Michael Rau

Learning Through Utility Optimization in Regression Tasks

accepted to the 4th IEEE International Conference on Data Science and Advanced Analytics
- DSAA2017 proceedings

M. Gatti, P. Vielzeuf, C. Davis, R. Cawthon, **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts - Methods and Systematics
Characterization

eprint arXiv:1709.00992, September 2017, submitted to MNRAS

B. Hoyle, D. Gruen, G. M. Bernstein, **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Redshift distributions of the weak lensing source galaxies
eprint arXiv:1708.01532, August 2017, submitted to MNRAS

M. A. Troxel, ..., **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Cosmological Constraints from Cosmic Shear
eprint arXiv:1708.01538, August 2017, submitted to PRD

C. Chang, ..., **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Curved-Sky Weak Lensing Mass Map
eprint arXiv:1708.01535, August 2017

J. Prat, ..., **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Galaxy-Galaxy Lensing
eprint arXiv:1708.01537, August 2017

A. Drlica-Wagner, ..., **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Photometric Data Set for Cosmology
eprint arXiv:1708.01531, August 2017, submitted to ApJS

DES Collaboration, ..., **M. M. Rau**, et al.

Dark Energy Survey Year 1 Results: Cosmological Constraints from Galaxy Clustering and
Weak Lensing

eprint arXiv:1708.01530, August 2017

Roman Zitlau, Ben Hoyle, Kerstin Paech, Jochen Weller,

Markus Michael Rau, Stella Seitz

Stacking for machine learning redshifts applied to SDSS galaxies

MNRAS, 460, 3152-3162, August 2016

Ben Hoyle, Kerstin Paech, **Markus Michael Rau**, Stella Seitz, Jochen Weller

Tuning target selection algorithms to improve galaxy redshift estimates

MNRAS, 458, 4498-4511, June 2016

Ben Hoyle, **Markus Michael Rau**, Kerstin Paech, Christopher Bonnett, Stella Seitz, Jochen Weller

Anomaly detection for machine learning redshifts applied to SDSS galaxies
MNRAS, 452, 4183-4194, October 2015

Ben Hoyle, **Markus Michael Rau**, Christopher Bonnett, Stella Seitz, Jochen Weller

Data augmentation for machine learning redshifts applied to SDSS galaxies
MNRAS, 450, 305-316, June 2015

Ben Hoyle, **Markus Michael Rau**, Roman Zitlau, Stella Seitz, Jochen Weller

Feature importance for machine learning redshifts applied to SDSS galaxies
MNRAS, 449, 1275-1283, May 2015

C. Sánchez, ..., **M. M. Rau**, et al.

Photometric redshift analysis in the Dark Energy Survey Science Verification data
MNRAS, 445, 1482-1506, December 2014