PROBABILISTIC PHOTOMETRIC REDSHIFTS IN THE ERA OF PETASCALE ASTRONOMY

BY

MATIAS CARRASCO KIND

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Astronomy
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Robert J. Brunner, Chair
Associate Professor Athol J. Kemball
Associate Professor Paul M. Ricker
Professor Jon J. Thaler

Abstract

With the growth of large photometric surveys, accurately estimating photometric redshifts, preferably as a probability density function (PDF), and fully understanding the implicit systematic uncertainties in this process has become increasingly important. These surveys are expected to obtain images of billions of distinct galaxies. As a result, storing and analyzing all of these photometric redshift PDFs will be non-trivial, and this challenge becomes even more severe if a survey plans to compute and store multiple different PDFs. In this thesis, we have developed an end-to-end framework that will compute accurate and robust photometric redshift PDFs for massive data sets by using two new, state-of-the-art machine learning techniques that are based on a random forest and a random atlas, respectively. By using data from several photometric surveys, we demonstrate the applicability of these new techniques, and we demonstrate that our new approach is among the best techniques currently available. We also show how different techniques can be combined by using novel Bayesian techniques to improve the photometric redshift precision to unprecedented levels while also presenting new approaches to better identify outliers. In addition, our framework provides supplementary information regarding the data being analyzed, including unbiased estimates of the accuracy of the technique without resorting to a validation data set, identification of poor photometric redshift areas within the parameter space occupied by the spectroscopic training data, and a quantification of the relative importance of the variables used during the estimation process. Furthermore, we present a new approach to represent and store photometric redshift PDFs by using a sparse representation with outstanding compression and reconstruction capabilities. We also demonstrate how this framework can also be directly incorporated into cosmological analyses. The new techniques presented in this thesis are crucial to enable the development of precision cosmology in the era of petascale astronomical surveys.

To the love of my life, my wife, the mother of our wonderful kids and my life partner Andrea.

Preface

This thesis is constructed as a compilation of articles in compliance with the rules for PhD thesis submission from the graduate research school at the University of Illinois at Urbana Champaign

The refereed publications arisen from this thesis are:

- **Carrasco Kind, M.**, & Brunner, R. J., 2013, *"TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests"*, MNRAS, 432, 1483 (Chapters 3 and 9)
- **Carrasco Kind, M.**, & Brunner, R. J., 2014, *"SOMz: photometric redshift PDFs with self-organizing maps and random atlas"*, MNRAS, 438, 3409 (Chapter 4)
- **Carrasco Kind, M.**, & Brunner, R. J., 2014, *"Sparse representation of photometric redshift probability density functions: preparing for petascale astronomy"*, MNRAS, 441, 3550 (Chapters 8 and 9)
- **Carrasco Kind, M.**, & Brunner, R. J., 2014, *"Exhausting the Information: Novel Bayesian Combination of Photometric Redshift PDFs"*, MNRAS, 442, 3380 (Chapters 6 and 7)
- Sanchéz, C., **Carrasco Kind, M.**, Lin, H., Miquel, R., et al. 2014, *"Photometric redshift analysis in the Dark Energy Survey Science Verification data"*, MNRAS accepted., arXiv: 1406.4407 (Chapter 9)
- Banerji, M., Jouvel, S., Lin, H., McMahon, R.G., Lahav, O., Castander, F., Abdalla, F., Bertin, E., Bosman, S., Carnero, A., **Carrasco Kind, M.**, et al. 2014, *"Combining Dark Energy Survey Science Verification Data with Near Infrared Data from the ESO VISTA Hemisphere Survey"*, MNRAS submitted., arXiv: 1407.3801 (Chapter 9)

Other non-refereed publications related to this thesis:

- **Carrasco Kind, M.** & Brunner, R. J. 2013, *"Implementing Probabilistic Photometric Redshifts"*, Astronomical Data Analysis Software and Systems XXII (ADASSXXII), ASPC, 475, 69C
- Newman, J., Abate, A., Abdalla, F., Allam, S., Allen, S., Ansari, R., Bailey, S., Barkhouse, W., Beers, T., Blanton, M., Brodwin, M., Brownstein, J., Brunner, B., **Carrasco Kind, M.**, Cervantes-Cota, J., Chisari, E., et al. Snowmass 2013 white paper, *"Spectroscopic Needs for Imaging Dark Energy Experiments:Photometric Redshift Training and Calibration"*, arXiv: 1309.5384
- Abate, A., et al. LSST-DESC white paper, 2012, *"LSST Dark Energy Science Collaboration"*, arXiv: 1211.0310

Acknowledgments

Personal acknowledgments

This thesis would certainly not have come to its completion without the help, support and trust of colleagues, friends and family. First, I would like to sincerely thank my advisor Robert J. Brunner for his help, guidance, advice, support and freedom during this work, for believing in me even before starting the thesis and for our extended brainstormings and interesting conversations about research, science, computing, family life and obviously soccer and for making this thesis a enjoyable experience. I am also grateful to the members of the thesis committee for their interest, comments and for taking the time to evaluate my work.

I also want to thank my friends and colleagues in the Astronomy department who contributed to create and maintain a friendly and stimulating working environment during these four years. To Gary for his friendship, his help, our conversations and for suffering with me through several classes during our coursework. To Rukmani for her friendship, support and for her coffee companion during these years and for making CosmoCoffee a real thing. To Manuel and David for our lunches and lively Spanish discussions and friendship. To Jessica, Dom, Nachi, Andy, Michael, Margaret, Ian, Ashley, Ricky, and many others for being very friendly, for their understanding that I wish I'd had more time to spend with them. Special thanks Yiran and Edward for their support, discussions and for making our office a very enjoyable and friendly place. I would also like to thank many other people in this department for their support, advice and help: Bryan Dunne, Mary Margaret, Sandie, Rebecca, Kevin, Paul Ricker, Athol Kemball, Tony Wong, You-Hua Chu, Brian Fields. Special thanks to Judy and in particular to Jeri for their substantial help and for making this department a home-like place to work.

I also want to acknowledge and thank my parents for their support from such a long distance, for doing the best in providing me with the best education and tools that helped to become the person I am today. Without their hard work, love, and support, none of this would have been possible. To my sisters for their love, support and for always being there for us. I would also like to sincerely thank my parents-in-law and sister-in-law who constantly traveled and helped us when we needed them. Their support and company

during the born and care of our three kids is unimaginable and without their help I couldn't have gotten through this thesis.

Last but definitely not least I would like to express my infinite gratitude to my wife Andrea, to whom this work is dedicated. Without her love, company, support, help, kindness, and advice I would not have been able to do this thesis. During these four years there was not a moment in which she wasn't supporting and encouraging me, especially on the toughest and stressful of times she was able to remain calm and to provide love and peace to overcome them. All of this while being responsible for the care of our three happy and active little monsters. For her admirable dedication, love and unconditional support I will be eternally grateful.

Technical acknowledgments

I gratefully acknowledge the use of the parallel computing resource provided by the Computational Science and Engineering Program at the University of Illinois. The CSE computing resource, provided as part of the Taub cluster, is devoted to high performance computing in engineering and science. This work also used resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. I also acknowledge the support from the National Science Foundation Grant No. AST-1313415. During two years of this thesis I have also been supported by the Computational Science and Engineering (CSE) fellowship at the University of Illinois at Urbana-Champaign.

Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G.

This work is also based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/IRFU, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This research used the facilities of the Canadian Astronomy Data Centre operated by the National Research Council of Canada with the support of the Canadian Space Agency. CFHTLenS data processing was made possible thanks to significant computing support from the NSERC Research Tools and Instruments grant program.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins Univer-

sity, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência e Tecnologia, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

Table of Contents

Chapter 1

Introduction

1.1 The need of distances to galaxies in cosmology

Two of the most important research topics in all of the physical sciences are dark energy and dark matter. Their complete picture remains one of the primary unsolved astrophysical problems (e.g., National Research Council, SMD 2010 Science Plan). Unraveling these mysteries is complicated because we can't observe dark energy or dark matter directly, only their effect on visible sources, like galaxies or supernovae. On the one hand, dark energy is responsible for the acceleration of the universe, with the first evidence observed from the study of distances to type Ia Supernovae (Riess et al., 1998; Perlmutter et al., 1999). On the other hand, dark matter provides the gravitational seed for the formation of astronomical objects like stars, galaxies and clusters of galaxies with several evidence for its existence (Trimble, 1987). The current cosmological picture is that these dark components dominate the formation and evolution of the large scale structures in the Universe, and many of the important parameters that describe the cosmological model of our Universe are well constrained by measurements of the temperature fluctuations in the Cosmic Microwave Background (CMB; Planck Collaboration et al., 2013). However, it is undeniable that there remains a need for these models to be constrained by the Universe at later times by studying how the mass and energy are distributed today within large scale structures. As a result, one fundamental probe in Cosmology is to measure the geometric and spatial three dimensional distribution of visible sources to infer the nature of these unseen components. In order to make significant progress in this manner, several hundred million galaxies are needed over a large area of the sky (Coil, 2013), therefore, galaxy surveys have become a major cosmological tool.

A large number of techniques have recently been developed to constrain parameters of the standard cosmological models by using observations from large photometric surveys, including primordial non-Gaussianity (e.g., Cunha et al., 2010), the distribution of dark matter in clusters (e.g., Simet et al., 2012) and its relationship to galaxies (e.g., Ross et al., 2010), and even constraining the neutrino mass (e.g., de Putter et al., 2012). The basic analyses that underlie these works can also be used to constrain dark energy via weak lensing tomography (e.g., Bernstein & Jain, 2004), photometric baryon acoustic oscillations (e.g., Zhan & Knox, 2006; Reid et al., 2010), the galaxy angular power spectrum (e.g., Ho et al., 2012), and the strong lensing of quasars (e.g., Coe & Moustakas, 2009).

All of these cosmological measurements are made by carefully measuring the spatial distribution of galaxies. But these methods have in common one very important aspect, they require the distance to the galaxies, which is a challenging task in astronomy. The distance between the galaxy and the observer is most accurately made by using a spectroscopic redshift, which is computed from the difference in the wavelengths of the emitted and detected light divided by the wavelength of the emitted light. Generally this is computed for a particular emission or absorption line (or a set of lines) which is a measurement of the recessional velocity of the galaxy due to the expansion of the Universe. Given a cosmological model, these redshifts can be translated to a physical distance. Therefore, spectroscopic galaxy surveys have played an important role in understanding the origin, composition, and evolution of our Universe. Surveys like the Sloan Digital Sky Survey (SDSS; York et al. 2000), WiggleZ (Drinkwater et al., 2010), and BOSS (Dawson et al., 2013) have imposed important constraints on the allowed parameter values of the standard cosmological model (e.g., Percival et al., 2010; Blake et al., 2011; Sánchez et al., 2013). However, spectroscopic measurements are considerably more difficult to obtain, and are, therefore, more expensive than photometric measurements, as they require long exposures in order to achieve sufficient signal-to-noise over a wide wavelength range. Also spectroscopic analyses are often limited by sample size, survey complexity (i.e., masks), and survey depth which has less effect on photometric ones.

As an example, while the Sloan Digital Sky Survey (SDSS; York et al. 2000) has taken millions of spectroscopic redshifts of galaxies to high precision (Aihara et al., 2011), the same survey has obtained detailed photometric measurements for a much larger sample of galaxies in considerably less time. This dichotomy will only grow with ongoing and planned surveys that are dominated by photometric observations.

1.2 Photometric redshifts

As a result, considerable attention has been focused on the estimation of redshifts by applying statistical techniques to the photometric observations of sources through different filters. These photometric redshift (hereafter photo-z) estimation techniques have become crucial for modern, multi-band digital surveys; and this need for fast and accurate photo-z estimation is becoming even more important for large photometric surveys like the Dark Energy Survey (DES[1]) and the Large Synoptic Survey Telescope (LSST[2]), which are

[1] http://www.darkenergysurvey.org/

[2] http://www.lsst.org/lsst/

probing galaxies that are often too faint to be spectroscopically observed. Adopting a photo-z approach allows cosmological measurements on galaxy samples that are currently at least a hundred times larger than comparable spectroscopic samples, that have relatively simple and uniform selection functions, and that extend to fainter flux limits and larger angular scales and thus probe much larger cosmic volumes. In summary, photo-z techniques provide a much higher number of galaxies with redshift estimates per unit telescope time than spectroscopic surveys (Hildebrandt et al., 2010).

With the growth of these large photometric surveys, the estimation of galaxy redshifts by using multi band photometry has grown significantly over the last two decades. The estimation of galaxy redshifts by using multi band photometry was first performed by Baum (1962), while Koo (1985) and Loh & Spillar (1986) were the first to compute galaxy redshifts by using digital photometric observations from charge coupled devices. Presently, there are many different methods for computing photometric redshifts (see, e.g., Hogg et al., 1998; Wang et al., 2008; Hildebrandt et al., 2010; Abdalla et al., 2011; Sánchez et al., 2014, for an updated comparison of current photometric redshift methods and public codes). These techniques can be broadly categorized as either template fitting algorithms or empirical training algorithms.

1.2.1 Template based techniques

The template fitting algorithms (e.g., Benítez, 2000; Bolzonella et al., 2000; Arnouts et al., 2002; Csabai et al., 2003; Coe et al., 2006; Ilbert et al., 2006; Feldmann et al., 2006; Brammer et al., 2008; Wolf, 2009; Assef et al., 2010; Sawicki, 2012) can either use empirical (e.g., Coleman et al., 1980; Kinney et al., 1996; Mannucci et al., 2001; Assef et al., 2010) or synthetic spectral templates (e.g., Bruzual & Charlot, 2003). These techniques estimate a photometric redshift by finding the best match between the observed magnitudes or colors and the synthetic magnitude or colors from the suite of templates that are sampled across the expected redshift range of the photometric observations.

The template fitting methods, which leverage model galaxy spectral energy distributions (SED), have been used extensively and are often preferred since once implemented they can be readily applied to new data by simply adopting the appropriate photometric filter transmission functions. This method is often preferred over empirical techniques as they can be applied without obtaining a high-quality spectroscopic training sample. Given a representative sample of template galaxy spectra, most of these techniques can reliably predict a photo-z, although the use of training data that includes known redshifts can improve these predictions (e.g., Benítez, 2000; Ilbert et al., 2006; Newman et al., 2013b). These techniques, however, are not exempt from uncertainties due to measurement errors on the survey filter transmission curves, mismatches when fitting the observed magnitudes or colors to template SEDs, and color-redshift degeneracies. Furthermore, template techniques generally become less reliable at high redshift where the uncertainties in galaxy SEDs increases, since the templates are often calibrated by using low redshift galaxies.

1.2.2 Machine learning techniques

Empirical training methods use a spectroscopic training data set to calibrate an algorithm that can be quickly applied to new photometric observations. Initially the training set was used to map a polynomial function between the colors and the redshift (e.g., Connolly et al., 1995; Brunner et al., 1997). More recently, this process has been extended to machine learning algorithms, including artificial neural networks (e.g., Collister & Lahav, 2004; Oyaizu et al., 2008b; Bonnett, 2013), boosted decision trees (e.g., Gerdes et al., 2010), random forest (e.g., Carliles et al., 2010; Carrasco Kind & Brunner, 2013c), nearest neighbors (e.g., Ball et al., 2007, 2008; Lima et al., 2008), spectral connectivity analysis (e.g., Freeman et al., 2009), Gaussian process (e.g., Way et al., 2009; Bonfield et al., 2010), support vector machines (e.g., Wadadekar, 2005), Quasi Newton Algorithm (e.g., Cavuoti et al., 2012; Brescia et al., 2014), and from analytical forms suggested by computational algorithms (e.g., Schmidt & Lipson, 2009; Krone-Martins et al., 2014). While only a few of these photo-z methods are publicly available, they all perform to a similar accuracy and provide only a single redshift estimate rather than a full redshift probability density function for each galaxy.

All of the aforementioned techniques can be categorized as supervised learning algorithms, where the input attributes (e.g., magnitudes or colors) are provided along with the desired outputs (e.g., redshift), which are all employed during the learning process. In this sense, the redshift information from the training set *supervises* the training phase. In Carrasco Kind & Brunner (2013a) we introduced a random forest technique to compute photo-zs by using prediction trees. In this supervised machine learning technique, the prediction trees use the values of the redshifts (from the spectroscopic sample) to determine the specific point and input dimension, which is an exact numerical value, at which the data will be divided into two branches. This process is repeated iteratively while building each tree in the forest.

On the other hand, an unsupervised machine learning photo-z technique does not use the desired outputs (e.g., redshifts from the spectroscopic sample) during the training process; thus no decisions are made based on this information. The only information used by the unsupervised algorithm are the input attributes themselves. A Self Organized Map (SOM): (Kohonen, 1990, 2001) is an unsupervised, neural network algorithm that is capable of projecting high-dimensional input data (e.g., the dimensions might represent the magnitudes, colors or other attributes of a galaxy) onto a low-dimensional (usually two dimensions are sufficient) map (Lawrence et al., 1999). Thus, a SOM corresponds to a nonlinear projection of the training data that attempts to preserve the topology of the input attributes from the multidimensional space.

Self organized maps have been utilized in several astronomical applications (e.g., Naim et al., 1997; Brett et al., 2004; in der Au et al., 2012; Fustes et al., 2013). Recently, Geach (2012) and Way & Klose (2012) have introduced the application of a SOM to compute a single photo-z estimator, providing strong evidence that this technique has distinct advantages and can also be extended to compute a photo-z PDF. The unsupervised nature of this approach provides a complementary tool to supervised algorithms, such as our previously mentioned work, thereby opening the possibility to develop a meta-classifier that uses multiple, complimentary approaches to improve the precision with which we can estimate photo-z PDFs. Another important characteristic of a SOM when applied to photo-z estimation is the ability to produce a structured ordering of the spectroscopic training data, since similar galaxies in the training sample are mapped to neighboring neural nodes in the trained feature map. The application of this technique for the classification of sources based on their location within the feature map, however, is still an underutilized tool.

When provided with a high quality spectroscopic training sample, empirical training techniques have been shown to have similar or even better performance (Collister & Lahav, 2004; Carrasco Kind & Brunner, 2014c). In addition, empirical techniques are generally simpler to apply to different data sets and frequently provide an improved quantification of any uncertainties, which can be encoded in a photo-z probability density function (PDF). They also have the additional advantage that it is easier to include extra information, such as galaxy profiles, concentration, angular sizes, or environmental properties, in addition to magnitudes or colors. These methods, however, are primarily reliable only within the limits of the training data, and sufficient caution must be exercised when extrapolating these algorithms beyond the limits of the training data.

1.2.3 Photometric redshift systematics

As the demand for more accurate photo-z methods has grown, techniques have branched out into new areas in order to improve the accuracy of photo-z estimation. While a complete understanding of the systematic uncertainties is needed for a reliable and accurate machine learning photo-z algorithm (see, e.g., Oyaizu et al., 2008a, for a discussion on photometric redshift errors), other issues have recently been recognized in the effort to generate the most accurate photometric redshifts. For example, Cunha et al. (2012a,b) analyzed the effect of systematics within the spectroscopic training data set that is used to estimate a galaxy photo-z. Likewise, other functionality that a modern photo-z algorithm should provide include an identification of outliers on the training set that lead to an incorrect estimation of a photo-z, an identification of the features within the training data that most strongly affect a photo-z estimate, and an identification of areas of parameter space (e.g., magnitudes, colors, and redshift ranges) that are under sampled by the training data.

The last two features are important to the design of photometric surveys, as they provide useful information to optimally and efficiently guide follow-up spectroscopy to generate the scientifically most useful training data set for these algorithms.

Sometimes systematics in the computation of photo-z are very hard to address due to their unknown nature, however we need to be able to identify those galaxies that might be contaminating our samples and develop an approach to identify, remove and minimize these outliers. These data play an important role in cosmological analyses as they can introduce biases in the model if they are not treated properly and a careful study needs to be done in order to handle them, which represents a challenging task by itself.

1.2.4 Probability density functions

Given the growth of photometric-only surveys, cosmological measurements will require the use of reliable photometric redshifts and a complete understanding of their uncertainties. As a result, photo-z methods will be most effective going forward if they can not only robustly provide a reliable redshift estimation but also a redshift probability density functions. Recently, particular attention has been focused on techniques that compute a full photo-z PDF for each galaxy. This is because a photo-z PDF contains more information than a single photo-z estimate, and the use of photo-z PDFs has been shown to improve the accuracy of cosmological measurements (e.g., Mandelbaum et al., 2008; Myers et al., 2009; Sheldon et al., 2012; Carnero et al., 2012; Jee et al., 2013) while not introducing any biases (e.g., Bordoloi et al., 2010; Abrahamse et al., 2011). For example, Myers et al. (2009) have shown that by using the full redshift PDF within a two-point angular quasar correlation function, as opposed to simply using a single redshift estimate, their measurement has been improved by a factor of nearly four, which is equivalent to increasing the survey volume by a similar factor. Likewise, Mandelbaum et al. (2008) discuss how the accuracy of photo-z and the inclusion of the photo-z PDF affect the calibration for weak lensing studies. Other recent studies (see, e.g., Sheth, 2007; van Breukelen & Clewley, 2009) have also demonstrated how a cosmological measurement can be improved by using a photo-z PDF. However, given the lack of reliable photo-z PDF estimation techniques, this areas remains relatively unexplored.

Given the importance of these photo-z PDFs, there is a present demand to compute them as efficiently and accurately as possible. Additional requirements include the need to understand the impact of systematics from the spectroscopic sample on the estimation of these PDFs (e.g., Oyaizu et al., 2008a; Cunha et al., 2012a,b), and to maximally reduce the fraction of catastrophic outliers (e.g., Gorecki et al., 2014). Considerable effort has, therefore, been put into both the development of different techniques and the exploration of new approaches in order to maximize the efficacy of photo-z PDF estimation. Yet, the combination of multiple, independent photo-z PDF techniques has not been exhaustively explored (e.g., Carrasco Kind & Brunner, 2013b; Dahlen et al., 2013).

1.2.5 Big data problem

One fact that all photometric surveys have in common is the need to efficiently handle an overwhelming quantity of imaging data. The reduction, analysis and storage of this data is a difficult problem; even with the growth of computational resources, efficiently handling these data remains a pressing problem. In particular, photo-z PDFs are currently computed on the summary catalogs that are produced by uniformly processing imaging data. But storing photo-z PDFs for billions of sources is a challenge in itself, which is further complicated if multiple, different photo-z techniques are desired or if different photo-z PDFs are generated by using different galaxy templates. This is a problem both for those managing the data archives and for the general community who desire to apply these photo-z PDF estimates to cosmological analyses. Thus the time is ripe to address this issue.

1.3 Thesis outline

In this thesis we address several of the aforementioned issues by using data from several spectroscopic and photometric surveys like SDSS, DEEP2, CFHTLens and DES that are introduced and detailed in Chapter 2.

In Chapter 3, we introduce TPZ (Trees for Photo-Z), a new, Python-based, machine learning, parallel code for estimating photometric redshift PDFs by using prediction trees and random forest techniques (Breiman et al., 1984; Breiman, 2001). Our approach is an ensemble learning method that generates several classifiers and combines their results into a final output. Prediction trees partition the multi-dimensional space recursively into smaller regions, which is terminated when a leaf only contains a few elements. Within these final leaves, our algorithm can leverage a simple model for the actual prediction, by using, for example, the mean value for a regression or the mode in a voting process as used in a classification scheme.

Likewise, the basic idea of a random forest method is to use bootstrap samples from the training data to build a set of prediction trees. These trees are constructed by selecting the best split point from a random subsample of the dimensions (e.g., magnitudes or colors) along which the data are subdivided. By aggregating the predictions from this forest of trees, we produce a more accurate estimate.

Later in Chapter 4, we extend the previous work of Geach (2012) and Way & Klose (2012) to use self organized maps to produce photo-z PDFs and to explore different configurations. This new work follows a similar approach presented in Chapter 3. Herein, we present a new ensemble learning method that generates multiple, different SOMs, and subsequently combines their results into a final output that is a probability distribution, which we call SOMz. In analogy to the random forest technique used by TPZ , we use bootstrap samples from the training data to build a set of unsupervised independent feature maps. By aggregating the predictions from this *atlas* of random maps, we produce a more accurate and robust final estimate.. Furthermore, we also explore the implementation of this algorithm by using three different two-dimensional topologies: a rectangular grid, a hexagonal grid, and a spherical surface corresponding to the 2D representation of the multidimensional training sample.

Given that the combination of multiple photo-z techniques is an underexplored area, we address this by using a third, and more standard technique, that employs template fitting to produced photo-z PDFs that is described in Chapter 5. In Chapter 6 we explicitly address this issue by presenting a novel Bayesian framework to combine and fully exploit different photo-z PDF techniques. In particular, we show that the combination of a standard template fitting technique with both a supervised and an unsupervised machine learning method can improve the overall accuracy over any single individual method. Finally, we show that this methodology can be easily extended to include additional, independent techniques and that we can maximize the complex information contained within a photometric galaxy sample. We also demonstrate how this combined approach can both reduce the number of outliers and improve the identification of catastrophic outliers when compared to the individual techniques, which is presented in Chapter 7.

Given the importance of photo-z PDFs as we enter the era of precision cosmology and the volume of data of current and future photometric surveys, we explore, in Chapter 8, different methods that allow us to manipulate and use photo-z PDFs in a more efficient manner by representing them as compactly as possible. We introduce the use of a sparse functional basis to represent a full photo-z PDF. This approach minimizes the data required to represent the photo-z PDF, while maximizing the accuracy of the PDF. This basis representation not only minimizes the storage requirements, but also allows us to manipulate PDFs in a more computationally efficient manner, thereby increasing the computational efficiency of resulting analyses. With this approach, each galaxy photo-z PDF is decomposed into an over determined basis system by minimizing the number of basis functions retained. We analyze how this approach compares with other representation techniques, in particular with a multi-Gaussian approach; and we demonstrate that, by using our proposed functional form, the integration and manipulation of photo-z PDFs is both easier and faster than when using either the original PDF or any other comparable technique.

We complete the thesis by showing applications that use photo-z PDFs and techniques described. First, we show how to compute the galaxy distribution $N(z)$, a very important measurement in cosmology, by using stacked photo-z PDF and using our sparse representation framework. Next, we show a recent application of our techniques in a photo-z code comparison analysis by using early data from the Dark Energy Survey. Finally we show a cosmological application on simulated data by computing the Angular Power Spectrum of the matter density. In Chapter 10 we present our general conclusions of this thesis, along with future research directions associated with this work.

Chapter 2

Photometric and spectroscopic data

Outline

During this thesis we have used data coming from several photometric and spectroscopic surveys that vary both in quantity and quality. In this chapter we briefly discuss these data sets and the specific data samples from each that we used during the training and testing process. It is also specified in which chapters these datasets are analyzed. The data sets used in this work, their abbreviations and chapters where is used are summarized in Table 2.

Table 2.1: Summary of the data used in this thesis and the specific chapters where it is used.

Name of subsample	Survey	No. of galaxies	Chapters where used
SL-1	SDSS	55,000	Chapter 3
SL-2	SDSS	1,147,397	Chapters 6, 7
PH-1	PHAT	1,984	Chapter 3
DP-1	DEEP2	20,227	Chapters 3, 4
DP-2	DEEP2	10,210	Chapters 6, 7
CF-1	CFHTLenS	49,868	Chapter 8
CF-2	CFHTLenS	1,000,000	Chapter 8
DS-1	DES	15,607	Chapter 9
DS-2	DES	25, 227, 559	Chapter 9

2.1 Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS; York et al., 2000) phases I, II and III conducted a photometric survey in the optical bands u, g, r, i, z that covered almost 14,000 square degrees, or approximately mor than one-fourth of the entire sky. The resultant photometric catalog contains photometry of over 10^8 galaxies, making the SDSS one of the largest surveys ever completed. The SDSS also conducted a spectroscopic survey of targets selected from the SDSS photometric catalog, obtaining spectra of about 10^6 low redshift galaxies.

In §3, we use a subset of the Main Galaxy Sample (MGS; Strauss et al., 2002) from the Data Release 7 catalog (Abazajian et al., 2009). Specifically, we selected 55,000 galaxies by using the online CasJobs website[1]. This spectroscopic data ranges from $z \approx 0.02$ up to $z \approx 0.3$ with a mean redshift of 0.1. From this sample, which we call SL-1, we randomly selected 15,000 galaxies to train the TPZ implementation, while holding the remaining 40,000 for testing. We note that this is a blind test, as the testing data are not used in any way to train or calibrate the TPZ algorithm. Of all the measured attributes in the SDSS photometric catalog, we have used only the four dimensions corresponding to the galaxy colors as derived by the extinction corrected model magnitudes : $u-g$, $g-r$, $r-i$, and $i-z$. We use the SDSS colors as opposed to the more commonly used magnitudes for this particular test to both demonstrate the flexibility of TPZ and to generate scientifically more interesting ancillary information.

In §6, we use a subset of the spectroscopic data contained within the Data Release 10 catalog (Ahn et al., 2013, SDSS-DR10), which includes over two million spectra of galaxies and quasars which include those taken as apart as the Baryonic Oscillation Spectroscopic Survey (BOSS) program (Dawson et al., 2013).

Specifically, we selected galaxies by using the online CasJobs website[2] and the following query from the DR10 data base:

```
SELECT spec.specObjID,
    gal.dered_u, gal.dered_g, gal.dered_r,
    gal.dered_i, gal.dered_z,
    gal.err_u, gal.err_g, gal.err_r,
    gal.err_i, gal.err_z,
    spec.z AS zs
INTO mydb.DR10_spec_clean_phot
FROM SpecObj AS spec
JOIN Galaxy AS gal
ON spec.specobjid = gal.specobjid,
    PhotoObj AS phot
WHERE spec.class = ‘GALAXY’ -- Spectroscopic class
                            -- (GALAXY, QSO, or STAR)
AND gal.objId = phot.ObjID
AND phot.CLEAN=1            -- Clean photometry flag
                            -- (1=clean, 0=unclean)
AND spec.zWarning = 0       -- Bitmask of warning
                            -- vaules; 0 means all
```

[1] http://casjobs.sdss.org/CasJobs/

[2] http://skyserver.sdss3.org/CasJobs/

```
-- is well
```

We also removed some additional bad photometric observations, such as the ones with larger photometric errors, ensured the redshift values were positive, and compute colors for the final catalog, which contains 1,147,397 galaxies. The spectroscopic data range from $z \approx 0.02$ up to $z \approx 0.8$. These data are dominated by the Main Galaxy Sample (MGS) at low redshifts, with mean redshift of $z \sim 0.1$, and by luminous red galaxies (LRG) at higher redshifts, with mean redshift of $z \sim 0.5$.

From this sample, which we call SL-2, we randomly selected 50,000 galaxies for training and hold the remaining 1,097,397 for testing. This training set corresponds to approximately 4.5% of the test set. We note that this is a blind test, as the testing data are not used in any way to train or calibrate the algorithms. Of all the measured attributes in the SDSS photometric catalog, we have only used the nine dimensions corresponding to the five galaxy, extinction corrected, model magnitudes and the four colors derived from these five magnitudes: u, g, r, i, z, $u-g$, $g-r$, $r-i$, and $i-z$.

2.2 PHoto-z Accuracy Testing Project

The PHoto-z Accuracy Testing (PHAT; Hildebrandt et al., 2010) project first compared the performance and systematics of different photo-z codes on synthetic data (PHAT0) that was specifically created for a contest, and also more recently used real data (PHAT1) in a similar manner; thereby providing a more realistic comparison by using real measurements. The PHAT project[3] provides filter responses for photo-z estimation by SED-fitting methods and a training data set for photo-z estimation by empirical methods. The true redshifts of the test data are not public, which provides a more reliable, blind comparison between different approaches (see Hildebrandt et al., 2010, for more details about the contest). In §3, we use the PHAT1 data, which consists of real observations selected from the Great Observatories Origins Deep Survey Northern field (GOODS-N; Giavalisco et al., 2004).

These data include photometry from the original ACS four-band data: F435W(B), F606W(V+R), F775W(i') and F850LP(z') that have been cross-matched with photometry from Capak et al. (2004), including U (from KPNO), B_J, V_J, R_C, I_C, $z^{'}$ (from SUBARU), and $HK^{'}$ (from QUIRC). In addition, the photometry of PHAT1 also includes Deep J and H bands (from ULBCAM; Wang et al., 2006), K_S (from WIRC; Bundy et al., 2005), and four Spitzer IRAC bands: 3.6, 4.5, 5.8, and 8.0μ. This photometric catalog was cross-matched with all available spectroscopic GOODS-N data (Cowie et al., 2004; Wirth et al., 2004; Treu et al., 2005; Reddy et al., 2006), producing a final data set of eighteen band photometry and spectroscopy for 1,984 galaxies, we will

[3] www.astro.caltech.edu/twiki_phat/bin/view/Main/WebHome

refer to this data set as PH-1.

For the contest, only 515 galaxy redshifts were published for use as training data; the remaining redshifts were unpublished and used internally by the PHAT project to conduct a blind comparison test. Despite the limited training data, multiple authors submitted the photo-z predictions and the results were published in Hildebrandt et al. (2010). As the contest had already been completed when we started this work, we were unable to participate. However, as discussed in chapter 3 we have tested TPZ on the PHAT1 training data in an analogous manner as the contest and have submitted our results to the official PHAT wiki.

2.3 Deep Extragalactic Evolutionary Probe

The DEEP survey is a multi-phase, deep spectroscopic survey that was performed with the Keck telescope. Phase I used the Low Resolution Imaging Spectrometer (LIRS) instrument (Oke et al., 1995), while phase II used the DEep Imaging Multi-Object Spectrograph (DEIMOS) (Faber et al., 2003). The DEEP2 Galaxy Redshift Survey is a magnitude limited spectroscopic survey of objects with $R_{AB} < 24.1$ (Davis et al., 2003; Newman et al., 2013a). The survey includes photometry in three bands from the Canada-France-Hawaii Telescope (CFHT) 12K: B, R, and I and it has been recently extended by cross-matching the data to other photometry databases. In this thesis, we use the Data Release 4 (Matthews et al., 2013), the latest DEEP2 release that includes secure and accurate spectroscopy for over 38,000 sources. The photometry for the sources in this catalog was expanded by using two u, g, r, i, and z surveys: the Canada-France-Hawaii Legacy Survey (CFHTLS; Gwyn, 2012), and the SDSS. For additional details about the photometric extension of the DEEP2 catalog, see Matthews et al. (2013).

To use the DEEP2 data in Chapters 3 and 4, we have selected sources with secure redshifts (ZQUALITY≥ 3), which were securely classified as galaxies, have no bad flags, and have full photometry. Even though the filter responses are similar, the u, g, r, i, and z photometry come from two different surveys and are thus not identical. We therefore treat those galaxies with SDSS photometry for fields 2, 3, and 4 of the DEEP2 target areas independently from those for field 1 with CFHTLS photometry. In the end, this leaves us with a total of 20,227 galaxies with eight band photometry and redshifts, from this data, which we call DP-1, we randomly select 10,000 for training and hold the rest for testing.

In chapter 6 we use a slight different data. Even though the filter responses are similar, the u, g, r, i, and z photometry originates from two different surveys and are thus not identical. We therefore only present in §6 the results from those galaxies that lie within field 1 that have CFHTLS photometry. Furthermore, we have corrected these observed magnitudes by using the extinction maps from Schlegel et al. (1998). In the end, this leaves us with a total of 10,210 galaxies each with eight band photometry and redshifts. From this data set, which we will refer as DP-2, we randomly select 5,000 galaxies for training and hold the remainder out for testing. The computation of photo-z PDFs was completed by using the magnitudes in the bands B, R, I, u, g, r, i, and z and their corresponding colors $B-R$, $R-I$, $u-g$, $g-r$, $r-i$, and $i-z$, providing a total of fourteen dimensions.

2.4 Canada-France-Hawaii Telescope Lensing Survey

We use data from the Canada-France-Hawaii Telescope Lensing Survey (Heymans et al., 2012; Erben et al., 2013), hereafter referred to as CFHTLenS[4], with the photometry presented in Hildebrandt et al. (2012). This galaxy survey includes more than twenty-five million objects observed in five photometric bands: u, g, r, i, and z, covering 154 square degrees that includes all five years worth of data from the Wide, Deep and Pre-survey components of the CFHT Legacy Survey (CFHTLS; Gwyn, 2012). To generate a spectroscopic training sample, we have cross matched the galaxies from the CFHTLenS with spectroscopic surveys whose sky coverage overlaps the four fields of the CFHTLenS survey.

In particular, we have selected high quality spectroscopic galaxies from the Deep Extragalactic Evolutionary Probe Phase 2 (DEEP2; Davis et al., 2007; Newman et al., 2013a), the VIMOS (VIsible imaging Multi-Object Spectrograph) VLT (Very Large Telescope) Deep Survey (VVDS; Le Fèvre et al., 2005; Garilli et al., 2008), the VIMOS Public Extragalactic Redshift Survey (VIPERS; Garilli et al., 2014), and the Sloan Digital Sky Survey Data Release 10 (Ahn et al., 2013, SDSS-DR10), which includes over two million spectra of galaxies and quasars taken as a part of the the Baryonic Oscillation Spectroscopic Survey (BOSS) program (Dawson et al., 2013). In the end, we have 49,868 high quality spectroscopic galaxies with a mean redshift of 0.6 to train our methods which we call CF-1. As the objective of chapter 8 is not to present photometric redshift PDFs for all of the CFHTLenS galaxies, we have randomly selected a subsample of 10^6 galaxies with no spectroscopic information from the survey that we use for the tests described in §8 which we call CF-2.

2.5 Dark Energy Survey

During the last part of this thesis we have also used data from the Science Verification (SV) period from the Dark Energy Survey[5] (DES; Flaugher, 2005) corresponding to observations carried out between late 2012 and early 2013 which provided science-quality images for almost 200 sq.deg at the nominal depth of the

[4] http://www.cfhtlens.org/

[5] http://www.darkenergysurvey.org/

survey. This galaxy survey is planned to cover approximately 5000 sq. deg. from the southern hemisphere to an unprecedented depth ($i_{AB} < 24$) and will include more then 300 million galaxies up to $z \sim 1.5$ in 5 photometric bands g, r, i, z and Y. The SV footprint was chosen to contain areas already covered by several deep spectroscopic galaxy surveys, including VVDS (Deep and Wide) (Le Fèvre et al., 2005, 2013; Garilli et al., 2008), ACES (Cooper et al., 2012), and zCOSMOS (Lilly et al., 2007, 2009) which together provide a suitable calibration sample for the DES photometric redshifts. Each one of the four overlapping fields (Sánchez et al., 2014) cover about the area of a single DES pointing, or about 3 sq. deg. The sample of photometric galaxies with spectroscopic information correspond to 15607 galaxies which is then separated in two for training and validations. We call this dataset DS-1. For a detailed description of this dataset as well as the photometric properties and reduction process, exposures times, please refer to Sánchez et al. (2014) where we not only describe the data in full detail but also performed a photometric redshift analysis on this data and address the current performance on DES data using several photo-z algorithms. In Banerji et al. (2014) we combine this data with infrared data from the ESO VISTA Hemisphere Survey survey (VHS; McMahon et al., 2013) to increase, among other improvements, the number of detected galaxies. The photometry of all objects observed during the SV period only are also used in this thesis and it correspond to approximately 25 millions of sources which will refer as DS-2 sample. This sample has shown to be of high quality as shown in Melchior et al. (2014) where this data is used to estimate the mass and the galaxy distribution of four galaxy clusters using weak lensing analysis. Both of these datasets are used in §9.

Chapter 3

Supervised machine learning for photo-z: TPZ

Outline

In this chapter, we present a new, publicly available, parallel, machine learning algorithm that generates photometric redshift PDFs by using prediction trees and random forest techniques, which we have named TPZ (Carrasco Kind & Brunner, 2013a) (Trees for photo-z). This new algorithm incorporates measurement errors into the calculation while also dealing efficiently with missing values in the data. In addition, our implementation of this algorithm provides supplementary information regarding the data being analyzed, including unbiased estimates of the accuracy of the technique without resorting to a validation data set, identification of poor photometric redshift areas within the parameter space occupied by the spectroscopic training data, a quantification of the relative importance of the variables used to construct the PDF, and a robust identification of outliers.

3.1 Prediction trees

Among the different non-linear methods that are used to compute photometric redshifts, prediction trees are one of the simplest yet most accurate techniques. Supervised learning methods using prediction trees, either classification or regression, have been shown to be one of the most accurate algorithms for low as well high multi-dimensional data (Caruana et al., 2008). They also are fast, can easily deal with missing data, and have similarities with other non-parametric techniques. For example, prediction trees are similar to k-nearest-neighbor (kNN) algorithms in that they both group data points with similar characteristics.

However, kNN use test data to identify similar points within the training set while keeping the parameter k fixed, even though some points might have a very different number of similar neighbors. On the other hand, prediction trees have terminal leaves that bound regions of the parameter space where the predictions (i.e., redshifts) and their properties (e.g., magnitudes) are similar. As both the quantity and identify of test data can vary between leaf (or terminal) nodes, prediction trees are known as *adaptive* nearest-neighbor methods (Breiman et al., 1984).

Prediction trees are built by asking a sequence of questions that recursively split the data, frequently into two branches, until a terminal leaf is created that meets a stopping criterion (e.g., a minimum leaf size). The small region bounding the data in the terminal leaf node represents a specific subsample of the entire data with similar properties. Within this leaf, a model is applied that provides a fairly comprehensible prediction, especially in situations where many variables may exist that interact in a nonlinear manner as is often the case with photo-z estimation. A visualization of an example tree generated by our technique is shown in Figure 3.1.

There are two classes of prediction trees (Breiman et al., 1984): classification and regression, both of which are implemented in TPZ .

(i) *Classification Trees (also called Decision Trees)*: As the name suggests, this type of prediction tree is designed to classify or predict a discrete category from the data. Each terminal leaf contains data that belongs to one or more classes. The prediction can be either a point prediction based on the mode of the classes inside that leaf or distributional by assigning probabilities for each category based on their empirically estimated relative frequencies. For example, in our photo-z technique we use the magnitudes or colors of galaxies to determine the probability that a galaxy lies either inside or outside a specific redshift bin (a detailed explanation of the algorithm is presented in §3.3.1).

The tree is built by starting with a single node that encompasses the entire data, and recursively splitting the data within a node into two or more branches along the dimension that provides the most information about the desired classes. Formally this is done by choosing the attribute that maximizes the *Information Gain* (I_G), which is defined in terms of the impurity degree index I_d:

$$I_G(T, M) = I_d(T) - \sum_{m \,\epsilon\, values(M)} \frac{|T_m|}{|T|} I_d(T_m) \tag{3.1}$$

where T is the training data in a given node, M is one of the possible dimensions (e.g., magnitudes) along which the node may be split, m are the possible values of a specific dimension M (in the case of magnitudes m might represent 2 or more magnitude bins), $|T|$ and $|T_m|$ are respectively the size of the total training data and the number of objects for a given subset m within the current node, and I_d is the function that represents the degree of impurity of the information.

There are three standard methods to compute the impurity index (I_d). The first method is by using the *information entropy*, which is defined in the expected manner (similar to Thermodynamics):

$$I_d(T) \equiv H(T) = -\sum_{i=1}^{n} f_i \log_2 f_i \tag{3.2}$$

where i is the class to be predicted (e.g., inside or outside a redshift bin) and the sum is over all n

possible classes (two in our example), and f_i is the fraction of the training data belonging to class i. The same definition applies for a subset of the data T_m.

The second option, is to measure the *Gini impurity* (G). In this case, a leaf is considered *pure* if all the data contained within it have the same class. The Gini impurity can be computed inside each node:

$$I_d(T) \equiv G(T) = \sum_{i=1}^{n} \sum_{j \neq i} f_i f_j \tag{3.3}$$

where f_i and f_j are the fractions of the training data of class i or j. The same equation applies for a subset of T along one particular dimension M. Since f_i are the fractions for all possible classes, we have that the $\sum_i f_i = 1$, and, therefore, $\sum_{j \neq i} f_j = 1 - f_i$. As a result, the expression for Equation 3.3 can be simplified to

$$I_d(T) \equiv G(T) = 1 - \sum_{i=1}^{n} f_i^2 \tag{3.4}$$

The third method is to simply measure the impurity degree by using the *classification error* (C_E):

$$I_d(T) \equiv C_E(T) = 1 - \max \{f_i\} \tag{3.5}$$

where the maximum values are taken among the fractions f_i within the data T that have class i. During the tree construction, the data are scanned over each dimension to determine the split point that maximizes the information gain as defined by Equation 3.1 and the attribute that maximizes this impurity index overall is selected. For example, Figure 3.2 shows these three impurity indices, for a node with data that are only categorized into two classes, as a function of the fraction of the data having a specific class. If all of the data belong to a specific class, the impurity is zero. On the other hand, if half of the data have one class and the remaining data all belong to the other class, the impurity is at its maximum. Our implementation can calculate any of these three different impurity indices, and any one of them can be selected for the construction of the prediction trees. Alternatively, the index providing the highest information gain at a given node can be selected.

(ii) *Regression Trees* A second type of prediction tree is used when the data to be predicted is continuous; since it does not use discrete classes, we instead fit a regression model to the data inside a leaf. The construction of a regression tree follows the same structure as the classification tree, and once again a node is generally divided into two branches (i.e., a binary tree). There are two primary differences, however, between regression and decision trees. First, each leaf has training data with different redshift values; the prediction value is based on a regression model covering these points. Usually, the mean of

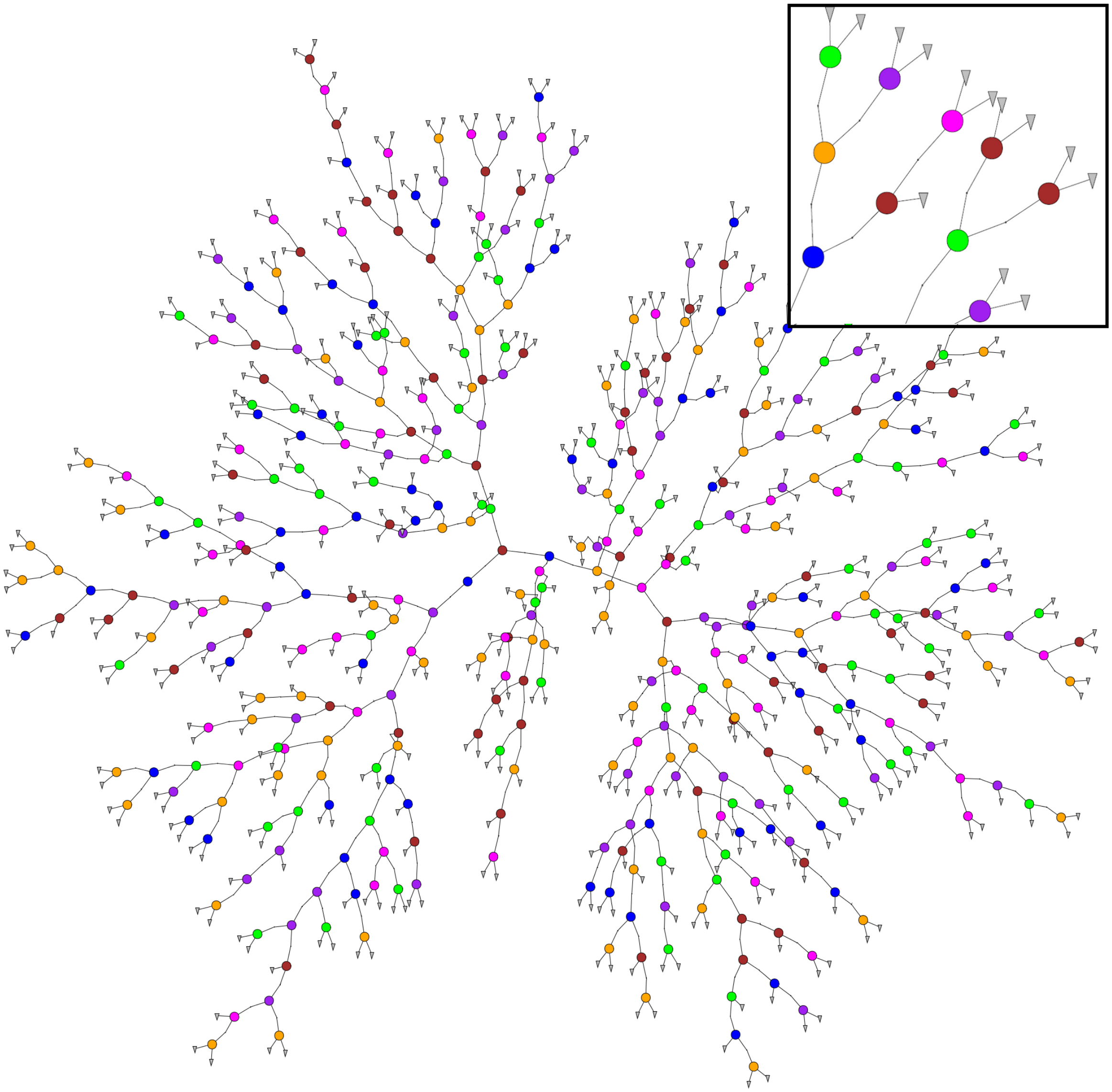


Figure 3.1: A simplified example of a binary prediction tree plotted radially. The initial node is close to the center of the figure. The splitting process terminates when a stopping criterion is reached. Individual colors represent the unique variable (e.g., fixed aperture g or r or magnitude colors) used for the splitting at each node. Each leaf provides a specific prediction based on the information contained within that terminal node (gray triangles in the figure). The subpanel corresponds to zoomed in region from the tree.

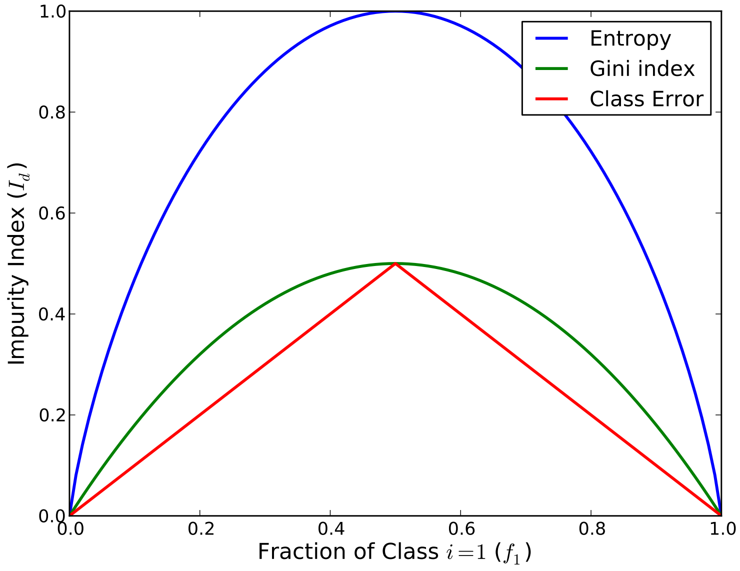


Figure 3.2: Impurity index I_d for a two-class example as a function of the probability of one of the classes f_1 using the information entropy(blue), Gini impurity (green) and classification error (red). In all cases, the impurity is at its maximum when the fraction of data within a node with class 1 is 0.5, and zero when all data are in the same category.

the training redshifts is returned, so each prediction is no longer a discrete classification, but is instead an estimation of a continuous variable. Second, the procedure used to select the best dimension to split for a regression tree is based on the minimization of the sum of the squared errors, which for a node T is given by

$$S(T) = \sum_{m \,\epsilon\, values(M)} \sum_{i \,\epsilon\, m} (z_i - \hat{z}_m)^2 \tag{3.6}$$

where m are the possible values (bins) of the dimension M, z_i are the values of the target variable on each branch/bin m, and $\hat{z}_m$ is the specific prediction model used. In the case of the *arithmetic mean*, we have that $\hat{z}_m = \frac{1}{n_m} \sum_{i \,\epsilon\, m} z_i$, where n_m are the members on branch m. This allows us to rewrite Equation 3.6 as

$$S(T) = \sum_{m \,\epsilon\, values(M)} n_m V_m \tag{3.7}$$

where V_m is the variance of the estimator $\hat{z}_m$.

At each node in our tree, we scan all dimensions to identify the split point that minimizes $S(T)$. The splitting dimension that has the lowest value of S is selected as the splitting direction, and this procedure is repeated until either some threshold in S is reached or any new nodes would contain less than the predefined minimum leaf size.

3.2 Random forest

Random forest is an *ensemble learning* algorithm that first generates many prediction trees and subsequently combines their predictions together. It is one of the most accurate empirically trained learning techniques for both low and high dimensional data (Caruana et al., 2008). The idea is simple, given a training sample T containing N objects that have M attributes (e.g., survey magnitudes), create N_T bootstrap samples of size N (i.e., N randomly selected objects with replacement). From these samples, we create the corresponding N_T prediction trees without pruning them back.

If all the variables are examined when deciding the best point to split, the method is called *bagging* (Breiman, 1996). An additional layer of randomness can be added to the bagging process by choosing the best split point from among a random subsample of $m_* < M$ variables at each node, where m_* is kept fixed during the process. The value of m_* is an adjustable parameter that is directly related to the *strength* of a tree (a strong tree has a low error rate) and the *correlation* between any two trees (the more correlated the trees, the higher the forest error rate). Increasing or reducing m_* has the same effect on both features. Of course we want to select the optimal value of m_*. A good starting point is to set $m_* \simeq \sqrt{M}$, although the accuracy of the algorithm is, in the end, not very sensitive to this parameter for a large number of trees and relatively small number of dimensions. After constructing all of the prediction trees, a final and robust prediction is calculated by combining all N_T estimates together.

Breiman (2001) first introduced this algorithm and showed that this technique performs very well when compared to many other learning techniques. This technique is robust against overfitting (i.e., there is no limit on the number of trees, N_T, in the forest), it runs efficiently on large data sets, it can generate an internal unbiased estimate of the error, and it can provide extra information about the relative importance of the input variables and the internal structure of the training data.

3.2.1 Ancillary information

Given a training set T, this extra, ancillary information can be calculated prior to the computation of the photo-z PDFs. As a result, we can use this *a priori* information to explore the efficacy of different parameter combinations while also obtaining an estimate of the bias and variance of the photo-z prediction. This is done by using *out-of-bag* (OOB) samples, which consist of a random sample of data that are left out of each tree. In the process of growing a forest, N_T trees are created using bootstrap samples of size N. In each of these samples, about one-third of the data are not used when constructing a tree, and are instead used as a test sample for the recently built tree. The test results created by using this OOB data are combined together to obtain estimators of the error, which, when built using a sufficiently large number of trees in the forest, has been shown to be unbiased and as accurate as using a validation set of the same size as the training set (Breiman, 1996). This removes, therefore, the need for a separate validation sample that can introduce a bias into the final result. This method also has the advantage of using the full spectroscopic data to compute PDFs.

The OOB data can also be used to estimate the relative importance of each attribute or dimension to the photo-z calculation. This provides an elegant method to identify and remove attributes that do not contribute significantly, thereby reducing the noise and dimensions of the problem. This also has the benefits of increasing the performance of the implementation, improving our understanding of the complexity in the interaction between different attributes, and improving the identification of new training data from, for example, follow-up observations. This relative importance is estimated for each attribute by first quantifying any variations in the prediction error when the OOB data are permuted only along the specific attribute, leaving the others unchanged. This process is repeated across all trees, and the end result is the average in the error increment when compared to the unperturbed variables for all the trees over the entire forest.

Another item we can construct is the proximity matrix, $Prox(i,j)$, which is a symmetric, positive definite matrix that gives the fraction of trees in the forest in which element i and j fall in the same terminal leaf. This matrix is constructed tree-by-tree by running all the data, both the OOB and the data used for growing, down each tree. When galaxy i and j are in the same leaf, their proximity is increased by one. At the end, all the proximities are normalized by the total number of trees; therefore, similar galaxies will tend to have higher proximities than dissimilar ones. This matrix can be computed for the training set, the test set, or both together. Since this matrix quantifies the relative similarity between galaxies, it can be used to identify outliers within a data set. For example, by computing the squared sum of all proximities for each galaxy, we can algorithmically identify galaxies with few neighbors by selecting sources with the lowest value, which can be flagged for further inspection.

To build or apply prediction trees, the data cannot have missing values for any of the attributes used to construct the trees (e.g., the most important survey magnitudes). To include more data into the classification process, we can use the proximity matrix to estimate any missing values or to replace highly uncertain values. We do this in an iterative process, by performing the forest growing step of the algorithm and replacing the missing attribute at each pass. We select the replacement value by computing the average parameter value from the k nearest galaxies; we can also inversely weight these galaxies by their respective distance. This process continues until we have obtained convergence or until a fixed number of iterations have been performed.

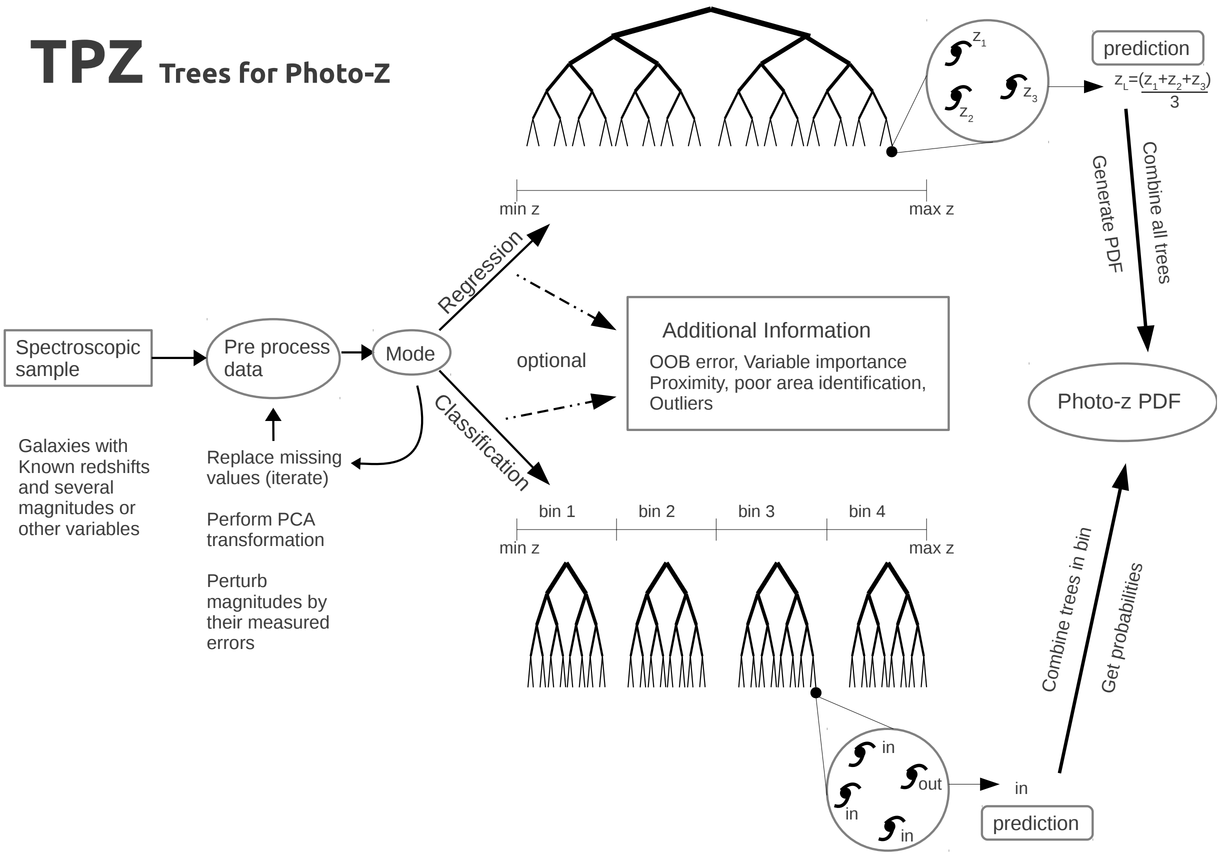


Figure 3.3: A simplified representation of TPZ , the details for each subprocess are described more fully within the text. Note that each tree drawn in the figure represents a full random forest with N_T bootstrap samples for every one of the N_R random perturbed samples. The big circles containing galaxies represent a terminal leaf, which are directly used to make a prediction for each new galaxy.

By using the proximity matrix, OOB error estimates, and the relevant importance of different attributes, we can also identify zones where the photo-z prediction is either poor or is loosely constrained by the training data. In either case, this knowledge is of vital importance when deciding what galaxies to target spectroscopically in order to optimally improve a training sample. One way this feature is implemented is by using the two most important attributes to map the areas of parameter space by their prediction error. This map can guide the identification of new data that increases the efficacy of the training sample by targeting those galaxies that minimize the prediction error in under sampled areas, thereby more effectively utilizing limited spectroscopic follow-up observations.

3.3 Previous work

Two previous works have utilized prediction trees for photo-z calculations. Carliles et al. (2008, 2010) predicted photometric redshifts and their errors by using a random forest built with the regression tree package for R by using the mean value as their leaf model. They used a subset of the main galaxy sample of the SDSS Data Release 6 (Adelman-McCarthy et al., 2008) catalog with colors as their attributes. They demonstrated that random forest methods are well suited to the photo-z estimation problem as they obtained comparable results to other machine learning methods, and they publicly released their R scripts. They did not, however, take full advantage of the ancillary information provided by the random forest technique, nor did they produce probability density functions.

Gerdes et al. (2010) have developed a new technique, called ArborZ, to compute photometric redshifts using boosted decision trees (BDT). These classification trees are constructed in a similar manner to our classification trees, as discussed in §3.1. In their approach, all data points start with equal weights, but after each tree is built, higher weights are assigned to points that were previously misclassified. This process iteratively combines weak classifiers into a single stronger one (Schapire et al., 1998); and, in the end, a weighted vote across the classifiers produces the final prediction. In their approach, they divide the redshift range into small bins and use an ensemble of BDTs to generate a probability distribution. A photometric redshift is estimated by determining the mean value of this distribution. They tested this algorithm on SDSS DR6 data as well as DES simulated data, finding similar performance to other empirical training methods, such as the photo-z estimates provided by Oyaizu et al. (2008b) in the case of the SDSS data, and by ANNz (Collister & Lahav, 2004) for the DES.

Our approach, detailed below, extends these previous results to create a new publicly available method that uses random forests to compute PDFs by using classification and/or regression trees. Our approach also uses extra information encoded within the measurement errors, generates extra, ancillary information describing the spectroscopic training sample, and provides a better control of the uncertainties. We also, therefore, are able to examine the importance of the attributes used to grow the trees, and identify areas in the attribute space where the training data are dominated by shot noise statistics.

3.3.1 TPZ Algorithm

Our implementation of prediction trees with random forest for photometric redshift PDF prediction, TPZ , is written in the Python[1] programming language and uses MPI for parallel communication to run efficiently on

[1] http://www.python.org/

distributed memory systems. As shown in Figure 3.3, our implementation is divided into three steps:

Data Pre-processing The *first step* prepares the data for the construction of the prediction trees. First, we optionally perform a principal component analysis (PCA) of the data in order to reduce strong correlations between attributes. This PCA transformation can reduce the dimensionality of the input data prior to the training, which can be important for large data sets with many attributes. This step also includes the replacement of missing values (explained later), which we do iteratively, finding that between 5–10 iterations leads to a convergence on the missing values. We next generate N_R training samples by perturbing the measured values according to the error on each variable, which we assume to be normally distributed. In this manner, we can incorporate the measurement error in the prediction tree construction, we reduce the bias on proximity matrices, and we introduce randomness into the construction of the trees in a controlled manner.

Random Forest Construction The *second step* is the actual construction of the random forest, where we generate fully grown prediction trees. We construct N_T trees by using bootstrapping with replacement for each perturbed sample in the set of N_R training samples we created in the first step generally the same size of the original training set. This step can be done several times with a smaller number of trees to both explore the parameter space and gain insight into the internal structure of the data prior to building the final prediction trees. Finally, this step can also produce the ancillary information that can characterize the performance of our code prior to estimating the final photo-z values.

Photo-z PDF Construction The *final step* uses the newly generated prediction trees to create individual photo-z PDFs for each source in the application data set. This process involves running each source down each tree, testing the source at each node until we arrive at a terminal leaf where we make a prediction. At the end, we combine all of the forest predictions into a probability density function.

3.3.2 Implementation modes

TPZ can use either type of prediction tree that uses random forests: classification or regression; the actual implementation details only differ after the first step.

Classification Mode: In this mode, the spectroscopic sample is divided into several redshift bins that either have a fixed width (or, alternatively, resolution), which allows a variable number of galaxies within each redshift bin, or have a fixed number of galaxies per redshift bin, which means our redshift bins are of variable width. Within each bin, we create a forest of classification trees, as described above, using the perturbed samples as well as the bootstrap samples. These trees classify an object as either lying *inside* or *outside* a bin. By using all of the training data within each bin, we both decrease the overall performance of our implementation due to the larger data volume and also increase the chance of catastrophic errors since most data will lie outside the bin of interest.

We address these issues by following a similar approach to that used by Gerdes et al. (2010). For each bin, we identify all sources that lie inside the bin. This number of galaxies with class *inside* is n_{in}. We next select a factor fn_{out} of the n_{in} galaxies that have spectroscopic redshifts that lie outside the bin by a factor of z_{out} times the width δz of the bin. This means that galaxies with class *outside* fall $z_{out} \times \delta z$ from the boundaries of the bin. This allows a better distinction between the class *inside* and the class *outside* as it would have if we include objects located very near to these boundaries. In the end, each bin will have $(1 + fn_{out})n_{in}$ galaxies available for training the forest.

If the training set is limited, wider bins can be used in order to have a sufficient number of training galaxies per bin. Furthermore, these bins can even be allowed to overlap by some value; this overlap can be taken into account when building the photo-z PDFs by normalizing by the fraction of wider bins that overlap with each other. After all of the forests are created for all of the bins, the test data are run down each tree in each forest, which assigns either the class *inside* or *outside* to the test source. After combining all of the assigned classes from the forest, we assign a probability for the source to belong to that redshift bin, which is simply the number of times the source was assigned the *inside* class divided by the total number of trees. By repeating this process for each bin and renormalizing the subsequent result, we generate a photo-z PDF for the source.

Regression Mode: In this mode, we use all available training data to fully grow each tree. For each perturbed sample, N_T trees are created using the methodology explained in §3.1(ii). At the end, there is one large random forest covering the entire spectroscopic range. The difference with the classification mode is that, after the tree has been constructed by splitting the nodes according to Equation 3.7, each terminal leaf only ends up with a few sources to make the prediction. In the simple case of obtaining a single estimate, this leaf can be replaced by the mean or the median of the values inside it; more generally, these values are kept for computing the PDF. To compute a photo-z, the test data are run down each tree in the forest. Each tree returns the set of spectroscopic redshift measurements that, after conversion to a given resolution, are converted into a PDF by normalizing to the total number of objects returned. All trees have the same weight when constructing the PDF, as well as the values of the terminal leaves identified in each tree. If a single value is desired, a mean value and its error can be returned via the standard methods by aggregating all of the relevant values as returned by the different trees.

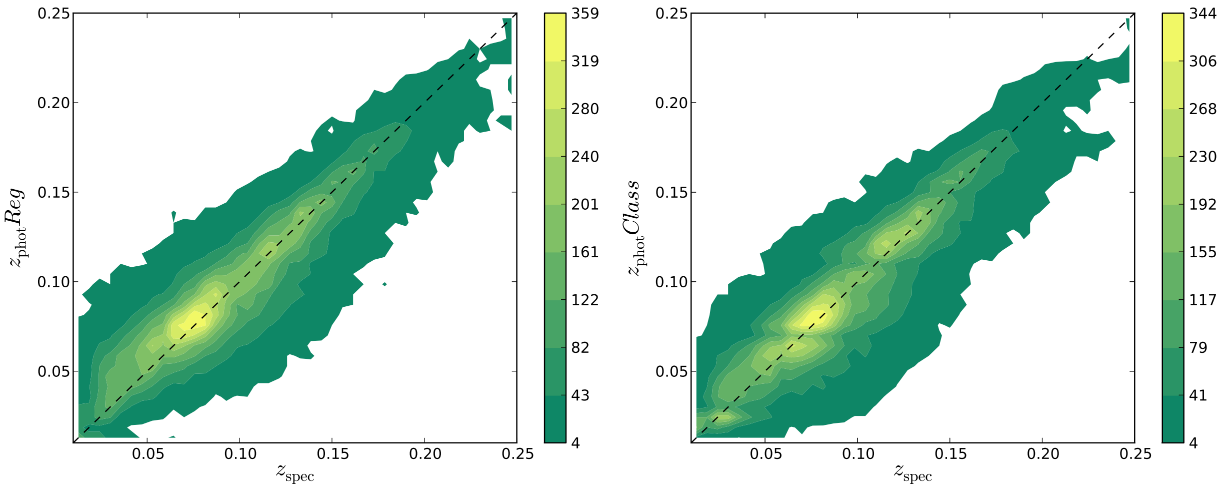


Figure 3.4: Photometric vs. spectroscopic redshift for all test SDSS MGS test galaxies using regression mode (*Left*) and classification mode (*Right*).

The choice of either of these modes will depend on the characteristics of the data being analyzed. On average, the regression mode runs faster than the classification mode for a specific accuracy, and is also better suited for data that are not uniformly distributed. The classification mode, on the other hand, provides a better characterization of the data as a function of redshift, since it creates its own random forest on each bin unlike the regression mode where a forest is created using the full range in redshift. The classification mode is also better suited for uniformly distributed data and can provide a reliable and robust prior probabilities in a Bayesian framework when using wider redshift bins. When faced with a high quality and rich training set, both modes will provide similar accuracies and error rates, but the regression mode, being faster, would generally be preferred.

Figure 3.3 shows a simplified workflow of our TPZ implementation. Each tree in this figure represents an entire forest, where the single tree results are averaged to get a final prediction. The classification mode predicts a probability that a source lies within each bin, thereby building up a photo-z PDF, while the regression mode keeps all sources found on a terminal leaf and combines their values to construct a photo-z PDF at the desired resolution. For both modes, ancillary information can be provided, and both modes share the same data pre-processing steps.

3.4 Application/Discussion

In this section, we apply the photo-z estimation technique presented in §3.3.1 to the SDSS main galaxy sample (SL-1), the PHAT1 blind test sample (PH-1), and the DEEP2 sample (DP-1), which were all introduced in §2.

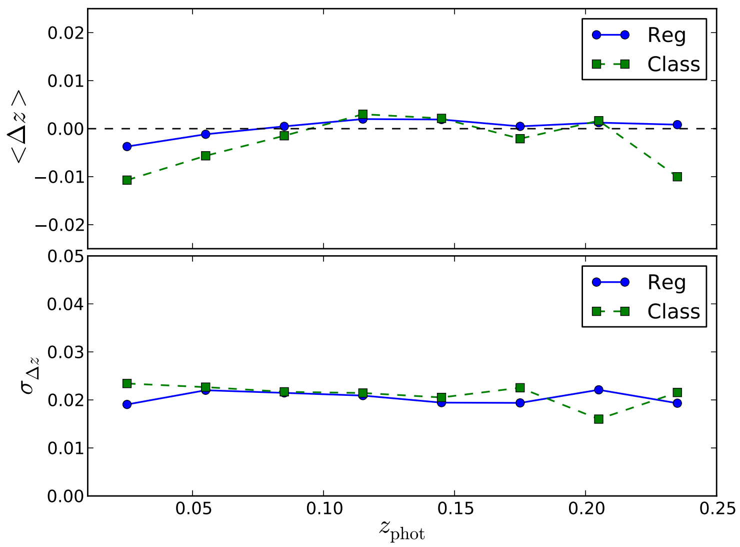


Figure 3.5: A comparison of the bias (upper panel) and the scatter (lower panel) as a function of redshift for the SDSS MGS data by using the regression mode (blue dots) and the classification mode (green squares).

Table 3.1: A comparison between the Regression Mode and the Classification mode for the SDSS MGS galaxies with different confidence level restrictions.

Implementation	$<\Delta z>$ $[10^{-3}]$	$\sigma_{\Delta z}[10^{-2}]$	Fraction[a]
Reg All	-0.08	2.25	100%
Class All	2.18	2.46	100%
Reg $zConf > 0.6$	-0.20	2.18	98.2%
Class $zConf > 0.6$	-2.15	2.34	94.2%
Reg $zConf > 0.75$	-0.33	1.97	91.0%
Class $zConf > 0.75$	-1.80	2.20	73.5%
Reg $zConf > 0.9$	-0.23	1.76	67.3%
Class $zConf > 0.9$	-0.92	1.82	34.7%

[a] Fraction of galaxies remaining after a cut on $zConf$.

Since the point of this chapter is to introduce the TPZ algorithm and our associated implementation, we use these three different data sets to highlight different features of the code. Thus we do not apply TPZ uniformly to each data set, and the three subsections herein are necessarily different.

3.4.1 SDSS Main Galaxy Sample: SL-1

We first apply TPZ to the SDSS SL-1, described in §2.1 using both the regression and the classification methods as explained in §3.3.2, and we present the results in Figure 3.4. The left and right panels compare the estimated photometric redshifts to the spectroscopic redshifts for all 40,000 galaxies held out for testing from the SL-1, for regression and classification modes, respectively. Both implementations show similar performance in the central part of the redshift distribution; however, there are differences at both the low and high redshift regions of this sample. Figure 3.5 shows both the mean of the bias, defined as $\Delta z = z_{\rm spec} - z_{\rm zphot}$, and its scatter for eight redshift bins. The regression mode performs slightly better at all redshift bins, but especially on the first and last bin, where the classification mode shows systematic errors in classification.

This error arises due to the lack of training data at those redshifts for the classification mode, where, though we allow some overlap between bins, we keep the bin size constant, which can result in large differences in the number of training objects per bin. This reduction is most pronounced in the lowest and highest redshift bins, which results in a lower accuracy and a higher scatter. We also are affected at the low redshift regime by the fact that a predicted redshift can not be negative, those introducing a positive skew to the predicted redshift values for very low redshifts.

Since both implementation modes produce photo-z PDFs, we can compute confidence levels, $zConf$, around the mean (or mode) for each individual PDF. To simplify comparisons with past results, we define $zConf$ as the integrated probability between $z_{\rm phot} \pm \sigma_{\rm TPZ}(1+z_{\rm phot})$. We select $\sigma_{\rm TPZ} = 0.03$ as an approximation to the intrinsic scatter of the algorithm when applied to the data, which can be computed by using the OOB data. Of course we could define $zConf$ in some other manner, but the results would be relatively unaffected. Figure 3.6 presents four different PDFs taken from the SL-1, each with different confidence levels that are shown as a bounded gray area under each PDF curve.

In this example, we measured $zConf$ around the mean of each PDF and the actual spectroscopic redshifts are shown as vertical dashed lines for reference. From this figure, we see that $zConf$ provides a reasonable summary of the concentration of the PDF, and can, therefore, be used to further restrict a photo-z sample by selecting only those PDFs with a $zConf$ value above some threshold. In general, as shown in this Figure, we see that lower confidence values are strongly correlated with less accurate predictions. Nevertheless, it is still possible to have a small fraction of galaxies with high $zConf$ PDFs that are estimated at the wrong redshift. We discuss the $zConf$ parameter and its use in identifying a clean galaxy sample in further detail in §3.4.3.

In Table 3.1, we present the mean value of the different performance metrics described in the previous paragraphs, as applied to the SDSS MGS, as well as the fraction of remain galaxies that remain in the sample after a cut on $zConf$. As before, we see that, on average, the regression mode outperforms the classification mode on this data set, although the difference is reduced when we apply a cut on the confidence level. Interestingly, at more restrictive $zConf$ cuts, the performance of both modes is similar; however, the number of galaxies remaining in the regression mode sample is higher. Note that since these are averaged values over the sample, any minor change implies a significant change on individual calculations.

As a result, we believe that making a cut on $zConf$ results in a cleaner sample, as shown by the improved

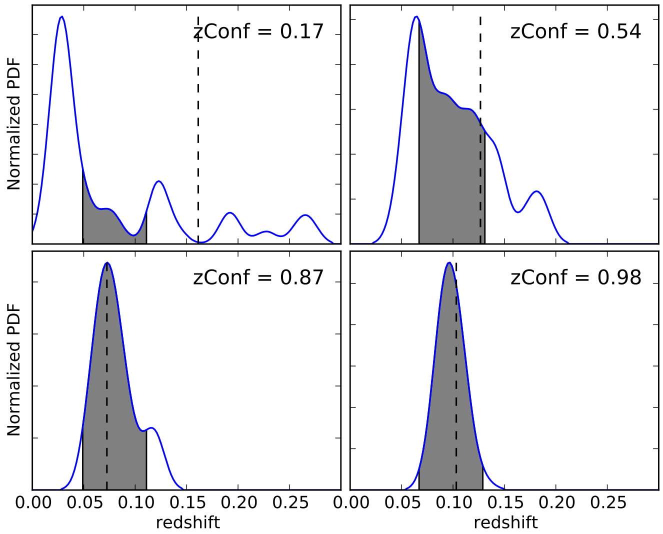


Figure 3.6: Four example PDFs produced by TPZ for the SDSS MGS selected with different values of $zConf$. The higher the value of $zConf$, the more narrowly concentrated the PDF is about the mean. The vertical dashed line corresponds to the spectroscopic value for the test galaxy and the gray area encloses the confidence level.

performance metrics for either implementation mode. The difference in the fraction of galaxies that remain in each sample indicates that, on average, PDFs generated by the classification implementation are broader than PDFs generated by the regression implementation. This result is reasonable, as the classification mode bins the redshift space and provides probabilities for all bins which can produce a more sparse distribution. In the classification mode the probabilities are computed individually for each redshift bin, which could be important and easily extended to build a prior distribution that can be used in a Bayesian method. Since the regression mode was shown to be more accurate for the SDSS (see, e.g., Figure 3.4 and Table 3.1), we use the mean of the PDF as calculated by the regression mode on the SDSS MGS data in the rest of this section, unless otherwise indicated.

We can broadly compare our use of $zConf$ to define clean galaxy samples to other published results; we note that a direct, one-to-one, comparison is problematic due to the different training sets and attributes used in computing photometric redshifts for the SDSS main galaxy sample. If we take a $zConf >= 0.75$, we keep 91% of the data and compute the fraction of galaxies with $|\Delta z| < z_i$, where $z_i =$ 0.001, 0.002 and 0.003 as 45.2%, 73.0% and 89.8%, respectively. These valued compare favorably to those from Laurino et al. (2011) who, even though they used an extended catalog, compute these same values to be 43.4 %, 72.4% and 86.9%, with a mean bias of $< \Delta z >= 15 \times 10^{-3}$ and $\sigma_{\Delta z} = 1.52 \times 10^{-2}$ (these latter values can be compared with our results shown in Table 3.1). Finally, we note that making a strict cut of $\Delta z > 0.006$ identifies an outlier fraction of 1.54%, while other groups, using extended catalogs as well, have reported

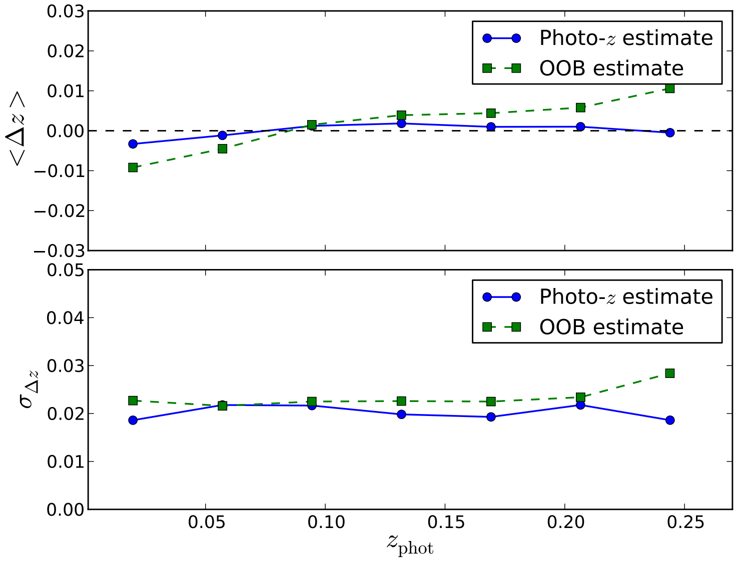


Figure 3.7: (*Top*) The averaged Δz as a function of redshift for all test galaxies from the SDSS MGS (blue circles) and from the OOB (Out-Of-Bag) data computed individually for each tree and subsequently averaged over the forest (green squares). (*Bottom*) The standard deviation of Δz as a function of redshift for the test set (blue circles) and the OOB data (green squares). In this case the OOB data provide a unbiased, upper-limit for these metrics.

values of 1.9% (Gerdes et al., 2010) and 2.6% (Oyaizu et al., 2008b).

Ancillary information

As detailed in §3.2.1, we can use the out-of-bag data to compute extra, ancillary information about the SL-1 dataset. For this purpose, we first select approximately one-third of the objects from each bootstrap sample. Using these data, we compute an unbiased indicator of the bias (i.e., Δz) and its standard deviation (i.e., $\sigma_{\Delta z}$) for each tree. Finally, we average these metrics over all trees. In Figure 3.7, we present in the top panel the mean bias as a function of redshift taken both from the test data (blue line) and from the OOB data used during the training process (green line). The bottom panel in this figure presents the standard deviation for each redshift bin. The RMS of these values provides an approximation to the intrinsic error and scatter of the TPZ code, which can be used to compute confidence levels. From the OOB data, we compute the RMS of the bias to be 0.0064, which can be compared to the value of 0.0017 obtained directly from TPZ for the SDSS MGS test sample. Likewise, we can approximate the scatter; for the OOB data we have 0.0235, while for the SDSS MGS test sample we have 0.0203. Thus, the OOB data provide upper limits for these metrics calculated by using only the training sample.

This OOB technique is unique due to the fact that the OOB data were not used to train a particular tree, yet the full data are used when building the forest by using the bootstrap samples. If we would have run all of the training data *after* the forest was constructed without using the OOB approach, we would have obtained biased (although lower values) for these metrics. This approach would thus not provide a prior estimation of the accuracy of TPZ . With the OOB data, we compute *a priori* these unbiased estimates exclusively from the training set, without the need for a validation set, allowing us to take full advantage of all available spectroscopic data.

The OOB data can also be used to compute the relative importance of each attribute, which can be done by permuting each of the attributes in the non OOB data when training the tree. The result of this process can be directly compared with the unperturbed case using the OOB data, as shown in Figure 3.8. In this figure, the left panel shows the relative importance factor, which is computed by using the absolute value of the OOB bias as a comparison metric, of the four colors used to build the regression trees for the MGS sample. In this plot, a factor of one implies that the attribute acts as a random variable, since a perturbation along that direction produces no changes. Any value greater than one produces a change in the bias, making it larger and therefore less accurate.

From this figure, we see that the $g-r$ color shows the largest relative importance factor, being close to four, meaning that the absolute bias, on average, changes by this same factor when this color is randomly perturbed. On the other hand, the $i-z$ color is the least, on average, relevant attribute in this context, with a relative importance factor less than 1.5. Due to the limited number of attributes in this test, however, removing this last color actually produces slightly worse results. In the general case when more attributes are present, removing less important variables will improve the results. While this result might seem counterintuitive, it results naturally from the random nature of the tree construction. Since only m attributes (e.g., three) are randomly selected to decide the split dimension, an attribute with overall low importance can be occasionally selected to split a node. By omitting attributes with lower importance, we force the trees to be built from attributes with greater information content, thereby improving the accuracy of the prediction.

Another interesting point is that the relative importance for both of the mentioned colors remain consistent, independent of redshift, while the other two colors show variation (i.e., $u-g$ and $r-i$ exchange importance ratings more than once), although they are overall consistent with each other. This behavior is mainly due to important spectral features, such as the 4000 Å break, passing between different filters, which TPZ identifies algorithmically, as important indicators of a galaxy's redshift. We see this result from another perspective in the central panel of Figure 3.8, which presents the RMS of the relative importance, sorted by their rank, for the four colors computed by using the absolute bias (blue line) and the variance (red line). Both metrics rank the attributes in the same order and either can be used to compute their importance to the data set. Perturbing the attributes produces a stronger effect on the absolute bias than on the scatter,

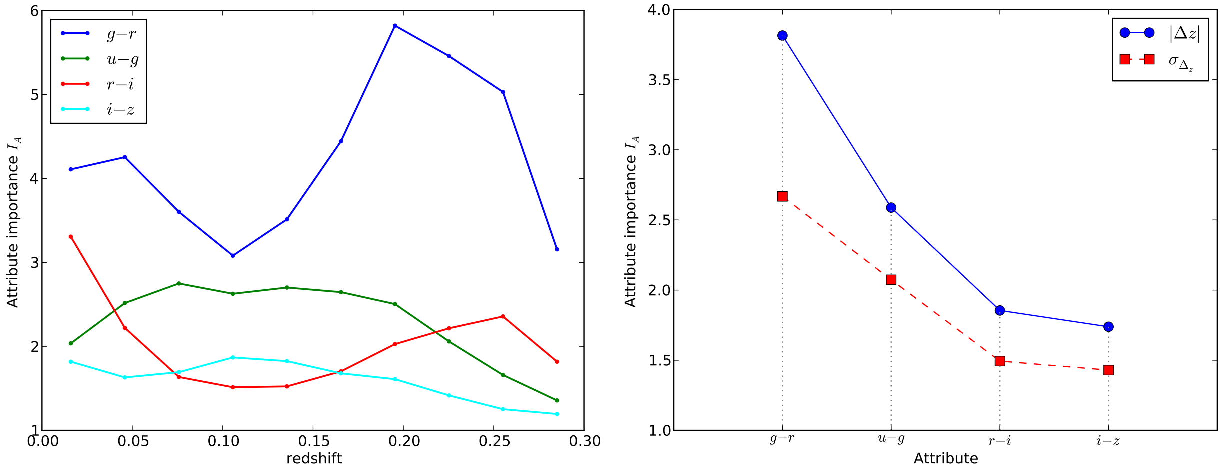


Figure 3.8: (*Left*): The attribute importance factor I_A as a function of redshift for the four attributes (SDSS colors) used in this analysis for the bias only. This factor quantifies how much the metrics decrease as we permute the attributes one at a time. (*Right*): RMS of the relative importance factor as a function of the attributes computed by using the bias (blue) and the scatter(red). (*Right*): A heat map constructed by using the two most important attributes, which indicates areas of parameter space where the photo-z prediction is poor. The higher the value (i.e., bluer) in a region, the more training data are needed to increase the accuracy of photo-z estimation within that region. These zones might also contain outliers or galaxies with bad photometry.

mainly because when perturbing one dimension, we lose information and thereby increase the likelihood that a galaxy will end up in a random branch of the tree, especially for an important attribute. This would likely lead to a misclassification, which directly affects the mean absolute bias.

Relative Importance

The importance rank can also be used to better understand the training data, to check whether it is possible to reduce the dimensionality of the problem, and to identify areas of the mapped parameter space where new training data can be most effectively incorporated. This latter point can be accomplished by identifying the leaf nodes, and the galaxies contained therein, for each tree and computing their accuracy on predicting for the OOB data along with their proximity matrices. By averaging over these results for all trees, we obtain the desired result.

For example, by using the two most important attributes previously identified for the SDSS MGS ($g-r$, and $u-g$), we present a heat map in Figure 3.9 that encodes the binned performance of these two attributes, where higher values indicate lower predictive success in that bin. In this plot, we see there are a few bins where performance is markedly lower (blue and light blue squares), and several areas that are lower than average (the yellow bins). On the other hand, there are two areas where the predictive power is quite high (deep orange-red), which are likely the result of the known color bi-modality of SDSS galaxies (Strateva

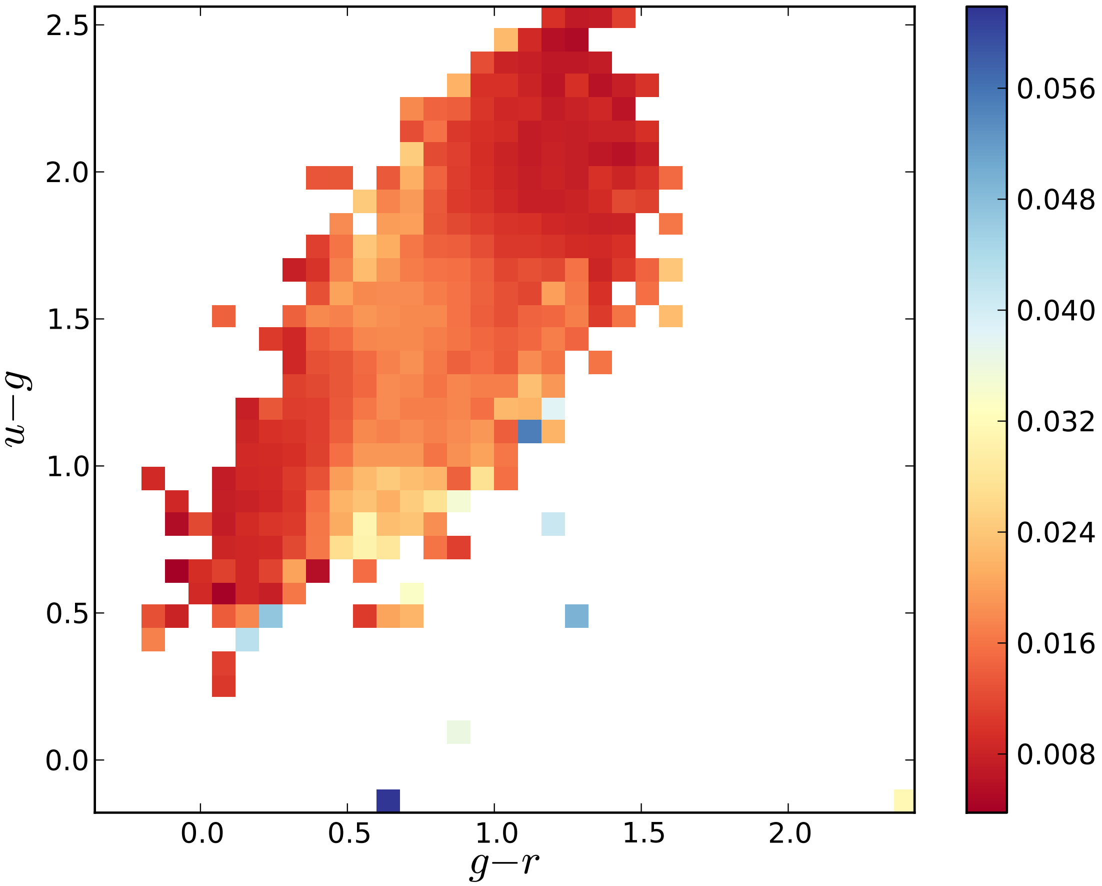


Figure 3.9: A heat map constructed by using the two most important attributes, which indicates areas of parameter space where the photo-z prediction is poor. The higher the value (i.e., bluer) in a region, the more training data are needed to increase the accuracy of photo-z estimation within that region. These zones might also contain outliers or galaxies with bad photometry.

et al., 2001) where early-type galaxies lie in the upper right part of this plot and late-type galaxies lie in the bottom left part of this plot. The areas in this heat map where the predictive performance is low can be caused by either a lack of training data, by galaxies with color degeneracies, or by galaxies with higher than normal magnitude errors. As a result, these areas can be prioritized for follow-up observations to improve the performance of the photo-z estimation.

Identifying new training data

Previously, we had stated that the relative importance of the different attributes, graphically shown in the heat map in Figure 3.9, could be used to optimally identify new training data. We test this assumption by first randomly selecting 1,000 galaxies as our training set, in order to simulate a poor training set, so that we can quantify the effects of both randomly adding new data and selectively adding new data by using the relative importance. We perform this test by first adding 1,000 new galaxies and second by adding 2,000 galaxies and computing the mean normalized bias, defined as $\Delta z' = (z_{\rm spec} - z_{\rm phot})/(1 + z_{\rm spec})$, and its standard deviation as we change the size of the training set by using the four color attributes from the SDSS MGS and and a forest with 100 prediction trees.

We summarize these test results in Table 3.2. As shown in the table, selecting galaxies from those zones with lower accuracy as indicated by the heat map produces more accurate predictions than adding galaxies randomly. In fact, even adding 1,000 galaxies by using the heat map produces a slightly better performance than adding 2,000 galaxies randomly. These results indicate that it is more important to selectively add galaxies to areas where the prediction is poor than to simply increase the size of the training set.

We continue this process, by continually adding either 1,000 or 2,000 new galaxies to the training set. As the bottom panel of Figure 3.11 for the SDSS MGS demonstrates, after about 5,000 galaxies (or at half the size of our full training set), the performance metric shows little variation, which is also reflected in the last row of Table 3.2 where the metrics for the 15,000 galaxy training set are presented for comparison. This test demonstrates how current and future photometric surveys can optimally construct training sets by either selectively using existing observations, or by obtaining new spectroscopic observations to improve the photo-z estimation.

Error distribution

After applying TPZ to the SDSS MGS, we can estimate photo-z errors directly from the estimated PDF by computing either the mean, the mode, or some other statistic from this distribution. As a demonstration, we calculate the error σ_{68} as the region of the photo-z PDF centered on the mean that contains 68% of the

Table 3.2: A comparison of the performance of TPZ for the SDSS MGS when extra data are added to the training set either randomly or by selectively using ancillary information.

Number of training galaxies	$< \Delta z' >$	$\sigma_{\Delta z'}$
1,000	-0.0043	0.042
1,000 + 1,000 from random	-0.0037	0.038
1,000 + 1,000 from map	-0.0033	0.032
1,000 + 2,000 from random	-0.0034	0.036
1,000 + 2,000 from map	-0.0022	0.025
15,000	-0.0018	0.021

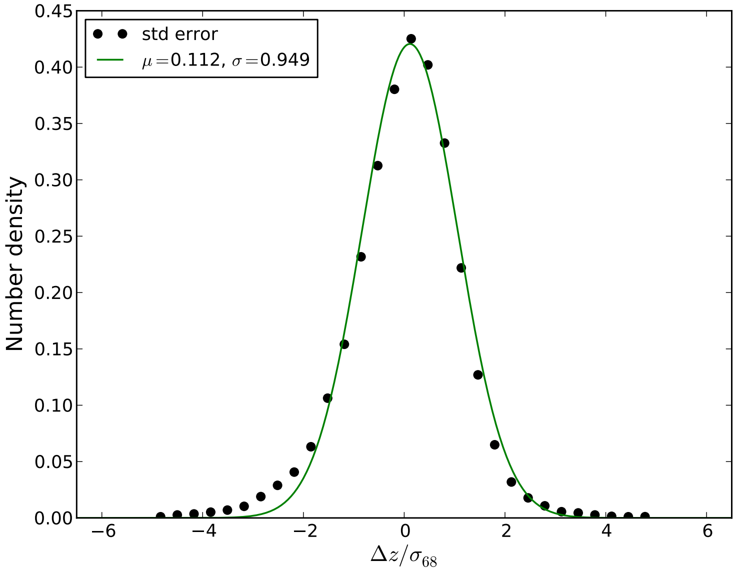


Figure 3.10: The photometric standardized error, $(z_{\rm phot} - z_{\rm spec})/\sigma_{68}$, for the MGS galaxies (black dots) using the mean of each individual PDF and the best fit Gaussian with $\mu = 0.112$ and $\sigma = 0.949$ (solid green curve).

cumulative probability. We next calculate the distribution of these standard errors by computing $(z_{\rm phot} - z_{\rm spec})/\sigma_{68}$ for each PDF, which is shown as the black points in Figure 3.10. For unbiased standard error estimates, this distribution should be normally distributed with zero mean and unit variance. When we fit our measured points, we obtain a Gaussian with mean equal to 0.112 and a width of 0.949, which is shown by the solid green curve.

This simple error estimate is quite close to the unbiased expectation, which is as we would expect for any reliable technique. The fit is not a perfect Gaussian due to a slightly extended tail on the left hand side of the distribution. We interpret this as a manifestation of the very narrow PDFs we have obtained and that the SDSS MGS is concentrated at lower redshifts where most photo-z techniques suffer from a small tendency to over-predict the photometric redshifts, as shown in the left panel of Figure 3.4.

Size of forest

When we construct a forest for prediction, one parameter that must be specified is the number of trees that should be constructed. This is important as the more trees in the forest, the higher the computational demands, which slows the training process and construction of photo-z PDFs. Thus, we test the performance of TPZ for the SDSS MGS by varying the number of trees built for our forest for a fixed-size training sample. As before, we compute the mean of the absolute bias and its standard deviation, and present how these quantities vary as the number of trees in our forest changes for a fixed training size of 10,000 galaxies.

These results are presented in the top panel of Figure 3.11, which shows that our algorithm does become more accurate as the number of trees increases. However, after around 100 trees, the predictive power of the forest shows little variation, indicating that this is a reasonable number of trees for this prediction process. Breiman (2001) demonstrated that, as the number of trees in a random forest increases, any margin function will converge to a limit value. Thus, as expected, we see our generalized error value converging. As a result, this implies that our technique does not over-fit the data as more trees are added in comparison to other machine learning methods.

Training size

Once we know the optimal number of trees that must be built for our forest, we next need to know the optimal size of our training set. By using 100 trees (as determined in the previous section), we vary the size of our training set and present the results in the bottom panel of Figure 3.11. As shown in this figure, the accuracy of TPZ for predicting photo-z does not change significantly after using around 70% of the galaxies. This is an interesting result, that our approach quantifies in an elegant manner, but which will obviously vary between different data sets. Fundamentally, as the training set increases in size, the prediction accuracy also increases until most of the multi-dimensional parameter space has been sampled and little extra information is added by new training galaxies.

Of course in this test we have not used the relative importance of our parameter attributes, as shown, for example, in the central panel of Figure 3.8. By manually selecting additional data, we should be able to reduce the values of these metrics significantly, which is discussed in the next section. But even in our current approach, we expect that some of our test data are not well represented in our training set, which will limit the accuracy of this approach. We see this as an opportunity, however, as we can compute a cross-data proximity matrix by using the trained forest to identify galaxies within the test data that are isolated with few neighbors in the parameter space. Once identified, these galaxies could be treated individually by using, for example, other photo-z estimation techniques (see, e.g., Carrasco Kind & Brunner (2014a)).

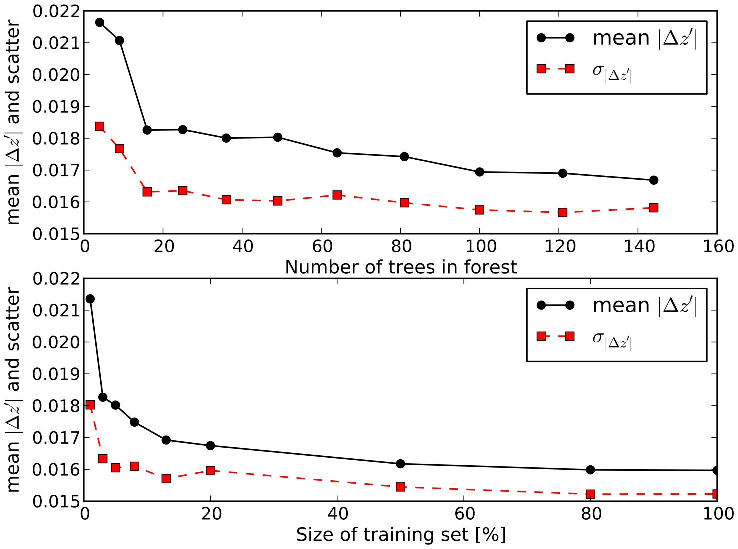


Figure 3.11: The absolute mean normalized bias defined as $|\Delta z'| = |(z_{\rm spec} - z_{\rm phot})|/(1 + z_{\rm spec})$, and its scatter as a function of the number of trees in the forest, keeping the training set fixed (top). The same two values as a function of the size of the training set keeping the number of trees fixed at 100 for galaxies in the SL-1 (bottom).

3.4.2 PHAT1 blind test: PH-1

We also tested TPZ on the PHAT1 dataset PH-1, described in §2.2, which is a blind contest where the test spectroscopic redshifts are unknown to the competitors. Therefore, this provides a reliable method to compare the performance of different photo-z techniques. In this contest, only a limited quantity of training data are provided; we have approximately 500 galaxies to train our algorithm for the approximately 1,500 galaxies that form the validation sample. These data also have a sparse redshift distribution, extending from $z \approx 0$ to $z \sim 6$. Despite these limitations, we applied TPZ to this training data, submitted our results to the contest, and obtained the resulting performance metrics from the PHAT leader (H. Hildebrandt, private communication). We present our specific results in Table 3.3, which can be compared directly with the results shown in Table 5 of the PHAT paper (Hildebrandt et al., 2010).

We computed validation results for four different photometric samples: by using all eighteen photometric bands, by omitting the Spitzer photometry and using only fourteen photometric bands, and by creating magnitude limited ($R < 24$) for each of these two galaxy samples. For these validation runs, we use the regression mode to create a forest of 150 trees with $m_* = 4$ (as described in §3.2). In all runs, we made no cuts on the $zConf$ parameter so that we could more directly compare our results to the other competitors. In the end, the TPZ results are among the most accurate photo-z predictions, especially when compared to other empirical training codes. Interestingly enough, TPZ even outperforms some template photo-z techniques,

Table 3.3: The TPZ results for the PHAT1 catalogue both with and without the IRAC bands, and for all galaxies and for a magnitude-limited sample with R < 24. Note that these are the same statistics presented in Table 5 of Hildebrandt et al. (2010) for other photo-z estimation techniques.

Run	bias[a]	scatter[b]	outlier rate[c]
18-band	-0.002	0.055	14.1 %
14-band	-0.007	0.055	12.6 %
18-band R $<$ 24	-0.004	0.055	11.1 %
14-band R $<$ 24	-0.009	0.054	9.6 %

[a] bias is defined as: $\Delta z' = \frac{z_{\rm spec}-z_{\rm phot}}{1+z_{\rm spec}}$

[b] RMS of the bias $\Delta z'$

[c] Outliers are defined as objects with $|\Delta z'| > 0.15$.

which are supposedly better suited for this particular challenge due to the dearth of training data and large redshift range covered by the validation sample. These results show that even in less than ideal conditions, TPZ provides a robust estimation of photometric redshifts. Note that due to the lack of training data and the extended redshift distribution of the validation sample, we did not generate ancillary information for the data by using the OOB approach.

3.4.3 DEEP2: DP-1

We have also tested TPZ by using the DEEP2 redshift survey data, which extends to much higher redshifts than the SL-1. As described in §2.3, we treat the galaxies with CFHTLS photometry independently from those with SDSS photometry, but in the end we merge the photo-z results. We follow a similar analysis to what we used with the SL-1, and after we compute the photo-z PDFs, we select only those galaxies with $zConf > 0.7$, which includes about 81% of the galaxies. We have that the average bias, using $\Delta z' = (z_{\rm spec} - z_{\rm phot})/(1 + z_{\rm spec})$, is -0.007 with $\sigma_{\Delta z'}$ = 0.059 and a outlier rate, defined as $|\Delta z'| > 0.15$ = 2.9%. We know of no previous photo-z analyses of these data (described in §2.3) with which to compare these results. The results are presented in Figure 3.12, which compares the photo-z computed by using the mean of each individual PDF with the spectroscopic redshift for the 7,856 galaxies. In this figure, we also compute the median, shown by the black dots, and the tenth and ninetieth percentiles, shown by the black error bars, within spectroscopic bins of width $\Delta z = 0.1$.

As this figure demonstrates, we see consistent results across all redshifts, and both the isodensity contours and the errors bars indicate that there are few outliers or catastrophic photo-z. However, at both ends of the distribution, we see several bins that show that the photo-z results are less accurate and are systematically higher for the first two bins and systematically lower for the last two bins. This effect is often seen with empirical techniques, as the spectroscopic training samples are often less complete at these redshifts, see, e.g.,

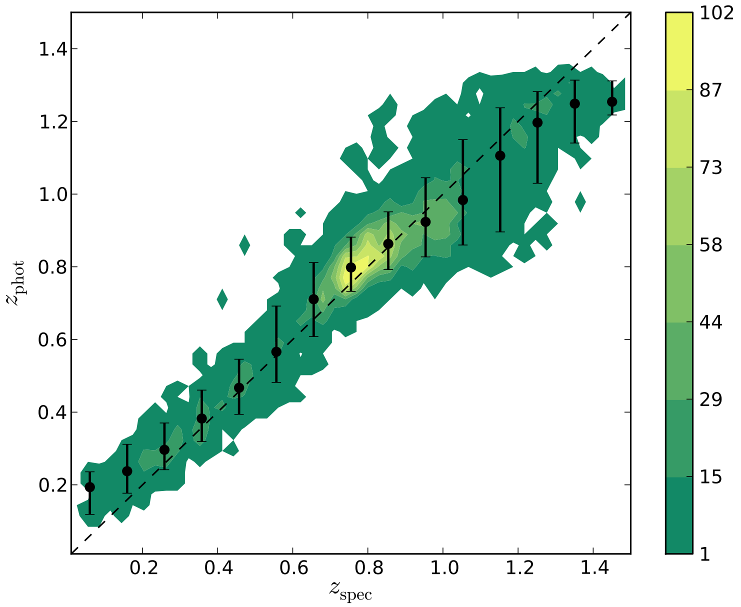


Figure 3.12: The TPZ photo-z with $zConf > 0.7$ versus the spectroscopic redshifts for 7,856 galaxies selected from the DEEP2 redshift survey. The black dots are the median values of z_{phot} and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic bin of width $\Delta z = 0.1$.

the redshift distribution in Figure 9.1. Another effect causing this skewness is that estimated photometric redshifts can not be negative, thus our probability distribution can not be symmetrical at the low redshift end. Another possible explanation for the low redshift systematic is the effect of galaxy inclination and the induced extinction on photo-z prediction as shown recently by Yip et al. (2011).

Likewise, the systematic underestimation at higher redshifts is likely affected by the fact that many of these galaxies are near the limit of the photometry and thus have higher than average magnitude errors. In combination with the lower density of training data, this will reduce the efficacy of our photo-z technique. To understand this effect, recall that our trees are built from objects whose photometry is sampled by assuming a normal distribution defined by the magnitude and magnitude error from the bootstrap samples. As the magnitude error increases, the range of possible values to sample increases, thereby producing a sparser sampling for this galaxy within our forest. Since there are few galaxies with redshifts above 1.3 in the training data, the branches on the forest for high-z galaxies are mainly dominated by training galaxies with redshifts closer to 1. As we build the PDF for the high-z galaxies, the PDFs will be positive skewed, and thus the mean value of each PDF will tend to be at lower redshift values.

We demonstrate this skewness in Figure 3.13, which shows the average skewness of the photo-z PDFs and the one-sigma error as a function of the spectroscopic redshift. These two quantities are computed as

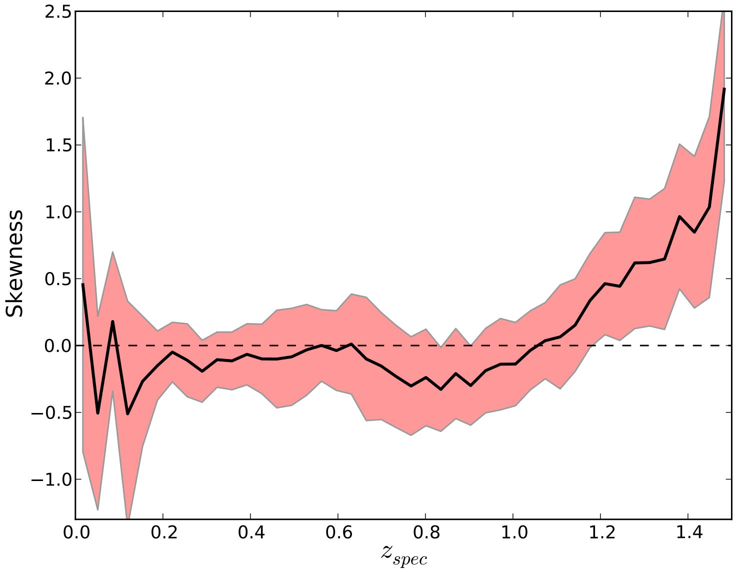


Figure 3.13: The skewness of the photo-z PDF as a function of spectroscopic redshift. The solid black line is the mean of the skewness and the pink shaded region corresponds to the one-σ interval. Positive skewness indicate a PDF skewed to lower redshifts.

the third standardized moment:

$$S_k = \int \left(\frac{z - \bar{z}}{\sigma_z} \right)^3 p(z)dz \tag{3.8}$$

with,

$$\bar{z} = \int zp(z)dz \qquad \text{and,} \qquad \sigma_z = \int (\mathrm{z} - \bar{\mathrm{z}})^2 \mathrm{p(z)dz} \tag{3.9}$$

where the integrals are computed over the redshift domain, and $p(z)$ is the photo-z PDF. We can see that for redshifts up to 1.1 the average skewness is very close to zero, showing a small trend to negative values, which will, on average, produce lower values for the mean photo-z. At higher redshifts, however, there is a clear increase in the average skewness, which will tend to produce lower values for the mean of the PDF. It is important to note that even though these PDFs may be (slightly) skewed, they still predict sufficient probability near the true redshift, information that is overlooked by other methods that use one point predictions. On the other hand, a catastrophic photo-z would have a symmetric PDF centered near the wrong redshift, which is not what we observe here.

Relative Importance

By using OOB data, we have computed the metrics from the training data that we compare in Table 3.4 to the metrics we obtained from the test data after the photo-z distributions were computed. The first two rows of this table show the complete results for all attributes for the DP-1 galaxies. From this we see

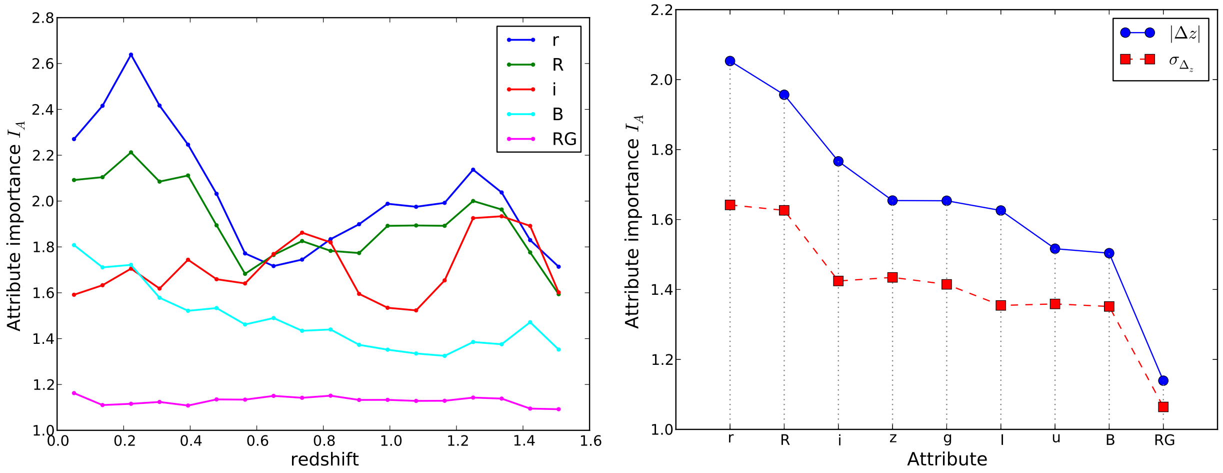


Figure 3.14: (*Left*): The variable importance factor, I_A, as a function of redshift for the three most and the two least important attributes (i.e., DEEP2 magnitudes) using the bias to quantify this importance. This index specifies how important an attribute is to the calculation of a metric when we permute the attributes one at a time. (*Right*): The RMS of the attribute importance factor as a function of the attributes computed by using the bias (blue) and its scatter (red). Both of these metrics capture the same relative attribute importance.

that there is strong agreement between the OOB and test data results for both the bias and the variance. We also computed the relative importance for the eight photometric bands and the RG attribute, which is the estimated R-band radius of an object in 0.207” pixels (i.e., the sigma of the Gaussian fit to the light distribution).

In the left panel of Figure 3.14, we present the attribute importance factor as a function of redshift for the three most and the two least important attributes. From this figure, we see that the R band and the r band are the most important attributes for making a photo-z prediction, similar to the $g - r$ color for the SDSS MGS. This demonstrates that by a pure statistical analysis, in the optical regime, the R band is the most effective attribute. These attributes show a peak in their importance between redshifts of 0.3 to 0.5. We interpret this increase to the presence of the 4000Åbreak being located at these redshifts within these filters. Likewise, the next two most important attributes are the i band and the z band, which are likely important for the same reason, albeit over a slightly higher redshift range.

On the other hand, the least two important attributes are the B band and the RG attribute. As shown in this figure, the RG attribute does not contribute to the photo-z prediction, instead acting like a random variable and thus is likely introducing extra noise into the calculation. We also see no clear evidence that the effect of this attribute changes with redshift. At low redshifts, this attribute could be affected by inclination angle or spectral type, while at higher redshifts, galaxies tend to be fainter and thus have smaller angular sizes. Presumably, these cumulative effects combine to erase any important information this attribute might provide to the photo-z calculation.

Table 3.4: A comparison in the accuracy of photo-z predication by using different attribute combinations from the DEEP2 data for all test galaxies. The first row are the metrics for TPZ using only OOB data, which are comparable to the values obtained from the full test data, shown in the second row. The remaining rows provide these metrics for training data that have had the indicated number of attributes removed from the calculation.

Attribute Selection	$< \lvert\Delta z'\rvert >$	$\sigma_{\Delta z}$
All attributes (OOB metrics)	0.052	0.053
All attributes	0.047	0.049
Remove 2 least important	0.044	0.046
Remove 2 most important	0.061	0.068
Remove 4 least important	0.044	0.048
Remove 4 most important	0.070	0.084

In the right panel of Figure 3.14, we present the mean relative importance for each attribute computed from the changes in the mean (blue circles) and the scatter (red squares) when this attributed is permuted, similar to the central panel of Figure 3.8. Once again, both metrics agree with the importance ranking. In order to characterize the attributes and their computed ranking of importance, we have made the following tests. First, we removed the two least important attributes, after which we remove the two most important attributes, reincorporating the previously removed attributes. We repeat this process, but now we remove alternately the four least and most important attributes. In each case, we test whether TPZ is able to correctly recognize the attribute importances. These results are summarized in Table 3.4, where we use the absolute mean value of $\Delta z'$ and its dispersion.

As is not surprising, we see that removing the two least important attributes does, in effect, improve the precision of TPZ while also making the code run faster since we have fewer dimensions to check when splitting nodes within the tree, less data to keep in memory when building the tree thus improving cache access, and random realizations from the input parameters will be faster since there are fewer dimensions to sample. Yet, removing four attributes shows a slight decrease in the overall performance, in this case we have removed too much information. While this decrease might seem rather small, since we are randomly selecting attributes when splitting nodes within the trees, by removing four we have increased the scatter since we are losing information. On the other hand, removing the most important attributes significantly affects the results, regardless of how many attributes we remove. As we would expect, the reason is clear. Since these attributes have the most information needed to subdivide the multidimensional parameter space in order to produce accurate photo-z, removing them negatively impacts the performance of TPZ .

In a further control test we added two extra artificial variables to the data set, one of which is strongly dependent on the source redshift, i.e. a function of redshift, while the second one is a uniformly distributed random variable. After computing their importance rankings, we can see from Figure 3.15 that TPZ recognized these two extra attributes and put them on the extreme limits of the importance ranking. The most important value ranks at about eight, the random variable ranks at one as expected, while the r magnitude is ranked with a value close to two. We notice that the variable RG is very close to one, and therefore to a random variable. As discussed above, we can safely remove this variable from our calculation as it does not provide any useful information. The legend on the plot indicates also the descending order in importance, in concordance with Figure 3.14.

Missing data

One interesting capability of TPZ is that it can be used to replace attributes in data that are either missing attributes or have attributes with large uncertainties. As discussed in §3.2.1, the replacement values can be computed from the proximity matrix, and we can apply this technique to data either in the training sample or in the application sample. In the former case, missing attributes would be replaced in order to maximize the size of the training set. The alternative would be to simply cull data with missing attributes from the training sample, which would decrease the robustness of our predictive power. In the latter case, missing attributes would be replaced in order to estimate a photometric redshift for a galaxy based on the incomplete but available information. In most cases, this will still result in a reliable prediction, without discarding any data, thereby increasing the overall statistical power of our approach.

To demonstrate this capability, we selected training and testing data sets that initially were complete and had relatively small errors (i.e., magnitude errors < 1 magnitude). We first randomly replaced 50% of the magnitudes in the training data with a bad value (e.g., 99), thus some galaxies in this sample have multiple bad attributes. From this new data set, we apply TPZ to generate a second training sample where the bad attributes have been replaced, using only six iterations (i.e, the replaced sample), and we also generate a third training sample where we simply remove any galaxies with missing or bad attributes (i.e., the cut sample). Likewise, we also generated a test sample with 50% of the attributes replaced by the bad value (i.e., the bad test sample)

We estimate photo-zs for the clean test sample by using all three training samples: the original, clean sample (i.e., the control), the replaced attribute sample, and the cut sample. Likewise, we use the clean training sample to replace missing attributes and estimate photo-zs for the bad test sample. We present the results of these tests in Table 3.5, where we compare the photo-z estimation for the clean sample with the replaced and cut samples. For this comparison, we use $\Delta z_{pp} = z_{\rm phot,clean} - z_{\rm phot,other}$ and $\Delta mag = mag_{\rm clean} - mag_{\rm other}$, along with their variances, where *other* can either be the replaced or cut samples. As shown in this Table, the replaced value sample produces, on average, superior photo-zs than the cut sample.

Table 3.5: Photo-z estimation metrics to demonstrate the robustness of our missing attribute technique. The first two rows show the average bias and its variance between the estimated photo-z , and replaced magnitude when either removing or recovering bad data in comparison with photo-zs predicted using the original, clean sample. The last row shows the the same metrics calculated by using the clean test sample, but for data missing in the test sample as compared to the clean original sample.

Recovered train	$< \Delta z_{pp} >$[a]	$\sigma^2_{\Delta z_{pp}}$[b]	Δmag	$\sigma^2_{\Delta mag}$
with removed data	-1.27	3.5	–	–
with recovered data	0.40	1.6	0.021	0.094
Recovered test	$< \Delta z_{pp} >$	$\sigma^2_{\Delta z_{pp}}$	Δmag	$\sigma^2_{\Delta mag}$
with recovered data	0.72	4.5	0.033	0.12

[a] in units of 10^{-3}
[b] in units of 10^{-3}

Likewise, we have estimated robust photo-zs for the bad test sample, which significantly increases the size of our resulting test data. Dealing with missing attributes is important , especially when a spectroscopic training sample is limited or when cross-matching between incomplete catalogs is carried out in order to develop a more complete catalog for photo-z estimation.

Photo-z PDFs and $zConf$

As discussed in §3.4.1, the $zConf$ parameter can be used to identify galaxies with narrow, concentrated photo-z PDFs, which ideally will result in galaxy samples that have the most accurate photo-z estimates. The $zConf$ parameter is demonstrated for DEEP2 galaxies in the left panel of Figure 3.16, which shows four representative photo-z PDFs selected with different values of $zConf$ as measured about the mean of each PDF. Both this figure and Figure 3.6, which presents four photo-z PDFs by using SDSS data, highlight the fact that wide and sparse distributions have low $zConf$ values while narrower PDFs have higher $zConf$ values.

The goal of a parameter like $zConf$ is to algorithmically identify galaxies that have, on average, the most accurate photo-z estimates. To test this hypothesis, we used all available DEEP2 training data to bud our prediction trees and estimate photo-zs for the DEEP2 test sample. From this sample, we applied three $zConf$ cuts: 0.5, 0.7, and 0.9, and calculated the bias and scatter as a function of redshift for the three resulting galaxy samples. We compare these results to the bias and scatter when no $zConf$ cut is applied in the right hand panel of Figure 3.16. As shown in this figure, both the mean absolute bias and the scatter are reduced as $zConf$ is increased, independent of redshift.

A simple, intuitive approach to select galaxies by their $zConf$ would be 0.5 as this selects galaxies that have a 50% probability that their photo-z redshift estimate lies within the limits imposed by $\pm\sigma_{\text{TPZ}}(1 + z_{\text{phot}})$. Furthermore, higher values would provide more accurate results at the expense of reduced statistical power

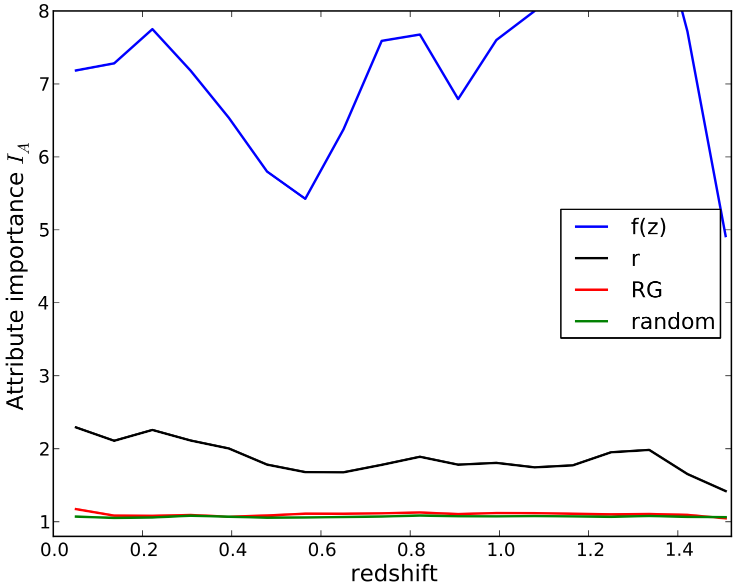


Figure 3.15: The variable importance factor, I_A, as a function of redshift for the most and least important attribute using the bias to quantify the importance ranking. As a control test, we added two artificial variables: an attribute that is a function of the spectroscopic redshift, and a uniformly distributed random attribute. TPZ is able to recognize these two extra attributes and rank them accordingly, as shown by the figure legend.

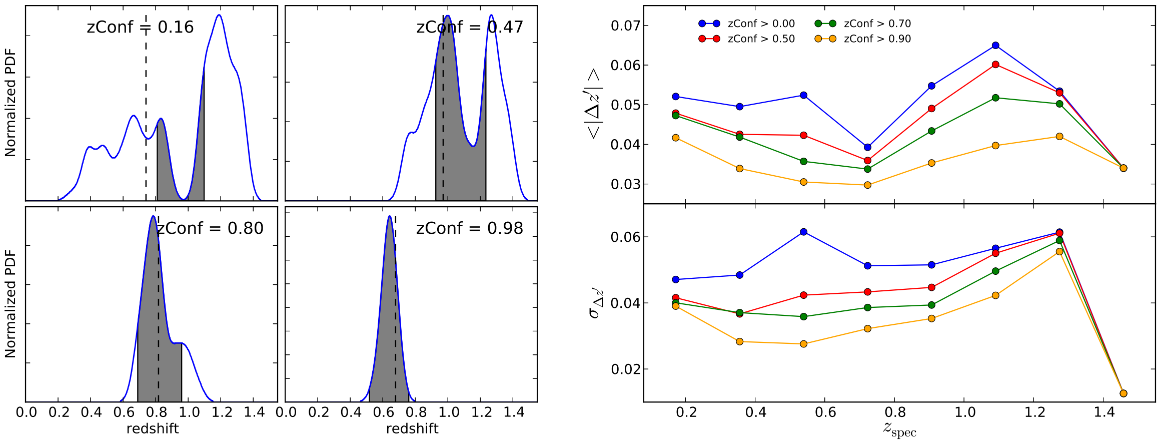


Figure 3.16: (*Left*): Same as Figure 3.6 but for four example galaxies taken from DEEP2. The vertical dashed line indicates the spectroscopic redshift and the gray area the $zConf$ value. (*Right*): The absolute normalized bias and the scatter for galaxy samples defined by different $zConf$ cuts by using the mean of the photo-z PDF as our estimate.

(i.e., a smaller, final catalog). In Figure 3.16, for example, cuts on $zConf$ at 0.5, 0.7 and 0.9 keep 90%, 76% and 38% of the galaxies from the original catalog. Alternatively, given the OOB data predictive results, a required accuracy or number density can be used to identify a suitable value of $zConf$.

3.5 Summary

In this chapter we introduced a supervised machine learning algorithm, TPZ , which is a three step algorithm that first preprocesses the data, completes galaxies with missing photometric values in an efficient manner, and also incorporates measurements errors. A photo-z PDF can be generated from the prediction trees in one of two modes: classification or regression. Both modes produces similar accuracies, but the regression mode is preferred when either the training data are either poorly sampled or not uniformly distributed. On the other hand, the classification mode provides a detailed synopsis of the redshift distribution that can be used to construct priors for use with other photo-z techniques.

We demonstrated the efficacy of the TPZ algorithm and its implementation by applying this new code to three different data sets described in §2: the SDSS main galaxy sample, the PHAT1 blind challenge, and the DEEP2 survey. As we will see in the next chapters, we also have successfully applied TPZ on data taken from the CFHTLens and also from the DES survey with remarkable results in both cases. With the SL-1 sample, we demonstrated that using confidence levels is important as they improve the overall accuracy of our photo-z sample by selecting those galaxies with narrow PDFs. This technique is unique in the sense that it does not need a separate validation test, yet provides ancillary information by using OOB data. We have shown that with these data, we obtain unbiased estimates of both the bias and the dispersion, which are very similar to the same values obtained from the test data for both the SL-1 and DP-1. Obviously, this result is extremely important when working with data that have unknown redshifts.

TPZ not only provides these prior metrics, but it also provides a ranking of the relative importance of the different photometric attributes that are used by the code. This completely statistical process recovers what is naturally expected from physical consideration of these different attributes. With this importance ranking, we can construct a heat map of the different locations in parameter space that produce poor photo-z estimations. Furthermore, we demonstrated that by adding new, manually selected data we can produce more accurate photo-z predication than by simply adding new galaxies randomly. This implies that we can optimally identify new training data for current and future photometric surveys, such as DES or LSST, in order to improve their photo-z predictions.

The attribute importance can also be used to remove those attributes that are least important, thereby improving the computational speed. In addition, we demonstrated that the performance metrics converge as the number of trees increases in the forest, providing a further method to reduce the computational time since we have a direct measure of the minimum forest size. Likewise, we also demonstrated by using the SL-1 that these same metrics also converge with the number of training galaxies for a fixed forest size. Thus, except for adding in manually selected training data to improve areas with poor photo-z prediction, we have an explicit limit for the number of training galaxies needed. Finally, with this technique we found that the error distribution was characterized by a Gaussian distribution with a mean very close to zero and variance very close to one, indicating that the source of errors is relatively unbiased.

We ran our code on the PH-1 data with excellent results; even with limited training data we were able to compute accurate photo-z 's that were comparable if not better to other empirical techniques as well as to some SED fitting techniques. By using the DP-1 redshift data, we tested TPZ over a large redshift range, obtaining very accurate results. In particular, we were able to identify the important attributes, which in this case was the R band magnitude followed by the I band magnitude, and the least important attributes, which in this case was the RG attribute and the B band magnitude. Despite these impressive results, we still have a slight systematically biased photo-z at very low and very high redshifts, which we primarily believe is caused by the low number of training data at these redshifts and also the fact that photo-z estimates can not be negative. We also see a positive skewness in the photo-z PDFs at high redshifts. We believe this result is due to the fact that these galaxies tend to be fainter and have larger magnitude errors. These larger magnitude errors produce a sparser forest at higher redshifts, which is manifested by having a lower photo-z PDF mean value at these same redshifts.

We have also demonstrated how the $zConf$ parameter can be used to select galaxy samples that have improved photo-z estimates with minimal outliers. A target value for this useful parameter can be set to a desired photo-z precision either by calculating the value expected by using OOB data or as required by a specific cosmological requirement. Likewise, we have demonstrated how TPZ can efficiently handle missing data within a catalog. By artificially generating bad or missing parameter values within both the training and the testing data sets, we were not only able to robustly recover the missing parameters but more importantly new photo-z estimates that are consistent with the photo-z estimates from the original, full data set. Therefore, this technique increases the power of photo-z estimation by recovering missing data from the training catalog as well as the power of our resulting sample statistics by recovering missing data from the application data set.

Since TPZ is an empirical algorithm, it is inherent dependent on the quality of its training data. Thus, as is the case with all empirical algorithms, TPZ is limited by the available spectroscopic training data.

Furthermore, the application of TPZ to regions of parameter space beyond the limits of the training data (i.e., extrapolation) will be less reliable. We do note, however, the TPZ does provide ancillary information that can be investigated to better understand the limitations imposed by the training set, to identify the optimal locations within the application data space where new training data will be most useful, and to quantify the possible errors associated with the extrapolation of this technique. In Chapter 6 we will discuss how to improve the photo-z solution by combining multiple approaches, this consider the use of an unsupervised machine learning approach which is fully described in the next Chapter.

Chapter 4

Unsupervised machine learning for photo-z: SOMz

Outline

In this chapter we explore the applicability of the unsupervised machine learning technique of Self Organizing Maps (SOM) to estimate galaxy photometric redshift probability density functions (PDFs). This technique takes a spectroscopic training set, and maps the photometric attributes, but not the redshifts, to a two dimensional surface by using a process of competitive learning where neurons compete to more closely resemble the training data multidimensional space. The key feature of a SOM is that it retains the *topology* of the input set, revealing correlations between the attributes that are not easily identified. We test three different 2D topological mapping: rectangular, hexagonal, and spherical. We also explore different implementations and boundary conditions on the map and also introduce the idea of a *random atlas* where a large number of different maps are created and their individual predictions are aggregated to produce a more robust photometric redshift PDF.

4.1 Self organized maps

Since their introduction (Kohonen, 1982), Self-Organized-Maps have been applied to a variety of scientific problems (see e.g., Kohonen, 2001, for a detailed description of the SOMs and some of their applications). A SOM is a type of an artificial neural network where the learning is unsupervised, there are no hidden layers, and a direct mapping is produced between the training set and the output network. Another important characteristic of a SOM is that the training phase of the algorithm is a competitive process, called vector quantization, where each node or neuron in the map competes with the other nodes or neurons to become more similar to the training data, i.e., each neuron tries to represent as closely as possible the galaxy training set within each timestep. This fact and the use of a neighbor function, which modifies a region of spatially close cells, make the SOM a unique tool that preserves very closely the topology of the multidimensional spectroscopic sample. As a result, similar nodes tend to be grouped together, where, for our purpose, each node represents galaxies with similar properties.

Figure 4.1 presents a schematic illustration of how a SOM is trained. During this phase, each node on the two-dimensional map can be represented by weight vectors of the same dimension as the size of the training

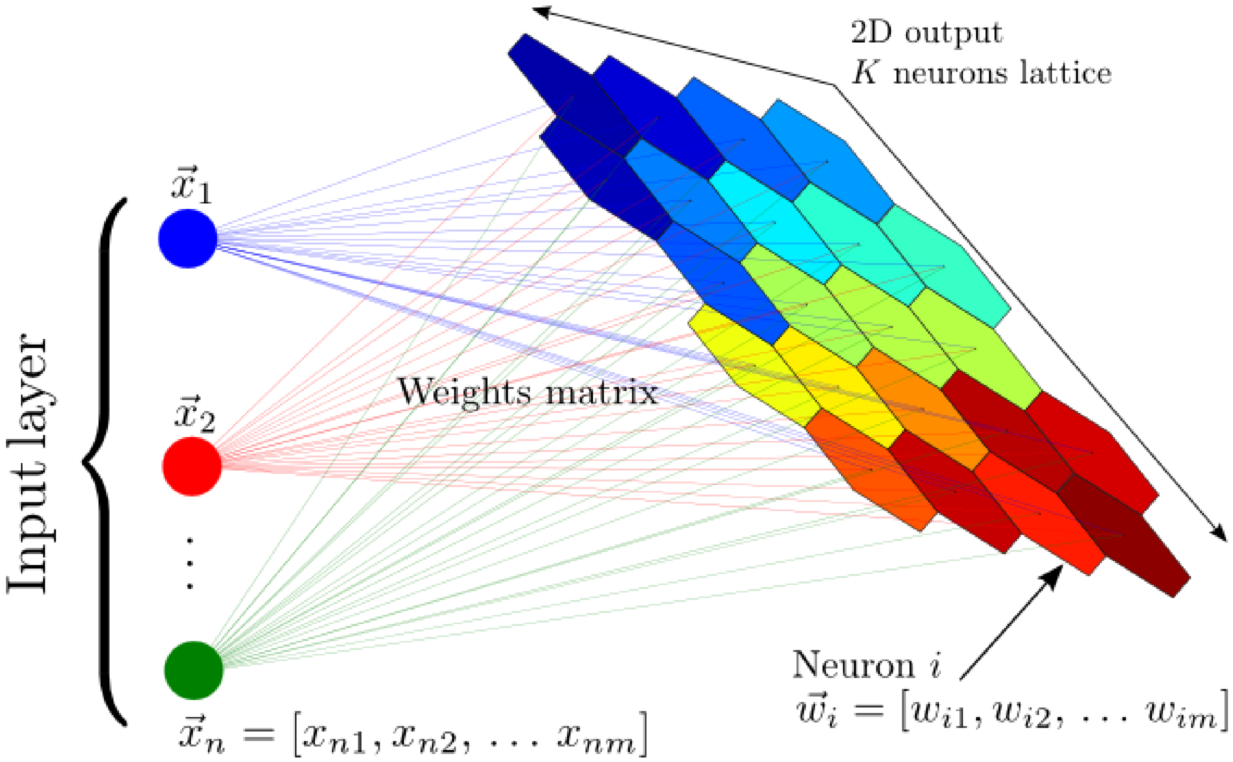

2D output
K neurons lattice
$\vec{x}_1$
$\vec{x}_2$
Input layer
Weights matrix
Neuron i
$\vec{w}_i = [w_{i1}, w_{i2}, \ldots w_{im}]$
$\vec{x}_n = [x_{n1}, x_{n2}, \ldots x_{nm}]$

Figure 4.1: A schematic representation of a self organized map. The training set of n galaxies is mapped into a two-dimensional lattice of K neurons that are represented by vectors containing the weights for each input attribute. Note that the galaxies and the weight vectors are of the same dimension m, and that one neuron can represent more than one training galaxy. The color of the map encodes the organization of groups of galaxies with similar properties. The main characteristic of the SOM is that it produces a nonlinear mapping from an m-dimensional space of attributes (e.g., magnitudes) to a two-dimensional lattice of cells or neurons.

galaxy sample. In an iterative process, the galaxies from the training set are individually used to correct the weight vectors so that the specific neuron (or node) that, at a given moment, best represents the input galaxy is modified, along with the weight vectors of its neighboring neurons, to become a better representation of the current input galaxy. This process is repeated for every galaxy; and the SOM generally converges within a few iterations to its final form where the training data is separated into *groups* of similar features, illustrated in Figure 4.1 by colors.

There are different versions of the basic SOM algorithm; however, all of them follow the same procedure when training a map. The differences arise in the method by which the weight vectors are updated. In this chapter, we present our results from testing two standard versions of the SOM algorithm mapped to three different topologies for a two-dimensional lattice.

4.1.1 SOMz Algorithm

We now present a more detailed discussion of the actual SOM algorithm. First, consider a set of n input vectors taken from the galaxy training sample, which we denote by $\mathbf{x} \in R^m$. These vectors are m-dimensional, where each dimension is a different, measured galaxy attribute, i.e., magnitudes, colors or any other information about the galaxy except the actual spectroscopic redshift. Second, consider a set of K weight vectors $\mathbf{w_k} \in R^m$ where $k = 1, ..., K$. These K weight vectors, which correspond to different neurons, are arranged in a two-dimensional lattice for a given topology. Initial values for the weight vectors are drawn from a uniform random distribution.

For every $n_{\rm it}$ iteration, all n galaxies from the training set are individually processed, and the weights are modified iteratively to optimally match each galaxy. This is the procedure which produces the self-organization of the maps and conserves the topology of the training space. When processing each training galaxy, the weight components of the neuron that most closely matches the current galaxy are updated, along with the weight components of the topologically closest neurons, to better represent this input entry within the featured map. The result of this direct mapping procedure is an approximation of the galaxy training probability distribution function, and it can be considered as a simplified representation of the attribute space of the galaxy sample. We have implemented two different techniques: on-line and batch, to update the actual weights of each cell.

1. On-line SOM: In this case, the weight vectors are updated recursively after processing each input galaxy. For each galaxy, the Euclidean distance between the galaxy's vector of attributes (denoted by $\mathbf{x}$) and

each neuron's weight vector from the map (denoted by $\mathbf{w_k}$) is computed at a given timestep t:

$$d_k(t) = d(\mathbf{x}(t), \mathbf{w_k}(t)) = \sqrt{\sum_{i=1}^{m} [x_i(t) - w_{k,i}(t)]^2} \tag{4.1}$$

From this list of distances, the best matching cell, or neuron, will be identified and denoted by the subscript b, as the cell that is the closest to the galaxy at timestep t:

$$d_b(t) = \min_k d_k(t) \tag{4.2}$$

With this technique, however, not only is the best-matching node updated but also that node's neighboring nodes. In this manner, the entire region containing the best-matching node is identified as being similar to the current training galaxy. This helps ensure similar nodes are co-located, which mimics how training galaxies that have similar properties tend to be co-located in the higher dimensional parameter space. To update the weights, we employ the following relation:

$$\mathbf{w_k}(t+1) = \mathbf{w_k}(t) + \alpha(t) H_{b,k}(t)[\mathbf{x}(t) - \mathbf{w_k}(t)] \tag{4.3}$$

where $\alpha(t)$ is the learning-rate factor, which is reduced monotonically for each timestep. This factor quantifies the magnitude of the correction for the cells as a function of time:

$$\alpha(t) = \alpha_s \left(\frac{\alpha_e}{\alpha_s}\right)^{t/(n_{\mathrm{it}} * n)} \tag{4.4}$$

where α_s is the starting value of α, usually close to unity, α_e is the ending value, and $n_{\mathrm{it}} \times n$ is the total number of timesteps. $H_{b,k}(t)$ is the neighborhood function that also decreases with time *and* with the distance between the nodes b and k. This function quantifies the physical extent to which nodes near to the best-matching node are also updated at every time step. The choice of the kernel's shape for this function does not significantly affect the results as the photo-z PDF estimation in this iterative process. However, the kernel must be smooth, it must be symmetric to avoid biases in any direction, and it must decrease monotonically away from the best matching node so that nodes closer to the best matching node are more strongly updated. The Gaussian Kernel is the simplest kernel that retains all of these features, therefore we use it in our photo-z PDF computations as:

$$H_{b,k}(t) = e^{-D_{b,k}^2/\sigma(t)^2} \tag{4.5}$$

where $D_{b,k}$ is the distance between the nodes b and k which depends on the topology used.

The parameter $\sigma(t)$ encodes the width of the neighborhood function that decreases with t, from a value comparable to the size of the map σ_0 to roughly the width of a single cell σ_f:

$$\sigma(t) = \sigma_0 \left(\frac{\sigma_f}{\sigma_0} \right)^{t/(n_{it} * n)} \tag{4.6}$$

This procedure is applied to all n training galaxies, which are processed in a random order during each iteration. This process is repeated for n_{it} iterations, where just a few iterations are sufficient. As a result, the weights are updated $n_{\text{it}} \times n$ times during the training process, but only the last updated weights are retained after the training process.

2. Batch SOM: This scheme is very similar to the on-line technique; however, in the batch method the weights are updated *only* after each iteration is completed and not after each training galaxy has been processed. As a result, the order in which galaxies are processed in this approach is irrelevant. The weights $\mathbf{w_k}(t_{it})$ are updated at the end of each iteration for a total of n_{it} times by using an accumulated sum:

$$\mathbf{w_k}(t_{it}) = \frac{\sum_{j=1}^{n} \widetilde{H}_{b,k}(t_{it}) \mathbf{x_j}}{\sum_{j=1}^{n} \widetilde{H}_{b,k}(t_{it})} \tag{4.7}$$

where the summation is over all n galaxies in the training sample, and t_{it} is the timestep representing a given iteration. $\widetilde{H}_{b,k}(t_{it})$ is computed by using Equation 4.5, but in this case the best-matching node is identified by using the weights computed at the end of the previous iteration:

$$\tilde{d}_k(t_{it}) = d(\mathbf{x}(t_{it}), \mathbf{w_k}(t_{it-1})) = \sqrt{\sum_{i=1}^{m} \left[x_i(t_{it}) - w_{k,i}(t_{it-1}) \right]^2} \tag{4.8}$$

and

$$\tilde{d}_b(t_{it}) = \min_k \tilde{d}_k(t_{it}) \tag{4.9}$$

Again, recall that the weight vectors $\mathbf{w_k}(t_{it})$ in the batch technique are computed at the end of the previous iteration and kept fixed during the current one. In this case, the update of the weight vectors is not recursive as in the on-line technique; therefore, the final map does not depend in any way on the order in which the training galaxies are sampled. In addition, the batch technique does not use the learning-rate $\alpha(t)$, which eliminates a potential source of poor convergence if this factor is not well determined.

Figure 4.2 illustrates the SOM algorithm and highlights the difference between the two techniques we have employed to update the weight vectors during the training process. Both techniques are initialized in the same manner, have common steps, and require similar running times. With the batch update technique, however, there is no dependency on α and only the neighborhood function is updated for each time step t.

4.1.2 2D Topologies

For each of the two SOM techniques discussed in the previous section, we have implemented three different, two-dimensional topologies: a rectangular grid with square cells, a hexagonal grid, and a grid of equal-area cells confined to the surface of a sphere. We also include the option to use periodic boundary conditions for the non-spherical case. Figure 4.3 presents the nodes for these three topologies constructed via the data described in §3. Each topology has roughly the same number of cells: 784 (rectangular), 756 (hexagonal), and 768 (spherical). For this figure we have employed the same training process using the online update scheme for each topology, and the cell colors encode the mean redshift of the galaxies represented by each cell after the last iteration has been completed. This simple visualization demonstrates how the SOM technique groups galaxies together via their input parameters, while the desired predictive attribute, in this case redshift, is only used at the end to visualize the map or to make photo-z estimations. The SOM technique, for all three topologies, clearly groups galaxies together in the map that have similar redshifts without any specific supervision, which is a major advantage of this method.

We now present the details of these three, two-dimensional topologies.

1. Rectangular grid: For this topology, each cell has eight direct neighbors. We calculate the distances $D_{b,k}$, which is used by $H_{b,k}(t)$, between the best-matching cell and the other cells by using the Euclidean distances $D_{b,k} = \sqrt{(x_b - x_k)^2 + (y_b - y_k)^2}$. This topology is the standard method used to create SOMs, and it has been extended by using periodic boundary conditions so the nodes are wrapped on one toroidal surface. This is functionally equivalent to folding a sheet of paper into a tube, and subsequently wrap the tube onto itself to form a torus.

2. Hexagonal grid: For this topology, each cell has six direct neighbors. We calculate the distances $D_{b,k}$ between cells by using the Euclidean distances between the centers of the cells as in the rectangular grid topology. It differs from the rectangular grid by the fact of all the neighbor's centers are located at the same distance which produces a smoother neighboring function. This topology can also be extended by using periodic boundary conditions so that the nodes are effectively wrapped on to one surface.

3. Spherical grid: This last topology naturally eliminates the problem of wrapping the nodes as the map

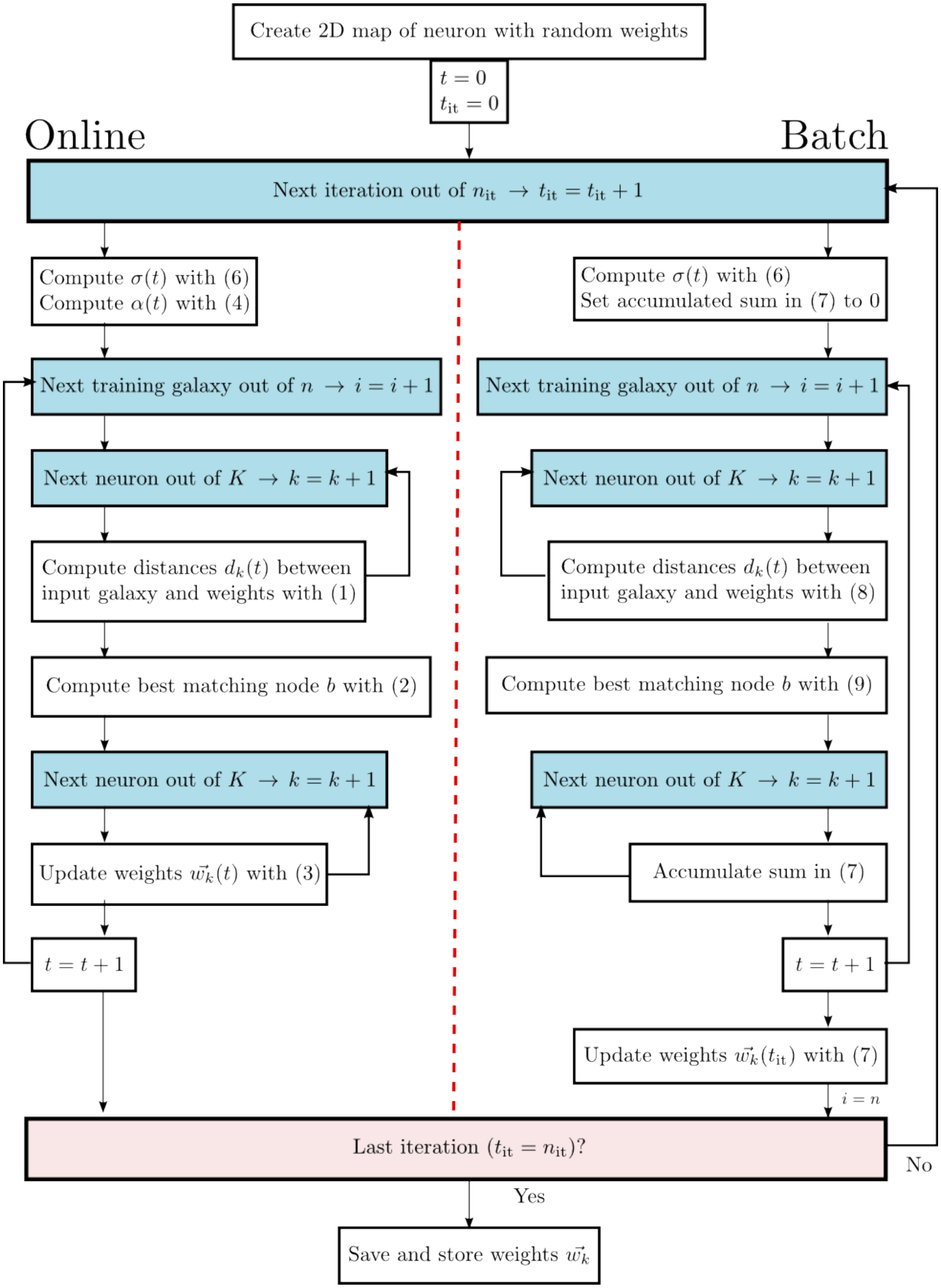

Create 2D map of neuron with random weights
$t = 0$
$t_{\rm it} = 0$
Online
Batch
Next iteration out of $n_{\rm it}$ → $t_{\rm it} = t_{\rm it} + 1$
Compute $\sigma(t)$ with (6)
Compute $\alpha(t)$ with (4)
Compute $\sigma(t)$ with (6)
Set accumulated sum in (7) to 0
Next training galaxy out of n → $i = i + 1$
Next training galaxy out of n → $i = i + 1$
Next neuron out of K → $k = k + 1$
Next neuron out of K → $k = k + 1$
Compute distances $d_k(t)$ between input galaxy and weights with (1)
Compute distances $d_k(t)$ between input galaxy and weights with (8)
Compute best matching node b with (2)
Compute best matching node b with (9)
Next neuron out of K → $k = k + 1$
Next neuron out of K → $k = k + 1$
Update weights $\vec{w_k}(t)$ with (3)
Accumulate sum in (7)
$t = t + 1$
$t = t + 1$
Update weights $\vec{w_k}(t_{\rm it})$ with (7)
$i = n$
Last iteration ($t_{\rm it} = n_{\rm it}$)?
No
Yes
Save and store weights $\vec{w_k}$

Figure 4.2: A flowchart illustrating our implementation of the SOM algorithm for photo-z estimation. Online and batch update schemes are presented on the left and right respectively.

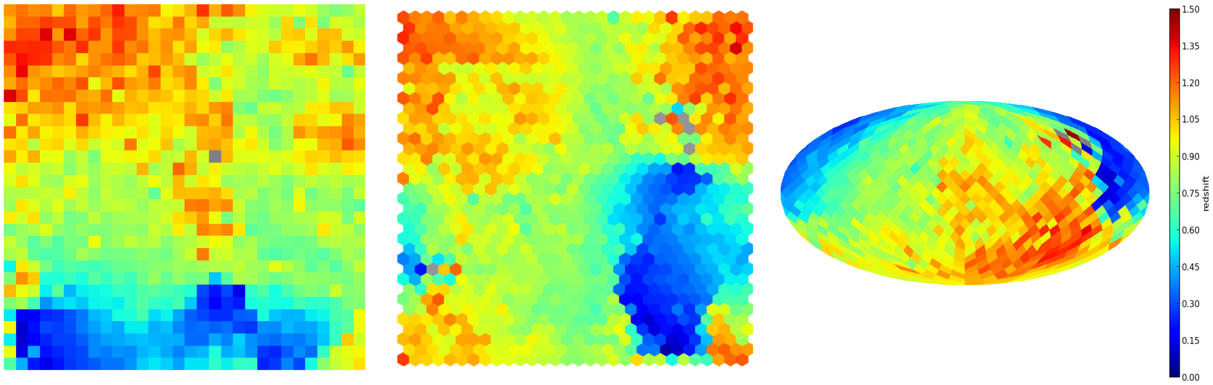


Figure 4.3: A comparison of the three different, two-dimensional topologies used in this work. Each topology employs equal-area cells, where the color encodes the mean redshift of all galaxies assigned to a cell after the training process is complete. The colorbar on the right applies to all three maps. (*Left*): Rectangular grid with 784 square cells. (*Central*): Hexagonal grid corresponding to 756 cells with periodic boundary conditions. (*Right*): Spherical grid using HEALPix with 768 cells.

is constructed directly on a continuous, two-dimensional surface. For this topology, we have used the HEALPix[1] (Górski et al., 2005) tools to construct the two-dimensional map where the cells are constructed to have the same area as the other topologies. We calculate the distances between cells by using the great-circle distance between the centers of each cell:

$$D_{b,k} = \cos^{-1}(\sin\phi_b \sin\phi_k + \cos\phi_b \cos\phi_k \cos(|\theta_b - \theta_k|) \tag{4.10}$$

where ϕ and θ are the latitude and longitude respectively of the best matching cell b and the k nearest cells.

4.1.3 Random Atlas

In machine learning, a random forest is an *ensemble learning* algorithm that first generates many randomized prediction trees and subsequently combines the predictions together into a meta-prediction. Random forests have been demonstrated (Caruana et al., 2008) to be one of the most accurate empirically trained learning techniques for both low and high dimensional data. Since we are using self-organized maps in this work, however, we can not construct a collection of trees as described in Chapter 3. Instead, we explore the construction of a collection of maps, which we aggregate and call a *random atlas* in a similar manner as a random forest.

Given a training sample of n galaxies that have m attributes (e.g., magnitudes), we create N_M bootstrap samples of size n (i.e., n randomly selected objects with replacement) to generate N_M different maps.

[1] http://healpix.jpl.nasa.gov

For each map, we can either use all available attributes and have weight vectors of the same dimensions or, alternatively, we can randomly select a subsample of attributes for each map that reduces possible correlations between maps. After all maps are built, a final and robust prediction can be calculated by combining all N_M estimates together. As we discussed in Chapter 3, this technique performs well when compared to other learning techniques and is also robust against overfitting (i.e., there is no limit on the number of maps, N_M, in the *atlas*)

4.1.4 SOM Implementation

In order to generate photo-z PDFs by using SOMs we have two major tasks. First, after preparing the training data, we generate N_R training samples by perturbing the measured training data attributes according to the measured uncertainty for that attribute, which we assume to be normally distributed. In this manner, we can incorporate the measurement error into the map construction. We also reduce the bias towards the data and introduce randomness into the construction of the maps in a systematic manner. Second, for each newly constructed training sample, we generate N_M new maps as described previously in §4.1.3 by using bootstrap samples.

In total, we produce $N_R \times N_M$ SOMs as described in §4.1.1. After all the final weights for each map are recorded, the galaxies for each sample are processed again by using those weights and are assigned to one of the K cells belonging to each map. This ensures that each cell in each map represents a subsample of galaxies that have similar characteristics. To compute a photo-z, we process each galaxy in the test sample (i.e., the photometric data) and determine which cell in each map best represents this galaxy. We repeat this procedure for all SOMs; and, when this is completed, we combine the predictions from all of the maps into a single probability density function that is normalized by the total number of predictions. In this manner, each map contributes equally to the final PDF.

This process is demonstrated for one example map in Figure 4.4, where the evolution of one SOM is sampled at different iterations using online updating. As before, the colors encode the mean redshift of the galaxies represented by each cell during each iteration. From this figure, we see that even at the first iteration there is a slight separation that quickly changes with time until convergence to the final distribution is achieved and galaxies with similar redshifts are spatially grouped in a self-organized manner.

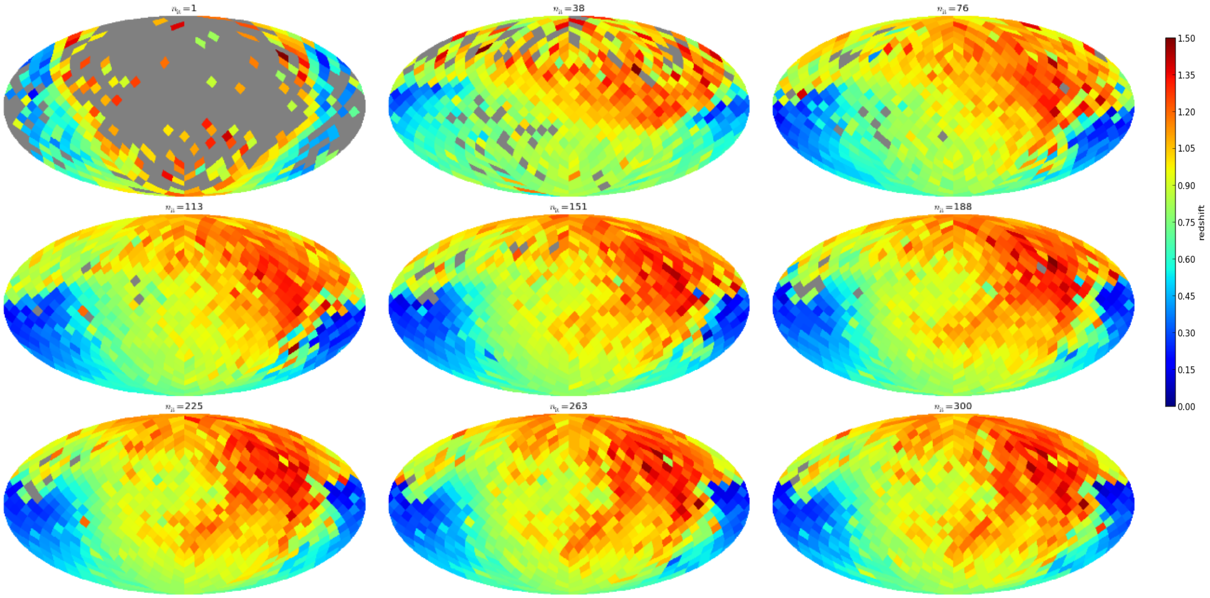


Figure 4.4: Evolution of the SOM at different iterations using the spherical topology and online updating. Colors encode the mean redshift of the galaxy being represented by each cell at each iteration as defined by the colorbar, similar to the one in Figure 4.3.

4.2 Results

In this section, we compare the results of our SOM implementation by using different parameter configurations with the DEEP2 data DP-1, introduced in §2.3. To compare different applications of this algorithm, we define the bias to be $\Delta z' = |z_{\rm phot} - z_{\rm spec}|/(1 + z_{\rm spec})$, and we present the standard metrics used to compare the accuracy of the different SOMs in Table 4.1. As shown in this table, we define several metrics to address the bias and the variance of the results (the first five rows) and also present three values to characterize the outlier fraction.

We have introduced the quantity KS, which represents the results of a Kolmogorov–Smirnov test to address whether the predicted photo-z distribution and the spectroscopic redshift distribution are drawn from the same underlying population. We present this new statistic since it provides one robust value to compare both distributions that does not depend on how we bin in redshift and it is defined as the maximum distance between both empirical distributions. For this statistic, we compute the empirical cumulative distribution function (ECDF) for both distributions. For the spectroscopic sample the ECDF is defined as:

$$F_{\rm spec}(z) = \sum_{i=1}^{N} \Omega_{z^i_{\rm spec} < z} \tag{4.11}$$

Table 4.1: Definition of the metrics used in the text to present and discuss the results,

Metric	Meaning
$< \Delta z' >$	mean of $\Delta z'$
$\|\Delta z'\|_{50}$	median of $\Delta z'$
$\sigma_{\Delta z'}$	Standard deviation of $\Delta z'$
σ_{68}	Sigma value at which 68% of $\Delta z'$ is enclosed
σ_{MAD}	Median absolute deviation = $\mathrm{median}(\|\|\Delta z' - \|\Delta z'\|_{50}\|\|)$
KS	Kolmogorov - Smirnov statistic for $N(z)$
$\mathrm{out}_{0.1}$	Fraction of outliers where $\Delta z' > 0.1$
$\mathrm{out}_{2\sigma}$	Fraction of outliers where $\|\Delta z' - < \Delta z' > \| > 2\sigma_{\Delta z'}$
$\mathrm{out}_{3\sigma}$	Fraction of outliers where $\|\Delta z' - < \Delta z' > \| > 3\sigma_{\Delta z'}$

where N is the number of galaxies in the redshift sample, and

$$\Omega_{z^i_{\mathrm{spec}} < z} = \begin{cases} 1, & \text{if } z_{\mathrm{spec},i} < z \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

The summation is carried out over all galaxies in the sample. Having computed the ECDF for both the photo-z and spectroscopic distributions, we compute the KS statistic as:

$$\mathrm{KS} = \max_z \left(||F_{\mathrm{phot}}(z) - F_{\mathrm{spec}}(z)|| \right) \tag{4.13}$$

As a result, as the KS statistic decreases, the two distributions become more similar.

All of the metrics listed in Table 4.1 are defined such that a lower value for the computed metric indicates a better overall photo-z solution. We have defined a new, meta-statistic, which we call I-score (symbolically represented by $I_{\Delta z'}$), to more easily compare different SOM parameter configurations (i.e., online or batch and a specific 2D topology) or different photo-z estimation techniques. For this new meta-statistic, we first must normalize each set of metrics across all different photo-z estimations so that we are not biased by different dynamic ranges. Thus, for example, we first compute the mean and standard deviation for $< \Delta z' >$, and subsequently rescale all individual $< \Delta z' >$ values so that this set of values has zero mean and unit variance.

We continue this process for all nine statistics listed in Table 4.1, and compute their weighted sum to obtain the I-score:

$$I_{\Delta z'} = \sum \frac{M_i}{w_i}, \tag{4.14}$$

where M_i is the rescaled metric and weight value for metric i out of the nine available. For simplicity, we use equal weights in the remainder of this thesis (and thus the I-score is simply the average of the nine rescaled metrics for each technique). As a result, the photo-z method (or parameter configuration) with the lowest I-score will be the optimal estimation technique. On the other hand, if we are looking for a technique or parameters configuration with, for instance, a lower outlier fraction, we could assign a higher weights accordingly to account for it. In this way, we can efficiently select the best method or configuration for specific needs.

4.3 Discussion

In order to explore the effects of different parameter configurations on the performance of our SOM photo-z implementation, we conducted twenty different tests and compare their I-score results in Table 4.2 by using six colors from the DEEP2 data: $B-R$, $R-I$, $u-g$, $g-r$, $r-i$, and $i-z$. These configurations include the use of all three topologies discussed in §4.1.2: Hexagonal (hex), Rectangular (rec) and Spherical (sph); the use of *online* or *batch* methods to update the weights as shown in Figure 4.2; and the use of a Random Atlas where different maps are built using random subsets (random = yes) of four colors or single maps where all six colors are used (random = no), which gives twelve different configurations. In addition, both rectangular and hexagonal topologies were used with both periodic and non-periodic boundary conditions (spherical is by definition wrapped), which gives us an additional eight configurations.

We determined the best values for the other parameters in our SOM photo-z implementation, which were then fixed for all twenty tests, by using an Out-Of-Bag data (similar to a validation sample) technique we presented in Chapter 3. For example, we set the values α_s and α_e in Equation 4.4 to be 0.9 and 0.5 respectively. In addition, each random atlas contains 100 different maps and each topology contained approximately 800 cells in a given map. All galaxies in the test sample were used for each run. The results, averaged over ten realizations, are presented in Table 4.2 for all the metrics, where we have used the mean redshift in place of each PDF for simplicity. Note that the last column is the I-score. For clarity, we highlight in red the best value for a particular metric.

We compare these different parameter configurations visually in Figure 4.5, where the twenty runs are plotted in terms of their bias and I-score values. In this figure, different symbols represent different topologies (squares for rectangular, diamonds for hexagonal and circles for spherical), colors represent the update method used (blue for online update and red for batch update).

The curves enclose all test results that either use a random subsample of attributes (purple) or all attributes (green) on each map inside the atlas. Note that the separation of these two groups of tests is a direct output from our SOM photo-z implementation. Finally, we highlight if periodic boundary conditions

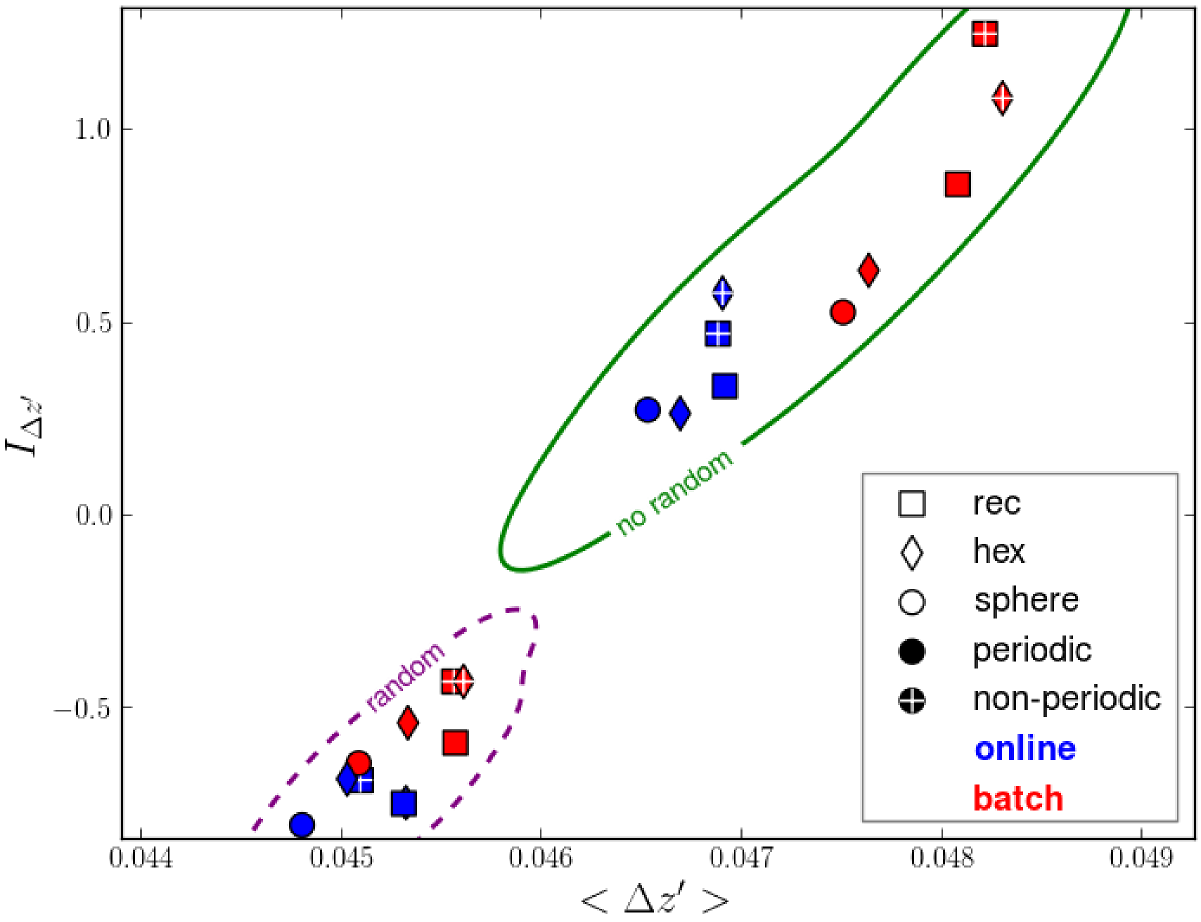


Figure 4.5: The I-score, $I_{\Delta z'}$ as a function of the bias, $< \Delta z' >$ for all twenty methods discussed in the text, averaged over ten different realizations for all galaxies (10,227) in the test sample. Enclosed by a green curve are the results of using all attributes on each map of the atlas and enclosed in purple the results when a random atlas was used. Blue symbols indicate an *online* update of the weights, while red symbols indicates a *batch* update. The symbols themselves represent different topologies and the white cross indicates periodic boundary conditions.

were used for rectangular or hexagonal topology by a white cross.

Overall, from both Table 4.2 and Figure 4.5, the best set parameter configuration is spherical topology with an *online* update using random atlas. This run has metrics that are the closest to the best values and it has the lowest I-score value. In the rest of this section, we explore the results of these different parameter configurations in more detail.

4.3.1 Random atlas

The first parameter configuration we examine is the use of a random atlas. As shown in Table 4.2 or Figure 4.5, there is a clear, albeit numerically small difference between the performance of our SOM algorithm with

Table 4.2: Results table for all the twenty combinations averaged after ten different realizations. The red entries show the best value on each column to aid the reading of the table.

Topology	Periodic[a]	update	random	$< \Delta z' >$	$\lvert\Delta z'\rvert_{50}$	$\sigma_{\Delta z'}$	σ_{68}	σ_{MAD}	KS	$\mathrm{out}_{0.1}$	$\mathrm{out}_{2\sigma}$	$\mathrm{out}_{3\sigma}$	$I_{\Delta z'}$
rec	NP	online	no	0.0469	0.0280	0.0685	0.0353	0.0194	0.0635	0.1042	0.0322	0.0152	0.4733
rec	NP	online	yes	0.0451	0.0273	0.0667	0.0338	0.0188	0.0719	0.0956	0.0298	0.0147	-0.6868
rec	NP	batch	no	0.0482	0.0291	0.0709	0.0358	0.0202	0.0636	0.1075	0.0321	0.0150	1.2525
rec	NP	batch	yes	0.0456	0.0277	0.0673	0.0340	0.0190	0.0742	0.0970	0.0295	0.0145	-0.4305
hex	NP	online	no	0.0469	0.0281	0.0685	0.0353	0.0195	0.0628	0.1045	0.0325	0.0153	0.5789
hex	NP	online	yes	0.0453	0.0275	0.0674	0.0339	0.0189	0.0728	0.0962	0.0289	0.0141	-0.7454
hex	NP	batch	no	0.0483	0.0290	0.0717	0.0359	0.0200	0.0630	0.1078	0.0311	0.0147	1.0855
hex	NP	batch	yes	0.0456	0.0278	0.0674	0.0340	0.0191	0.0734	0.0970	0.0292	0.0145	-0.4302
sph	P	online	no	0.0465	0.0277	0.0685	0.0351	0.0193	0.0626	0.1031	0.0324	0.0150	0.2753
sph	P	online	yes	0.0448	0.0272	0.0669	0.0337	0.0187	0.0718	0.0961	0.0296	0.0143	-0.8034
sph	P	batch	no	0.0475	0.0287	0.0696	0.0356	0.0198	0.0625	0.1057	0.0311	0.0143	0.5293
sph	P	batch	yes	0.0451	0.0274	0.0664	0.0338	0.0189	0.0732	0.0970	0.0300	0.0144	-0.6428
rec	P	online	no	0.0469	0.0281	0.0695	0.0352	0.0195	0.0651	0.1041	0.0314	0.0145	0.3370
rec	P	online	yes	0.0453	0.0273	0.0674	0.0338	0.0188	0.0738	0.0962	0.0292	0.0141	-0.7474
rec	P	batch	no	0.0481	0.0293	0.0708	0.0358	0.0201	0.0655	0.1075	0.0308	0.0140	0.8618
rec	P	batch	yes	0.0456	0.0277	0.0673	0.0339	0.0191	0.0745	0.0975	0.0291	0.0139	-0.5900
hex	P	online	no	0.0467	0.0279	0.0691	0.0351	0.0194	0.0615	0.1038	0.0319	0.0148	0.2663
hex	P	online	yes	0.0450	0.0272	0.0670	0.0337	0.0189	0.0725	0.0948	0.0296	0.0147	-0.6838
hex	P	batch	no	0.0476	0.0289	0.0704	0.0357	0.0199	0.0619	0.1070	0.0310	0.0141	0.6378
hex	P	batch	yes	0.0453	0.0275	0.0672	0.0339	0.0190	0.0739	0.0970	0.0296	0.0144	-0.5375

[a] P stands for periodic boundary topology and NP for non-periodic

and without the use of random subsampling of attributes. This finding is remarkably similar to the result we discussed in Chapter 3 where a random forest was shown to be superior to prediction trees that used the full set of attributes. The likely explanation is the random sampling of attributes when building the maps (trees) for a random atlas (forest) more completely explores the set of attribute combinations than when using all attributes.

These maps are constructed using Bootstrap sampling; thus by definition all maps are different although they are likely to be highly correlated, which will yield stable results after a certain number of maps have been generated. When using random sampling of the attributes, however, we are by definition introducing extra variation into the algorithm. This can reduce the noise variables that will always contribute when all attributes are included, and will on average yield better statistics when a large number of maps are generated so that all variables are used multiple times in different combinations for different maps. For example, if we construct 100 maps where each map is constructed by randomly selecting four color attributes out of six possible colors, we can be sure all attribute combinations (in this case 15) are sufficiently covered.

The appropriate number of attributes to be randomly selected when constructing a random atlas can be determined either by testing the algorithm using Out-Of-Bag (Carrasco Kind & Brunner, 2013a) data on previous runs or by selecting a value somewhere between the total number of attributes and the number of dimensions of the SOM (in this case we have a two-dimensional topology). Alternatively, in Chapter 3 we discussed using $\sqrt{M}$, where M is the set of attributes, although this is likely too small for lower dimensional problems. A reasonable compromise might be to simply use $2/3$ of the attributes when constructing each map.

We test the dependence of the random atlas on the number of input attributes by constructing two hundred maps using a spherical topology with online updating and only changing the number of attributes that are used for the random sampling. The results are presented in Table 4.3, where our I-score statistic indicates that four attributes are optimal (as well as two other metrics). However, it is interesting to note that three attributes perform only moderately worse than four, and that two attributes show comparable performance to either five or six attributes. We found similar results by using other parameter configurations (i.e., varying the topology and update method), suggesting the optimal number of attributes is dependent on the data themselves.

One last observation from the data presented in Table 4.2 is that all metrics have their lowest values when using random sampling, expect for the KS statistic. This means that, on average, using all attributes produces an $N(z)$ from the training sample that seems to be a better match to the spectroscopic sample than when using random subsamples (i.e., the $N(z)$ ECDFs are more similar). This is most likely a result of the

Table 4.3: Performance of the SOM algorithm by using a spherical topology with an online update for different number of attributes used in the construction of the random atlas

Attributes	$< \Delta z' >$	$\sigma_{\Delta z'}$	KS	$\mathrm{out}_{0.1}$	$I_{\Delta z'}$
1	0.0903	0.1023	0.1852	0.3153	1.6830
2	0.0558	0.0659	0.1229	0.1379	-0.2452
3	0.0446	0.0607	0.0846	0.0885	-0.4636
4	0.0432	0.0610	0.0767	0.0784	-0.4754
5	0.0422	0.0633	0.0614	0.0824	-0.2726
6	0.0436	0.0653	0.0583	0.0886	-0.2262

fact that a random atlas prediction produces a photo-z PDF that has a smaller bias and scatter (as shown in Table 4.2) and is thus more strongly peaked about the mean value than a photo-z PDF that does not use random sampling. When simply using the mean value from a PDF, the $N(z)$ ECDF will thus be more strongly concentrated about the mean leading to a higher KS statistic. As we will show, by using the full photo-z PDF when constructing the sample $N(z)$, we generate a more realistic redshift distribution that reduces the KS statistic by a factor of a few, reinforcing this interpretation.

4.3.2 Weights updating

The second parameter we explore is the method used to update the weights that control how cells in the final topology are modified. Both Table 4.2 and Figure 4.5 demonstrate that the online weight updating method consistently performs better, when all other parameters are kept fixed, than the batch updating method. We interpret this difference as a manifestation of the dynamic nature of online updating, where the weights are updated after analyzing each galaxy, as opposed to the once an iteration update that is performed with the batch method. As a result, the online method will easily converge given a sufficient number of iterations and will produce a more accurate topological mapping. On the other hand, the batch method is nearly parameter free, while the online method depends on the parameter α. In addition, the batch method is computationally faster since the number of weight updates is considerably smaller than the online method, and the batch method is easier to scale to big data since the processing is inherently parallel.

4.3.3 Topologies

The next parameter we tested is the type of two-dimensional topology used for the SOM. Although not as obvious as the previous two parameters, the results suggest that spherical topology is superior than either the rectangular or the hexagonal grids, when given approximately equal number of cells and when using a random atlas (we note that when using the full attribute set the hexagonal topology slightly outperforms the spherical topology). Given the nature of the spherical topology, a direct comparison is only realistic when we compare to periodic boundary conditions for rectangular and hexagonal topologies.

Although we explore the effect of the map size on the performance of this algorithm in subsection 4.3.6, we note that since we use the HEALPIX scheme to characterize the spherical topology, we are implicitly constrained in the number of cells in our final map, which is given by $n_{\rm cells} = 12 \times {\rm nside}^2$ where ${\rm nside}$ is a power of 2. For these tests, we used ${\rm nside}$ = 8 which corresponds to 768 cells. By using this relation, the next map size for a spherical topology would be 3072 cells, which nearly equals the number of galaxies in our training sample, and is, therefore, too large for this particular problem. This is a limitation of the spherical topology, as both the rectangular and hexagonal topologies are not restricted in this manner; thus we might be able to fine tune the number of cells in these alternative topologies in order to outperform the spherical topology. We do not, however, test this hypothesis in this thesis.

Since we use the HEALPIX representation for spherical topology, we consider that the natural periodicity of this topology is superior to the forced periodicity with the rectangular and hexagonal topologies. HEALPIX generates equal-area cells and the cell centers are naturally aligned along the same latitude. Thus, it is reasonable to expect that the HEALPIX scheme produces cell weights that more closely match the spectroscopic data set. On the other hand, we have no natural driver to choose between rectangular or hexagonal, and, depending on the value of the other algorithm parameters one may outperform the other.

4.3.4 Periodicity

Next, we look at the use of periodic boundary conditions, which from the results presented in Figure 4.5 and Table 4.2 appear to, in general, outperform the non-periodic boundary conditions (with the understanding that the spherical topology is implicitly periodic). Specifically, when comparing periodic (solid colors) and non periodic cases (white crosses) in Figure 4.5 for the hexagonal and rectangular topologies, the periodic case performs slightly better, although this is not universally true.

While the redshift distribution of our training sample data are limited to the range $0 \leq z \leq 1.5$, the SOM mapping process has no such restrictions when optimizing the topological mapping, for example when processing the colors of the galaxies. On the other hand, non-periodic conditions might work better for classification problems where clear separation is desired between classes of objects (e.g., star versus galaxy); but in a regression problem, like photo-z estimation, a clear separation is not necessarily desired as we do not want to bias the mapping either away from or specifically towards any particular region of the parameter space.

4.3.5 Other parameters

Besides the previously discussed parameters, our SOM algorithm does not depend on many other parameters (and of course the SOM can be applied in a non-parametric manner). One parameter that must be specified when using online updating, however, is the learning rate factor, α. This parameter quantifies the correction applied to each cell at each time step, and can take values from 1.0 (maximal correction) to some minimum value, often close to 0.5. In the end, the SOM algorithm is not extremely sensitive to this parameter as the neighborhood function exerts more control over the corrections applied to the neighboring cells. We do acknowledge that, if the number of iterations is limited, this factor might become more important since fewer corrections will be applied.

Another parameter to specify is the number of iterations to use when constructing the SOM, as we need a sufficient number to generate a map that truly represents the data appropriately. This number will depend on both the size of the input data set and the number of cells in the map. For the example discussed herein, we found that 100 iterations were sufficient. For a larger data problem, the number of iterations should be increased, with the exact value determined empirically by, for example, terminating the iterative process if the map changes by some value (e.g., 1%) or if the map evaluation does not change beyond some small tolerance.

As an example, Figure 4.4 demonstrates how a spherical topology map changes in nine different steps during 300 iterations using the online updating scheme. In each map, the color of a cell encodes the mean redshift of all galaxies within that cell after each evaluation. After the first evaluation, the map is not fully populated, thus only some of the cells are populated. The iteration process, however, quickly begins to populate the cells and by iteration 113 (middle left image) the map becomes fairly stable with only a few empty cells. The last three maps (iterations 225, 263, and 300 left-to-right in the bottom row) are nearly identical, demonstrating how the iterative process has essentially converged and the map can be used for photo-z predictions.

4.3.6 Size of map and size of atlas

The two algorithm parameters that remain to be identified are the number of cells to use within an individual map, and the number of maps to use within a random atlas. We explore the effects of changing one of these parameters while keeping the other fixed in Figure 4.6, where the top panel identifies the dependence on the number of maps used in a random atlas, keeping the size of each map fixed at 756 cells, while the bottom panel highlights the dependence on the number of cells in an individual map, keeping the number of maps fixed at 100. In both panels, we use four metrics to quantify the performance of the SOM: the bias $< \Delta z' >$

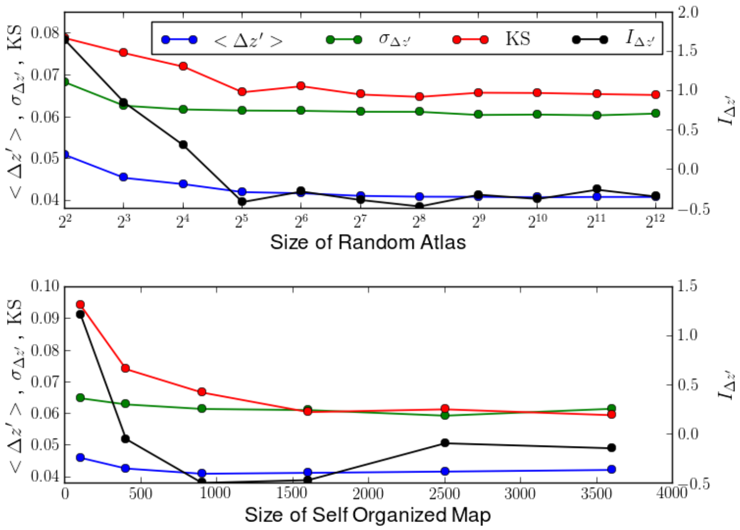


Figure 4.6: Bias $< \Delta z' >$ (blue), scatter $\sigma_{\Delta z'}$ (green), KS (red) and $I_{\Delta z'}$ (black) as a function of the number of maps contained in the random atlas with fixed map size of 756 cells (top panel), and as a function of map size keeping fixed the number of maps (100) in the random atlas (bottom).

(blue), the scatter $\sigma_{\Delta z'}$ (green), the KS statistic (red), and the I-score $I_{\Delta z'}$ (black).

As shown in the top panel, where we have constructed a SOM using hexagonal topology with online updating and a fixed number of cells in each map, increasing the number of maps in the random atlas does improve the performance and reduces the value of these metrics. At some point, however, adding more maps does not produce any improvements as all possible parameter combinations have been included in the atlas a few times and new maps become redundant. On the other hand, as shown in the bottom panel, increasing the number of cells with a single map will also improve the value of the metrics.

Eventually, however, the mean number of sources per cell decreases to the point where we have empty cells with no predictive power, and we also suffer from over-fitting. This primarily affects the fraction of outliers and subsequently the I-score, which depends on all of the metrics. Thus we find an optimal size (for this particular data set) of approximately 1500 cells. This confirms the results presented by Way & Klose (2012) who also found that increasing the number of cells in a map produces better results until over-fitting affects the metrics.

4.3.7 Photo-z PDF using SOMs

For simplicity, to this point we have compared the different SOM configurations and overall performance by simply using a single predictive value (in this case, the mean value of the photo-z PDF). As we discussed in §4.1.1, however, the SOM technique generates a full probability distribution function for the photometric

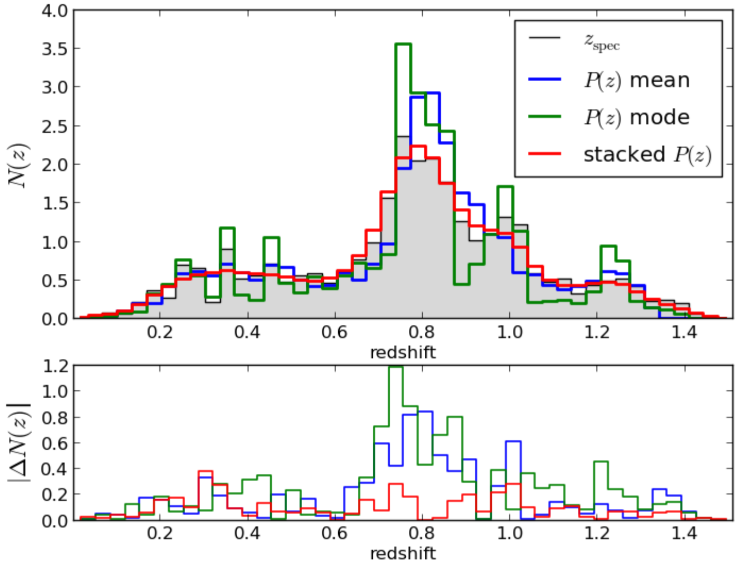


Figure 4.7: $N(z)$ (top) and absolute error (bottom) for the galaxies used to compute Figure 4.8, showing the difference by using the mean, the mode and a stacked photo-z PDF in which the full photo-z PDF of individual galaxies are summed together.

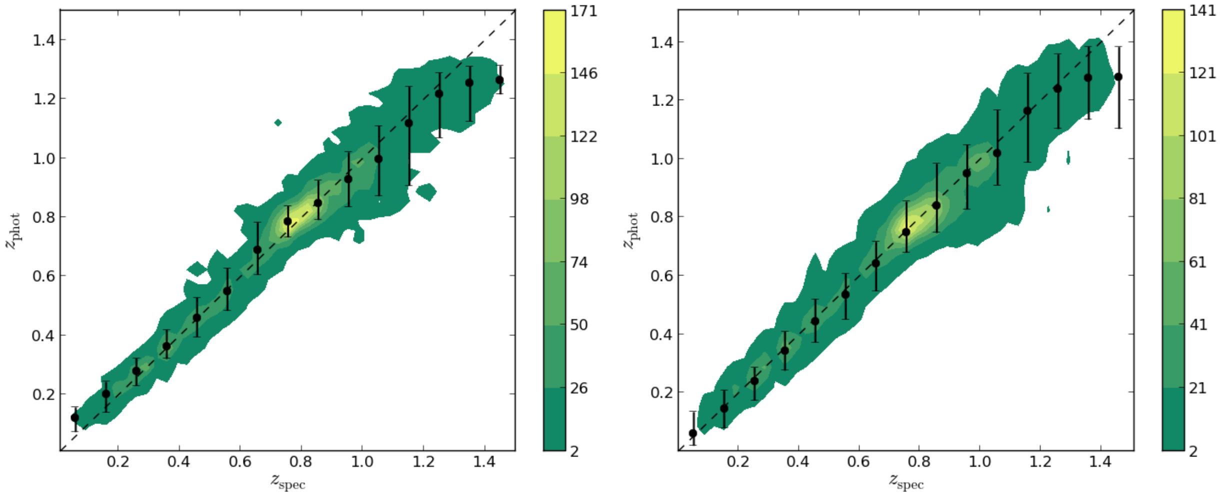


Figure 4.8: Spectroscopic redshift versus photometric redshift estimated by using a SOM for (left) the mean of the photo-z PDF and (right) the full photo-z PDF. In both panels, we use an identical number of pixels to construct the image and also use the same number of contours to present the color-mapping. The galaxies used to make this image were selected in an identical manner with $zConf > 0.7$ (with a total of 8387 galaxies) from the DEEP2 survey. The black dots are the median values of $z_{\rm phot}$ and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic bin of width $\Delta z = 0.1$.

redshift of each individual galaxy. We can use all of the information encoded in these PDFs when making cosmological measurements. For example, we can more accurately compute the sample redshift distribution, $N(z)$, which is used in a variety of cosmological measurements, by including the full photo-z PDF.

This can be seen from the data in Table 4.2, where the best KS statistic, which is a measurement of how well the true $N(z)$ is recovered, is 0.0615, which was computed by using the mean of the PDF. If on the other hand we use the full photo-z PDF, this same metric value is 0.0221, which is almost a factor of three better. This value of a KS statistic is traditionally interpreted in that we cannot reject the null hypothesis (that both the spectroscopic and the photometric distributions are the same) at a 5% level.

We explicitly compare the spectroscopic $N(z)$ to the measured $N(z)$ distributions computed by using the mean of the photo-z PDF (blue line), the mode of the photo-z PDF (green line), and the full photo-z PDF (red line) in the top panel of Figure 4.7. In addition, the bottom panel displays the absolute error between the spectroscopic $N(z)$ and these three different measured $N(z)$ distributions. In both panels, the full photo-z PDF is clearly shown to more closely match the spectroscopic distribution, a result that we also saw in Chapter 3 with photo-z PDFs generated by using TPZ . This simple test highlights the power of using the full information provided by a photo-z PDF.

In general, computing other metrics, such as the bias or scatter, by using the photo-z PDF will produce slightly larger values than simply using the mean of the photo-z PDF, since for these simpler metrics the mean is a sufficient estimator of the full PDF, while the full PDF adds information from other bins, decreasing the precision to which these metrics are computed. While these metrics are primarily useful in merely characterizing the approximate accuracy of the algorithm, it is still important that these metrics are symmetric and unbiased as a function of redshift (which confirms the lack of any systematic biases in the algorithm).

Following our previous definition of $zConf$ from Chapter 3 (i.e., the integrated probability between $z_{\rm phot} \pm \sigma_{SOM}(1 + z_{\rm phot})$, where we have set the expected scatter $\sigma_{SOM} = 0.075$), we compare spectroscopic versus photometric redshift in Figure 4.8 by using the mean of the PDF (left panel) and full PDF (right panel) for exact same 8387 galaxies selected to have $zConf > 0.7$. Both panels are constructed from the same galaxy sample, share the same number of pixels, and have the same number of contours (although the dynamic range of the contours varies). The over plotted black dots and error bars convey the median and tenth and ninetieth percentiles, respectively, for spectroscopic bins of width $\Delta z = 0.1$

By construction, the galaxies used for Figure 4.8 all have a concentrated photo-z PDF, and as shown in Figure 4.8, by using the mean of the photo-z PDF we have a tight, symmetric relationship. This reaffirms the conclusion found in Chapter 3 —but this time for a SOM photo-z PDF— that the $zConf$ value can be used to identify galaxies with accurate photo-z estimates. Although the photo-z PDF provides a more accurate $N(z)$

relationship, by using the mean of the photo-z PDF we generate a slightly tighter correlation. On the other hand, the full photo-z PDF generally produces a more symmetric distribution, which can be seen both from this figure and the median values, except for the last two bins that suffer from low numbers as seen in Figure 4.7. As a result, the final choice of using the full PDF or a particular statistic characterizing the full photo-z PDF should be empirically quantified as it will likely depend on the particular problem under study.

4.3.8 Comparison with TPZ

As SOMz used herein is an unsupervised learning method, it can be illustrative to compare this new method to an existing, supervised learning method. As we borrow many techniques in this chapter from our random forest technique outlined in Chapter 3, in this section we compare the performance of SOMz with TPZ , specifically focusing on the results computed by using both methods for the DEEP2 dataset compiled by Matthews et al. (2013). The SOM results were produced by using a random atlas with spherical topology and online updating, while the TPZ results were produced by using the regression mode with 100 trees and $m_* = 3$ (explained in Chapter 3) generating PDFs of the same redshift resolution $\Delta z = 0.012$. As the SOM results were generated by using galaxy colors, we ran TPZ by using the same colors and the same training set used to generate the SOM results.

We present a summary of key statistics from these two estimation methods in Table 4.4. From the end results of each technique, we create three subsamples by splitting on the $zConf$ value. The first observation from these values is the similar performance of both techniques, which is somewhat surprising given the differences between the two algorithms. On the other hand, it seems likely that the randomness feature in our implementation of the random forest and the random atlas algorithms both improve the performances of these algorithms to a similar degree. When constructing a photo-z PDF, the full multi-dimensional space is subdivided (TPZ is a supervised process while SOMz is an unsupervised process) into smaller volumes, which, in both cases, contain galaxies with similar properties, that are subsequently used to make redshift predictions.

The second observation from this table is that for some metrics the random forest implementation is superior, while for others the random atlas implementation wins. Given this observation, and the inherent differences between the two approaches, it seems reasonable to want to explore the combination of the predictions from disparate learning methods. We have already started to address this issue by exploring how the performance of the photo-z approach is improved when combining techniques (Carrasco Kind & Brunner, 2013b). We defer further discussion of this topic to chapter 6 and Carrasco Kind & Brunner (2014c), where we will explore the development of meta-classifiers that combine supervised, unsupervised,

Table 4.4: A summary of key metrics that were computed by using the same datasets for solutions provided by the SOM algorithm and TPZ . The number in parenthesis is the $zConf$ value used on each case.

Method	$< \Delta z' >$	$\sigma_{\Delta z'}$	KS	KS_{PDF}	$\mathrm{out}_{0.1}$
SOM (0.5)	0.0417	0.0608	0.0659	0.0311	0.0803
TPZ (0.5)	0.0408	0.0640	0.0352	0.0175	0.0808
SOM (0.7)	0.0382	0.0586	0.0621	0.0307	0.0660
TPZ (0.7)	0.0374	0.0594	0.0320	0.0162	0.0664
SOM(0.9)	0.0318	0.0520	0.0620	0.0304	0.0427
TPZ (0.9)	0.0306	0.0516	0.0294	0.0157	0.0430

and template-fitting techniques to make more accurate photo-z PDF estimations.

4.4 Summary

We have presented a new approach that computes a photo-z PDF with similar performance as other machine learning techniques that we call SOMz. This new approach is an unsupervised machine learning algorithm that uses Self-Organized-Maps, which project the muti-dimensional space of attributes (magnitudes or colors) to a 2 dimensional map that attempts to conserve the topology of the higher dimensional data. Each neuron or cell in the map is updated after each galaxy is processed by means of weights that are iteratively corrected in order to better represent the training data. The spectroscopic target information is not used at all in the process of building the maps, although it is used to identify the galaxies that belong to a cell in order to make predictions from the two-dimensional map. In this Chapter we introduce the concept of a *random atlas*, in analogy to a random forest for decision trees, in which a number of different maps are created whose individual predictions are subsequently aggregated to produce a final photo-z PDF.

We also explored the different configurations that can be used to build a SOM, and introduce a new metric, the I-score, which efficiently takes into account different metrics indicators of the overall performance, such as, the bias, the scatter or how well the photometric redshift distribution matches the spectroscopic one, in order to differentiate these configurations. We found that by using a random subsample of attributes to build different maps we can produce a significantly better solution than by using all the attributes, which works similarly to the random forest for prediction trees. We explored two approaches to updating the weights for each cell in the final two-dimensional map. The first technique is called online updating, in which all weights are updated dynamically after processing each galaxy. The second is called batch updating, in which the galaxy weights are applied *en masse* after each iteration. Our testing indicates that online updating produces more accurate photo-z estimates, but is also harder to parallelize, which can be a limitation when applied to very large data sets. On the other hand, the batch method is easier to parallelize as the cumulative work can be done in blocks on different cores, but is slightly less precise since the weights are updated less frequently.

The SOM process constructs a two-dimensional topology; thus we also explored three different representations for this final map: rectangular, hexagonal, and spherical grids. While the rectangular and hexagonal grids showed similar performance, we found that the spherical grid performed slightly better, likely due to its natural lack of boundary conditions that avoid biases near the edge of the grid. For the other two, flat grids we also imposed wrapped, periodic boundary conditions, but they still perform slightly worse than the spherical topology when using approximately the same number of cells. Overall, however, we do see that wrapping the two-dimensional grids (either naturally, or with imposed periodic boundary conditions) provides a better two-dimensional representation of the multi-color space occupied by the galaxies in our analysis.

On the other hand, other SOM parameters had less of an effect on the final photo-z calculation. First, the number of iterations must be large enough to allow convergence. Second, if using the online method, the degree of correctiveness needs to be close to unity, while the batch method lacks this parameter. We also explored the effect of the number of cells used to construct a given map and the number of maps in a random atlas. For the former, we find that an improvement in the photo-z estimation can be achieved, but the process is limited, depending on the training data volume, as eventually we can suffer from over–fitting. For the latter, we found that after a few hundred maps, the random atlas has effectively been populated by all possible parameter configurations and no new information is added with additional maps. For our demonstration example, we found an ideal combination of approximately 1500 cells and 200 maps produced the optimal photo-z predictions.

While our new, unsupervised approach presented herein performs to a similar degree of accuracy as previous, supervised techniques, including our own TPZ algorithm presented in Chapter 3, there are different strengths and weaknesses of each approach. As a result, we have explored how to optimally combine different photo-z estimation techniques, including the use of supervised, unsupervised, and template fitting techniques (which is described in the next Chapter) into a meta-algorithm to both produce more accurate photo-z estimates as well as an improved identification of prediction outliers. This will be discussed in the next two Chapters.

Chapter 5

Spectral energy distribution fitting for photo-z: BPZ

Outline

In this Chapter we review and discuss some new work on a template fitting approach using Bayesian techniques. This is an independent approach with the advantage that it doesn't need a training set as long a high-quality and representative library of SED is available. We explore the applicability of BPZ (Benítez, 2000) on our datasets and demonstrated that when a spectroscopic training exists we can build a prior probability distribution using a functional form or a Naive Bayesian Classifier we can improve the overall photo-z solution.

In the last two chapters we have discussed two empirical techniques to compute photometric redshifts ans we have discussed the limitations of these training methods. One critical aspect is that these methods can perform remarkably well provided that a high quality and representative training set is available. This training set must contain galaxies with properties similar to those whose redshift are computed. This, by itself, represents a challenge especially if the redshift or the color distributions span a large range of galaxies. For example, computing photo-z 's at higher redshifts than the spectroscopic sample or in areas of the magnitude space poorly sampled is less reliable. We studied this issue in Chapter 3 by computing ancillary information to understand the data, its limitations and the sample variance, but this doesn't necessarily solve the problem. Instead, we can use standard techniques to compute photo-z by fitting the observed magnitudes using a set of calibrated galaxy templates where not training data is necessary. This method, however, relies on how representative is the template library used. In this chapter we introduced the template fitting approach to complement what we have developed to make more accurate and reliable photo-z predictions.

5.1 Bayesian Photometric Redshift

Using spectral templates to estimate galaxy photo-zs from broadband photometry has a long history (Baum, 1962); and this approach is, not surprisingly, one of the most utilized techniques. A primary advantage of this technique is the fact that a training sample is not required, thus this approach can be considered unsupervised. On the other hand, this technique has the disadvantage that a complete and representative library of spectral energy distributions (SEDs) are required. Thus any incompleteness in our knowledge of the template SEDs that fully span the input galaxy photometry will lead to inaccuracies or mis-estimates in the computation of a galaxy photo-z.

A number of different groups have published template fitting photo-z estimation methods, all of which are roughly similar in nature. For this thesis, we have modified and parallelized one of the most popular, publicly available template fitting algorithms, BPZ (Benítez, 2000). BPZ uses Bayesian inference to quantify the relative probability that each template matches the galaxy input photometry and determines a photo-z PDF by computing the posterior probability that a given galaxy is at a particular redshift. One advantage of using Bayesian statistics is its simplicity. From Bayesian statistics we know that the product rule states that the joint probability of a variable x and a variable y (for both to be true) is given by:

$$P(x, y) = P(y \mid x)P(y) \tag{5.1}$$

where $P(y \mid x)$ is the conditional probability of the variable y given the information about variable x. Since $P(x, y) = P(y, x)$ by construction we can easily derive Bayes theorem in the following form:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} \tag{5.2}$$

where in this example, $P(x \mid y)$ is called the *posterior* distribution of variable x (Usually the parameter space), $P(y \mid x)$ is the *likelihood* of getting y given the variable x, $P(x)$ is the *prior* probability which encodes the information on the distribution of the variable x and $P(y)$ is called the *evidence* which acts as a normalization factor, useful when multiple hypothesis are being tested. Another important rule in this Bayesian framework is the *sum* rule in which we can write $P(x)$ as:

$$P(x) = \sum_{y} P(x, y) \tag{5.3}$$

where we marginalize over all possible values of the variable y. In the case we have a set of variables in n dimensions $\mathbf{y} = \{y_i\}$ we can write the above as:

$$P(x) = \sum_{y_1, y_2, \ldots, y_n} P(x, y_1, y_2, \ldots, y_n) \tag{5.4}$$

On the other hand if the set $\mathbf{y_i}$ is mutually exclusive then we have:

$$P(x, y_1, y_2, \ldots, y_n) = P(x) \prod_{i=1}^{n} P(y_i \mid x) \tag{5.5}$$

5.2 Likelihood estimation

When applying the previous set of rules to the photo-z inference we can write this probability as $P(z \mid \mathbf{x}, vecI)$ for a specific template t, where $\mathbf{x}$ represents a given set of magnitudes (or colors) and $\mathbf{I}$ represents any kind of extra information for computing the posterior probability of the hypothesis. If the identification of a specific template is not required, we can later marginalize over the entire set of templates $\mathbf{T}$.

By using Bayes theorem, we have:

$$P(z \mid \mathbf{x}, \mathbf{I}) = \sum_{t \in \mathbf{T}} P(z, t \mid \mathbf{x}, \mathbf{I}) \propto \sum_{t \in \mathbf{T}} \mathcal{L}(\mathbf{x} \mid z, t, \mathbf{I}) P(z, t \mid \mathbf{I}) \tag{5.6}$$

$\mathcal{L}(\mathbf{x} \mid z, t, \mathbf{I})$ is the likelihood that, for a given redshift z and spectral template t, a specific galaxy has the set of magnitudes (or colors) $\mathbf{x}$. All the extra knowledge given by $\mathbf{I}$ will be affect only in the prior but not in the likelihood, therefore we can simply write this likelihood as: $\mathcal{L}(\mathbf{x} \mid z, t)$

$P(z, t \mid \mathbf{I})$ is the prior probability of a specific galaxy is at redshift z and has spectral type t given the information $\mathbf{I}$. This information can be computed from physical models of the galaxy distribution or can be refined if a spectroscopic sample if one is available. The photo-z PDF is, therefore, either the posterior probability, if a prior is used, or the likelihood itself if no prior is used. This last point arises since the likelihood only depends on the collection of template SEDs; and, if this collection is representative of the overall galaxy sample, the likelihood can be used by itself as a photo-z PDF even without a spectroscopic training sample.

As the goal of a template fitting method is to minimize the difference between observed and theoretical magnitudes (or colors), this approach is heavily dependent on both the library of galaxy SED templates that are used for the computation and the accuracy of the transmission functions for the filters used for particular survey. SED libraries are generally built from a base set of SED templates. These base templates broadly cover the Elliptical, Spiral, and Irregular categories, and a template library can be constructed by interpolating between the base spectral templates to create new spectra. One of the most widely used set of base templates are the four CWW spectra (Coleman et al., 1980), which include an Elliptical, an Sba, an Sbb, and an Irregular galaxy template. When extending an analysis to higher redshift, these temples are often augmented with two star bursting galaxy templates published by Kinney et al. (1996). One additional

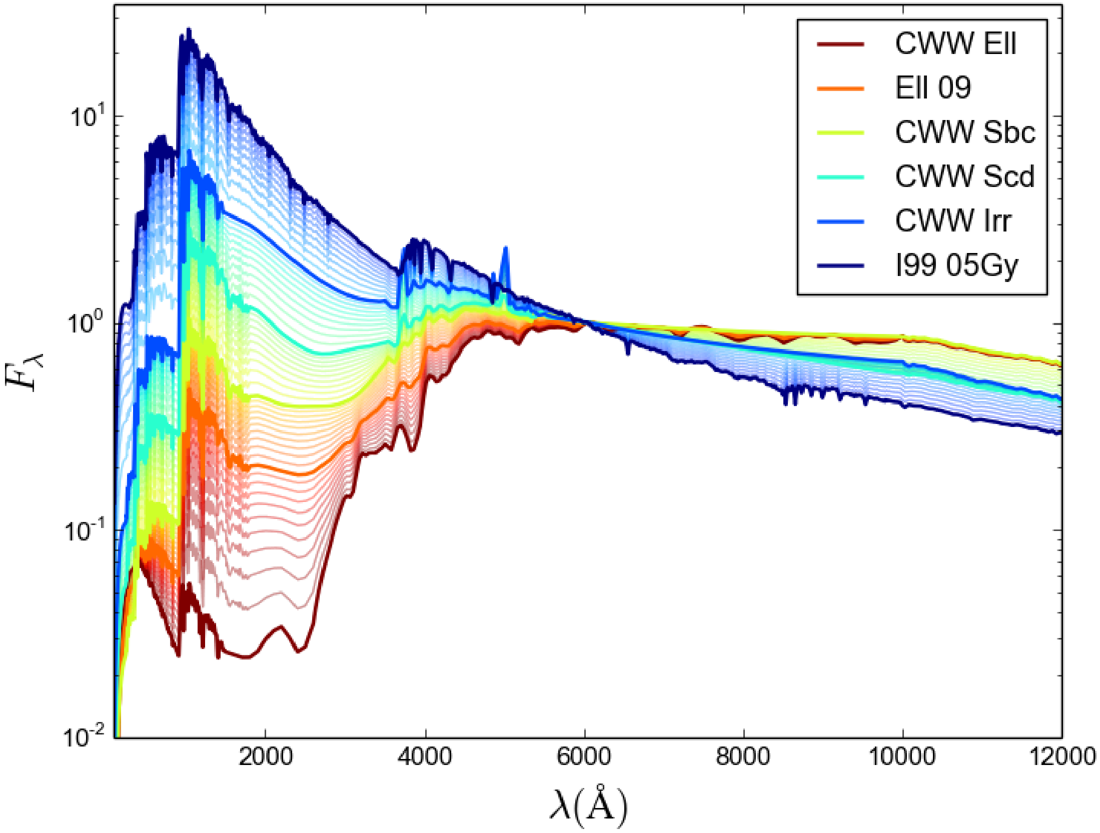


Figure 5.1: The SED templates in the rest frame extracted from the CWW library used in this thesis. For reference only it also shown the interpolation between the given templates (on thicker lines) from Elliptical galaxies (red) to Irregular galaxies (blue).

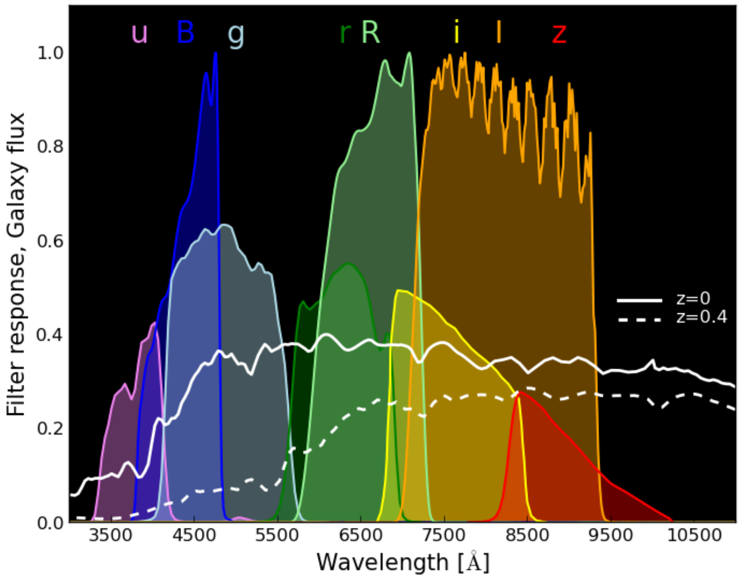


Figure 5.2: An Elliptical galaxy spectrum at z=0 and redshifted to z = 0.4 overlaid by the eight photometric filters from the DEEP2 galaxy survey (3 from the original survey and $ugriz$ from a matched catalog (Matthews et al., 2013)) as also described in §2.3.

effect some template approaches consider is the presence of interstellar dust, which will introduce artificial reddening. Figure 5.2 shows an example of a library of galaxy SED taken from Coleman et al. (1980) which have been interpolated to expand it when computing the likelihood and the prior by the algorithm. We can observe from this Figure the different spectral features expected to be observed in our galaxy sample like the 4000 Åbreak on elliptical galaxies or the [OII] and [OIII] emission lines in case of a Irregular galaxy.

Once the library of galaxy SED templates has been constructed, the templates are convolved with the transmission functions for a particular survey to generate synthetic magnitudes as a function of redshift for each galaxy template. For the most accurate results, these transmission functions should include the effects of the Earth's atmosphere (if the observations are ground-based), as well as all telescope and instrument effects. This convolution process is demonstrated visually in Figure 5.2, which presents an example Elliptical galaxy spectral template at redshift zero and at a redshift 0.4. Overplotted on this figure is the filter set (B, R, and I) used by the DEEP2 survey, which is the data analyzed in this paper, along with the five extra filters: u, g, r, i, z presented in the DEEP2 photometry catalog compiled by Matthews et al. (2013) as described in §2.3.

When this process is completed we will have a multidimensional space of magnitudes (or colors) and redshifts which can be explored, by maximizing the Likelihood (or minimizing the χ^2 statistics) to find the combination that best represent each galaxy at the time of being processed. Figure 5.3 shows an example on how, by maximizing the Likelihood $\mathcal{L}(\mathrm{x} \mid z, t)$, is it possible to obtain a reasonable good fit based on the

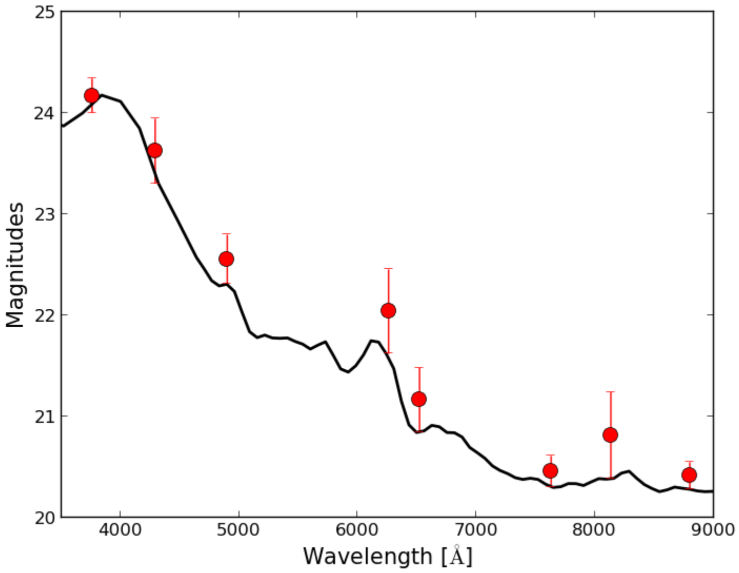


Figure 5.3: Example of one galaxy magnitudes and their errors (red dots) being fit by maximizing the likelihood using a library of galaxy templates at a several redshifts, the best fit is shown in solid black.

magnitudes if the galaxy. In this case the galaxy was taken from the DP-1 sample (from Chapter 2) and from a set of libraries which are convolved to transform them to magnitudes and redshifted the template that best represent the observations (red points) at a particular redshift $z = 0.51$ is shown as a black line. Note that we are able to fit a template to this data regardless of extra information about how likely is one template over another at this redshift or how likely is a template give a particular set of colors and/or magnitudes.

5.3 Prior information

As discussed before, it is possible to use only the maximization of the Likelihood in order to get a photo-z PDF if not extra knowledge is available. The use of a prior in a Bayesian analysis, however, is recommended as it can provide more accurate probabilities and can help to easily discriminate, in this application between two templates of galaxies or two similar redshift values with similar likelihood values.

5.3.1 Empirical function as a prior

For the photo-z computations, the prior probability can be computed directly from physical assumptions, from an empirical function calibrated by using a spectroscopic training sample (e.g., Benítez, 2000), or from an empirical function calibrated by using machine learning techniques. For example, Benítez (2000) propose the following prior function for the information $\mathbf{I}$ given by a single magnitude m_0, usually the reference magnitude:

$$P(z,t \mid m_0) = P(t \mid m_0)P(z \mid t, m_0) \propto f_T e^{-k_t(m-m_0)} \times z^{\alpha_t} \exp\left(-\left[\frac{z}{z_{mt}(m)}\right]^{\alpha_t}\right). \tag{5.7}$$

where $z_{mt}(m) = z_0 t + k_{mt}(m - m_0)$. The five parameters of this function: f_T, m_0, α_t, z_{mt}, and k_{mt} can be constrained either by using direct fitting routines, or by using Markov Chain Monte Carlo methods to sample these parameters from a given representative spectroscopic set. These five parameters are dependent on the template t and can be quantified independently. For additional details on the underlying Bayesian approach, we refer the reader to the original paper by Benítez (2000).

5.3.2 Using Machine Learning to compute the prior

We developed a Random Naïve Bayes Classier (RNBC) prior which learns from a training sample and produces individual prior for the galaxies (Carrasco Kind & Brunner, 2013b). A Naïve Bayes method is a supervised learning algorithms based on applying Bayes' theorem with the ”naive” assumption of independence between every pair of features (Zhang, 2004). It assumes that every variable can provide information about classes independently of the other variables, simplifying its framework but with high predictive power (Caruana et al., 2008).

By using available training data and RNBC, we can use the set of magnitudes $\mathbf{m}$ from the training set as our information vector $\mathbf{I}$, in this case the prior defined in Equation 5.6 $P(z, t \mid \mathbf{I})$ as :

$$P(z, t \mid \mathbf{m}) = \frac{P(z,t)P(\mathbf{m} \mid z,t)}{P(\mathbf{m})} \tag{5.8}$$

where $P(z,t) = P(z)P(t \mid z)$ which is computed from the training sample, if the template for the training galaxies is not available they can be computed easily using the Likelihood defined before $\mathcal{L}(\mathbf{x} \mid z,t)$, but for a fixed z corresponding to the spectroscopic redshift. The normalization term $P(\mathbf{m})$ is a constant which can be taken as unity. The term $P(\mathbf{m} \mid z,t)$ can be written, using that $\mathbf{m}$ is n-dimensional, as:

$$\begin{aligned}
P(\mathbf{m} \mid z,t) &= P(m_1, m_2, \ldots, m_n \mid z,t) \\
&= P(m_1 \mid z,t)P(m_2, m_3, \ldots, m_n \mid z,t,m_1) \\
&= P(m_1 \mid z,t)P(m_2 \mid z,t,m_1)P(m_3, \ldots, m_n \mid z,t,m_1,m_2) \\
&\vdots \\
&= P(m_1 \mid z,t)P(m_2 \mid z,t,m_1)\cdots P(m_n \mid z,t,m_1,m_2,\ldots m_{n-1})
\end{aligned} \tag{5.9}$$

Assuming ”naively” that magnitudes are independent (even though they are not) they contribute individually to every source information, i.e. $P(m_i \mid m_j) = P(m_i)$ for $i \neq j$ we can write the above as:

$$\begin{aligned} P(\mathbf{m} \mid z, t) &= P(m_1 \mid z, t) P(m_2 \mid z, t) P(m_3 \mid z, t) \cdots P(m_n \mid z, t) \\ &= \prod_{i=1}^{n} P(m_i | z, t) \end{aligned} \tag{5.10}$$

Therefore we can express the prior from Equation 5.8 as:

$$P(z, t | \mathbf{m}) \propto P(z) P(t|z) \prod_{i=1}^{n} P(m_i | z, t) \tag{5.11}$$

where the term $P(m_i|z,t)$ is modeling assuming a normal distribution with mean and variance computed from the training set. Priors are built by aggregating the results of several hundred bootstrap samples. We use a random subset of magnitudes in each bootstrap, similar to the construction of a random forest (Breiman, 2001; Carrasco Kind & Brunner, 2013a). As illustration on how the prior provides extra information to obtain more accurate results, Figure 5.4 shows the probabilities obtained for one example galaxy taken from the DP-1 sample. The top panel shows the likelihood obtained for three different galaxy types (Elliptical, Spiral and Irregular). The middle panel shows the prior as defined in Equation 5.11 for the different galaxy types. The bottom panel shows the comparison when using the prior or not for the same galaxy. In this case, if no prior is used, a misclassification will occur without warning as the likelihood gives a smooth and well peaked PDF at the wrong redshift for the wrong type (Spiral). By using the prior calibrated on a spectroscopic set we can get more accurate results otherwise would have been unnoticed, just using the likelihood in this example we had no reason to suspect we had a wrong galaxy type at the wrong redshift besides the fact that the PDF it is wider than the usual galaxy but not rare enough to put a flag on it. In the next chapter we will discuss how this extra information from a template fitting technique can get us closer to the limit of the information that we can extract from the galaxy samples.

5.4 Summary

We introduced the basic Bayesian framework that describe the process of computing photo-z PDFs using a template based approach. We showed with examples how the Likelihoods are computed and how this information can be combined with prior information about the galaxy distribution to obtain better constrains on the galaxy photo-z . The prior information can be obtained from a separate training set where the redshift and the morphological types is known (which can be computed by fixing z and by obtaining the best fitting

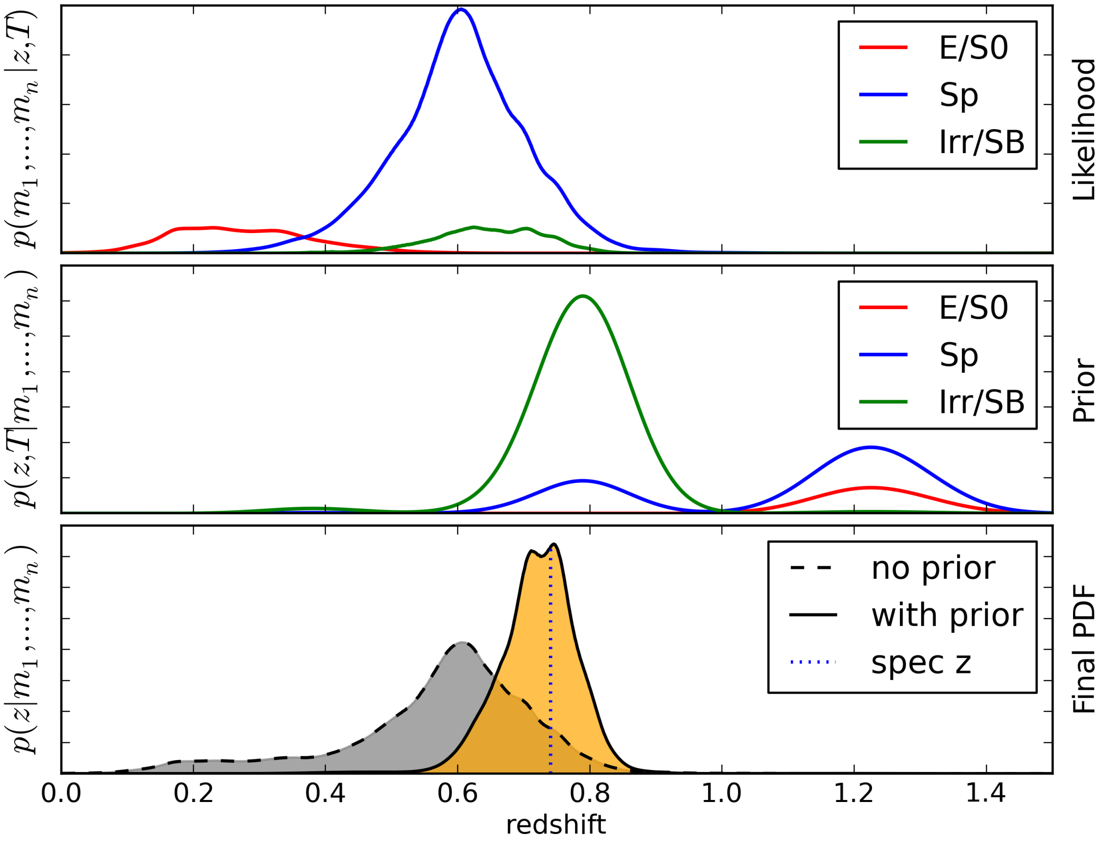


Figure 5.4: A demonstration of our the prior process by using a galaxy from the DP-1 data. (top) The likelihood for a single galaxy to be one of the different templates. (middle) The computed prior for this galaxy by using an RNBC on training data. (bottom) The final PDF computed with (solid line, orange area) and without (dashed line, gray area) the use of prior information. The vertical line shows the true redshift for this galaxy.

among the spectral types). We discussed that this prior can take a functional form or can be estimated using a Random Naive Bayes Classifier which has the advantage of computing prior for individual galaxies improving the photo-z estimation.

In the next Chapter we will discuss how to efficiently combine the photo-z techniques described so far in order to maximize the information available in the galaxy catalogs to push the accuracy of the estimated photo-z PDF to its limit, given the quality of the spectroscopic and photometric sets.

Chapter 6

Exhausting the Information: Bayesian Combination of Photo-z PDFs

Outline

In this chapter, we present a novel and efficient Bayesian framework that combines the results from different photo-z techniques into a more powerful and robust estimate by maximizing the information from the photometric data. To demonstrate this we use the supervised machine learning technique based on random forest descriped in Chapter 3, an unsupervised method based on self-organizing maps described in Chapter 4, and a standard template fitting method as shown in Chapter 5 but can be easily extend to other existing techniques. By using different performance metrics, we demonstrate that we can improve the accuracy of our final photo-z estimate over the best input technique.

6.1 Photo-z PDF Combination Methods

In the last three chapter we have described the different techniques we have developed and used in this thesis that span most of the current techniques described in the literature. We now turn our attention to the different methods with which we can combine distinct photo-z PDF estimation techniques (see e.g., Carrasco Kind & Brunner, 2013b, where we first discussed combining Bayesian and machine learning predictions). In the statistics and machine learning communities, this topic is known as *ensemble learning* (Rokach, 2010). Recently, Dahlen et al. (2013) have demonstrated that, on average, an improved photo-z estimate can be realized by combining the results from multiple template fitting methods. In this section, we build on this previous work to identify how Bayesian techniques can be used to construct a combined photo-z PDF estimator.

We can frame the problem mathematically by writing the set of photo-z PDFs for a given galaxy as a set of models $\mathbf{M}$, where each individual model M_k (e.g., TPZ, SOMz, or modified BPZ) provides a distinct photo-z PDF or posterior probability. A photo-z PDF can be written as $P(z \mid \mathbf{x}, \mathbf{D}, M_k)$, where $\mathbf{x}$ is the set of magnitudes or colors (note that without loss of generality we can use other attributes in this process) used to make the prediction and $\mathbf{D}$ corresponds to the training set which consists of N_d galaxies. We can also abbreviate this photo-z PDF as $P_k(z)$. These photo-z PDFs are each subject to the following constraint:

$$\int_{z_1}^{z_2} P_k(z)dz = 1 \tag{6.1}$$

for every model M_k, where z_1 and z_2 are the lower and upper limits, respectively, for the redshift range spanned by the galaxy sample. In the following subsections, we introduce different methods to aggregate these photo-z PDFs and show the results of these different methods in §6.2.

Given the variety of photo-z PDF estimation methods we are using (i.e., supervised, unsupervised, and model-based), we fully expect the relative performance of the individual techniques to vary across the parameter space spanned by the data. For example, supervised methods should perform the best in areas populated by high quality training data, while unsupervised or model-based methods should perform better where we have little or no training data. As a result, we can bin a specific subspace of our multi-dimensional parameter space and apply an individual combination method to each bin separately. This technique is demonstrated later in more detail with the Bayesian Model Averaging method (although it is more generally applicable).

6.1.1 Weighted Average

The simplest approach to combine different photo-z PDF techniques is to simply add the individual PDFs and renormalize the sum. In this case the final photo-z PDF is given by:

$$P(z \mid \mathbf{x}, \mathbf{M}) = \sum_k P(z \mid \mathbf{x}, M_k). \tag{6.2}$$

We can improve on this simple approach by including weights in the previous equation:

$$P(z \mid \mathbf{x}, \mathbf{M}) = \sum_k \omega_k P(z \mid \mathbf{x}, M_k). \tag{6.3}$$

These weights, ω_k, can be estimated for each input method by using the cross validation or OOB data, or from an intrinsic characteristic of the photo-z PDF, such as $zConf$ that we introduced in Chapter 3. In this work we use three weight schemes in addition to the uniform case:

PDF shape weights

In this case, ω_k is given by the the $zConf$ parameter, which is similar to the *odds* parameter presented in Benítez (2000) $zConf$ is defined as the integrated probability between $z_{\rm phot} \pm \sigma_k(1 + z_{\rm phot})$, where $z_{\rm phot}$ is a single estimated value for the photo-z PDF. This single photo-z estimate can be either the mean or the mode of the photo-z PDF. Likewise, we can estimate σ_k for each input method either by using the OOB data, by selecting a constant value across all input methods, or by selecting these values separately so that all photo-z PDFs have the same cumulative $zConf$ distributions. $zConf$ quantifies the sharpness of the PDF and can take values from zero to one. In Chapter 3 and Chapter 4, we demonstrated that there is a correlation between this value and the accuracy of the overall photo-z. Specifically, we observed that, on average, galaxies with higher $zConf$ have more accurate photo-z PDFs than galaxies with lower $zConf$ values.

Best fit weights

An alternative method to compute the values of ω_k is to use the cross-validation data to first determine the weight values that minimize the difference between $z_{\rm phot}$ and $z_{\rm spec}$; and, second to apply these best fit values to the test data. This method seeks the optimal linear combination of each individual PDF, thus it allows the values of ω_k to be negative. After the combination is completed, we renormalize according to Equation 6.1. This method can be applied to a binned sub-sample to take advantages of the performance of each method in different areas of the attribute space.

Oracle scheme

As mentioned, when the input, multi-dimensional data have been binned (c.f. Figure 6.7), we can use the cross-validation data to select only one model from among all available input models to only be used with the test data located within that specific bin. Since we are allowed to only select one input model, this will result in an assigned weight value of one for the chosen model and zero otherwise, however the chosen model is allowed to vary between bins.

The primary disadvantage of these simple, additive models is that incorrect estimates for the errors for the selected input model can bias the final result. On the one hand, if a technique has underestimated errors, the final result will be biased towards this one input method. On the other hand, overestimation of the errors will bias the final result away from this particular method. One approach to address this issue, as discussed by Dahlen et al. (2013), is to either smooth or sharpen the photo-z PDFs estimated by each method by using the OOB data until their error distributions are approximately Gaussian with unit variance. We can generalize this approach to transform a photo-z PDF as $P_k(z) = P_k(z)^{\alpha_k}$, where we adjust the value of α_k by using either the cross validation data when errors are over estimated or use a Gaussian smoothing filter when they are under estimated.

6.1.2 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is an ensemble technique that combines different models within a Bayesian framework. BMA accounts for any uncertainty in the correctness of a given model by integrating over the model space and weighting each model by the estimated probability of being the *correct* model. As a result, BMA acts as a model selection procedure that handles the uncertainty in selecting the best model by using a combination of models instead. This is because BMA considers the uncertainty in selecting the best model while working under the assumption that only one model is actually the best (Monteith et al., 2011). BMA has been used for astrophysical problems (see e.g., Gregory & Loredo, 1992; Trotta, 2007; Debosscher et al., 2007) in, for example, the determination of cosmological parameters and variable star classification (see, Parkinson & Liddle, 2013, for a review on using BMA in astronomy).

When using BMA, the training data are used to characterize each of the models that will be combined. For each galaxy, the final PDF, $P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M})$, is given by:

$$P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}) = \sum_{k} P(z \mid \mathbf{x}, M_k) P(M_k \mid \mathbf{D}) \tag{6.4}$$

$P(M_k \mid \mathbf{D})$ is the probability of the model M_k given the training data $\mathbf{D}$, which can be viewed as a simple, model dependent weighting scheme. This probability can be computed by using Bayes' Theorem:

$$P(M_k \mid \mathbf{D}) = \frac{P(M_k)}{P(\mathbf{D})} P(\mathbf{D} \mid M_k) \propto P(M_k) \prod_{i=1}^{N_d} P(d_i \mid M_k) \tag{6.5}$$

We have omitted the $P(\mathbf{D})$ term as it is merely a normalization factor and we use the same data for all models. d_i is the i^{th} element from the training data $\mathbf{D}$, which are assumed to be independent.

For each model, we assign the value ϵ_k as an average error for the estimation process. ϵ_k can be computed as the fraction $N_k^{(b)}/N_d$, where $N_k^{(b)}$ is the number of galaxies considered to be misestimated or *bad* for the particular photo-z PDF method k. To quantify when a specific galaxy is a bad prediction we compute

$$N_{k,i}^{(b)} = \begin{cases} 1 & \text{if } \int_{z_s-\delta_z}^{z_s+\delta_z} P(z \mid \mathbf{x}, d_i) dz \leq \pi_z, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

In this equation, z_s is the spectroscopic redshift for the i^{th} training set galaxy. The first parameter, δ_z, controls the width of a window centered on z_s within which we accumulate photo-z probability for the i^{th} training galaxy around the true redshift. The second parameter, π_z, is the minimum probability within this window for which we consider the model prediction to be good. We find that $\pi_z = 0.5$ and $\delta_z = 0.05$ provides a good discriminant between good and bad photo-z model estimates.

Given the individual good/bad predictions for each training set galaxy, we can compute the total number of bad predictions, $N_k^{(b)}$, by summing over the individual predictions, $N_{k,i}^{(b)}$, for the entire training data, $\mathbf{D}$. The total number of good prediction will naturally be $N_d - N_k^{(b)}$. As a result, we can rewrite Equation 6.5:

$$P(M_k \mid \mathbf{D}) \propto P(M_k)(1-\epsilon_k)^{N_d - N_k^{(b)}}(\epsilon_k)^{N_k^{(b)}}, \tag{6.7}$$

where $P(M_k)$ is the probability of each model k, which we can assume to be unity for all models. Therefore, the final PDF for each galaxy is given by

$$P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}) \propto \sum_k P(z \mid \mathbf{x}, M_k) P(M_k) \times (1-\epsilon_k)^{N_d - N_k^{(b)}}(\epsilon_k)^{N_k^{(b)}}. \tag{6.8}$$

We applied the BMA technique to individual bins within the multi-dimensional parameter space occupied by a given data set. We demonstrate this binned BMA technique in Figure 6.7, where we use a Self Organized Map to project our entire input parameter space to a two-dimensional map. In this manner, all magnitudes or colors are used to form the binned regions within which the parameters of the ensemble learning approach can vary. After computing photo-z PDFs for all galaxies with each method, we use BMA to determine the relative weights for these input techniques within each bin; we can visualize these weights as different colors across the two-dimensional map, as shown in Figure 6.7. This figure graphically displays how the *accuracy* of each photo-z PDF estimation varies across the parameter space, and thus how the different weights themselves vary.

6.1.3 Bayesian Model Combination

As discussed, Bayesian Model Averaging tries to select the best model among the ones introduced to the algorithm. Alternatively, we can modify BMA to produce an more optimal model combination technique (Monteith et al., 2011) known as Bayesian Model Combination (BMC). With BMC, instead of directly combining the three different photo-z PDF estimates as was the case with BMA, the Bayesian process is used to explore different combinations of the individual photo-z PDF techniques. Thus, an ensemble of different photo-z PDF combinations are generated and we directly compare different model combinations.

As a simple example, we could first generate hundreds different random weights for all three of our photo-z PDF estimation techniques, and second use these to compute hundreds of new *sets* of PDFs by computing a simple weighted average by using Equation 6.3. Finally, we could apply BMA to this PDF ensemble to determine the final PDF. In this case, we could write Equation 6.4:

$$P(z \mid \mathbf{x}, \mathbf{D}, \mathbf{M}, \mathbf{E}) = \sum_{e \in \mathbf{E}} P(z \mid \mathbf{x}, \mathbf{M}, e) P(e \mid \mathbf{D}), \tag{6.9}$$

where e is an element from the set $\mathbf{E}$ of these hundreds combined models. Here we need to compute the performance of each combination e and apply the BMA formulation, shown in Equations 6.5 and 6.6, to those models by using the model e instead of M_k, i.e.,

$$P(e \mid \mathbf{D}) \propto P(e) \prod_{i=1}^{N_d} P(d_i \mid e). \tag{6.10}$$

Fundamentally, with BMC we are marginalizing over the uncertainty in the correct model combination, where in BMA we marginalized over the uncertainty in identifying the correct model from the entire ensemble.

The number of model combinations $\mathbf{E}$ is, in principle, infinite, and in practice can be very large. To overcome this, we can use sampling techniques over a reasonable, finite number of models. Naively we might use randomly generated weights, however, this approach can be costly to fully span the allowed range of weights and convergence towards a satisfactory solution might be slow. Thus, instead of assigning weights randomly or using incremental steps within a regular grid, we sample the weights from a Dirichlet distribution where the *concentration* parameters are modified until they converge to stable values. We require that the set of weights, w_k, for each of the three models, M_k, satisfy $\sum w_k = 1$ and also $w_k > 0$.

For a concentration parameter $\boldsymbol{\alpha}$ of the same dimension as $\mathbf{w}$, we have that the probability distribution for $\mathbf{w}$ is given by:

$$P(\mathbf{w}) \sim \mathcal{D}\text{ir}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k w_k^{\alpha_k - 1}, \tag{6.11}$$

where $\mathcal{D}\text{ir}(\boldsymbol{\alpha})$ is the *Dirichlet* distribution, $\Gamma(\alpha_k)$ is the *gamma function* and k are the base models, which in this thesis are `TPZ`, `SOM`z, and our modified `BPZ`. In order to generate a set $\mathbf{E}$ of combined models, we first set α_k to unity for all values of k. Second, we sample from this distribution n_s times (n_s is a fixed number, generally between 2 and 5, which we fixed at 3) to get a set of n_s weights and n_s new model combinations. Next, we compute $P(e \mid D)$ by using Equations 6.5 and 6.6 for each model in the set of n_s models. We, temporarily, select the best model among the set n_s, i.e, the one with highest $P(e \mid \mathbf{D})$, and update the α_k parameters by simply adding the weights from the corresponding model to the current values of $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{t} + \max_{\mathbf{w}_e \in n_s} P(e \mid \mathbf{D}) \tag{6.12}$$

where t is just a symbolic reference to the fact that $\boldsymbol{\alpha}$ is being updated every 3 steps.

We use the latest values for $\boldsymbol{\alpha}$ to continue the sampling process to obtain the next set n_s of model combinations. As a result, we continually (by adding n_s new models at each step) extend our set of model combinations $\mathbf{E}$. As the chain of models in this set is constructed iteratively, the process can be terminated either when a predefined number of model combinations has been reached or when new model combinations have started to converge. This process behaves similarly to a Markov Chain Monte Carlo process, and we have an analogous phase to the *burn in* step, where we can omit some number of model combinations at the start of our set $\mathbf{E}$ of model combinations. Thus, our final photo-z PDF prediction is the application of BMA over the remaining elements in $\mathbf{E}$, we have set for this work the size of E to be 800. Finally, we note that, as was the case with BMA, we can develop a binned version of our BMC technique, where we develop different model combinations for different region of the magnitude (color) space by using a SOM.

6.1.4 Hierarchical Bayes

A Hierarchical Bayesian (HB) method provides a different approach to combine the individual photo-z PDFs. In a manner similar to BMA, we include the uncertainty that a given photo-z PDF for a specific galaxy might be incorrectly predicted as a set of nuisance parameters over which we later marginalize.

Adopting our previous notation, we follow a similar approach to Fadely et al. (2012) and Dahlen et al. (2013), and we write the photo-z PDF for an individual galaxy for each base method k:

$$P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k) = \sum_j P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_{kj}) \times P(\theta_{kj} \mid \mathbf{D}, M_k), \tag{6.13}$$

where we have introduced the *hyperparameter* θ_k, a nuisance parameter that characterizes our uncertainty in the prior distribution of model k. The parameter θ_k can be quantified in different forms, but essentially is the misclassification probability of the k^{th} method. Thus, we quantify this mis-prediction probability with $P(\theta_k)$; and we drop the dependence on $\mathbf{x}$, the measured galaxy attributes, as it does not directly affect the parameter θ_k. Since we will marginalize over θ, we keep the term $\mathbf{D}$ as we can use the training data to place limits on θ_k by using the cross-validation data. We note that these probabilities are subject to:

$$\sum_j P(\theta_{kj} \mid \mathbf{D}, M_k) = 1. \tag{6.14}$$

If we consider the case where galaxies are predicted correctly or are outliers, j is a binary state. In this model, if we assume that γ_k is the fraction of galaxies that are mispredictions or are labeled as outliers for method k, we have: $P(\theta_{k0} \mid \mathbf{D}, M_k) = \gamma_k$ and $P(\theta_{k1} \mid \mathbf{D}, M_k) = (1 - \gamma_k)$. In this case, Equation 6.13 becomes:

$$P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k) = P_{def}(z \mid M_k, \theta_k)\gamma_k + P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k)(1 - \gamma_k), \tag{6.15}$$

where $P_{def}(z \mid M_k, \theta_k)$ is the default PDF that should be used for the k^{th} method when the original PDF for that method has been determined to be mis-predicted or wrong. In the second term, we use the original PDF for the method k, which is multiplied by the fraction of well predicted objects $1 - \gamma_k$.

The final PDF after we combine the different photo-z PDFs from our base methods in the HB approach is given by:

$$P(z \mid \mathbf{x}, \mathbf{D}, \theta) = \prod_k P(z \mid \mathbf{x}, \mathbf{D}, M_k, \theta_k)^{1/\beta}. \tag{6.16}$$

Here, following Dahlen et al. (2013), we have introduced an extra parameter β, which is a constant value that quantifies the degree of covariance between the different base methods. $\beta = 1$ corresponds to complete independence between the base methods, while $\beta = 3$ (or, more generally, the total number of methods) would correspond to full covariance between them. We can compute β from the OOB sample in such way the final error distribution follows a normal distribution with zero mean and unit variance, as we have done in this chapter. Alternatively, we can marginalize over all possibles values of β when no cross validation data is available and we can integrate over the uncertainty of this parameter.

Finally, by marginalizing over θ we have our final PDF: $P(z \mid \mathbf{x}, \mathbf{D})$, or simply $P(z)$ given by:

$$P(z) = \int_0^1 P(z \mid \mathbf{x}, \mathbf{D}, \theta) P(\theta) d\theta, \tag{6.17}$$

where $P(\theta)$ is a constant which in the simple case is equal to unity. If OOB data is available, we can narrow down the range of allowed values for θ (or effectively γ_k), so we can set up a limited range for γ_k based on the performance of each method k on this data. In this case, $P(\theta)$ will act as a top-hat window function. In any case, the final $P(z)$ is subject to Equation 6.1. As discussed before, we can either apply the HB approach to the entire data set, or we can partition the input space and apply the HB approach independently to the binned regions of the parameter space.

6.2 Results/Discussion

We now turn to the actual application of the ensemble learning approaches described in §6.1 to the data introduced in §2.3. We present the seven combination methodologies we use in this section in Table 6.1, which also includes an abbreviated name that we will use to refer to a specific technique. We follow a

Table 6.1: The photo-z PDF combination methods, their weights and abbreviations presented in this chapter.

Method	Weights[a]	Abbreviation
Weighted Average	Uniform	$\mathrm{WA}_{\mathrm{flat}}$
Weighted Average	$zConf$	$\mathrm{WA}_{\mathrm{shape}}$
Weighted Average	best fit	$\mathrm{WA}_{\mathrm{fit}}$
Weighted Average	oracle predictor	$\mathrm{WA}_{\mathrm{oracle}}$
Bayesian Model Averaging		BMA
Bayesian Model Combination		BMC
Hierarchical Bayes		HB

[a] if applicable

similar approach to Chapter 4 in order to compare different combination methods, and define the bias to be $\Delta z' = |z_{\mathrm{phot}} - z_{\mathrm{spec}}|/(1 + z_{\mathrm{spec}})$. We also present the standard metrics we use to compare the performance of the different combination techniques in Table 6.2. As shown in this table, we define five metrics to address the bias and the variance of the results (the first five rows) and we present three values to characterize the outlier fraction.

We also use the KS metric, which represents the results of a Kolmogorov–Smirnov test that quantifies the likelihood that the predicted photo-z distribution and the spectroscopic redshift distribution $N(z)$ are drawn from the same underlying population. This metric provides a single, robust value to compare both distributions that does not depend on how the results are binned by redshift, and it is defined as the maximum distance between both empirical distributions.

To determine this statistic, we compute the empirical cumulative distribution function (ECDF) for both distributions. For the spectroscopic sample, the ECDF is defined as:

$$F_{\mathrm{spec}}(z) = \sum_{i=1}^{N} \Omega_{z^i_{\mathrm{spec}} < z} \tag{6.18}$$

where N is the number of galaxies in the redshift sample, and

$$\Omega_{z^i_{\mathrm{spec}} < z} = \begin{cases} 1, & \text{if } z_{\mathrm{spec},i} < z \\ 0, & \text{otherwise} \end{cases} \tag{6.19}$$

The ECDF for the photo-z distribution is simply the accumulation of the probability presented in the photo-z PDF. The summation is carried out over all galaxies in the sample. Given the ECDF for both the photo-z and spectroscopic distributions, we compute the KS statistic as:

$$\mathrm{KS} = \max_z \left(||F_{\mathrm{phot}}(z) - F_{\mathrm{spec}}(z)|| \right) \tag{6.20}$$

Table 6.2: The definition of the metrics used to compare different photo-z combination methods.

Metric	Meaning
$< \Delta z' >$	mean of $\Delta z'$
$\lvert\Delta z'\rvert_{50}$	median of $\Delta z'$
$\sigma_{\Delta z'}$	Standard deviation of $\Delta z'$
σ_{68}	Sigma value at which 68% of $\Delta z'$ is enclosed
σ_{MAD}	Median absolute deviation = $\mathrm{median}(\lvert\lvert\Delta z' - \lvert\Delta z'\rvert_{50}\rvert\rvert)$
KS	Kolmogorov - Smirnov statistic for $N(z)$
$\mathrm{out}_{0.1}$	Fraction of outliers where $\Delta z' > 0.1$
$\mathrm{out}_{2\sigma}$	Fraction of outliers where $\lvert\Delta z' - < \Delta z' > \rvert > 2\sigma_{\Delta z'}$
$\mathrm{out}_{3\sigma}$	Fraction of outliers where $\lvert\Delta z' - < \Delta z' > \rvert > 3\sigma_{\Delta z'}$
$I_{\Delta z'}$	I-score, a weighted combination of all other metrics.

Thus, as the KS statistic decreases, the two distributions become more similar.

All of the metrics listed in Table 6.2 are positive and characterized by the fact that lower metric values indicate a more accurate photo-z PDF. In Chapter 4 we defined a new, meta-statistic called I-score (symbolically represented by $I_{\Delta z'}$) that provides a single statistic to simplify the comparison of different photo-z techniques. To compute this metric, we first normalize each set of metrics across all different photo-z estimation techniques so that we are not biased by different dynamic ranges. Thus, for example, we first compute the mean and standard deviation for $< \Delta z' >$ for each combination technique, and subsequently rescale all individual $< \Delta z' >$ values so that this set of values has zero mean and unit variance.

We continue this process for all nine statistics listed in Table 6.2, and compute their weighted sum to obtain the total I-score:

$$I_{\Delta z'} = \sum w_i M_i, \tag{6.21}$$

where M_i is the rescaled metric and weight value for metric i out of the nine available. For simplicity, we use equal weights in the remainder of this chapter (and thus the I-score is simply the average of the nine rescaled metrics for each technique). As a result, the photo-z method (or parameter configuration) with the lowest I-score will be the optimal estimation technique. On the other hand, if we were looking for the technique or the specific parameter configuration with, for instance, the lower outlier fraction, we could assign higher weights accordingly to select the best technique. In this way, we can efficiently select the best method or configuration for specific research requirement.

6.2.1 Cross validation data

In Chapter 3, we introduced OOB data and demonstrated its use as a cross-validation data set that provided error quantification and overall performance similar to what could be expected when applying an algorithm directly to the test data set. When building a tree with TPZ or a map with SOMz, a fraction of the overall training data, usually one-third, is extracted and not used during the tree/map construction process. The resultant tree/map is subsequently applied to this unused data to make a photo-z prediction, and this process is repeated for every tree/map. These photo-z predications are aggregated for each galaxy to make a photo-z PDF; and by construction a galaxy can never be used to train any tree/map that is subsequently used to predict that galaxy's photo-z. Thus, as long as the OOB data remains similar to the final testing data, the OOB data provide results that will be similar to the final test data results and can be used to guide expectations when applied blindly to other data.

As an illustration of this process, Figure 6.1 compares the photometric (as computed by using SOMz) and spectroscopic redshifts for galaxies in the training (5,000 in total) and testing (5,210) samples as selected from field 1 of the DEEP2 data set, DP-2 set described in Chapter 2. As shown in this Figure, the performance on both the OOB and the testing data are visually similar and there is no indication of overfitting. In addition, general features in the result, like the spread of the data or the slight tilt of the distribution of points relative to the diagonal, are observed in both samples.

A similar conclusion is observed with the SDSS data, as shown in Figure 6.2 where the photometric (as computed by using TPZ) and spectroscopic redshifts for 50,000 galaxies from the training set are compared to 50,000 randomly selected galaxies from the test set from the sample SL-2 from §2.1. Both distributions show similar behavior and global trends, thus we conclude that, as expected, the OOB data can be used to predict the performance of an PDF combination algorithm on real data.

Another method to contrast the results from these data is to compute the correlation between each of the three photo-z estimation techniques discussed earlier as a function of redshift. For this, we use the photo-z PDFs for all galaxies, and we calculate the Pearson correlation coefficient R_{ik} within each redshift bin. Even if the three input methods are completely independent, we should expect a positive correlation between them if their predictions are similar. In fact, we desire a positive correlation (but not necessarily a perfect correlation) between the techniques as this will indicate the different techniques are all performing well.

We present the Pearson correlation coefficient for the three photo-z PDF estimation techniques for the DEEP2 data (top panel) and the SDSS data (bottom panel) in Figure 6.3. In this figure we display these correlation coefficient computed from the cross-validation (OOB) data (dashed line) and the test data (solid line). The global agreement between these lines further demonstrates the importance of the OOB data as a predictor of the performance of a given technique. This figure also demonstrates a tighter correlation between the two machine learning algorithms than between any machine learning algorithm and the template technique, which is not surprising given the similarities in the methods. While not shown, the shape of the

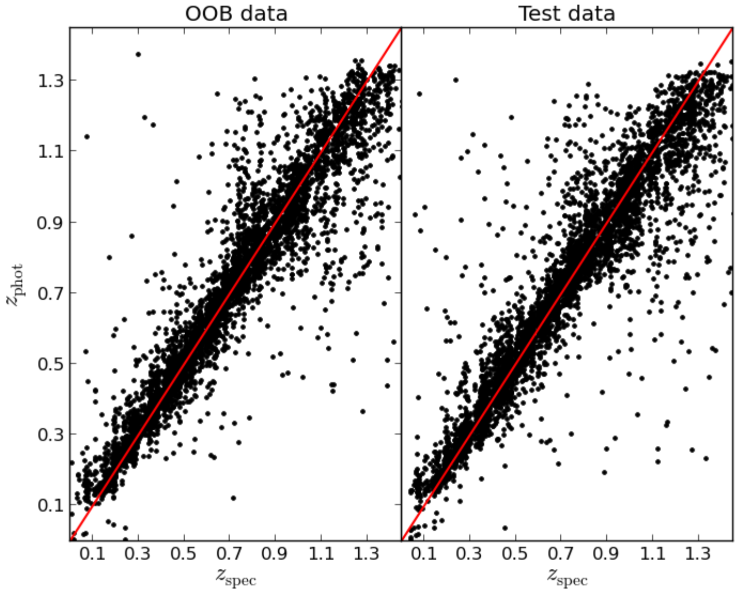


Figure 6.1: A comparison of the photometric (computed by using SOMz) and spectroscopic redshifts for training set (left) and test set (right) galaxies from field 1 of the DEEP2 survey (DP-2 set).

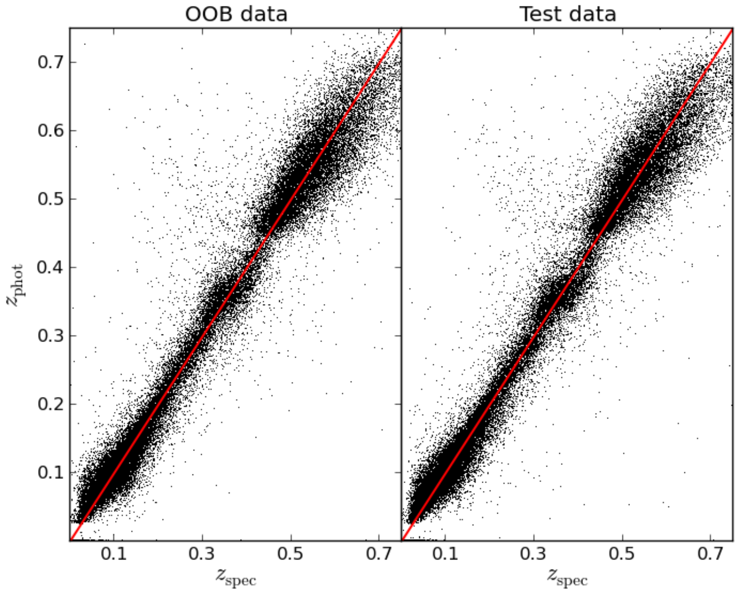


Figure 6.2: A comparison of the photometric (computed by using TPZ) and the spectroscopic redshift from the SDSS-DR10 for the 50,000 training set galaxies (left) and 50,000 galaxies randomly subsampled from the 1,097,397 galaxies in the test set (right). Data corresponding to the SL-2 data set from Chapter 2

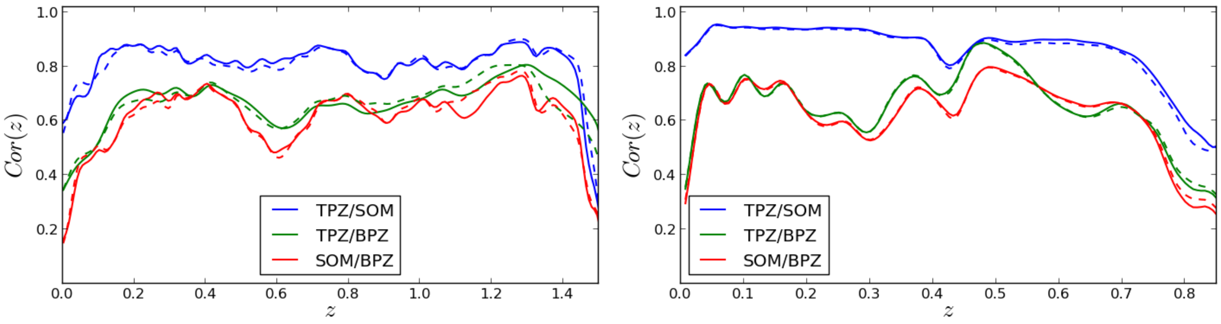


Figure 6.3: The Pearson correlation coefficient between the individual photo-z PDF estimation methods as a function of redshift for the DEEP2 (top) and SDSS (bottom) data. The coefficients measured from the cross-validation (OOB) data (dashed line) and from the test data (solid line) are nearly identical, indicating the utility of the OOB data in predicting the performance of an algorithm on blind test data. Note that a positive correlation is beneficial since this measures the relative performance of different techniques in predicting redshifts.

covariance matrices resemble the spectroscopic $N(z)$ distributions presented in 6-combo/Figures 6.9 and 6.13. We conclude that this is expected since a larger number of galaxies can naturally produce a greater chance for divergent photo-z estimates.

As mentioned previously, a concern when combining photo-z PDFs from different methods is to reduce the likelihood of being biased by methods that might under- or overestimate their errors. To further demonstrate the importance of the cross-validation data, we compare the normalized error distribution between the cross-validation (OOB) and test data in Figure 6.4 for both DEEP2 (top panel) and SDSS (bottom panel) data, where the photo-z PDFs were generated by TPZ . In both cases, the two curves are nearly identical, and we confirmed the same result with both SOMz and BPZ. Thus we can use the OOB data error estimate to rescale the PDF for the test data by using the results computed from the OOB data.

6.2.2 Photo-z PDF Combination for DEEP2

To combine the three photo-z PDF techniques discussed in previous chapters, we employ a binning strategy to allow different method combinations to be used in different parts of parameter space. We first create a two dimensional, 10×10 SOM representation of the full 14-dimensional space (eight magnitudes and six colors, note that we do not compute a color between the two different photometric input surveys) by using a rectangular topology to facilitate visualization. With this map we can perform an analysis of all galaxies that lie within the same cell, in a similar process to that described in Chapter 4, but now instead of predicting a photo-z, we are computing the optimal model combination. We apply all seven combination methods presented in Table 6.1 to all galaxies within each cell by using the OOB data that are also contained within the same cell. We note that the $\mathrm{WA}_{\mathrm{flat}}$ and $\mathrm{WA}_{\mathrm{shape}}$ methods do not depend on this binning, and can,

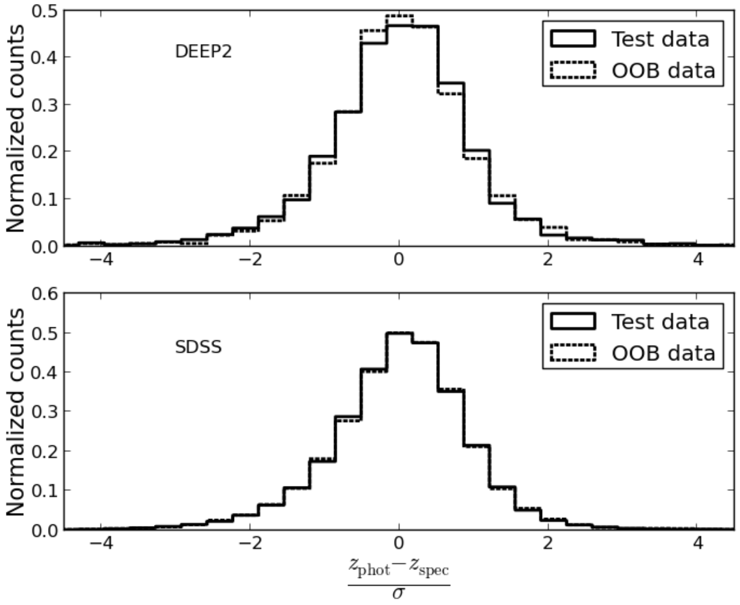


Figure 6.4: The normalized error distributions for galaxies in DEEP2 (top) and SDSS (bottom). The error distribution computed from the test data is shown in red, while the error distribution for the cross-validation (OOB data) is shown in black. The excellent agreement highlights the importance of the OOB data in predicting the results of blind test data predictions.

therefore, be used without OOB data. We also could employ the HB approach without using this map, but in this case we would need to define $P_{def}(z \mid M_k, \theta_k)$ and perform the marginalization over the entire range of θ_k without any prior on this value.

We present a summary of the results obtained by applying the seven different combination techniques to all the galaxies within the DEEP2 data in Table 6.3. The bold entries in this Table highlight the best technique for any particular metric. The first three rows in this Table show the individual photo-z PDF estimation techniques, of which TPZ generally performs the best and is thus shown in the first row as the benchmark. This Table also clearly indicates that the seven different combination techniques generally have a similar performance, and, as shown in the last four rows, often perform better than TPZ.

We observe that the last four methods: $\mathrm{WA}_{\mathrm{fit}}$, BMA, BMC, and HB all use the binned model combination approach, and thus can take advantage of the different performance characteristics of individual codes. In this case, BMC provides the best performance as measured by the I-score $I_{\Delta z'}$, the bias $< \Delta z' >$, the scatter $\sigma_{\Delta z'}$, and the outlier fraction $\mathrm{out}_{0.1}$. Overall, the differences are close to 5% for many of the metrics, which, while small, are still significant since these are averaged metrics over the full test galaxy sample.

In Figure 6.5, we present a visual comparison between the ten different photo-z estimation techniques for five different metrics: bias, scatter, outlier fraction, KS test, and the I-score. In each panel, the horizontal

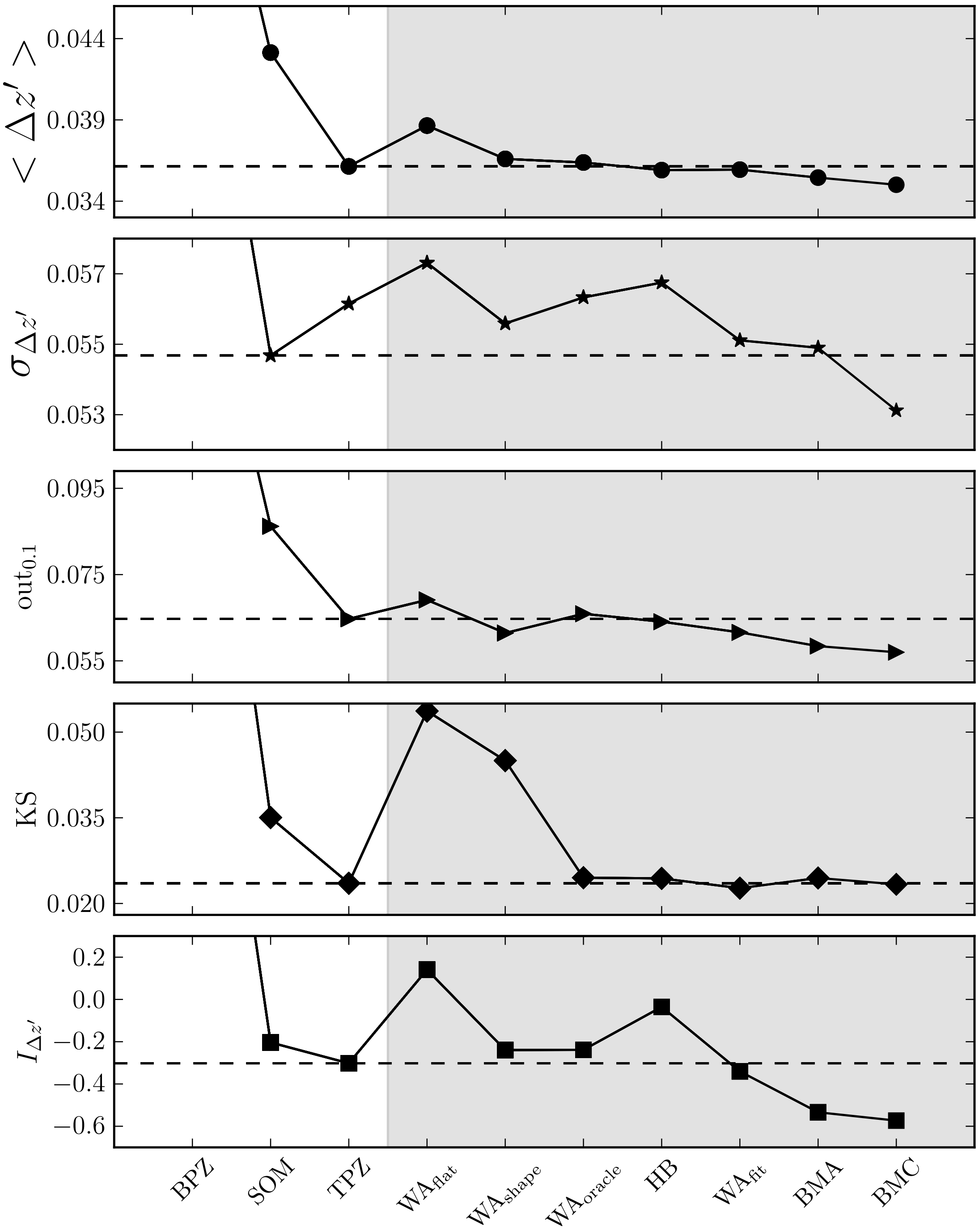


Figure 6.5: A comparison of the average performance for the three individual photo-z PDF estimation methods and the seven different photo-z PDF combination approaches for five different metrics as defined in Table 6.2 for the DEEP2 data. The horizontal dashed line indicates the best result for a given statistic among the three individual methods (note, BPZ is not always shown at the provided scale), and the shaded area separates the individual methods from the combined approaches. All values are presented in Table 6.3.

Table 6.3: A summary of the performance results for the three individual methods and the seven different photo-z PDF combination methods as applied to the DEEP2 data, no magnitude cut was applied during the training phase. The bold entries highlight the best value within each column to aid in the interpretation of the table (c.f. Figure 6.5).

Combination method	$< \Delta z' >$	$\|\Delta z'\|_{50}$	$\sigma_{\Delta z'}$	σ_{68}	σ_{MAD}	KS	$\mathrm{out}_{0.1}$	$\mathrm{out}_{2\sigma}$	$\mathrm{out}_{3\sigma}$	$I_{\Delta z'}$
TPZ	0.0361	0.0205	0.0561	0.0257	0.0139	0.0235	0.0647	0.0307	0.0184	-0.3021
SOM	0.0431	0.0291	0.0547	0.0325	0.0188	0.0350	0.0862	**0.0284**	**0.0150**	-0.2035
BPZ	0.0635	0.0476	0.0679	0.0428	0.0273	0.1342	0.1636	0.0338	0.0170	2.3255
$\mathrm{WA}_{\mathrm{flat}}$	0.0386	0.0231	0.0573	0.0285	0.0155	0.0537	0.0691	0.0313	0.0192	0.1409
$\mathrm{WA}_{\mathrm{oracle}}$	0.0364	0.0206	0.0563	0.0260	0.0139	0.0245	0.0659	0.0313	0.0184	-0.2385
$\mathrm{WA}_{\mathrm{shape}}$	0.0366	0.0217	0.0556	0.0268	0.0146	0.0450	0.0614	0.0297	0.0186	-0.2392
$\mathrm{WA}_{\mathrm{fit}}$	0.0359	0.0208	0.0551	**0.0253**	**0.0137**	**0.0227**	0.0616	0.0318	0.0178	-0.3404
BMA	0.0355	0.0211	0.0549	0.0257	0.0140	0.0245	0.0584	0.0289	0.0178	-0.5339
BMC	**0.0350**	0.0208	**0.0531**	0.0255	0.0140	0.0233	**0.0570**	0.0297	0.0176	**-0.5734**
HB	0.0359	**0.0199**	0.0568	0.0259	**0.0137**	0.0244	0.0641	0.0329	0.0196	-0.0354

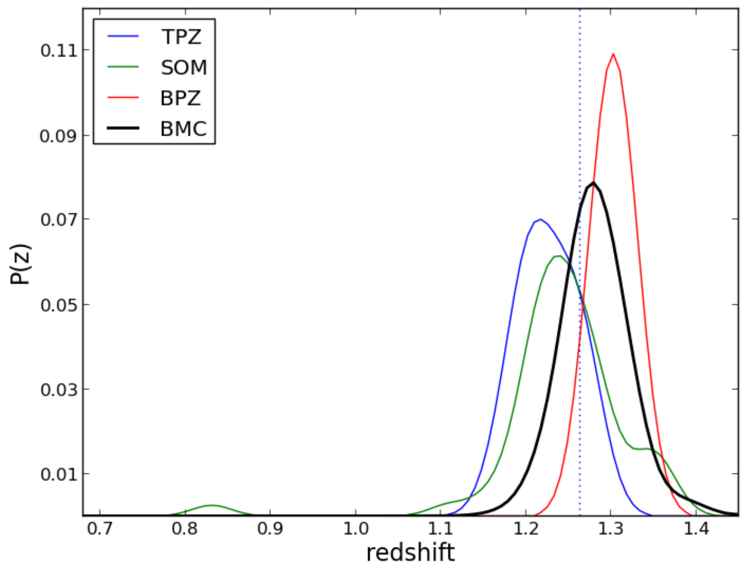


Figure 6.6: An comparison between the three individual photo-z PDF estimation techniques and a combined PDF computed by using BMC and Equation 6.8 for a single example galaxy taken from the DEEP2. The vertical line indicates the true source redshift.

dashed line shows the best value from the individual photo-z PDF estimation methods and the shaded area separates the individual from the combined methods. This Figure demonstrates that the Bayesian modeling techniques provide better performance than the best individual method over all five metrics, and also that by employing the binning scheme to optimize the combination approach we achieve better performance than for the best individual technique.

We compare the actual photo-z PDF for a single galaxy selected from the DEEP2 survey as estimated by the three individual techniques with the photo-z PDF estimated by the BMC method in Figure 6.6. This Figure clearly shows how the re-normalized combined PDF from the three individual photo-z PDF estimation techniques has been improved as the BMC result is closer to the true galaxy redshift, shown by the vertical line. These combination techniques identify which individual method works best in different cells, and can use that information to either weight the individual photo-z PDFs accordingly, or in the case of BMC to marginalize over the uncertainty in the correct weights to produce the best combination.

We apply a SOM to the DEEP2 field 1 data in order to construct a two-dimensional, binned combination of the three individual photo-z PDF estimation methods. We use this SOM to determine the weights for the three individual methods for each cell, and present the results in Figure 6.7 when using the BMA approach as it is easy to interpret. We also show the mean DEEP2 R-band magnitude for all galaxies in a given cell in the lower right panel, which clearly indicates the ability of the SOM to preserve relationships between galaxies when projecting from the higher dimensional space to the two-dimensional map. Of course, the

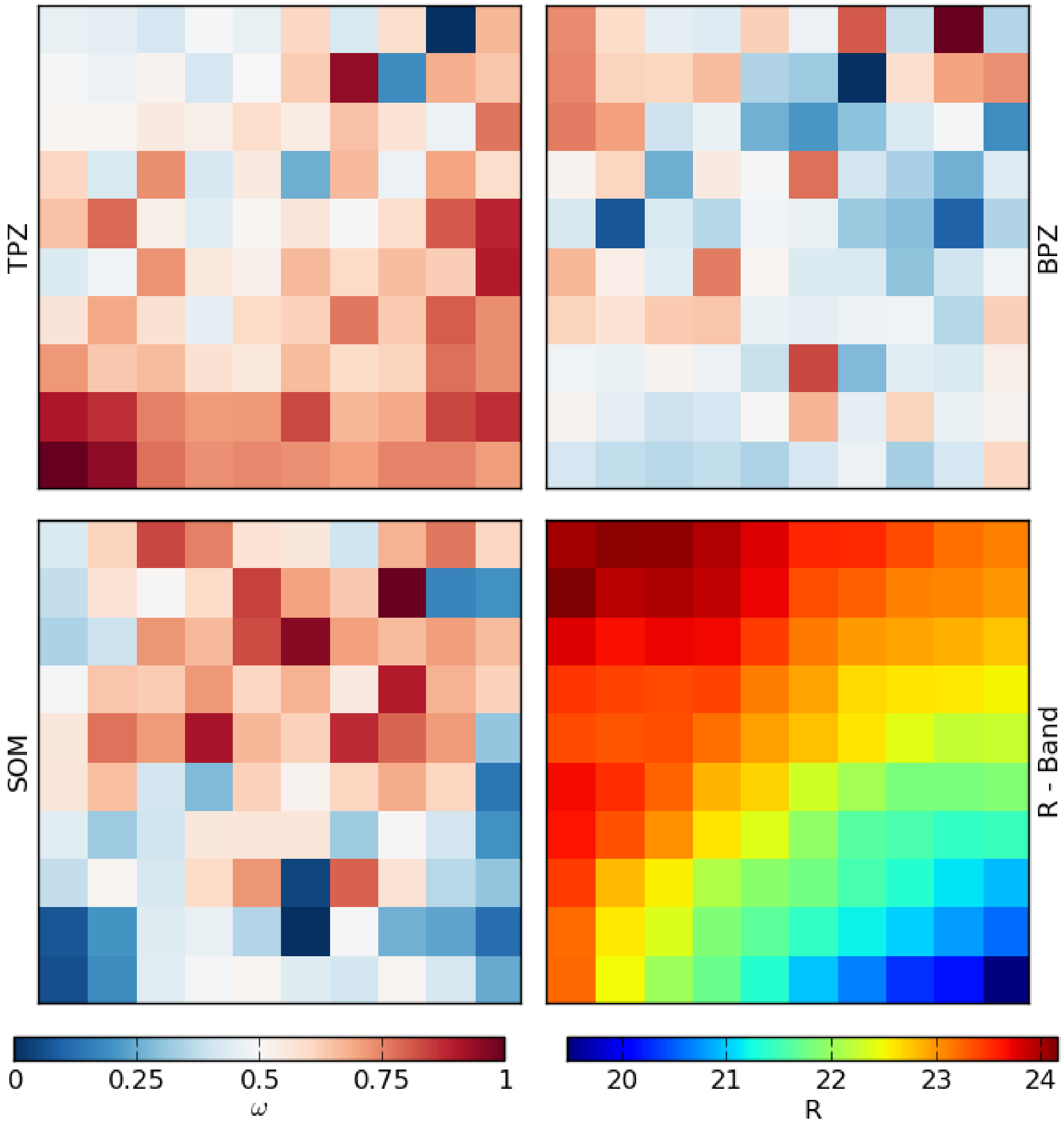


Figure 6.7: A two-dimensional SOM showing the relative weights for the BMA combination scheme applied to the three individual methods for the DEEP2 field 1 data (TPZ is top left, BPZ is top right, and SOMz is bottom left). In each panel, the color map indicates the value of the weight relative to the other cells in the map. The bottom right panel shows the same cells colored by the mean R-band magnitude for the cross validation galaxies.

SOM mapping is a non-linear representation of all magnitudes and colors, thus the DEEP2 R-band map should only be used to provide guidance.

In the three weight maps, a redder color indicates a higher weight, or equivalently that the corresponding method performs better in that region. These weight maps demonstrate the variation in the performance of the individual techniques across the two-dimensional parameter space defined by the SOM. For example, BPZ performs the best, as expected, in the upper left corner of the map, which is approximately where the faintest galaxies, at least in the DEEP2 R-band, are stored. On the other hand, TPZ performs better in the lower sections of the map, which approximates to brighter DEEP2 R-band magnitudes. Interestingly, SOMz performs relatively better in the upper middle of the map, which corresponds to the middle range $21 \lesssim R \lesssim 23$. The overall variation in weights across the map reflects the performance differences between the individual methods, which are exploited by the combination algorithms in order to identify the optimal combined performance.

We can also compare the global performance of the BMC method with the three individual photo-z PDF methods as a function of the spectroscopic redshift as shown in Figure 6.8. In this Figure, the photometric redshifts are the computed as the mean of each PDF, and the median is shown as black points along with the tenth and ninetieth percentiles as vertical error bars, enclosing 80% of the distribution on each redshift bin. The performance of the BMC method is generally more accurate, resulting in a tighter distribution that suffers fewer outliers when compared to the benchmark TPZ method. Interestingly, the SOMz performance is similar to TPZ, while BPZ is worse, with wider spread and several discontinuities. Nevertheless, the combined method still uses BPZ, as shown in the weight maps, as appropriate to generate an overall improved performance, especially for the faintest galaxies as discussed previously. We note, however, that the number counts in the last few bins are very low for the DEEP2 training and testing sets as shown in Figure 6.9. Therefore, although on average BPZ has better performance statistics over those bins (with large error bars), the photo-z results remain subject to Poissonian fluctuations (which is important when constructing a SOM to subdivide the galaxies when applying the combination models), thus the BMC results do not emphasize the BPZ results in the highest redshift bins.

Of all of the ten different metrics presented in Table 6.3, only the KS test does not show a marked improvement over the benchmark TPZ method. This metric does not depend on the redshift binning and it is computed by using the stacked PDF for each method. As a result, this metric is expected to be less sensitive to a combination approach, since stacking the PDF smooths out little discrepancies between the models. After integrating over a large number of galaxies PDFs, the individual methods will not differ significantly from one another and the final $N(z)$ distribution will resemble the one from the benchmark method.

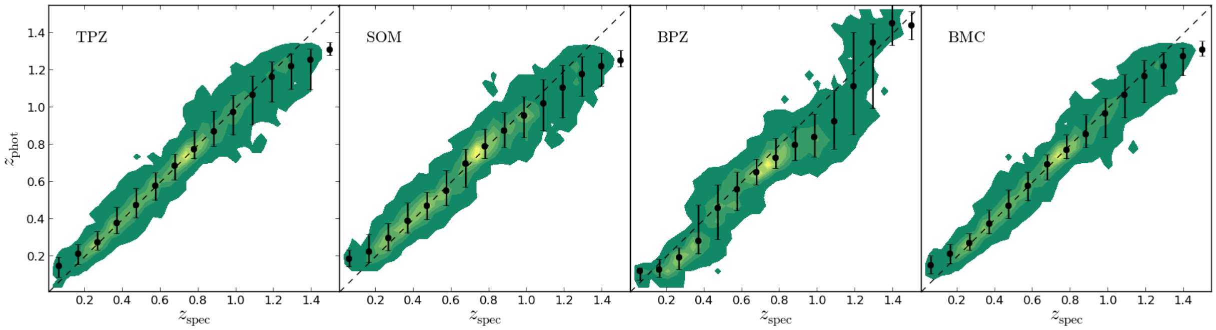


Figure 6.8: A comparison of the photometric and the spectroscopic redshifts for all DEEP2 field1 galaxies. From left to right, the comparison is for the TPZ, SOMz, BPZ, and the BMC techniques.The black dots are the median values of $z_{\rm phot}$ and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic redshift bin of width $\Delta z = 0.1$

Figure 6.9 shows the final $N(z)$ produced by stacking the PDFs from the BMC technique for galaxies from the DEEP2 (in solid black) and the corresponding DEEP2 spectroscopic $N(z)$ for the same galaxies (in gray). As also seen in Chapter 3 and Chapter 4 for TPZ and SOMz respectively, both distributions match exceedingly well.

6.2.3 Photo-z PDF Combination for the SDSS

We now change our focus to the analysis of the SDSS galaxy sample, which consists of 1,097,397 galaxies taken from the SDSS-DR10 data; we now retain 50,000 galaxies for training purposes. We apply the same three photo-z PDF estimation methods and seven different combination methods. We construct a SOM-defined, 10×10 two-dimensional map to subdivide the multi-dimensional magnitude and color space by using a rectangular topology to facilitate visualization. As before, we use cross-validation data to identify the best set of model parameters within each individual cell in our two-dimensional map. As shown in 6-combo/Figures 6.3 and 6.4, the photo-z PDFs computed by using the cross-validation and testing data sets are comparable and unbiased.

We present in Table 6.4 the same ten metrics for each method, and in bold we highlight the best method for each metric. Overall, the results obtained for this data set are remarkable, especially for the outlier fraction and the dispersion. We once again treat TPZ as the benchmark method; but note that, interestingly enough, in two cases, including the KS metric, TPZ does provide the best result. In addition, both BMA and BMC have very similar results, with the latter being slightly better.

After these two models, $\rm WA_{shape}$, which is OOB data independent, shows good performance, especially when looking at the $I_{\Delta z'}$ score. For any given individual metric, however, it does not perform better than other combination methods. For this data, BPZ provides good results; thus we expect that the set of template

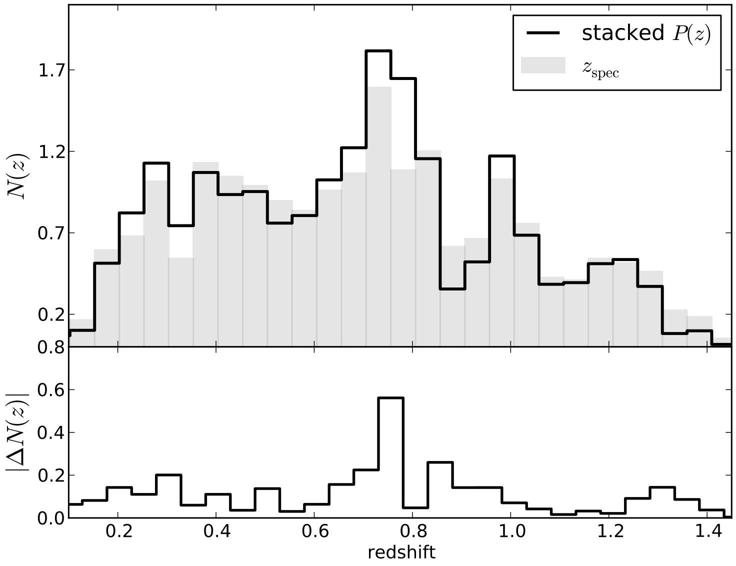


Figure 6.9: Top panel: The $N(z)$ for the DEEP2 sample computed directly from the spectroscopic redshifts (gray) and by stacking the photo-z PDF estimates from the BMC method (black). Bottom Panel: The absolute difference between these two $N(z)$ distributions.

described in §5 are a good representation of the galaxies in the SDSS photometric data. In particular, this seems true of the LRGs that dominate this sample for $z \gtrsim 0.3$.

We present the performance of the three individual and seven combination methods when applied to the SDSS data for five of the most common metrics in Figure 6.10. As was the case with the DEEP2 data, the Bayesian combination methods provide good performance. We also see the same variation in the KS metric, especially when comparing the combination methods to TPZ. However, TPZ is not always the best performer among the individual techniques, for example SOMz displays the best performance as measured by $\sigma_{\Delta z'}$ and $\mathrm{out}_{0.1}$.

As we discussed in Chapter 4, SOMz performs quite well when using a spherical topology; in the current application to the SDSS data, we have used a random atlas containing 300 maps that use spherical topology each with 3072 total cells. Interestingly, the $\mathrm{WA}_{\mathrm{oracle}}$ method, which selects the best method within each binned cell, often selects the SOMz result as we can infer from Figure 6.10. Although in general the *oracle* combination method is not the best possible combination, as shown by the overall performance of the BMA and BMC combination methods on this data.

We also display the SOM-defined, 10×10 two-dimensional map used to determine the weights for the three individual methods for each cell in Figure 6.11. In this map, we identify galaxies within the OOB and test data to determine the parameters for the combination models. One of the benefits of using an unsupervised learning method for this mapping is that we can use any property from the galaxies within this

Table 6.4: A summary of the performance results for the three individual methods and the seven different photo-z PDF combination methods as applied to the SDSS-DR10 data, with no magnitude cut applied to the training data set. The bold entries highlight the best value within each column to aid in the interpretation of the table (c.f. Figure 6.10).

Combination method	$< \Delta z' >$	$\|\Delta z'\|_{50}$	$\sigma_{\Delta z'}$	σ_{68}	$\sigma_{\rm MAD}$	KS	$\rm out_{0.1}$	$\rm out_{2\sigma}$	$\rm out_{3\sigma}$	$I_{\Delta z'}$
TPZ	0.0188	0.0137	0.0219	0.0139	**0.0082**	**0.0260**	0.0078	0.0297	0.0121	-0.2875
SOM	0.0201	0.0149	0.0209	0.0152	0.0094	0.0381	0.0070	0.0334	0.0125	0.7836
BPZ	0.0230	0.0164	0.0289	0.0167	0.0103	0.0367	0.0134	**0.0228**	0.0111	1.7143
$\rm WA_{flat}$	0.0195	0.0139	0.0235	0.0145	0.0088	0.0292	0.0082	0.0251	0.0104	-0.2507
$\rm WA_{oracle}$	0.0193	0.0141	0.0220	0.0145	0.0089	0.0373	0.0067	0.0266	**0.0100**	-0.1495
$\rm WA_{shape}$	0.0192	0.0136	0.0236	0.0143	0.0086	0.0297	0.0081	0.0243	0.0102	-0.4114
$\rm WA_{fit}$	0.0200	0.0141	0.0242	0.0149	0.0090	0.0274	0.0090	0.0255	0.0107	0.0244
BMA	0.0183	0.0133	0.0209	0.0139	0.0084	0.0261	0.0060	0.0296	0.0110	-0.6384
BMC	**0.0183**	**0.0133**	**0.0203**	**0.0138**	0.0084	0.0267	**0.0059**	0.0296	0.0109	**-0.6873**
HB	0.0198	0.0143	0.0237	0.0147	0.0090	0.0271	0.0084	0.0251	0.0106	-0.0975

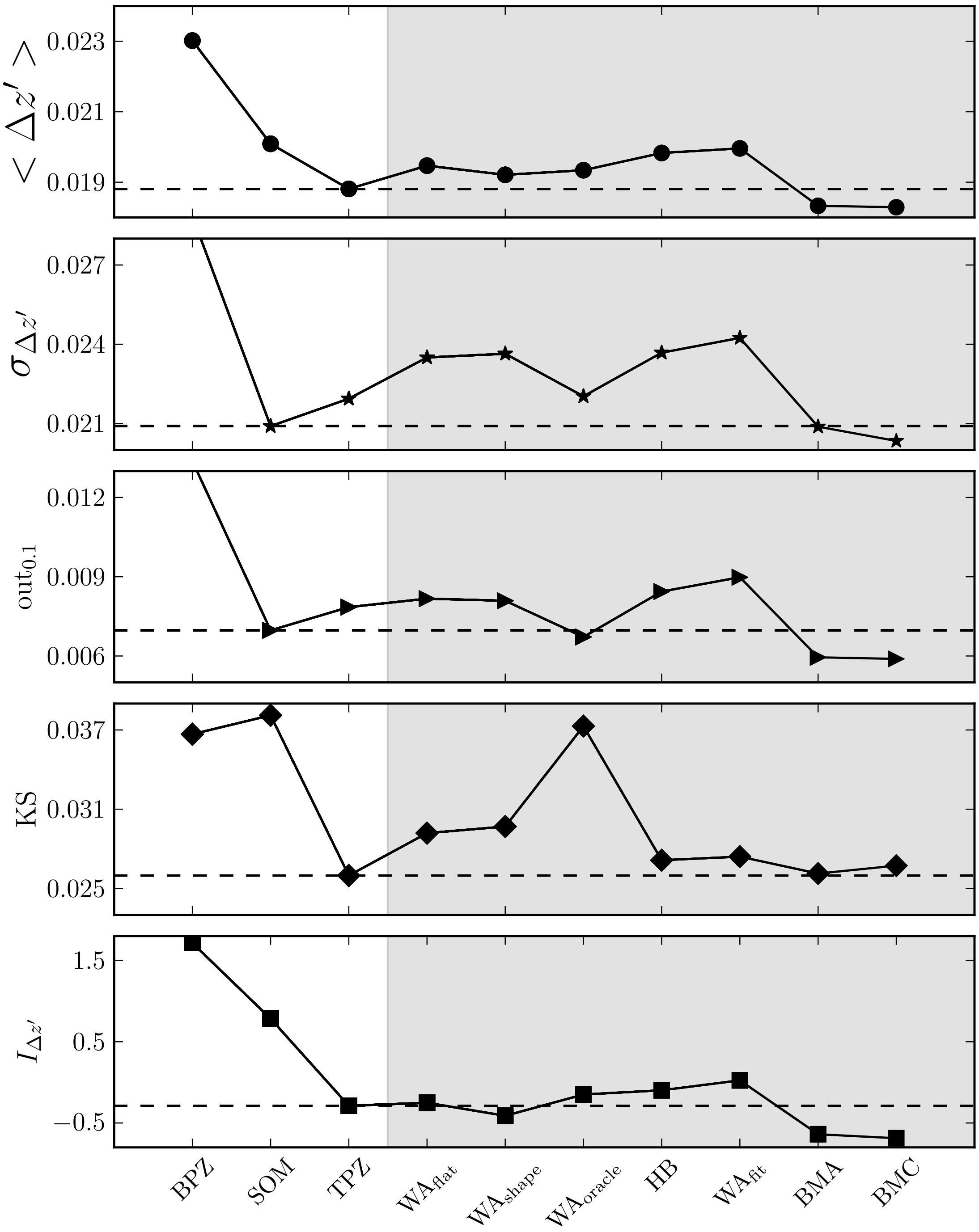


Figure 6.10: A comparison of the average performance for the three individual photo-z PDF estimation methods and the seven different photo-z PDF combination approaches for five different metrics as defined in Table 6.2 for the SDSS data. The horizontal dashed line indicates the best result for a given statistic among the three individual methods, and the shaded area separates the individual methods from the combined approaches. All values are presented in Table 6.4.

map to construct a representation, such as the mean SDSS r-band magnitude map shown in the bottom right panel of Figure 6.11. In this panel the brighter galaxies are generally on the right while the fainter galaxies are on the left, even though all five magnitudes and four colors were used to construct the SOM-defined, two-dimensional map.

The weighting for the three individual methods show interesting patterns, and TPZ and SOMz seem complimentary in that TPZ is weighted most strongly at fainter r-band magnitudes (the left side of the map) while SOMz is weighted most strongly at brighter r-band magnitudes (the right side of the map). This result is most likely an artifact from the bi-modality of the training data, which is dominated at low redshift by the SDSS main galaxy sample and at high redshifts by the SDSS-III LRG sample. At brighter magnitudes and lower redshifts, the SOMz approach where a high-dimensional space is projected to two-dimensions does a better job of maintaining complex relationships within the data. At fainter magnitudes and higher redshifts, however, the data are dominated by the homogeneous LRG sample. The TPZ approach performs better for this sample, since the high-dimensional space is recursively sub-divided by TPZ to maximize the information gain, which may only require one or two dimensions.

Another interesting observation from these weight maps is that BPZ performs well over much of the parameter space, with a particular strong weighting in a narrow vertical band on the extreme left of the map and again in the center of the map. Given the nature of the input galaxy sample, it seems reasonable to expect that these areas of the map are dominated by Elliptical galaxies. Another interesting observation is that there are six cells in the second column from the left that all have the same value in each weight map (pink for TPZ, white for BPZ, and light blue for SOMz). These cells are primarily empty, i.e., they contain weights and training data but they lack test galaxies and thus have a constant value, which illustrates how strongly the galaxies (i.e., MGS or LRG) are concentrated in this SOM-defined, two-dimensional topology.

The number of galaxies, either for training or testing, within each cell can vary significantly, which is simply due to the fact that we used a fixed number of cells (in this case 100) to represent the higher dimensional space when fewer cells would have been sufficient. However, the empty cells do not affect the performance of the photo-z combination methods, they are simply not used during the analysis. It is the fact that these individual methods perform differently across these cells that makes the combination approach a powerful technique to maximally extract information from the available data.

We next provide a comparison between the photo-z PDFs computed by the three individual techniques and the BMC technique and the SDSS spectroscopic redshift for all 1,097,397 galaxies in Figure 6.12. The first observation from the figure is the bi-modality of the sample, which is the result of the two primary sub-populations (i.e., MGS and LRGs). Overall, the results are quite good with a very tight correlation,

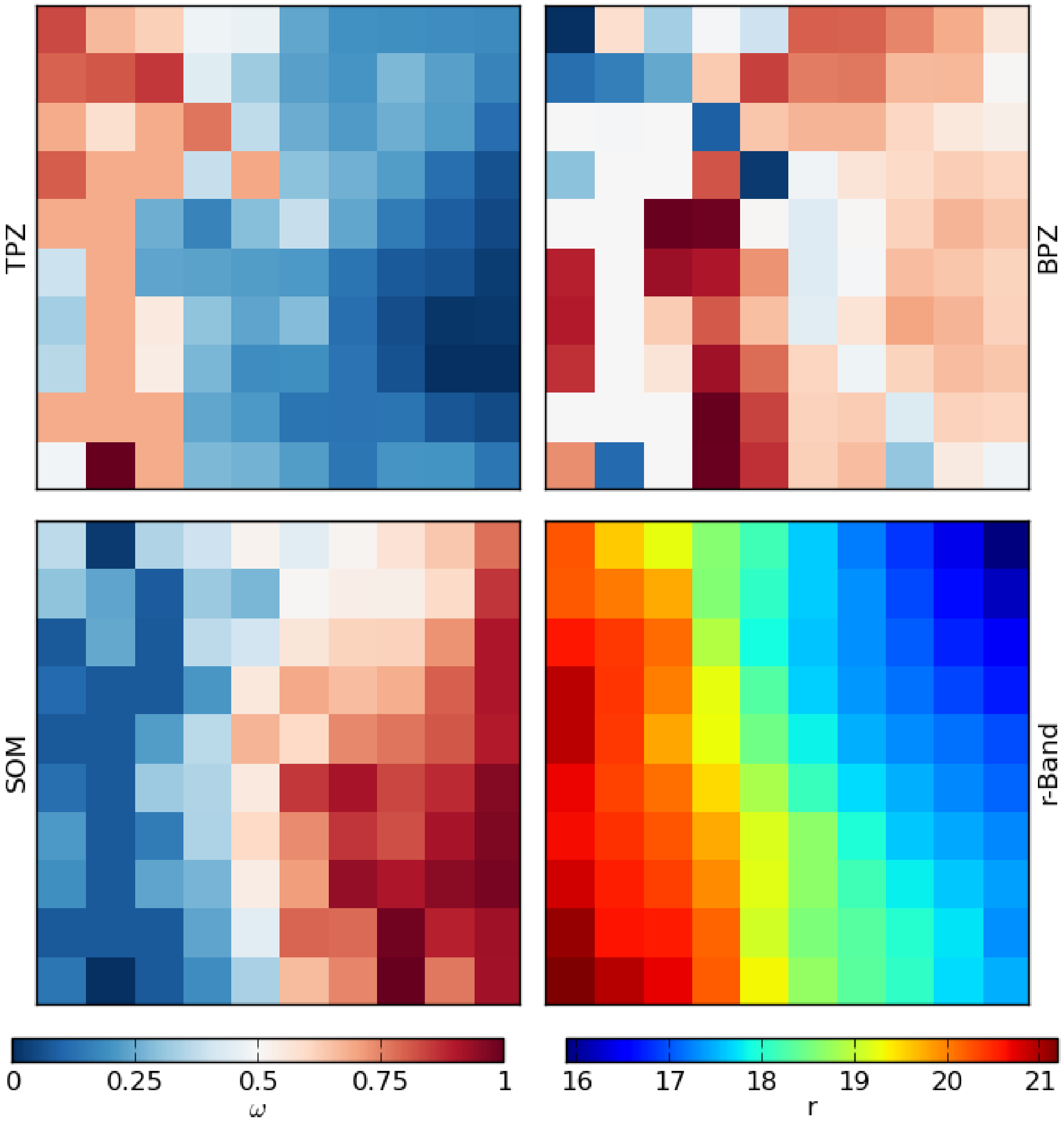


Figure 6.11: A two-dimensional SOM showing the relative weights for the BMA combination scheme applied to the three individual methods for the SDSS data (TPZ is top left, BPZ is top right, and SOMz is bottom left). In each panel, the color map indicates the value of the weight relative to the other cells in the map. The bottom right panel shows the same cells colored by the mean SDSS r-band magnitude for the cross validation galaxies.

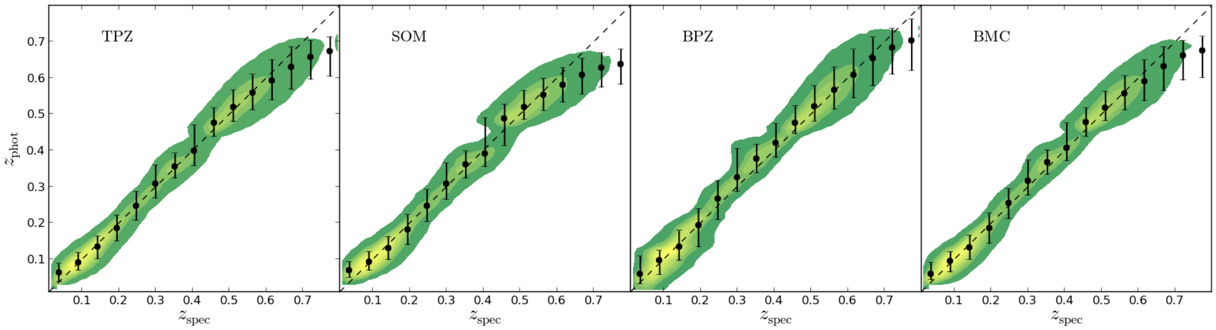


Figure 6.12: A comparison of the photometric and the spectroscopic redshifts for all SDSS galaxies. From left to right, the comparison is for the `TPZ`, `SOM`z, `BPZ`, and the BMC techniques.The black dots are the median values of $z_{\rm phot}$ and the errors bars correspond to the tenth and ninetieth percentiles within a given spectroscopic redshift bin of width $\Delta z = 0.05$

especially in areas of high source density areas. The main exception is at the highest redshifts where there is a slight underestimation; and, as seen before, we can observe how these different approaches provide similar results, which are therefore correlated, while still differing in other areas where one method may outperform the others. The most right panel is the $\rm BMC$ which shows a slightly tighter distribution in comparison to the others.

Finally, in Figure 6.13 we present the galaxy redshift distribution for both the spectroscopic sample (in gray) and the photometric redshift distribution, computed by stacking the individual galaxy PDFs (in black). This Figure highlights that the underestimation of the photo-z at high redshifts in Figure 6.12 coincides with the strong decline in the number of galaxies after $z = 0.75$. More importantly, however, this $N(z)$ figure shows the excellent agreement between the photometric and spectroscopic galaxy redshift distributions. Given the fact that the SDSS galaxy sample contains two distinct populations, this agreement is remarkable.

6.3 Summary

We have presented and analyzed different techniques for combining photo-z PDF estimations on galaxy samples from the DP-1 and SL-2 sets. In particular, we use three independent photo-z PDF estimation methods: `TPZ`, a supervised machine learning technique based on prediction trees and a random forest; `SOM`z, an unsupervised machine learning approach based on self organizing maps and a random atlas; and `BPZ`, a standard template-fitting method that we have slightly modified to parallelize the implementation.

We developed seven different combination methods that employ ensemble learning with cross-validation data to maximize the information extracted. Of these seven methods, four employ a weighted average where the weights can either be selected to be uniform across the input methods, to be determined from the shape

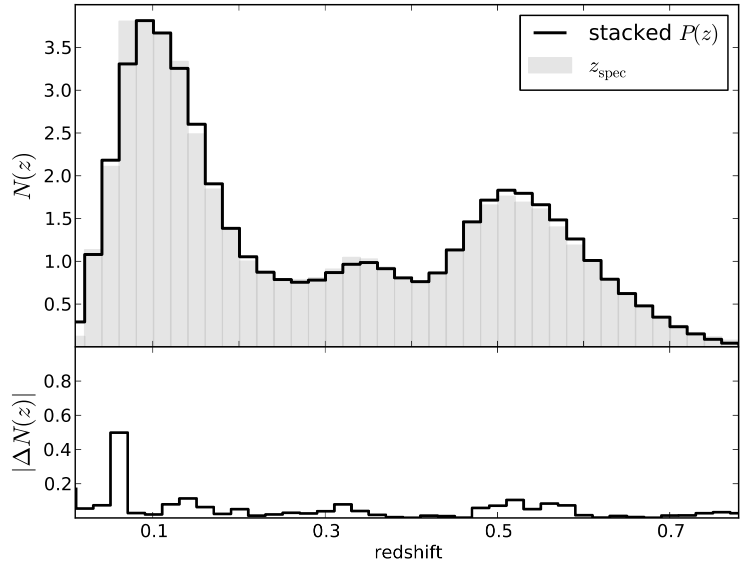


Figure 6.13: Top panel: The $N(z)$ computed directly from the spectroscopic redshifts (gray) and by stacking the photo-z PDF estimates from the BMC method (black). Bottom Panel: The absolute difference between these two $N(z)$ distributions.

of the photo-z PDF (e.g., by using the $zConf$ parameter), to be determined by an *oracle* estimator where one (ideally the best) method is preferentially selected, and where the weights are obtained by a fitting procedure applied to the OOB data. Three of the combination methods were Bayesian techniques: Bayesian Model Averaging (BMA), Bayesian Model Combination (BMC), and Hierarchical Bayes (HB).

We expect the individual photo-z PDF estimation techniques to perform differently across the parameter space spanned by our galaxy samples; for example, template-fitting techniques are expected to work better at higher redshifts than machine learning methods, which perform optimally when provided high-quality, representative training data. Thus we construct a two-dimensional, 10×10 self-organizing map (SOM) to subdivide the high-dimensional parameter space occupied by the galaxy samples. We apply different photo-z PDF estimation techniques within each cell in this map, since each cell should contain galaxies with similar properties. A visual inspection of these maps indicates that the two machine learning methods: `TPZ` and `SOM`z are generally complementary, and that in combination with a model based technique such as `BPZ` we are able to maximize the coverage of this multidimensional space efficiently.

We also verified that by using the OOB data, as introduced in Chapter 3, we can an obtain an accurate, unbiased and *honest* estimation of the performance of a photo-z PDF estimation technique on the test data. We also computed the correlation coefficient and the error distribution and showed they also behave similarly for the cross-validation (i.e., the OOB data) and the test data. These computations are extremely important when combining photo-z PDF techniques as we can learn from the OOB data the optimal parameters needed for a specific ensemble learning approach, and thereby maximize the performance of that combination technique when applied to *blind* test data.

Overall, we found that the BMA and BMC are the best photo-z PDF combination techniques as they have better performance metrics when compared to the individual photo-z PDF estimation techniques, especially when unbiased cross-validation data is available. This result is true for both datasets . When OOB data is not available, we can instead use the $zConf$ parameter as a weight for each method after first renormalizing the individual photo-z PDFs. We can also use the Hierarchical Bayes method to combine these predictions, which we demonstrated can also lead to better results.

The computational cost to apply these Bayesian models to galaxy samples will depend directly on the size of the data set, the number of photo-z estimation techniques used, and the resolution of the given photo-z PDFs. In Chapter 8 we demonstrate how a sparse basis representation can reduce the storage significantly and that manipulation of these PDFs can be improved within the bases framework thereby reducing computational costs.

Finally, we have demonstrated that even when a photo-z PDF technique is very accurate, we can still make improvements by extracting additional information about the distribution of galaxies in the higher dimensional parameter space and the individual performance of the photo-z PDF algorithms. There are currently a large number of published algorithms to compute photo-z 's, many of which also compute photo-z PDFs. Even if their performance is similar, these techniques will all have their own advantages and disadvantages. Thus we believe the combination of different techniques is the future of photo-z research, and we expect additional research to be forthcoming in this area.

In the next chapter we show how we can extend our analysis on combinig techniques to make improvements not only on the photo-z computation but also on the identification of outliers in the data set.

Chapter 7

Photo-z Outliers

Outline

In this chapter we discuss better techniques to identify outliers for the photo-z computation and how combining multiple techniques can improve not only the accuracy of the photo-z solution but also the identification of outliers in the data. Removing outliers from the galaxy sample is an extremely important task when putting constrains on the cosmological parameters, therefore a proper identification of those is fundamental for today's cosmology.

7.1 Outliers identification

As we have discussed previously, aggregating information from multiple photo-z PDFs estimation techniques can improve the overall photo-z solution. In this section, however, we explore how this information can be combined to improve the identification of outliers within the test data. In particular, we attempt to use all possible information in order to identify these objects, from the shape of each photo-z PDF as computed by all individual methods to the differences in their predicted photo-z. We adopt a Naïve Bayes Classifier (NBC) (Zhang, 2004) to identify these two groups, a technique that has found widespread adoption to identify spam email messages. The advantage of this approach is that it is easy to implement, is fast and efficient for large dimensional data, and can be very competitive with other classifiers (Domingos & Pazzani, 1997; Frank et al., 2000). We have previously introduced this methodology in Chapter 5 to compute priors in a Bayesian inference for photo-z , in this case we take a slightly different approach as we will use it in this classification problem.

Let $\boldsymbol{\theta}$ be the set of N_θ parameters, θ_i, we will use to identify the outliers. By using the Bayes Theorem, we can compute the probability for an object to be an outlier, given $\boldsymbol{\theta}$ as:

$$P(\text{out} \mid \boldsymbol{\theta}) = \frac{P(\text{out})P(\boldsymbol{\theta} \mid \text{out})}{P(\boldsymbol{\theta})} \tag{7.1}$$

where the *evidence*, $P(\boldsymbol{\theta})$ is given by

$$P(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \text{out}) + P(\boldsymbol{\theta} \mid \text{in}) \tag{7.2}$$

and *out* refers to outliers and *in* refers to inliers, the only two classes we identify in this analysis. The Naïve Bayes Classifier assumes that all θ_i variables are independent, even if their independence is weak or even if there is a strong dependence between any of them. Each variable provides information about these two classes, and this information can be combined to make a stronger classifier (Zhang, 2004). For instance, in Chapter 3 we showed that outliers tend to have a broader (larger values of $zConf$) and multi-peaked PDFs, and herein we treat these values as independent data even though multi-peaked PDFs are indeed generally broader.

By using this assumption, similarly to what we did in Chapter 5, we have:

$$P(\boldsymbol{\theta} \mid \text{out}) = P(\theta_1, \theta_2, \ldots, \theta_{N_\theta} \mid \text{out}) = \prod_{i=1}^{N_\theta} P(\theta_i \mid \text{out}) \tag{7.3}$$

and similarly,

$$P(\boldsymbol{\theta} \mid \text{in}) = \prod_{i=1}^{N_\theta} P(\theta_i \mid \text{in}) \tag{7.4}$$

We can now rewrite Equation 7.1:

$$P(\text{out} \mid \boldsymbol{\theta}) = \frac{P(\text{out}) \prod P(\theta_i \mid \text{out})}{\prod P(\theta_i \mid \text{out}) + \prod P(\theta_i \mid \text{in})}, \tag{7.5}$$

which is similar to the method used by Gorecki et al. (2014), who demonstrated the potential of this approach to identify photo-z outliers. Here, however, we use a different set of variables that are generated for all three individual photo-z PDF methods.

In our case we use N_{peak}, the number of peaks in each photo-z PDF; r_{peak}, the logarithm of the ratio between the height of the first peak and the height of the second peak; z_{mean}, the mean of each photo-z PDF; z_{mode}, the mode of each PDF;$zConf$, measured with respect to the mean and the mode of the photo-z PDF; and the difference in the photo-z , as enumerated by the mean and the mode between each of the three methods. Thus, we have six metrics computed individually for each of our three photo-z PDF estimation techniques, and an additional six metrics for the difference in photo-z mean and mode between the three techniques. As a result, we have a total of twenty-four metrics, to which we can add the input data for each survey.

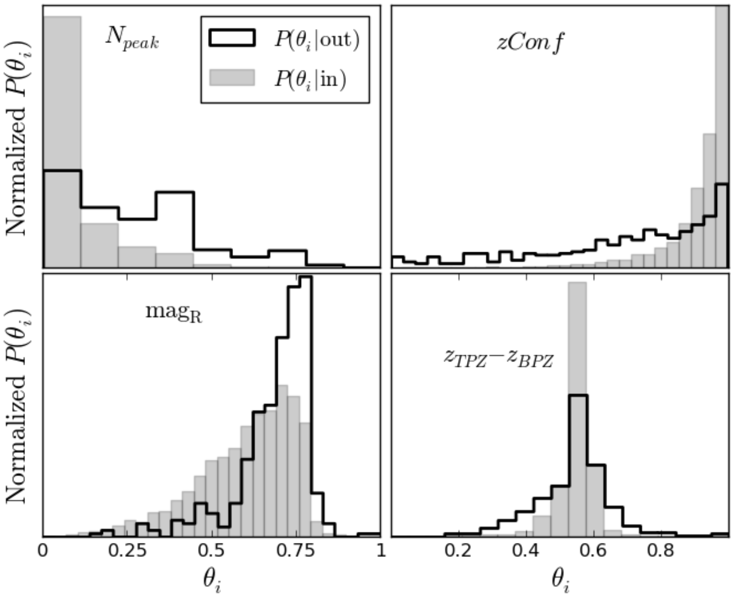


Figure 7.1: The normalized distributions of four of the set of thirty-eight (rescaled) $\boldsymbol{\theta}$ variables from the DEEP2 data that are used for outlier detection. The variables are binned as outliers (black line histograms) or inliers (gray histogram). From the top left and following in a clockwise direction: N_{peak}, the number of peaks in the TPZ PDF; $zConf$, as computed from TPZ, the R-band magnitude, and the difference between the photo-z computed by using the mean of the TPZ and BPZ PDFs.

We, therefore, have a total of thirty-eight variables for the DEEP2 survey, while for the SDSS we have a total of thirty-three variables to use for outlier detection. For convenience, we rescale each of these variables to lie between zero and one. $P(\theta_i \mid \mathrm{in})$ and $P(\theta_i \mid \mathrm{out})$ are evaluated by using the OOB or cross-validation data, which we have shown can reliably predict the results on the test data. Once computed, these distributions are evaluated for the test data, where $P(\mathrm{out} \mid \boldsymbol{\theta})$ is evaluated separately for each galaxy in the test data.

Figure 7.1 presents the normalized distributions of four rescaled variables (i.e., θ_i) taken from the DEEP2 test data. Note that the inlier and outlier distributions are normalized to have unit area, thus these distributions illustrate how these two populations differ and not how the relative numbers between the inlier and outlier populations vary. The four variables shown in this Figure include the number of peaks in the TPZ PDFs, $zConf$ computed by TPZ, the R-band magnitude, and the difference between the mean of the TPZ and BPZ photo-z PDFs. In just these four distributions, there is clear separation between the galaxies labeled as outliers (black line) and inliers (gray shaded area), where the outlier identification metrics are defined by using Table 4.1. In particular, for this Figure we use $\mathrm{out}_{0.1}$, i.e., galaxies for which $\Delta z' > 0.1$. While not shown, a similar result is seen for the other distributions. The result that outliers and inliers follow distinct distributions is what makes this a powerful approach. In effect, all information is assumed to be independent,

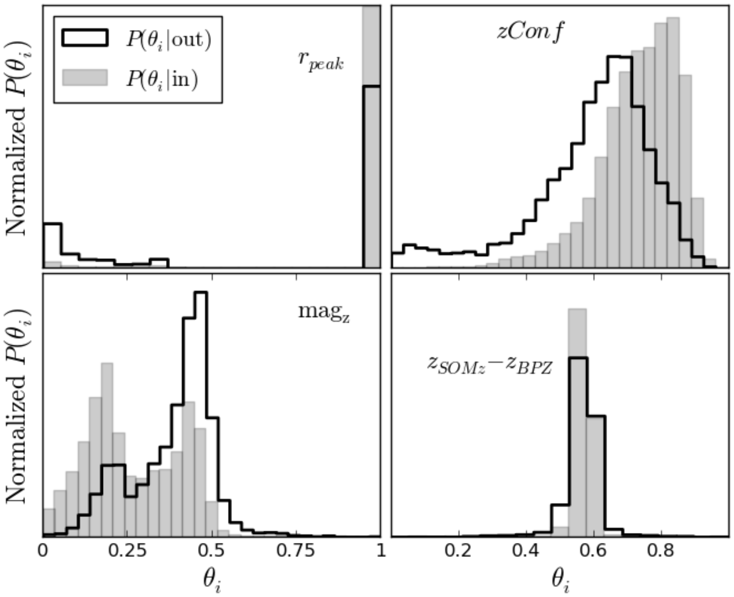


Figure 7.2: The normalized distributions of four of the set of thirty-three (rescaled) $\boldsymbol{\theta}$ variables from the SDSS data that are used for outlier detection. The variables are binned as outliers (black line histograms) or inliers (gray histogram). From the top left and following in a clockwise direction: r_{peak}, the logarithmic ratio of the first two peaks in the TPZ PDF; $zConf$, as computed from SOMz, the SDSS z-band magnitude, and the difference between the photo-z computed by using the mode of the SOMz and BPZ PDFs.

and when combined allows an efficient identification of catastrophic outliers.

We see a similar trend in Figure 7.2, but now for galaxies in the SDSS test data. In this Figure, we have selected four different rescaled variables; namely, the logarithmic ratio between the first and the second peaks of the TPZ PDF (note that if the PDF has one peak, we fix this value to be four), the $zConf$ computed from SOMz, the SDSS z-band magnitude, and the difference between the mode of the SOMz and BPZ photo-z PDFs. Once again, this Figure highlights that in each of these distributions there is a separation between the outliers and inliers, and that in combination we obtain an even better discriminant between these two classes.

By using Equation 7.5, we can combine the values of all of the rescaled variables (i.e., θ_i) to compute $P(\text{out} \mid \boldsymbol{\theta})$ for each galaxy in the DEEP2 and SDSS, both for the OOB and the test data. We present these $P(\text{out} \mid \boldsymbol{\theta})$ distributions for the DEEP2 in Figure 7.3 and for the SDSS in Figure 7.4. Both 7-outliers/Figures are similar, showing a clear separation between the outliers and inliers in both data sets. The probability ranges between zero and one, and the outliers are generally concentrated near one, while the inliers are concentrated near zero. While some mis-classifications remain, the contamination has been greatly reduced, meaning we can successfully identify a majority of the outlier population. Lastly, while there are a few galaxies with probabilities lying somewhere between zero and one, these distributions are highly bimodal,

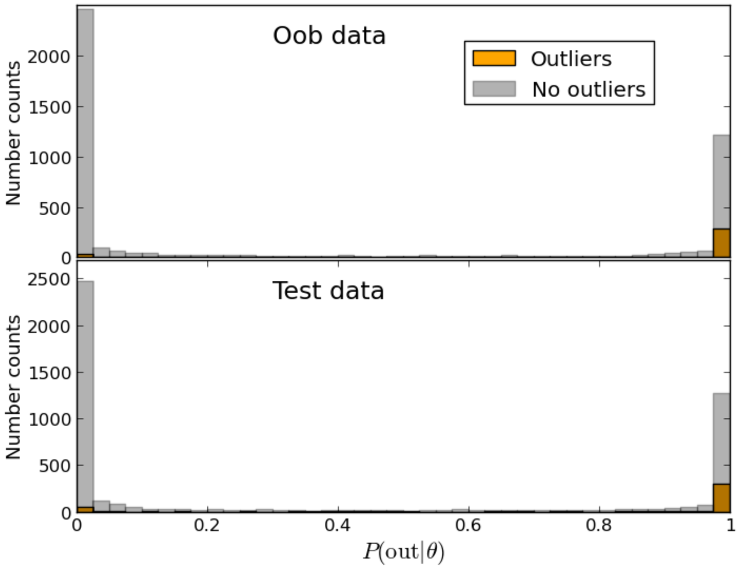


Figure 7.3: The count distribution of $P(\text{out} \mid \boldsymbol{\theta})$ for the DEEP2 OOB data (top) and test data (bottom) showing both the outliers (orange) and inliers (gray).

which reinforces the belief that this method provides a remarkably good discriminant between these two populations.

Once again, in both 7-outliers/Figures 7.3 and 7.4, the OOB and test data distributions show strong similarities. As a result, we can expect that any cut we make on the OOB data will produce similar results in the test data, allowing us to make a robust classification of outliers in potentially blind test data.

7.2 Comparison with $zConf$

In Chapter 3 we introduce the parameters $zConf$ (Carrasco Kind & Brunner, 2013a) which quantifies the shape of the PDF by computing the area under the photo-z PDF around the mean (or any other single estimate) given some window. We demonstrated that this approach also provides a better way to characterize outliers and that it can be used to clean the galaxy sample after the photo-z computations. We showed in Figure 3.16 that there is a correlation between $zConf$ and the accuracy of the results. To compare this outlier approach with that parametrization, we show in Table 7.1 the effects of selecting outliers by using this NBC approach and by using the $zConf$ approach for the DEEP2 data. To simplify the comparison, we first select inlier galaxies by using the $P(\text{out} \mid \boldsymbol{\theta})$ to cut the test data sample, and subsequently choosing those galaxies in the test data that have the highest $zConf$ so that we have the same number of galaxies selected via both techniques.

The information in this Table demonstrates that the NBC approach produces a sample of galaxies that

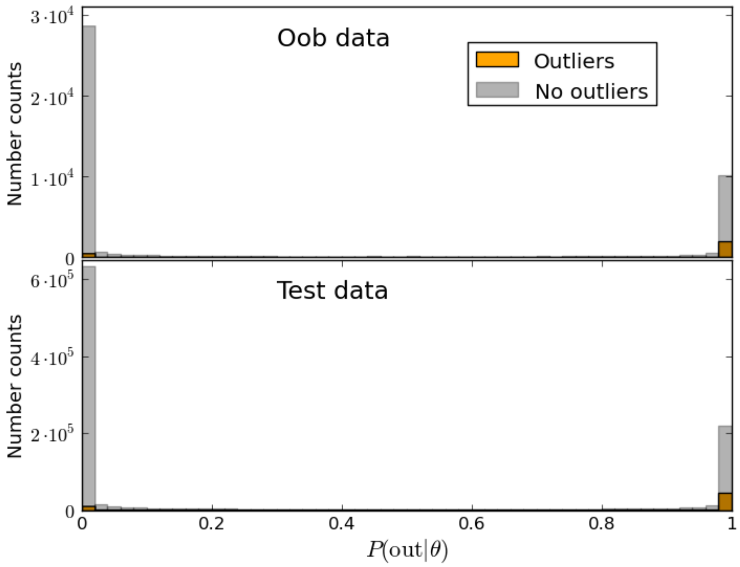


Figure 7.4: The count distribution of $P(\mathrm{out} \mid \boldsymbol{\theta})$ for the SDSS OOB data (top) and test data (bottom) showing both the outliers (orange) and inliers (gray).

Table 7.1: The effect of removing outliers from the DEEP2 test data on several, select performance metrics by using the Naïve Bayes Classifier and the $zConf$ cut approach. The two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$<\Delta z'>$	$\sigma_{\Delta z'}$	$\mathrm{out}_{0.1}$
NBC	$<$ 0.998	83.0 %	0.02819	0.03948	0.0362
$zConf$	$>$ 0.854	83.0 %	0.02868	0.04186	0.0371
NBC	$<$ 0.894	72.0 %	0.02616	0.03548	0.0304
$zConf$	$>$ 0.893	72.0 %	0.02721	0.03895	0.0330
NBC	$<$ 0.174	56.0 %	0.02565	0.03470	0.0251
$zConf$	$>$ 0.918	56.0 %	0.02595	0.03575	0.0289

Table 7.2: The effect of removing outliers from the SDSS test data on several, select performance metrics by using the Naïve Bayes Classifier and the $zConf$ cut approach. The two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$< \Delta z' >$	$\sigma_{\Delta z'}$	$\mathrm{out}_{0.1}$
NBC	$<$ 0.999	83.0 %	0.01560	0.01533	0.0022
$zConf$	$>$ 0.7018	83.0 %	0.01589	0.01704	0.0035
NBC	$<$ 0.802	72.0 %	0.01473	0.01411	0.0012
$zConf$	$>$ 0.755	72.0 %	0.01475	0.01549	0.0026
NBC	$<$ 0.001	56.0 %	0.01387	0.01309	0.0006
$zConf$	$>$ 0.807	56.0 %	0.01366	0.01410	0.0020

have a smaller spread in $\Delta z'$ along with a smaller number of outliers than the $zConf$ method, which was previously shown to be beneficial in this regard (Chapter 3). We interpret this result as suggesting that a $zConf$ cut can potentially remove *good* galaxies whose photo-z PDF happens top be broad, while retaining some *bad* galaxies that have a well-localized photo-z PDF. By using a Naïve Bayes approach, we collect all information from photo-z PDFs predicted by using different, semi-independent methods, allowing a more robust discriminant between outliers and inliers. Finally, we notice that as always there is a trade-off between completeness, whereby we try to retain as many *good* galaxies, and contamination, whereby we try to minimize the inclusion of *bad* galaxies. The final choice in this conflict should be determined by the scientific application, but by producing a probabilistic value, subsequent researchers can make these cuts more easily.

We performed a similar analysis on the SDSS galaxy sample and present the results in Table 7.2. As was the case with the DEEP2 galaxies, we see that the NBC approach once again does better in identifying outliers within the sample, as the NBC cuts have a smaller scatter and the fraction of remaining outliers is remarkably small. We also notice that the mean bias is similar between the two approaches, but the number of outliers, defined as $\Delta z' > 0.1$, is significantly reduced when we adopt the Bayesian approach. This is yet another piece of evidence supporting the benefits of aggregating information to make decisions.

We can also test how the definition of an outlier affects this approach. Previously we identified an outlier as a galaxy that had $\Delta z' > 0.1$; but for the purpose of this test, we apply a much more restrictive cut of $\Delta z' > 0.05$. We apply the NBC cut and produce a matched sample by imposing a $zConf$ cut to both the DEEP2 and the SDSS galaxies, presenting the information in Table 7.3. We find, once again, that even for this more restrictive approach we produce a cleaner catalog (of the same size) as compared to using only the $zConf$ parameter. Interestingly, even after removing almost 30% of the galaxies from the DEEP2 galaxy sample, we still have over a 10% outlier contamination. On the other hand, this tight cut applied to the SDSS galaxies produces a very small contamination of $\sim$ 2%, for both methods, albeit the NBC approach is still slightly better.

Table 7.3: The effect of removing outliers, defined as $\Delta z' > 0.05$, from the DEEP2 and SDSS test data on several, select performance metrics by using the Naïve Bayes Classifier and the $zConf$ cut approach. For each data set, the two techniques are applied to ensure equal numbers of galaxies are selected, which is indicated by the *Fraction* column.

Method	Criteria	Fraction	$< \Delta z' >$	$\sigma_{\Delta z'}$	$\mathrm{out}_{0.05}$
DEEP2					
NBC	$<$ 0.996	72.0 %	0.02780	0.03934	0.138
$zConf$	$>$ 0.878	72.0 %	0.02809	0.04244	0.141
SDSS					
NBC	$<$ 0.85	72.0 %	0.01461	0.01407	0.0247
$zConf$	$>$ 0.75	72.0 %	0.01479	0.01554	0.0278

While producing galaxy samples that are less affected by outliers than competing techniques, the NBC approach has an additional advantage in that it can easily be extended to other variables and to other photo-z algorithms. In effect, any information that might increase the efficacy of outlier identification can be included in order to improve this discriminant while still maximizing the overall galaxy sample size.

7.3 Summary

Within our Bayesian Framework, we developed a novel, Naïve Bayesian Classifier (NBC) that efficiently identifies outliers within the galaxy sample. The approach we present gathers all available information from the different photo-z PDF estimation techniques regarding the shape of the PDF, the location of the mean and mode, and the magnitudes and colors, which are all *naively* assumed to be independent, in order to compute a Bayesian posterior probability that a certain galaxy is an outlier. The distribution of these probabilities for an entire galaxy sample indicate that this is a very powerful method to separate outliers from inliers (i.e., *good* galaxies), and we further demonstrated that this approach can produce a more accurate and cleaner sample of galaxies than competing techniques, such as the use of the $zConf$ parameter. An important takeaway point is that all information provided by the catalogs and the photo-z PDF methods, no matter how redundant the information might appear, helps in building this discriminant probability. Given the probabilistic nature of this computation, the final application of this technique can be chosen to maximize the scientific utility of the resulting galaxy data for a specific application.

Chapter 8

Photo-z PDF representation and storage

Outline

In this chapter we introduce the use of a sparse basis representation to fully represent individual photo-z PDFs. By using an Orthogonal Matching Pursuit algorithm and a combination of Gaussian and Voigt basis functions, we demonstrate how our approach is superior to a multi-Gaussian fitting, as we require approximately half of the parameters for the same fitting accuracy with the additional advantage that an entire PDF can be stored by using a 4-byte integer per basis function. By using data from the CFHTLenS described in 2.4, we demonstrate that only 10 to 20 points per galaxy are sufficient to reconstruct both the individual PDFs and the ensemble redshift distribution, $N(z)$, to an accuracy of 99.9% when compared to the one built using the original PDFs computed with a resolution of $\delta z = 0.01$, reducing the required storage of 200 original values by a factor of 90%.

We compute the photo-z PDF for all one million galaxies in our test sample CF-2, described previously, by using our spectroscopic training sample CF-1 using TPZ . We used all colors and magnitudes, which results in a total of nine attributes, and construct 600 trees to make the predictions. TPZ also uses the attribute errors during the prediction process, in part to deal with missing attributes in the catalog(see Chapter 3 for a detailed description). We also have computed photometric redshifts for galaxies in the training sample by using a cross validation technique called Out-Of-Bag (Breiman, 2001; Carrasco Kind & Brunner, 2013a) in which a photo-z PDF is obtained for all galaxies in the training set by using all the trees that do not contain that particular galaxy. This approach, therefore, avoids over-fitting; and we have shown that this method is reliable and also unbiased (Carrasco Kind & Brunner, 2014a).

For illustration, we present a sample of forty photo-z PDFs randomly selected from the CFHTLenS galaxies in Figure 8.1, presented in increasing order by the computed mean value of their photo-z PDF. The redshift range for the galaxies are the same and the PDFs have all been normalized to unity. From this figure, it is clear that these photo-z PDFs are not simple functions, often having multiple peaks; and they are, therefore, poorly represented by a single Gaussian, which has often been used for simplicity in the past. In Figure 8.2, we present a summary of the results on the training data determined by using the OOB cross-validation technique. The top panel compares $z_{\rm phot}$, computed by using the mean value of each photo-z PDF, with $z_{\rm spec}$

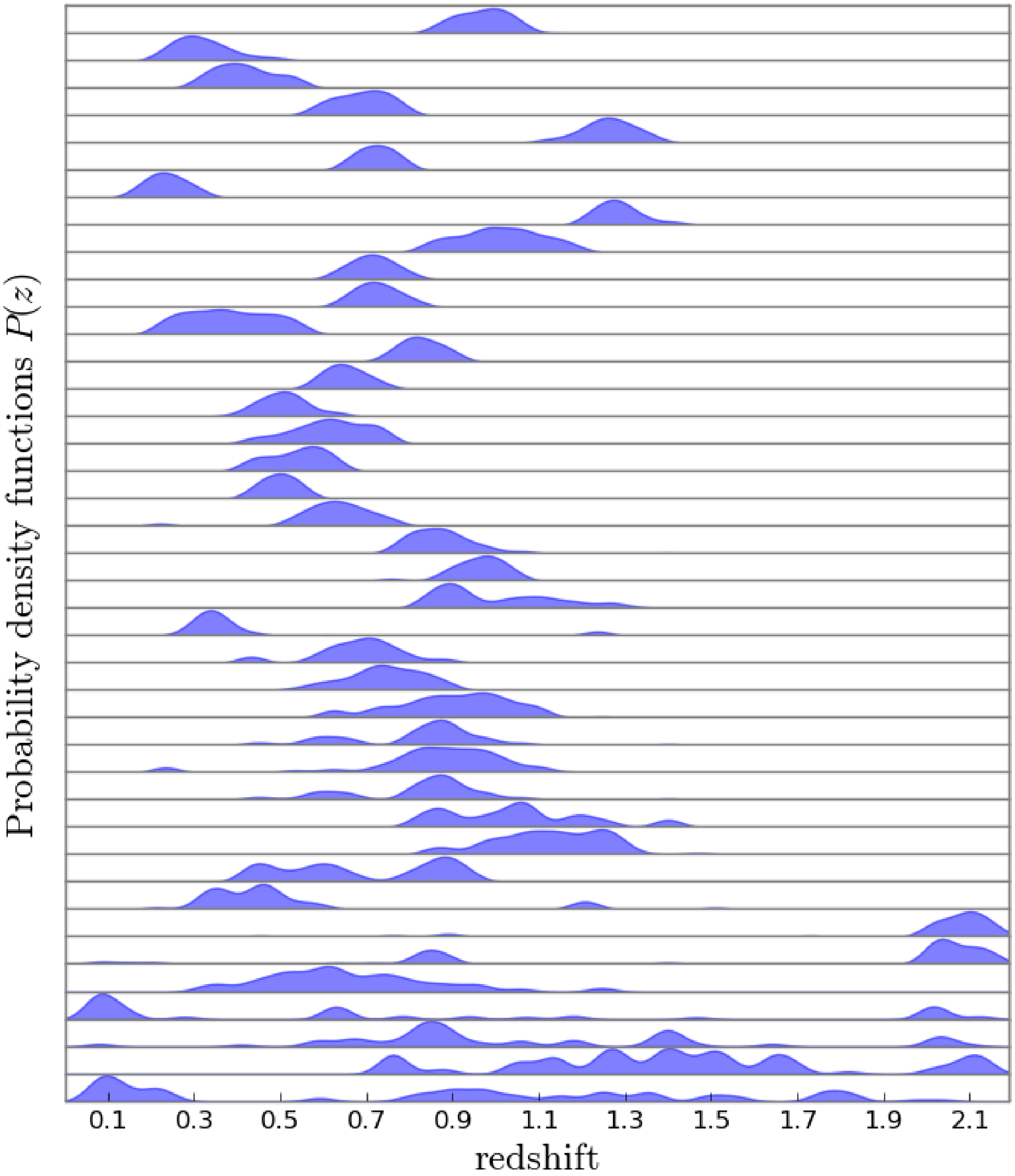


Figure 8.1: Forty representative, randomly selected photo-z PDFs computed for the CFHTLenS data by using TPZ, each normalized to unity. In each subplot, the horizontal axis is redshift and the vertical axis is the probability density. The PDFs are sorted in the vertical axis by the number of peaks.

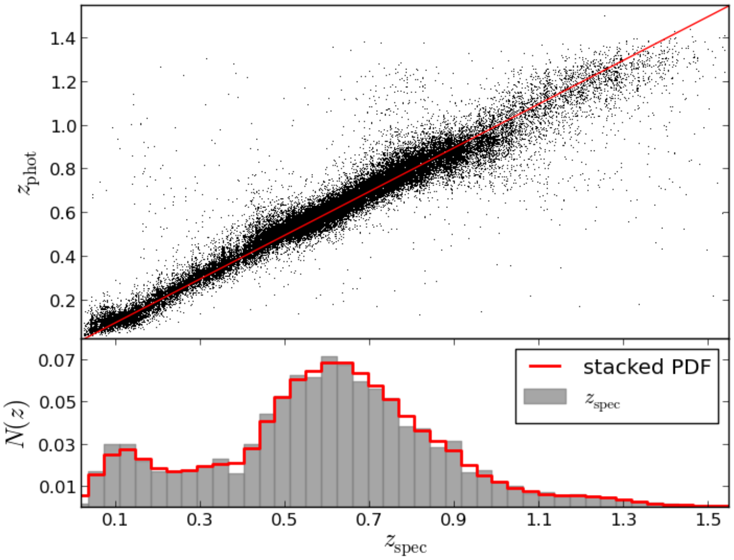


Figure 8.2: *Top:* A comparison of the photometric and spectroscopic redshift for all galaxies in the training sample computed by using the OOB cross-validation technique. The red line shows the one-to-one line for guidance. *Bottom:* The $N(z)$ distribution computed by using the spectroscopic redshifts (gray area) and by stacking the photo-z PDFs (red) for the training sample galaxies.

for all 49,868 galaxies in the training sample. This indicates the approximate performance of TPZ on the real dataset within the limits of the training sample. The bottom panel shows the distribution of the galaxies as a function of redshift in terms of their spectroscopic values (in gray) compared with the $N(z)$ obtained by stacking the PDFs (red line). We can see a remarkable agreement between these two distributions, as we have shown previously (Carrasco Kind & Brunner, 2013a).

8.1 PDF Representation

In this section we present the different methods that we use to represent the full photo-z PDF. For the rest of this discussion, we will make the following assumptions. First, we have N total galaxies in our sample with individual photo-z PDF estimates. Second, we can represent the photo-z PDF, $P_k(z)$, for the $k^{\rm th}$ galaxy in the sample by $\mathbf{pz}_k$. Finally, we have n sample points in the original galaxy photo-z PDF. Thus $P_k(z)$ is sampled at a resolution δz, given by $\delta z = \Delta z/n$, where Δz is the redshift range spanned by the photometric data.

8.1.1 Statistical Representation

The simplest representation for a full photo-z PDF is to use a summary statistic. We will consider five different, summary statistics, four of which are single values: the mean, the mode, the median of the PDF, and a Monte-

Carlo sampling from the original cumulative PDF (Wittman, 2009). This fourth approach involves sampling a random number from the range 0–1, and determining the cumulative probability to this numerical value. Finally, the last representation is a single Gaussian fit to the original photo-z PDF, which provides a two value summary: the mean and variance of the Gaussian. As a result, we require either N, for the first four approaches, or $2N$, for the last approach, statistics to represent the full photo-z PDF catalog.

8.1.2 Multi-Gaussian Fitting

The second representation for a full photo-z PDF we explore is the application of a multi-Gaussian fit to each photo-z PDF (see, e.g., Bovy et al., 2011, 2012). In this approach, each Gaussian (when more than one is used) included during the fitting process will require three parameters: the amplitude, the mean, and the variance. In this approach, we first determine the number of peaks, Np_k, in the photo-z PDF. We increase this value by one and use the result as the number of Gaussians to be used in the fitting process for that specific photo-z PDF. The extra Gaussian improves the fit to extended wings in the photo-z PDF distribution, which often arise from the residuals of the Gaussian fits to the individual PDF peaks.

To determine the best fit values, we use a Levenberg-Marquardt minimization algorithm. In this case, each $P_k(z)$ can be represented by:

$$\mathbf{pz}_k = \sum_{i=1}^{Np_k+1} \alpha_{k,i} e^{-\frac{(\mathbf{z}-\mu_{k,i})^2}{2\sigma_{k,i}^2}} \tag{8.1}$$

where $\boldsymbol{\alpha}_k$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$ are vectors of dimension $Np_k + 1$ containing the amplitude, mean, and variance for each Gaussian included in the fitting process. As a result, the total number of values needed to represent the full photo-z PDF catalog is $\sum_k 3(Np_k + 1)$.

8.1.3 Sparse Basis Representation

The final technique that we will use to represent a photo-z PDF is the sparse basis representation. In this case, we will adopt a set of basis functions to represent a PDF by using the following model:

$$\mathbf{pz}_k = \mathbf{D}\boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k \tag{8.2}$$

where $\mathbf{D}$ is a dictionary or basis matrix of dimension $n \times m$, where $m > n$. Thus, we have an over-determined problem as the number of basis functions, m, is much larger than the dimension, n, of each photo-z PDF. In this case, each column, $\mathbf{d}_j$, of the dictionary matrix, $\mathbf{D}$, is a basis function that must be ℓ_2 normalized, i.e., $||\mathbf{d}_j||_2 = 1$ for $j = 1, 2, \ldots, m$.

We want to find, for each galaxy k, the optimal vector solution $\boldsymbol{\delta}_k$, which is determined such that its pseudo-norm $||\boldsymbol{\delta}_k||_0$ is minimized. Alternatively, this can be equivalently stated that we want to minimize the number of non-zero entries in the vector, $\boldsymbol{\delta}_k$, given the residual error $\boldsymbol{\epsilon}_k$. In this case, we can either use a predefined number of basis functions or we can define a fixed residual for every galaxy in the sample. Either way, the total number of points required to represent the entire catalog is given by $\sum_k 2(Nb_k)$, where Nb_k is the number of basis functions used for each galaxy. Note that in this case we only need two numbers for each functional basis: the functional coefficient (a floating point number), and the index number of the function within the basis set (a integer number). This already corresponds to a potentially large reduction in the total data volume required to archive photo-z PDFs. We will see in subsequent sections that we can also represent the basis function coefficients by using integers; and that, in addition, we can combine both terms into a single thirty-two bit integer, thereby reducing the total number of values required to $\sum_k(Nb_k)$.

Finding $\boldsymbol{\delta}_k$ in this over determined problem can be challenging. For this analysis, we have selected to use Orthogonal Matching Pursuit (OMP), an iterative algorithm that finds, at each step, the column, $\mathbf{d}_j$, of the dictionary matrix, $\mathbf{D}$, that best represents the current residuals. This process is repeated until a predefined criteria is reached, either a residual threshold or the total number of basis functions used. Fundamentally, this approach is similar to the well known `CLEAN` algorithm, which is used to analyze interferometric radio observations (Högbom, 1974). The advantage of OMP over the standard Matching Pursuit algorithm (Mallat & Zhang, 1993) is that a specific basis function can only be selected once. Since the residuals are orthogonalized during the selection of the basis functions for the current galaxy, we generate an independent set of basis functions to represent each galaxy's photo-z PDF.

Conceptually, the OMP algorithm that we apply to all galaxies can be enumerated[1]:

1. Initialize all variables. First, define the residual vector to be the original photo-z PDF, $\boldsymbol{\epsilon}_k^0 = \mathbf{pz}_k$. Second, create an empty set of cumulative selected basis functions, $\mathbf{B}_k$. Finally, set $\boldsymbol{\delta}_k = 0$, and define $i = 0$ as the number of the current iteration.

2. Compute the current set of basis functions. First, find the column vector, $\mathbf{d}_b$, from the dictionary matrix, $\mathbf{D}$, where b is the index position that maximizes the projection of $\boldsymbol{\epsilon}_k^i$:

$$\mathbf{d}_b^i = \max_{\mathbf{d}_j \in \mathbf{D}} |\mathbf{d}_j^T \cdot \boldsymbol{\epsilon}_k^i| \tag{8.3}$$

   Second, add this selected basis function to the set $\mathbf{B}_k$, i.e., $\mathbf{B}_k = (\mathbf{B}_k, \mathbf{d}_b^i)$.

3. Orthogonally project the original photo-z PDF onto the linear space spanned by the columns of all

[1] Note, the superscript T indicates transposition

previously selected basis functions:

$$\mathbf{w}_k^i = \mathbf{B}_k^T \cdot \mathbf{pz}_k \tag{8.4}$$

where $\mathbf{w}_k^i$ is a temporary vector corresponding to the coefficients of the currently used basis functions in $\mathbf{B}_k$.

4. Complete the projection by updating the residuals by using the temporary vector $\mathbf{w}_k^i$:

$$\boldsymbol{\epsilon}_k^{i+1} = \mathbf{pz}_k - \mathbf{B}_k \cdot \mathbf{w}_k^i \tag{8.5}$$

5. Check the stopping criteria: $||\boldsymbol{\epsilon}_k^{i+1}||_2 < \epsilon_{th}$, where ϵ_{th} is the threshold residual or $i > i_{lim}$, where i_{lim} is the number of required basis functions. If the pre-selected stopping criteria is met, the calculations are completed: $\boldsymbol{\delta}_k = \mathbf{w}_k^i$ and $\mathbf{pz}_k = \mathbf{D} \cdot \boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k^{i+1}$, where $\boldsymbol{\delta}_k$ is sparse. Finally, the photo-z PDF representation is defined:

$$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k \tag{8.6}$$

On the other hand, if the predefined stopping criteria is not met, the iteration step is increased, $i = i + 1$, and steps 2–5 are repeated by using the current residual vector. This process is repeated over all galaxies k, where $k = 1, 2, \ldots, N$.

Dictionary Selection

Given the nature of the shape of the photo-z PDFs (see, e.g., Figure 8.1), it is natural to select a set of Gaussian-like basis functions that span the redshift range of our photometric galaxy sample. We can use the the original resolution and redshift range spanned by the generated photo-z PDF to determine the dictionary to use for the sparse basis representation. One of the primary advantages of this method is that these dictionary entries are composed of analytic functions that can be combined with other functional forms. There are no restrictions, other than computational time, on how large of a dictionary we can use, as there is no requirement for the dictionary to be permanently stored. Furthermore, a photo-z PDF can be restored even without reconstructing the dictionary, as long as the indices and coefficients are efficiently stored.

We select N_μ Gaussian functions, whose mean values span the redshift range of our galaxy sample, which has a redshift resolution δz. Thus, we can compute:

$$N_\mu = \left\lceil \frac{\Delta z}{\delta z} \right\rceil \tag{8.7}$$

where $\Delta z = z_2 - z_1$ and z_2 and z_1 are, respectively, the upper and lower limits of the redshift range spanned by our galaxy sample. We select, at each N_μ location, N_σ values for the standard deviation that linearly span the range from a minimum value of $\sigma_{\rm min}$ to a maximum value $\sigma_{\rm max}$. The minimum value is selected in such a way that we will approximately have a single Gaussian that fills a single redshift bin of width δz. In practice, a Gaussian vanishes at approximately 3σ from the mean; therefore, we can select $\sigma_1 = \delta z/6$.

On the other hand, we select the broadest basis function to approximately cover half of the full redshift range Δz at each position; therefore, we select $\sigma_{\rm max} = \Delta z/12$. Although the extreme basis functions are not frequently used, they ensure that we cover all possibilities. Finally, we set the resolution between different values of σ to be $\delta z/2$ in order to make sure the difference between two consecutive Gaussian basis functions is on the order of δz. Setting $\Delta\sigma = \sigma_{\rm max} - \sigma_{\rm min}$ we have that N_σ is given by:

$$N_\sigma = \left\lceil \frac{2\Delta\sigma}{\delta z} \right\rceil \tag{8.8}$$

which can be simplified to

$$N_\sigma = \left\lceil \frac{\Delta z}{6\delta z} - \frac{1}{3} \right\rceil \approx \frac{N_\mu}{6} \tag{8.9}$$

As some photo-z PDFs have extended wings, we also generate N_γ basis functions for each Gaussian basis function with extended profiles by using a Voigt profile. Voigt profiles are widely used in spectral line fitting, and are defined as the convolution between a Gaussian distribution and a Lorentzian distribution. A Voigt profile can be written as the real part of the Faddeeva function (Abramowitz & Stegun, 1972):

$$V(x;\sigma,\gamma) = \frac{1}{\sigma\sqrt{2\pi}} \,\mathrm{Re}\left[e^{-z^2}\left(1 - \mathrm{erf}(-iz)\right)\right] \tag{8.10}$$

where $\mathrm{erf}(-iz)$ is the complex *error function*. $z = \frac{(x-\mu)+i\gamma}{\sigma\sqrt{2}}$ is a complex variable, where μ is the center of the function, σ is the standard deviation from the Gaussian, and γ determines the strength of the extended wings and is a parameter from the Lorentz distribution. As a result, if $\gamma = 0$, we have a Gaussian distribution with parameters μ and σ.

We present examples of different Voigt profiles in Figure 8.3 given a fixed $\mu = 0.3$ and $\sigma = 0.01$, but with γ varying from zero (Gaussian) to one σ. We do not, however, select pure Lorentzian profiles, as they produce distributions that are too extended to be practical for this analysis. In practice, we find that an upper limit of $\gamma = 0.5\sigma$ is sufficient to accurately model any extended wings. Thus, including the Gaussian case with $\gamma = 0$, we fix $N_\gamma = 6$ and allow γ to vary linearly from 0 to 0.5σ in steps of 0.1σ. Thus, in the most simple case we would only consider basis functions with $\gamma = 0$ and $N_\gamma = 1$. On the other hand, Figure 8.4

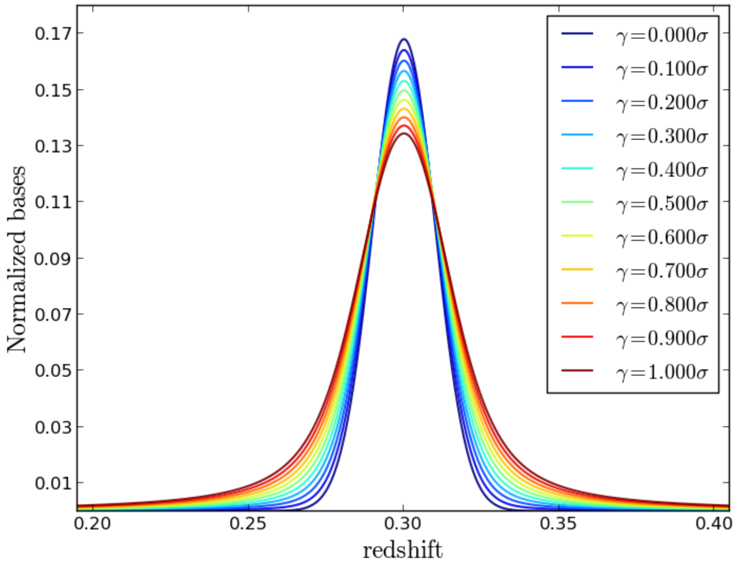


Figure 8.3: Different normalized $||\mathbf{d}_j||_2 = 1$ Voigt profile basis functions with the same mean, $\mu = 0.3$, and sigma, $\sigma = 0.01$, for different values of γ, which ranges from 0 (blue) to 1σ (red). Note that for $\gamma = 0$, we recover the standard Gaussian distribution. In a full dictionary, we create these profiles over the entire redshift range of the galaxy sample for different values of σ.

shows an example of a dictionary of basis functions described in the text which are shown with a vertical shift for illustration. We observe that the basis span all the redshift range. On each filled region we will have several different spreads for each bases which are not shown here but the solid colored area shows the span of basis at that given location. In practice the resolution in redshift and sigma is higher.

In total, the dictionary is composed of $N_{\text{total}} = N_\mu \times N_\sigma \times N_\gamma$ bases, which all have ℓ_2 norm equal to unity. By using our previous definitions, we have the following approximate rule of thumb for creating a dictionary:

$$N_{\text{total}} \approx N_\mu^2 = \left(\frac{\Delta z}{\delta z}\right)^2 \tag{8.11}$$

Although this is an estimate, it provides a very good approximation to the total number of bases needed given the resolution of the original photo-z PDF. Additional bases are not necessary and little is gained by using a finer resolution. Photo-z codes generally provide photo-z PDFs by using roughly two to three hundred points. According to Equation 8.11, we notice that for 250 sample points in a PDF, we would need approximately 62,500 bases. Thus, we can use a 2-byte integer to express the indices into our basis function dictionary, which has important ramifications in the compact storage of photo-z PDFs as discussed in the next section.

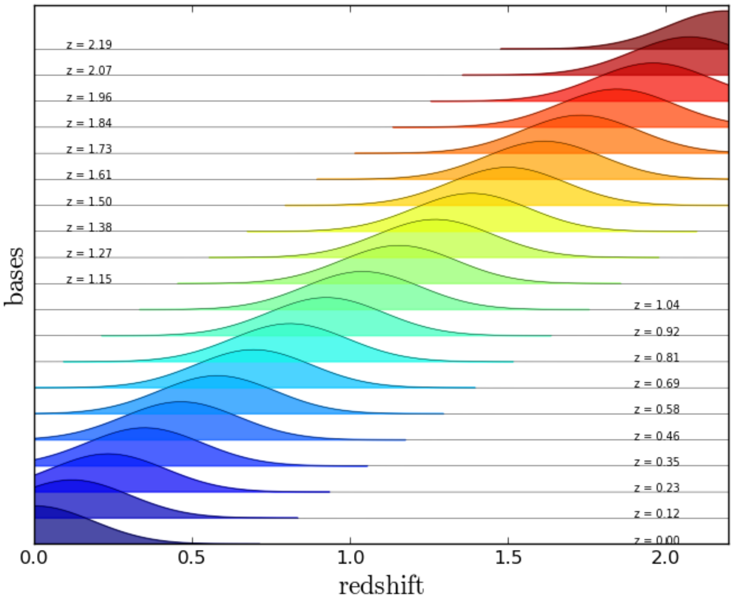


Figure 8.4: A illustration of the bases used in the dictionary, each shaded region is composed by several functions of different widths, those colored shaded regions are full of bases. In a real dictionary the spacing between these functions is much tighter (200 points in the redshift range)

8.2 Discussion

We have applied the previously discussed photo-z PDF representation techniques to the CFHTLenS data introduced in Section 2. We have computed a photo-z PDF for each galaxy in the one million test sample by using the TPZ software to compute a PDF with two hundred sampled points at a resolution of $\delta z = 0.011$.

We display one such photo-z PDF in Figure 8.5 where the original distribution is shown in green, a multi-Gaussian representation is shown in blue, a sparse basis representation is shown in red, and a single Gaussian model is shown with a blue dashed line. We can see that both the sparse basis representation and the multi-Gaussian agree remarkably well with the original photo-z PDF, to the point where it is hard to see the original PDF. As one would expect in this multi-peak PDF, the single Gaussian model does not reproduce this photo-z PDF very well. The inset panel provides a zoomed-in view showing the sparse basis representation of the photo-z PDF and the actual basis functions used in the representation. As the number of bases is increased, we expect some of them to have a negative coefficient, as shown in the inset, which aids in the reconstruction of the residuals from the previous bases. Given the iterative nature of this process, we select the new basis function that optimally corrects the residuals of previous bases in order to best reconstruct the photo-z PDF by using the minimum number of functions.

In order to quantitatively compare the reconstruction of the photo-z PDF by using the three methods as shown in Figure 8.5, we compute the multi-Gaussian fitting for all 10^6 galaxies from our CFHTLenS test

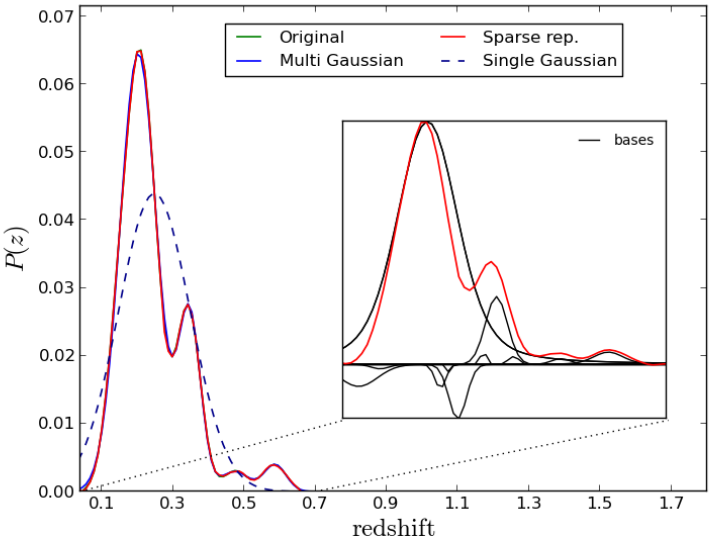
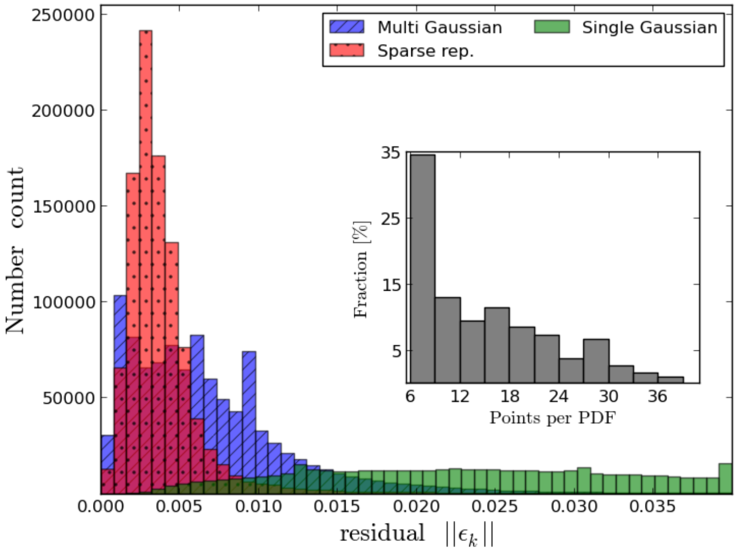


Figure 8.5: The representation of an original photo-z PDF (green) given by three techniques: multi-Gaussian (blue), single Gaussian (blue dashed line), and sparse basis representation (red). The inset panel shows the final bases (in black) used to represent the photo-z PDF while the recovered distribution is shown in red.



centering

Figure 8.6: The residual distribution for all CFHTLenS galaxies computed by using the multi-Gaussian representation (blue) and the sparse basis representation (red). In each case, we use the same number of representation values for each galaxy photo-z PDF. For comparison, the Single-Gaussian representation is shown in green. *Inset*: The distribution of points (bases or fitting parameters) per galaxy photo-z PDF. The number of peaks, Np_k, per photo-z PDF is the same, but divided by 6. Thus, there are $3(Np_k+1)$ parameters per galaxy.

sample. For each galaxy we record the number of values (parameters) required to accurately reconstruct the original PDF. Note that in this fitting approach, we are not fixing the number of Gaussian functions used in the reconstruction, but are instead defining the number of Gaussians as the number of peaks in the photo-z PDF plus one extra Gaussian to compensate for the residuals and extended profiles. In addition, we also compute, for each galaxy, the optimal sparse representation by using a variable number of basis functions that are constrained to match the number of points used in the multi-Gaussian fitting. We also compute the best single Gaussian fit to each PDF to demonstrate the importance in using the information contained within the full PDF as opposed to simply treating each photo-z estimate as a Gaussian PDF.

After computing the different representations for each galaxy, we next compute the norm of the residuals between each representation and the original photo-z PDF for each galaxy and accumulate the results. We compare the resulting distributions in Figure 8.6. First, we notice the broad shape of the single Gaussian distribution (green). In fact, the width of the distribution exceeds the plot boundaries as the median of the single Gaussian distribution of residuals is 0.043, which is outside the range of the Figure. Second, we observe that when using the same number of values to represent the photo-z PDF, the sparse basis representation produces much smaller residuals with a more concentrated distribution than the multi-Gaussian fitting. Specifically, the median of the sparse representation residual distribution is 0.0033, which is almost half of the value (0.0058) for the multi-Gaussian fitting. Both of these results indicate that either method provides a good representation of the photo-z PDF by using a small number of values. We also show the distribution of values required to reconstruct the photo-z PDF of each galaxy in the inset panel of Figure 8.6. This subplot indicates that approximately 35% of the galaxies are single peaked (six values are required for two Gaussians, which in our implementation means a single peak plus an extra Gaussian for the extended wings). The distribution extends up to thirty-nine values for roughly 1-2% of the sample, which corresponds to twelve peaks in a photo-z PDF. The average number of values per galaxy is fourteen, which, in itself, implies a large compression ratio when compared to the original two hundred values while still providing a very good reconstruction of the full photo-z PDF.

While a natural number of basis functions can be determined for the multi-Gaussian representation, the sparse basis representation is more general and thus does not have a simple, natural number of basis functions. In order to better understand the optimal number of basis functions for photo-z PDFs, we compute the sparse basis representation for all galaxies in the test sample by using a different number of fixed bases. We combine the residuals, and plot the median value of the distribution as a function of the number of values used to represent the PDF as blue dots in Figure 8.7. As shown in the figure, as we increase the number of bases, the residuals decrease monotonically. This decrease is quite rapid at first, as expected, and slowly

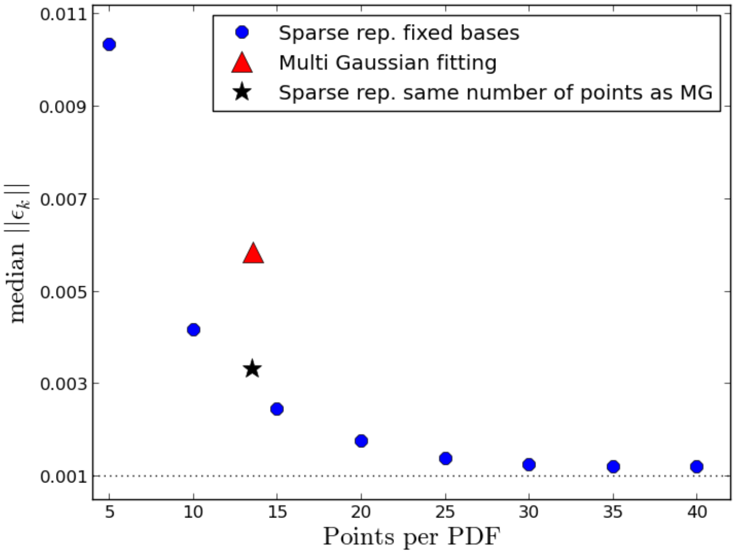


Figure 8.7: The median of the residual distribution as a function of the number of fixed bases used to reconstruct each galaxy's photo-z PDF when using the sparse representation technique (blue dots). For reference, the median of the multi-Gaussian residual distribution (red triangle) and the median of the sparse representation with variable number of bases (black star) are also shown, where on average both techniques need fourteen points per photo-z PDF.

decreases until approximately twenty-five bases are used. For comparison, we also show the multi-Gaussian residuals for fourteen values (red triangle) and the corresponding sparse basis representation residuals for approximately the same number of values (black star), demonstrating the superiority, in terms of precision, of the sparse representation over the multi-Gaussian. If we restrict the number of values to twenty, we have a median residual of 0.018, which corresponds to a median reconstruction of all one million test galaxies at 99.82% at a resolution of $\delta z = 0.011$. Since the original photo-z PDF contained two hundred points, this implies a compression ratio of ten.

Clearly these results will vary depending on the galaxy sample. In particular, the data we use in this analysis are from the CFHTLenS, which is a representative deep survey with galaxies that have photo-z PDFs with up to twelve peaks. The performance of the sparse representation also depends directly on the number of peaks in each PDF when we globally fix the number of bases. In Figure 8.8, we display the median of the residual distribution as a function of the number of peaks in the photo-z PDF, with different curves corresponding to different numbers of globally fixed bases. For a fixed number of bases, the residual increases as the number of peaks increase. Thus, a galaxy sample that consistently has a low number of peaks will have increased performance when using a smaller number of bases.

For example, we achieve a 99.5% reconstruction by using only ten values for galaxies with four or fewer peaks. In Carrasco Kind & Brunner (2014b), we discussed the relationship between the number of peaks and

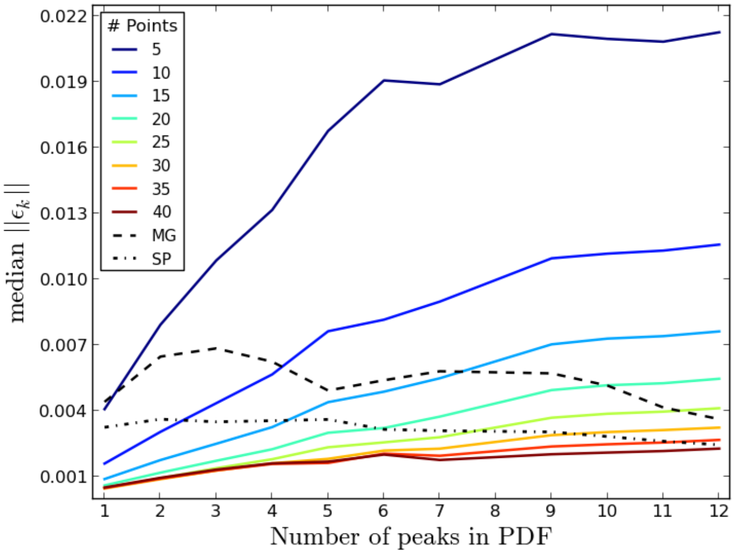


Figure 8.8: The median of the residual distribution as a function of the number of peaks in the photo-z PDF when using (solid color lines) a different number of fixed bases in the sparse basis representation, (black dashed line) when using the multi-Gaussian fitting technique, and (black dashed-dotted line) when using the sparse representation when the number of bases is equivalent to the number of multi-Gaussians.

the shape of the photo-z PDFs with the outlier fraction. With this in mind, we could reduce the number of bases used to reconstruct a sample and flag those with a high number of peaks, where the reconstruction is less reliable, for further investigation. In fact, we achieve a reconstruction of 99% for photo-z PDFs with three or fewer peaks when using only five bases for the sparse representation. This produces a compression ratio of forty when the original photo-z PDF has two hundred points.

For comparison, we also show the fitting residuals for the multi-Gaussian (black dashed line) and sparse representation (black dashed-dotted lines) where the variable number of bases matches the number of multi-Gaussians. The performance of the multi-Gaussian fitting is less dependent on the number of peaks simply because the number of parameters dynamically changes for each photo-z PDF. Overall, the multi-Gaussian performance is fairly consistent at around 0.005, even as we increase the number of peaks. The sparse representation with a variable number of bases, on the other hand, is less dependent on the number of peaks and has residuals that are nearly 50% smaller than the multi-Gaussian fitting at an approximately constant value of 0.003.

PDF Storage

In the previous section, we discussed how the sparse representation and the multi-Gaussian fitting can accurately represent a photo-z PDF by using only a few dozen values with a reconstruction level of 99%. In the case of the multi-Gaussian fitting, the number of parameters to be stored will depend on the number of peaks in each individual PDF. As discussed previously, we will have $3(Np_k + 1)$ parameters, which are all floating point numbers. For this dataset we found that the average number of values (or floating point parameters) required is fourteen; but to store these data for all galaxies, we would need to combine the results from different galaxies in order to take advantage of the galaxies that require fewer values so that we can also store those galaxies that require a larger number of parameters. Varying the number of values to store galaxy photo-z PDFs in this manner might not be practical, as it will likely depend strongly on the archival and storage system while also increasing the computational difficulty in dealing with a varying number of parameters for different photo-z PDFs. The practical solution would be to use thirty-nine fixed values (the maximum required for this dataset) for all galaxies and store them independently. This result is also true for the varying sparse representation, which we have demonstrated has a better performance in comparison to the multi-Gaussian when representing a photo-z PDF.

On the other hand, requiring a fixed number of basis functions per galaxy alleviates this issue and also has the additional benefit that there is no need to pad with zeros since having more points for single peaked galaxies simply provides a more accurate representation. We have shown that by using ten to twenty values we are able to produce a residual on the order of 0.1%, where all galaxies are stored independently. One additional (and very important) advantage of the sparse basis representation is that all bases in the dictionary have ℓ_2 norm equal to unity. Furthermore, when bases are computed by using the OMP algorithm, the absolute values of all coefficients are, by definition, less than unity. They can be negative, however, as seen in Figure 8.5. Since the PDFs are probability distributions, by definition the integral of the PDFs over the redshift range must also be unity. As a result, we can rescale all coefficients; and, as long as their relative amplitudes are the same, we can always impose the integral normalization at the end of the reconstruction.

If we continue this line of reasoning, we can rescale the coefficients of every basis function for a given galaxy so that the coefficients have absolute values between zero and one. When doing this we will be sure that the first basis function has unit amplitude without loss of accuracy on the very first basis. We can discretize this range by using approximately 32,000 sampling points (specifically 2^{15}) between zero and one, and store the corresponding integer from this range, and its sign, in a single sixteen-bit value. The error introduce by this discretization is very small, on the order of 10^{-5}, which is almost always negligible for most applications. In this approach, the most important basis is always first, and since it defines the scale, is always stored with no rounding errors.

We have, in fact, used this discretization throughout this chapter; the difference introduced by using this discretization and the real values is less than 0.0005% and thus it does not directly affect the representation

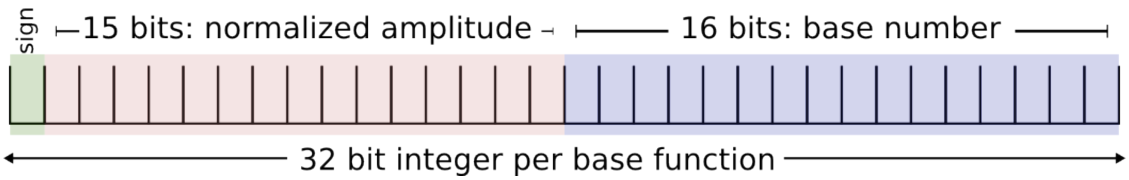

sign
15 bits: normalized amplitude
16 bits: base number
32 bit integer per base function

Figure 8.9: A single four-byte integer scheme to store a single basis function in the sparse representation method. The first sixteen bits store the coefficients (including sign), while the second sixteen bits store the location of the bases in the dictionary.

accuracy. This allows us to only use one value per basis function in our sparse representation. Since our dictionary contains fewer than 65,000 bases, which can be completely represented by a sixteen-bit integer, we can use a single four-byte integer, as shown graphically in Figure 8.9, to store both the base function and its amplitude and sign. More specifically, if we have a two hundred point photo-z PDF, which corresponds to a resolution of $\delta z = 0.011$ over the range $z = 0$ to $z = 2.2$, and we fix our representation to use ten bases, we can achieve an average reconstruction accuracy of 99.5% by using only 40Bs per photo-z PDF. Given a million galaxies that we treat in this manner, we will only need approximately 38MBs to store all of their PDFs. In addition, since we are only storing four-byte integers to represent the full photo-z PDF, we can potentially reduce the overall disk storage requirements by employing existing bit compression techniques (Lemire & Boytsov, 2012) which will be important for relational database systems.

The representation and data encoding scheme we have proposed is, of course, even more flexible than we have demonstrated. If our photo-z PDFs employ either a different redshift resolution, span a larger redshift range, or simply have been sampled at a higher number of points, we can still use a four-byte integer representation. For example, if the original PDF is sampled at a finer resolution, we can double N_μ and reduce N_γ by one-half and still retain the same number of bases, recall we simply need the number of bases to be less than 2^{16} (or 65,536) in order to still have 2^{15} bits to encode the basis function index. In an extreme case, we can revert to a purely Gaussian set of basis functions and allow N_μ and N_σ to vary while keeping the total number of bases below the 2^{16} limit. In this case, we likely would need to increase the number of fixed bases in order to accurately represent the photo-z PDF.

If the number of required bases exceeds the 2^{16} limit, because, for instance, our photo-z PDFs are sampled at an extremely high resolution or span a large redshift range, we can always increase the size of the dictionary beyond this two-byte limit. In this case, we simply have a very dense dictionary, where fewer fixed bases would be necessary; thus, each basis function would be stored in either a six-byte or an eight-byte integer, depending on the details of the computational system. Another alternative would be to fix the number of bits used to encode each type of basis function; for example, to use two-bits for N_γ, six-bits for N_σ and eight-bits for N_μ, resulting in four, sixty-four, and two hundred and fifty-six possible values for each basis function. As a fixed framework, this technique could also simplify the storage and functional indexing. Finally, as we have mentioned earlier, there is no need to store the entire dictionary since it is simply defined over a functional basis. Instead, we only need to store the parameters required to regenerate the dictionary so that we can either regenerate the dictionary or generate the individual functions themselves as needed.

8.3 Summary

In this Chapter, we have presented different techniques to represent and efficiently store photo-z PDFs, which have been shown to convey significantly more information than a single photo-z estimate. As we enter the era of precision cosmology, the growth of large, dense photometric surveys has created an unmet need to quantify and manage these probabilistic values for hundreds of millions to billions of galaxies. Specifically, we have introduced the use of a sparse basis representation that uses a dictionary of Gaussian functions and Voigt profiles, which have extended wings, to accurately and efficiently represent each photo-z PDF. We minimize the number of required bases while maintaining a high accuracy by using an Orthogonal Matching Pursuit algorithm, which provides a unique set of bases for each photo-z PDF while minimizing the residual between the original and final photo-z PDF.

We use the CF-1 and CF-2 data to compute photo-z PDFs by using our TPZ code, producing PDFs with two hundred points and a redshift resolution of $\delta z = 0.011$. By using these PDFs, we demonstrate the our proposed sparse basis representation reconstructs a more accurate PDF than other techniques, include a multi-Gaussian fitting approach with a flexible number of parameters based on the number of peaks in each PDF. If we use the exact same number of parameters with our sparse representation as used by the multi-Gaussian fitting, we found that the sparse basis representation results are superior with the additional benefit that each basis or parameter can be stored using a single integer. We also showed that, with a fixed number of bases, we could achieve both a highly accurate PDF that also has a large compression ratio. As a specific example, we found that by using only ten (twenty) values per photo-z PDF, we could reconstruct a photo-z PDF at over a 99.5% accuracy with a compression ratio of twenty (ten), providing a significant storage reduction without a loss of information.

We quantified the number of bases required within the sparse representation dictionary, specifically finding that $(\Delta z/\delta z)^2$ bases are sufficient to represent the galaxy photo-z PDFs in our CFHTLenS test sample, where Δz is the overall redshift range and δz is the photometric redshift PDF resolution. If the number of points in the original PDF is approximately 200–250, we can use a dictionary with fewer than 2^{16} bases, which results in an accurate PDF reconstruction while only requiring a single sixteen-bit integer to store the basis index in the dictionary. Furthermore, since the bases themselves are normalized, all basis coefficients are less than unity by definition; and, since the photo-z PDFs are also normalized, we only need to retain the relative amplitudes of each basis function.

Therefore, we can independently rescale the coefficients for each galaxy to their maximum value and subsequently represent them by using a discretized range containing 2^{15} values. This will provide a resolution less than 10^{-5}, and since we set the maximum value from the most significant basis function, it is always correctly represented. As a result, we can also store the coefficients (sign included) in a separate sixteen-bit integer without losing information. Taken together, we can completely encode a single basis function, both dictionary index and coefficient, in a single four-byte integer, simplifying the data management and significantly reducing the data storage and reconstruction computational requirements.

Of course the results we have presented will depend on the quality of the photo-z PDFs to which they are applied, which themselves depend on the details of the photo-z algorithm that generated them. As would naively be expected, single peaked photo-z PDFs are most accurately reconstructed by using either a multi-Gaussian fitting or a sparse basis representation, where only five points per photo-z PDF is sufficient to achieve a 99% accurate reconstruction. Overall, these results are very promising, as current and future photometric surveys will produce up to tens of billions of photo-z PDFs. Our proposed approach will either allow a reduction in the overall storage requirements or increase the number of photo-z PDFs that can be persistently maintained for each galaxy without increasing the required amount of storage. In the next Chapter we will see one direct application to this sparse representation as well as other general applications of the work presented so far.

Chapter 9

Applications

Outline

In this chapter we will discuss some of the applications of what it has been exposed on previous chapters that were not previously covered. We show how the use of photo-z PDFs provides a better reconstruction of the galaxy distribution as a function of the redshift in comparison to single photo-z estimators and how our sparse representation can help in its computation by introducing a new framework. We present a summary of the results of applying TPZ on early data taken from the Dark Energy Survey and the main results we obtained during a photometric redshift analysis on this data by comparing with other photo-z codes. We also show how the photo-z PDFs are also incorporated to carry out clustering studies using the Angular power Spectrum that is easily applicable to other clustering measurements like the Angular correlation function by using simulated data that mimics current photometric surveys.

9.1 N(z) and the galaxy distribution

In the previous chapters we have stated the importance of photo-z PDF in the analysis of clustering of galaxies and other cosmological measurements. We provided a detailed analysis on how we compute photo-z PDFs using our own methods and how we can improve these computations and also introduced a novel way to represent them and to store them. All these points are essential in order to reduce all possible systematics and to improve the metrics and accuracy of the PDFs which will enhance the cosmological measurement and will help to reduce the error in the parameter estimation.

Although we have already discussed the computation of $N(z)$ as one direct outcome from the photo-z computation, here we will discuss its computation and used in more detailed based on previous discussion. Most of the results we have presented within this thesis have been based on the estimation of a single metric computed from the photo-z PDF, for example the mean or mode. Obviously, using a single value to represent the PDF wastes significant information, but since many photo-z applications mimic spectroscopic redshift applications, new approaches must be developed to capitalize on the full information content of a photo-z PDF. Furthermore, several recent works have shown (e.g., Mandelbaum et al., 2008; Cunha et al., 2009;

Wittman, 2009; Bordoloi et al., 2010; Abrahamse et al., 2011), the use of a single number to represent the photo-z leads to biases. As a result, we present a simple, yet very important application that uses the full photo-z PDFs—estimating the galaxy redshift distribution, N(z). This function is a fundamental measurement and is very important to a number of cosmological applications including weak lensing tomography (e.g., Mandelbaum et al., 2008; Jee et al., 2013; Chisari et al., 2014) and projecting three-dimensional theoretical power spectra to angular clustering measurements (Blake et al., 2007; Myers et al., 2009; Hayes et al., 2012; Wang et al., 2013).

To illustrate this point, We compute the normalized galaxy redshift distribution, $N(z)$, for all the galaxies in DP-1 sample without $zConf$ cuts using the full photo-z PDF. This is the same data introduced in Chapter 2.3 and used in Chapter 3 and 4, shown in Figure 9.1 as the shaded gray area. As demonstrated by this figure, in this spectroscopic survey, most galaxies were selected to have redshifts between 0.6 and 1.2. Next, we compute the binned photometric redshift distribution by using the mean value from each photo-z PDF, shown by the red curve. While this curve does trace the gross features of the underlying spectroscopic redshift distribution, it fails to capture the full detail and can be significantly different at certain redshifts, including at the mode. For comparison, we show in black the photo-z PDF redshift distribution that we obtain by simply stacking the individual PDFs together. With this simple approach, we obtain a more accurate representation of the true sample redshift distribution. Here we have used all the galaxies, without selecting galaxies by their confidence level. This demonstrates that all individual PDFs computed with TPZ carry important information about the underlying distribution.

These differences are more clearly exposed in the bottom panel of Figure 9.1, where we show the absolute fractional error, $(N_{phot} - N_{spec})/N_{spec}$, as a function of redshift, using the same color scheme as before. From this figure, we see that the stacked PDF has a smaller error for almost all redshifts. In addition, the photo-z PDF redshift distribution is considerably smoother and looks more like a fit to the spectroscopic sample, which is another benefit of using the full photo-z PDF. For this particular demonstration, the photo-z PDF presented used a bin size of 0.002, while the spectroscopic and photometric redshift distributions used a bin size of 0.03. Of course, we can generate smoother distributions for either the spectroscopic or photo-z mean value redshift distributions by reducing the bin size, however, the trade off is that we run the risk of increasing the shot noise in the resulting distribution.

We came to similar conclusions on the other Chapters as well, Figures 4.7, 6.9, 6.13 and 8.2 also show examples where the $N(z)$ distributions of the spectroscopic sample (for different data sets) is compared to the one obtained by stacking the photo-z PDF together with remarkably agreement validating this discussion.

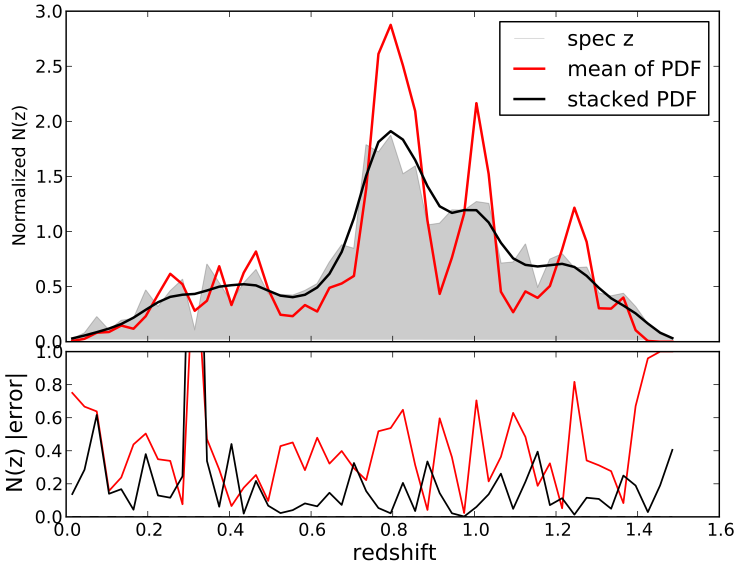


Figure 9.1: (*Top*): The redshift distribution for the all DEEP2 spectroscopic sample of galaxies (shaded gray histogram), computed from the mean value of individual photo-z PDFs (red curve), and computed by stacking individual photo-z PDFs (black curve). (*Bottom*): The residual absolute error between the spectroscopic redshift distribution and the two photo-z redshift distributions shown using the same color scheme.

9.1.1 $N(z)$ using sparse representation

We now focus on how to compute $N(z)$ and how to incorporate the tools presented in Chapter 8 to reduce the computational cost. Usually, this function is computed by binning spectroscopic observations of galaxies as a function of redshift; but for a photometric survey, this distribution is optimally computed by integrating over all individual photo-z PDFs at a given resolution. This approach is indeed more computational challenging than using individual photo-z when a large number of galaxies is available, but as discussed before is more accurate. Therefore, even with this simple application, however, we benefit from the use of a sparse representation for our photo-z PDFs, since we can transform our theoretical framework to use our basis functions. Thus we can operate directly over the dictionary and use the sparse basis indices and coefficient parameters to calculate the true and reconstructed values taking into account the normalization.

As a demonstration we use the same data used in Chapter 8 and detailed in §2.4. We can derive the framework to compute the galaxy redshift distribution directly over the basis functions. For this, we start by writing the definition of $N(z)$ and we use the same notation introduced in the previous chapter:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz \tag{9.1}$$

where the sum is over all N galaxies and $P_k(z)$ is the photo-z PDF of a given galaxy k. z is the midpoint

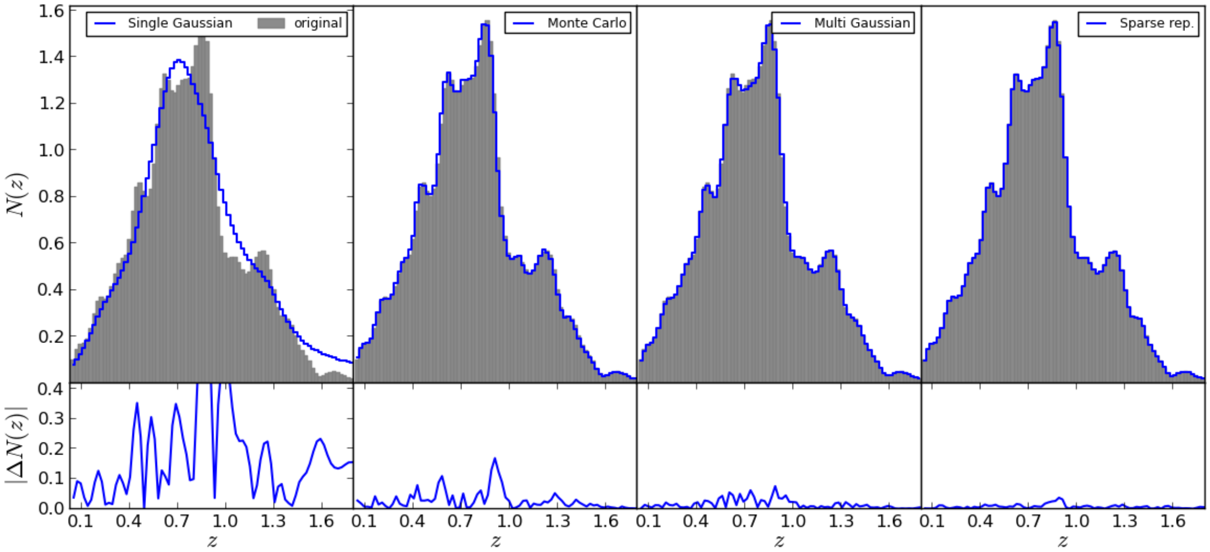


Figure 9.2: The $N(z)$ distribution for all 10^6 galaxies from the CFHTLenS data (CF-2) computed by using the four photo-z PDF representation techniques. Within each panel, the original $N(z)$ computed by stacking the full photo-z PDFs is shown in gray and a different representation method is shown in blue. From left to right we have the single Gaussian model, the Monte Carlo sampling described in Section 8.1.1, the multi-Gaussian fitting method, and the sparse representation method that uses the same number of bases as the multi-Gaussian method. The bottom panels show the absolute difference between the original and reproduction at the same scale.

of each redshift bin, which have a fixed width Δz. We can rewrite this equation in terms of the PDF representation for $P_k(z)$, which we previously defined as $\mathbf{pz}_k$. Thus, in the sparse basis representation, we can express each PDF as:

$$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k \tag{9.2}$$

where $\boldsymbol{\delta}_k$ is a sparse vector, which might contain ten to twenty elements, that contains the amplitudes for each functional basis and $\mathbf{D}$ is an $n \times m$ dictionary, where n is the number of points in the original PDF and m is the total number of bases. By using this result, we can rewrite Equation 9.1:

$$N(z) = \sum_{k=1}^{N} \boldsymbol{\delta}_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{D} dz \tag{9.3}$$

where $\boldsymbol{\delta}_k$ is independent of redshift so that we only need to integrate once over each basis function in the dictionary; thus, we only have m integrations instead of N.

Furthermore, we can precompute this integral over $\mathbf{D}$, which we denote by $\mathbf{I_D}(z)$. This integral corresponds to a vector of length m, where each entry is the integral over each one of the basis functions $\mathbf{d}_j$ in

D:

$$\mathbf{I}_{\mathbf{D}}(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \qquad j = 1, 2, \ldots, m \tag{9.4}$$

Since all N galaxies are expressed in terms of the m bases, we can also pre-factorize the coefficients (or amplitudes) in a vector $\boldsymbol{\delta}_N$:

$$\boldsymbol{\delta}_N = \sum_{k=1}^{N} \boldsymbol{\delta}_k \tag{9.5}$$

Therefore, after these precomputations we can simply express $N(z)$ as:

$$N(z) = \mathbf{I}_{\mathbf{D}}(z) \cdot \boldsymbol{\delta}_N \tag{9.6}$$

reducing the computation to a simple dot product of precomputed quantities. For each bin, we need to compute $\mathbf{I}_{\mathbf{D}}(z)$, but $\boldsymbol{\delta}_N$ is computed only once and can be used both for all bins in the computation of $N(z)$ and in other cosmological applications. This result is also true for other linear operations that might be involved in another cosmological analysis. Thus, by working directly in the space defined by the basis functions, we can reduce computational memory and processing times significantly.

We compare the original $N(z)$ for 10^6 test galaxies from the CFHTLenS sample to different $N(z)$ distributions reconstructed by using Equation 9.6 for different representation formats in Figure 9.2. Each original galaxy photo-z PDF has a resolution of $\delta z = 0.011$ and contains two hundred values. We restrict the comparison in Figure 9.2 to four techniques: a single Gaussian model, the Monte Carlo estimator described in Section 8.1.1, a multi-Gaussian fitting technique, and the sparse basis representation. However, we compute the fractional percentile error between the original $N(z)$ and all eight techniques and compare the results in Table 9.1. Before discussing the performance of individual techniques, we note that the lower panels in Figure 9.2 are all shown at the same scale to facilitate direct comparisons.

In the first panel, we see that the single Gaussian model clearly shows a significant difference, which is visible both from the distribution itself and in the bottom panel from the absolute error between these two distributions. Next, we see that the single point photo-z estimation computed by using a Monte Carlo sampling shows a surprisingly good agreement with the original distribution. This result was discussed by Wittman (2009), who demonstrated that this technique does provide a fair statistical representation of the sample's galaxy redshift distribution. This approach, where the $N(z)$ distribution is computed as a random sample drawn from the cumulative PDF of each galaxy, statistically compensates for the photo-z errors for an individual galaxy and thus produces a reliable $N(z)$ distribution. This approach does, however, introduce much larger errors on the estimation of individual galaxy photo-zs. While one might be tempted to store a

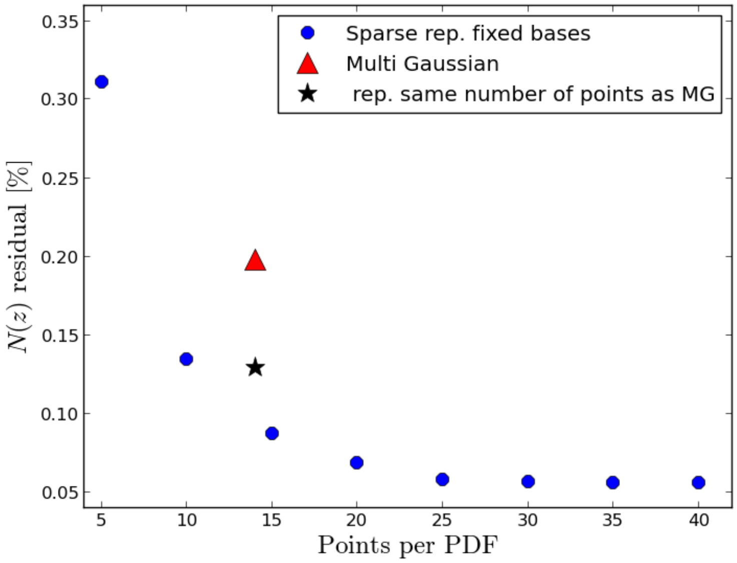


Figure 9.3: The fractional percentile error between the original $N(z)$ and the reconstructed $N(z)$ computed by using Equation 9.6 when fixing the number of bases used to represent the photo-z PDFs for all galaxies with the sparse representation technique (blue dots). For comparison, we also show the multi-Gaussian (red triangle) and the sparse representation with a variable number of bases (black star) residuals. On average, both of these latter techniques require fourteen points per photo-z PDF.

photo-z PDF by using this approach in order to accurately recover an $N(z)$ by storing a minimum quantity of new data, it would be easier to simply compute and store the actual $N(z)$. Furthermore, since this technique is dependent on using a large number of galaxies to generate a more accurate $N(z)$, if one is interested in computing the redshift distribution for galaxy subsets, the reconstruction accuracy might suffer as the number of galaxies in the subsamples is decreased.

In the third panel, we see that the multi-Gaussian fitting technique has a smaller error than the first two methods. As discussed previously, this method provides an accurate representation of a photo-z PDF, thus it would be expected to also yield an accurate representation for $N(z)$. Finally, in the last panel we have the results for the sparse basis representation where the number of bases used is defined to be the same as required for the multi-Gaussian fitting method. As seen previously with the distribution of residuals, we see that, with this direct cosmological application, we recover the original $N(z)$, by using the same number of values to represent the photo-z PDF, more accurately than with other techniques.

We present the fractional percentile error between the original $N(z)$ and the reconstructed $N(z)$ computed by using Equation 9.6 for different photo-z PDF representation techniques in Figure 9.3. As also seen in Figure 8.7, we see that as the number of bases increases for the sparse representation (shown in blue dots) the accuracy of the reconstruction also improves, but here we focus on the error in the reconstruction of $N(z)$. Since additional bases will produce a more accurate photo-z PDF representation, we also expect a more accurate $N(z)$ reconstruction when the number of bases increases. For comparison, we also show the multi-Gaussian fitting (red triangle) and the sparse representation (black star) where the number of bases matches the multi-Gaussian fitting value.

We observe that, by using only fifteen values, we can reconstruct the $N(z)$ distribution to an accuracy of 99.9% as measured with respect to the original distribution. In addition, this result changes only slightly when we limit our representation to ten bases. We also see that the error values are slightly better than we saw when reconstructing the individual photo-z PDFs, because computing the $N(z)$ smooths over the individual photo-z PDFs, thereby reducing the impact from small discrepancies in individual photo-z PDFs that might result from using a specific functional basis. If we increase our representation to use forty bases, we can reconstruct the $N(z)$ distribution to nearly 99.95%, but the decrease in the error, however, does not change significantly once we have used approximately twenty-five bases, suggesting there are diminishing returns.

In Section 8.1.1, we introduced several different individual photo-z estimates that are widely used, including the mean, the mode, and the median of the photo-z PDF, and Monte Carlo sampling from the cumulative photo-z PDF. These single estimates show an even larger fractional error than visible on the vertical axis shown in Figure 9.3, and are thus presented in Table 9.1, which summarize the results from all of the methods presented herein, including the number of values required by the representation method and the fractional percentile error for that method in reconstructing the original $N(z)$.

The entries in Table 9.1 are presented in ascending order by the size of this fractional percentile error. From these entries, we see that the single Gaussian model has, on average, a reconstruction error of 2.2% while the single value estimates all have reconstruction errors over 6%. The Monte Carlo sampling method provides the best reconstruction results when using a single photo-z estimate with an error of about 0.4%, which is comparable to a sparse representation that uses five bases. As mentioned previously, however, this technique does not provide accurate individual photo-z estimates. We also observe that the difference when using thirty, thirty-five, or even forty bases is very small, although it is bigger than the resolution in the discretization scheme; thus our proposed discretization method does not impact these results and we can safely represent each basis function by using a single four-byte integer.

The integration over the dictionary of bases, as shown in Equation 9.4, can also be used to compute $N(z)$ over different redshift bins. In this case, the integration can be performed by using the bases and subsequently applying Equation 9.6 when using the sparse basis representation. Furthermore, we can extend this approach to analyze multiple photo-z PDFs for each galaxy, where they are each represented by the same dictionary. This would prove useful when a survey has stored photo-z PDFs for the same galaxy by using different galaxy

Table 9.1: The fractional percentile error between the original $N(z)$ and a reconstructed $N(z)$ computed by using the sparse basis representation, single and multi-Gaussian fitting, and all of the single point photo-z techniques described in Section 8.1.1. We also list the number of values required to represent the photo-z PDF. The table is sorted in ascending order by the percentile error.

Method	Values per PDF	Error [%]
sparse rep. fixed	40	0.05545
sparse rep. fixed	35	0.05551
sparse rep. fixed	30	0.05611
sparse rep. fixed	25	0.05750
sparse rep. fixed	20	0.06829
sparse rep. fixed	15	0.08729
sparse rep. same MG	14	0.12930
sparse rep. fixed	10	0.13440
multi-Gaussian	14	0.19779
sparse rep. fixed	5	0.31113
Monte Carlo	1	0.37294
single Gaussian	2	2.19095
Median PDF	1	6.63550
Mean PDF	1	7.47077
Mode PDF	1	13.24271

spectral templates. Thus, a scientist could either compute an $N(z)$ by using a single, per-galaxy *best* template or compare different $N(z)$ that are computed by using different template combinations. Since the integrals could all be precomputed, the only new computation is for the basis coefficients for each galaxy, dramatically reducing the overall computational demands.

For example, we might have different (or even updated) priors for different galaxy types in a survey. We can quickly apply these new priors to the precomputed dictionary integrals and recover the results for each galaxy given their basis coefficients in an efficient manner. Alternatively, one might want to minimize over the galaxy type under certain restrictions, which can be applied over the precomputed integrals of the dictionary of bases. The minimization problem subsequently becomes a simple task of selecting the minimum or maximum sum over the coefficients. As should be evident from these examples, there exist a number of different applications where our proposed sparse basis representation not only reduces the overall storage requirements, often significantly, but also reduces the computational requirements for cosmological analyses.

9.2 Dark Energy Survey Science Verification data

In Sánchez et al. (2014) we present results of a study of the photometric redshift performance on early data coming from the Dark Energy Survey which is briefly described in Chapter 2.5 and fully detailed in Sánchez et al. (2014). We study the performance of 13 photo-z codes and a detailed study on 4 particular codes including TPZ . Within the spectroscopic sample (DS-1) matched to the DES data, we make 2 subsamples;

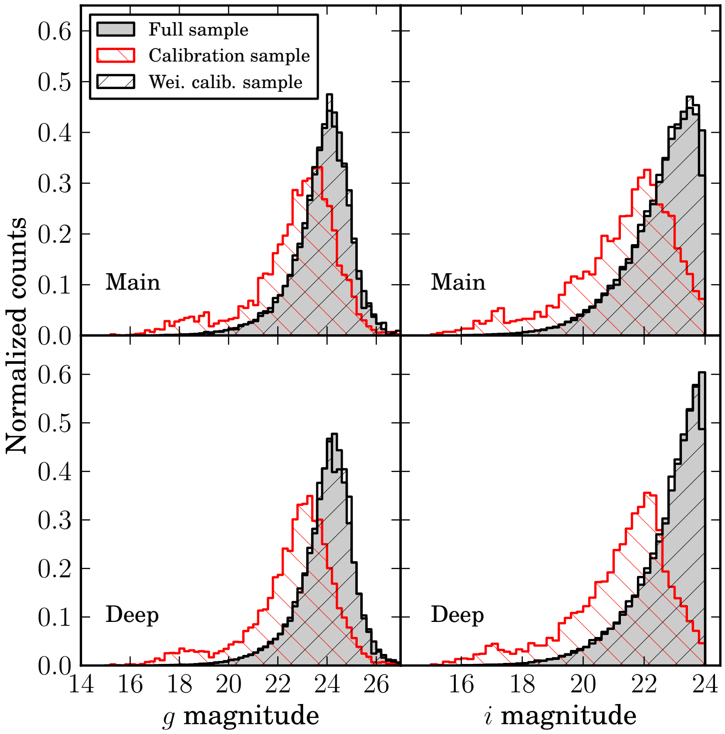


Figure 9.4: g and i magnitude distributions for the full, calibration and weighted calibration sample. The difference between the full and the calibration samples is apparent, the latter being significantly brighter. After applying the weighting procedure described in Lima et al. (2008), the weighted calibration distributions agree very well with the corresponding DES-SV distributions.

Main and Deep, the former refers to a sample that has the same depth in magnitude as the main depth of the DES survey. The Deep sampler refers to a deeper depth observations taken from data selected from supernova fields which are observed multiple times. In this section we present a summary of the main results from that study, where some of the text and figures are extracted from Sánchez et al. (2014).

9.2.1 The weighting procedure

In order to assess the photo-z performance of the DS-1 data we would ideally need a calibration sample being representative of the DES-SV full sample (DS-2 data), i.e. having exactly the same photometric properties (magnitude and colour distributions). However, spectroscopic galaxy samples are shallower, and suffer from selection effects. A weighting procedure, which assigns a weight to each of the galaxies in the calibration sample so that the distributions of their photometric observables reproduce the distributions of the same observables in the full sample, can be used provided there is enough overlap between the photometric spaces of the calibration and full samples (Lima et al., 2008; Cunha et al., 2009).

Different algorithms can be used to compute the weights, but basically all compare local densities in

Table 9.2: Definition of the metrics used in the text to present the main results. These are computed in the whole redshift range as well as in bins of width 0.1 in photometric redshift. Detailed definitions can be found in the appendix.

Metric	Description	Requirement
$\overline{\Delta z}$	mean of the Δz distribution	-
$\sigma_{\Delta z}$	standard deviation of the Δz distribution	-
Δz_{50}	median of the Δz distribution	-
σ_{68}	half width of the interval around Δz_{50} containing 68% of the galaxies	< 0.12
$\mathrm{out}_{2\sigma}$	fraction of galaxies with: $\lvert\Delta z - \overline{\Delta z}\rvert > 2\sigma_{\Delta z}$	< 0.1
$\mathrm{out}_{3\sigma}$	fraction of galaxies with: $\lvert\Delta z - \overline{\Delta z}\rvert > 3\sigma_{\Delta z}$	< 0.015
$\overline{\Delta z'}$	mean of the $\Delta z' = \Delta z/\epsilon_{\mathrm{phot}}$ distribution	-
$\sigma_{\Delta z'}$	standard deviation of the $\Delta z'$ distribution	-
$\mathrm{N}_{\mathrm{poisson}}$	difference between $N(z)^{\mathrm{phot}}$ and $N(z)^{\mathrm{spec}}$ normalized by Poisson fluctuations	-
KS	Kolmogorov - Smirnov statistic for $N(z)^{\mathrm{phot}}$, $N(z)^{\mathrm{spec}}$	-

the photometric spaces of the two samples (calibration and full) and assign a weight to each photometric region of the calibration sample equal to the ratio between the densities of galaxies in the full sample and the calibration sample in a given region. In this study we use a nearest neighbour algorithm to compute the weights that we use extensively throughout the section. A detailed description of the method can be found in Lima et al. (2008).

We apply the weighting technique within a region in the multidimensional space defined by $18 < i_{AB} < 24$; $0 < g - r < 2$; $0 < r - i < 2$. In Fig. 9.4 one can check how the weighting procedure is efficiently applied for the sample used in this study. The figure shows, for two DES bands, and the Main and Deep samples, the magnitude distributions for the full sample, the calibration sample and the weighted calibration sample, whose distributions agree very well with those of the full sample.

9.2.2 Metrics used to compare and assess the different methods

To carry out a the photometric study and compare different approaches we use an extensive set of metrics that are shown in Table 9.2 together with the DES science requirements for photo-z , defined before the start of the survey.

The photo-z metrics we consider are intended to measure the quality of the photometric redshifts in terms of their bias, scatter, and outlier fraction statistics, and also in terms of the fidelity of the photo-z errors and of the agreement between the photo-z and true redshift distributions. For each photo-z code and each galaxy we have either the photo-z estimation and its associated error or a probability density function $P(z)$. As described in the text, a vector of weights were computed in order to match the spectroscopic and photometric samples in multi-color and magnitude space. On each sample we have a vector $\boldsymbol{\omega}$ of weights corresponding to the N galaxies on each test set, where $\sum_{i=1}^{N} \omega_i = 1$. If no weights are used, the default value $\omega_i = \frac{1}{N}$ is assigned to each galaxy. We define the individual bias as $\Delta z_i = z_{\text{phot},i} - z_{\text{spec},i}$ and the statistics used in this work as follows:

1. mean bias($\overline{\Delta z}$):
$$\overline{\Delta z} = \frac{\sum \omega_i \Delta z_i}{\sum \omega_i} \tag{9.7}$$

2. $\sigma_{\Delta z}$:
$$\sigma_{\Delta z} = \left(\frac{\sum \omega_i \left(\Delta z_i - \overline{\Delta z} \right)^2}{\sum \omega_i} \right)^{\frac{1}{2}} \tag{9.8}$$

3. median (Δz_{50}), the median of the Δz distribution, fulfilling:
$$P_{50} = P(\Delta z \leq \Delta z_{50}) = \int_0^{\Delta z_{50}} \omega(\Delta z) d(\Delta z) = \frac{1}{2} \tag{9.9}$$

4. σ_{68}, half of the width of the distribution, measured with respect to the median, where 68% of the data are enclosed. This is computed as:
$$\sigma_{68} = \frac{1}{2} \left(P_{84} - P_{16} \right) \tag{9.10}$$

5. $\text{out}_{2\sigma}$, the fraction of outliers above the $2\sigma_{\Delta z}$ level:
$$\text{out}_{2\sigma} = \frac{\sum W_i}{\sum \omega_i} \tag{9.11}$$
where,
$$W_i = \begin{cases} \omega_i, & \text{if } |\Delta z_i - \overline{\Delta z}| > 2\sigma_{\Delta z} \\ 0, & \text{if } |\Delta z_i - \overline{\Delta z}| \leq 2\sigma_{\Delta z} \end{cases}$$

6. $\text{out}_{3\sigma}$, the fraction of outliers above the $3\sigma_{\Delta z}$ level:
$$\text{out}_{3\sigma} = \frac{\sum W_i}{\sum \omega_i} \tag{9.12}$$
where,
$$W_i = \begin{cases} \omega_i, & \text{if } |\Delta z_i - \overline{\Delta z}| > 3\sigma_{\Delta z} \\ 0, & \text{if } |\Delta z_i - \overline{\Delta z}| \leq 3\sigma_{\Delta z} \end{cases}$$

7. $\overline{\Delta z'}$, the mean of the distribution of Δz is normalized by their estimated errors. Ideally this distribution should resemble a normal distribution with zero mean and unit variance. We define $\Delta z_i' = \Delta z_i / \epsilon_{\text{phot},i}$

where $\epsilon_{\mathrm{phot},i}$ is the computed error of the photometric redshift for galaxy i. Then:

$$\overline{\Delta z'} = \frac{\sum \omega_i \Delta z'_i}{\sum \omega_i} \tag{9.13}$$

8. $\sigma_{\Delta z'}$:

$$\sigma_{\Delta z'} = \left(\frac{\sum \omega_i \left(\Delta z'_i - \overline{\Delta z'} \right)^2}{\sum \omega_i} \right)^{\frac{1}{2}} \tag{9.14}$$

9. N_{poisson}, a metric that quantifies how close the distribution of photometric redshifts $N(z_{\mathrm{phot}})$ is to the distribution of spectroscopic redshifts $N(z_{\mathrm{spec}})$. For each photometric redshift bin j of width 0.1, we compute the difference of $N(z_{\mathrm{phot}}) - N(z_{\mathrm{spec}})$ normalized by the Poisson fluctuations on $N(z_{\mathrm{spec}})$:

$$n_{\mathrm{poisson},j} = \frac{\left(\sum\limits_{z_{\mathrm{phot},i} \,\epsilon\, \mathrm{bin}_j} \omega_i N - \sum\limits_{z_{\mathrm{spec},i} \,\epsilon\, \mathrm{bin}_j} \omega_i N \right)}{\sqrt{\sum\limits_{z_{\mathrm{spec},i} \,\epsilon\, \mathrm{bin}_j} \omega_i N}}$$

Then N_{poisson} is computed as the RMS of the previous quantity:

$$N_{\mathrm{poisson}} = \left(\frac{1}{n_{bins}} \sum_{j=1}^{n_{bins}} n^2_{\mathrm{poisson},j} \right)^{\frac{1}{2}} \tag{9.15}$$

10. KS is the Kolmogorov-Smirnov test that quantifies whether the two redshift distributions ($N(z_{\mathrm{phot}})$ and $N(z_{\mathrm{spec}})$) are compatible with being drawn from the same parent distribution, independently of binning. It is defined as the maximum distance between both empirical cumulative distributions. The lower this value, the closer are both distributions. The empirical cumulative distribution function is calculated as:

$$F_{\mathrm{spec}}(z) = \frac{\sum\limits_{i=1}^{N} \Omega_{z_{\mathrm{spec},i} < z}}{\sum \omega_i}$$

where,

$$\Omega_{z_{\mathrm{spec},i} < z} = \begin{cases} \omega_i, & \text{if } z_{\mathrm{spec},i} < z \\ 0, & \text{otherwise} \end{cases}$$

Similarly, the empirical cumulative distribution function $F_{\mathrm{phot}}(z)$ is computed for $N(z_{\mathrm{phot}})$. Then the KS statistic is computed as:

$$\mathrm{KS} = \max_{z}\left(|F_{\mathrm{phot}}(z) - F_{\mathrm{spec}}(z)|\right) \tag{9.16}$$

For the submissions with a $P(z)$ for each galaxy, these cumulative distributions are computed taking into account the $p(z)$ distribution for each galaxy

9.2.3 Photo-z methods

Before presenting a summary of the results we list in Table 9.3 the algorithms used during this analysis. For more relevant information, regarding the details at the time of running these codes see Sánchez et al. (2014) and for an exhaustive description of them see the references listed in Table 9.3. For template-based methods, a standardized set of filter throughput curves has been used. Most of the codes have been run in standalone mode, while a fair fraction of them has been run within the DES Science Portal, with compatible results. Due to the large number of codes used, the study, other than showing the DES-SV photo-z capabilities, also serves as a helpful reference to compare different photo-z codes using real data from a deep galaxy survey.

Table 9.3: List of methods used to estimate photo-z's in this section. Code type and main references are given.

Code	Reference
Training-based	
DESDM, Artificial Neural Network	Oyaizu et al. (2008b)
ANNz, Artificial Neural Network	Collister & Lahav (2004)
TPZ , Prediction Trees and Random Forest	Carrasco Kind & Brunner (2013a, 2014c)
RVMz, Relevance Vector Machine	Tipping (2001)
NIP-kNNz, Normalized Inner Product Nearest Neighbor	de Vicente et al., in preparation
ANNz2, Machine Learning Methods	Sadeh et al., in preparation
ArborZ, Boosted Decision Trees	Gerdes et al. (2010)
SkyNet, Classification Artificial Neural Network	Bonnett (2013); Graff et al. (2014)
Template-based	
BPZ , Bayesian Photometric Redshifts	Benítez (2000); Coe et al. (2006)
EAZY, Easy and Accurate Redshifts from Yale	Brammer et al. (2008)
LePhare	Arnouts et al. (2002); Ilbert et al. (2006)
ZEBRA, Zurich Extragalactic Bayesian Redshift Analyzer	Feldmann et al. (2006)
Photo-Z	Bender et al. (2001)

9.2.4 Results

During this study we carried out several tests including some ones like train on a deep depth sample, test on a main depth sample, add u band when available, among others. We will summarize the results of the *Test 1* which is the most important of the test configurations and defer the reader for a complete overview of the results to Sánchez et al. (2014). We also check the differences in the results under variations in the calibration data and the weights used. Note that the results presented in this subsection are those considering all the galaxies (with quality cuts), which are represented by one single statistic, later in the section we analyze some of these results in more detail.

Test 1: *Main-Main*

This test is the most representative of the results shown in this study, the default case. We use here the Main training sample to train and calibrate the photo-z algorithms and the Main testing sample to validate them, therefore, the test represents the real situation for most of the data collected in the DES survey.

In order to display the performance of all codes, in Figure 9.5 we show the z_{phot} vs. z_{spec} scatter plot for all the codes listed in Table 9.3. Furthermore, we compute all the metrics presented in Table 4.1 and described in Section 9.2.2. The results, using all the objects in the testing sample except for the 10% quality cut (based on photo-z errors and allowed by the DES requirements), are shown in Table 9.2.4.

The left panel of Figure 9.6 shows σ_{68}, related to the precision of the photometric redshifts versus the mean bias of the photo-z's. The black dashed line sets the DES science requirement on σ_{68}, and one can check how most of the codes presented in this work are below this line, thus fullfilling this important requirement on precision. Also, among the codes satisfying the σ_{68} requirement, there is a subgroup having very low bias as well. In the left panel of Figure 9.6 we show a zoomed-in of this region of interest, where we can see how training-based codes, either producing a single photo-z estimate or a probability density function, $P(z)$, are the ones showing best performance (all the codes in the zoomed-in region belong to the training-based category), among these we note that our TPZ and SkyNet provide the lowest metrics among them all.

One crucial aspect of photo-z studies, which we have discussed along this thesis, is the estimation and calibration of the true galaxy redshift distributions $N(z)$. In this section we use two metrics to compare the reconstruction of the true redshift distribution by the different photo-z algorithms: the $\mathrm{N}_{\mathrm{poisson}}$ and KS statistics, defined before. In both cases, the smaller the value, the closer are the true redshift distribution and its reconstruction through photo-z's. The right panel of Figure 9.6 shows these values for all the codes analyzed in Test 1. As expected, the two metrics are strongly correlated. It can also be seen how having a redshift PDF for each galaxy, instead of a single-estimate photo-z, helps a given code to have a better redshift reconstruction. This can be inferred looking at the cases where both the PDF and the single-estimate are displayed (TPZ , ANNz2, BPZ): in all these cases the PDF version of the code obtains better results in terms of these two metrics. As for the results, TPZ and the nearest-neighbor code, NIP-kNNz, show the best performance in this regard.

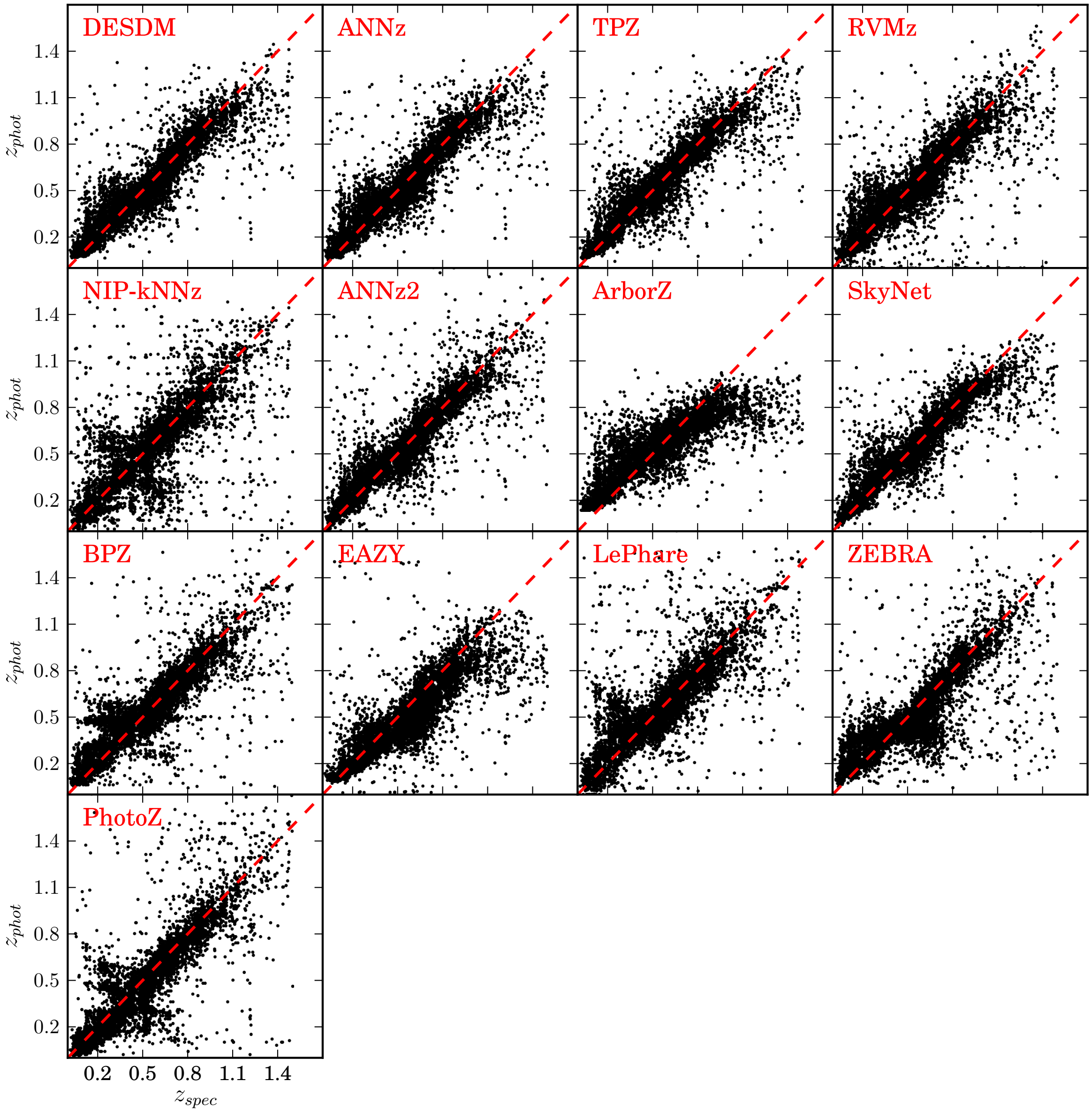


Figure 9.5: z_{phot} vs. z_{spec} scatter plot for all the codes analyzed in Test 1 and listed in Table 9.3.

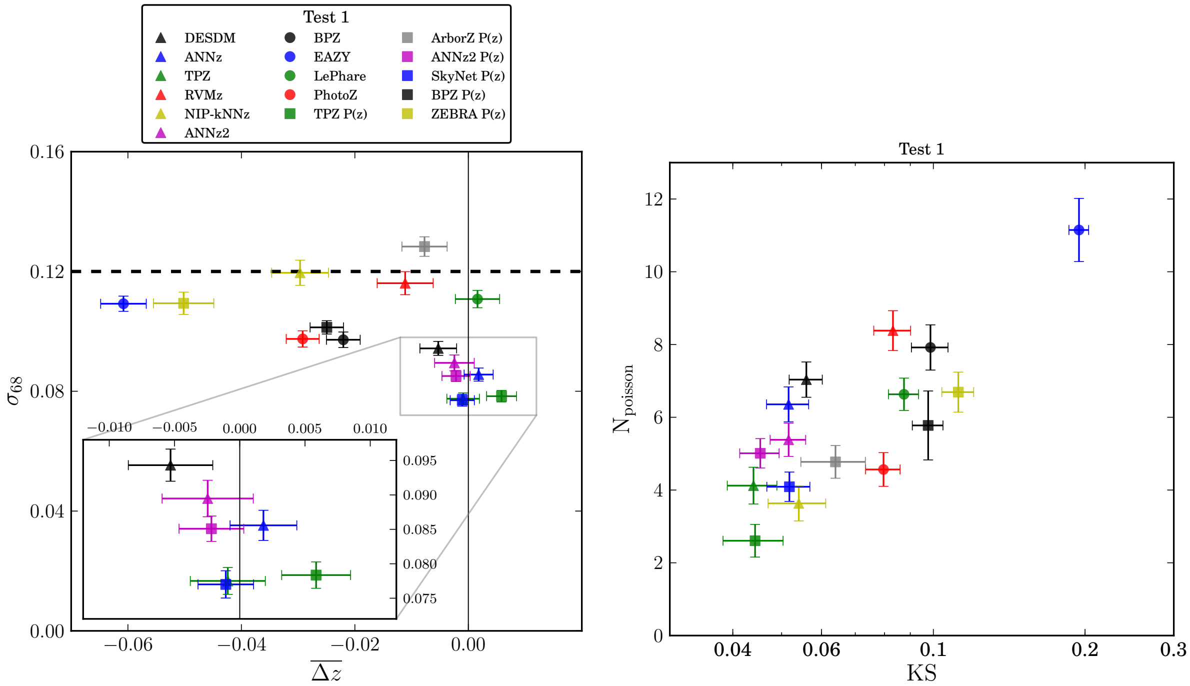


Figure 9.6: *(Left)*:σ_{68} vs. bias for all the codes analyzed in Test 1. Black dashed lines represent the DES science requirements in this and subsequent figures. Training-based codes have triangles as markers, template-based have circles, and codes producing a probability density function (PDF) for the redshift are marked with a square. Training-based codes, either producing a single photo-z estimate or a PDF, are the only ones present in the region of best performance (zoom-in). *(Right):* $N_{\rm poisson}$ vs. KS statistics for all the codes analyzed in Test 1. Both metrics show how the true galaxy redshift distribution is reconstructed through photo-z's, for each code. The smaller the value of the metric, the better the reconstruction. A strong correlation between the two metrics is observed, as expected.

As pointed out in Carrasco Kind & Brunner (2014c) these results might vary for different regions on the multidimensional photometric space or within the redshift range. Usually, training-based algorithms perform better on areas well populated with training galaxies and poorly on those less dense regions (as in high redshift bins), fact that we can observe from Figure 9.5 where training-based methods tend to have tighter distributions at the center while some template-based methods can compute photo-zs for galaxies at higher redshift more efficiently.

9.2.5 Results for DESDM, TPZ, SkyNet and BPZ photo-z codes

So far we have compared a large number of photo-z codes in a variety of situations and configurations. Next we look in greater detail at four photo-z codes: DESDM, TPZ , SkyNet and BPZ . The DESDM photo-z code, a regression artificial neural network, is integrated within the DES Data Management service, so its results will be made available together with all the DES data products, making it a clear choice to be studied in

Table 9.4: Results of all the photo-z metrics listed in § 9.2.2 for all the codes analyzed in Test 1. The errors are computed from bootstrap resampling with 100 samples. The weighting procedure has been applied, together with a cut on the 10% of the galaxies with the highest estimated photo-z errors for each code.

Test 1	$\overline{\Delta z}$	Δz_{50}	σ_{68}	$\sigma_{\Delta z}$	$\mathrm{out}_{2\sigma}$	$\mathrm{out}_{3\sigma}$	$\overline{\Delta z\prime}$	$\sigma_{\Delta z\prime}$	$\mathrm{N}_{\mathrm{poisson}}$	KS
DESDM	-0.005 ± 0.003	-0.003 ± 0.002	0.094 ± 0.002	0.135 ± 0.005	0.053 ± 0.005	0.018 ± 0.003	-0.047 ± 0.032	1.479 ± 0.052	7.035 ± 0.486	0.056 ± 0.004
ANNz	0.002 ± 0.003	-0.001 ± 0.003	0.086 ± 0.002	0.118 ± 0.004	0.049 ± 0.004	0.015 ± 0.002	0.096 ± 0.046	3.341 ± 0.134	6.355 ± 0.480	0.052 ± 0.005
TPZ	-0.001 ± 0.003	0.004 ± 0.002	0.078 ± 0.002	0.122 ± 0.006	0.046 ± 0.004	0.019 ± 0.002	0.019 ± 0.032	1.529 ± 0.063	4.122 ± 0.505	0.044 ± 0.005
RVMz	-0.011 ± 0.005	-0.004 ± 0.002	0.116 ± 0.004	0.180 ± 0.008	0.060 ± 0.005	0.023 ± 0.003	-0.084 ± 0.041	1.371 ± 0.098	8.382 ± 0.548	0.083 ± 0.007
NIP-kNNz	-0.030 ± 0.005	-0.011 ± 0.002	0.120 ± 0.004	0.197 ± 0.009	0.058 ± 0.005	0.018 ± 0.003	-0.186 ± 0.030	1.116 ± 0.117	3.633 ± 0.482	0.054 ± 0.007
ANNz2	-0.002 ± 0.003	-0.003 ± 0.002	0.089 ± 0.003	0.151 ± 0.009	0.042 ± 0.003	0.021 ± 0.002	0.063 ± 0.071	2.280 ± 0.255	5.381 ± 0.458	0.052 ± 0.004
BPZ	-0.022 ± 0.003	-0.021 ± 0.002	0.097 ± 0.003	0.137 ± 0.006	0.049 ± 0.003	0.018 ± 0.002	-0.194 ± 0.032	1.750 ± 0.075	7.919 ± 0.622	0.099 ± 0.008
EAZY	-0.061 ± 0.004	-0.063 ± 0.003	0.109 ± 0.003	0.153 ± 0.010	0.035 ± 0.005	0.015 ± 0.002	-0.331 ± 0.074	3.982 ± 0.804	11.148 ± 0.868	0.195 ± 0.009
LePhare	0.002 ± 0.004	-0.007 ± 0.003	0.111 ± 0.003	0.171 ± 0.008	0.047 ± 0.002	0.024 ± 0.002	1.177 ± 0.379	42.883 ± 6.603	6.632 ± 0.444	0.087 ± 0.006
PhotoZ	-0.029 ± 0.003	-0.029 ± 0.002	0.097 ± 0.003	0.142 ± 0.006	0.058 ± 0.004	0.017 ± 0.002	-0.268 ± 0.020	1.003 ± 0.030	4.565 ± 0.464	0.080 ± 0.006
TPZ P(z)	0.006 ± 0.003	0.011 ± 0.002	0.078 ± 0.002	0.119 ± 0.006	0.049 ± 0.005	0.018 ± 0.003	0.125 ± 0.030	1.484 ± 0.059	2.607 ± 0.449	0.044 ± 0.006
ArborZ P(z)	-0.008 ± 0.004	0.001 ± 0.004	0.128 ± 0.003	0.153 ± 0.005	0.056 ± 0.005	0.016 ± 0.003	-0.028 ± 0.024	0.962 ± 0.025	4.774 ± 0.449	0.064 ± 0.009
ANNz2 P(z)	-0.002 ± 0.002	-0.002 ± 0.002	0.085 ± 0.002	0.118 ± 0.004	0.051 ± 0.004	0.016 ± 0.003	-0.004 ± 0.024	1.315 ± 0.042	5.010 ± 0.404	0.045 ± 0.004
SkyNet P(z)	-0.001 ± 0.002	0.001 ± 0.002	0.077 ± 0.002	0.104 ± 0.003	0.072 ± 0.006	0.015 ± 0.002	-0.006 ± 0.014	0.829 ± 0.027	4.091 ± 0.404	0.052 ± 0.005
BPZ P(z)	-0.025 ± 0.003	-0.025 ± 0.003	0.101 ± 0.002	0.132 ± 0.006	0.046 ± 0.004	0.014 ± 0.002	-0.224 ± 0.033	1.750 ± 0.080	5.776 ± 0.948	0.098 ± 0.007
ZEBRA P(z)	-0.050 ± 0.005	-0.030 ± 0.002	0.109 ± 0.004	0.177 ± 0.012	0.043 ± 0.006	0.018 ± 0.002	-0.383 ± 0.047	1.906 ± 0.166	6.692 ± 0.550	0.112 ± 0.008

detail here. TPZ and SkyNet are state-of-the-art training-based methods using, respectively, random forests and artificial neural networks to compute photo-zs , and yielding the best performance among all the codes utilized in this analysis. Finally, BPZ is the template-based photo-z code showing best performance in the tests previously shown, and it has been widely used by other galaxy surveys such as CFHTLenS (Heymans et al., 2012; Hildebrandt et al., 2012). All these four codes are public.

A very important issue, which is actually the most important result needed from photo-z studies in order to perform many cosmological analyses, is the estimation of true redshift distributions $N(z)$. In Figure 9.7 we observe how the full redshift distribution reconstructed from the four photo-z codes compares to the spectroscopic distribution. The DESDM code produces one single value for the photo-z of each galaxy in the testing sample while the other three are $P(z)$ codes, so that they return a probability density function (PDF) for each galaxy to be at a given redshift. This is the reason why the $N(z)$ reconstruction looks smoother for TPZ , SkyNet and BPZ , since these are computed from stacking all individual photo-z PDFs. Quantitatively, one can measure how good an $N(z)$ reconstruction is by looking at the $\mathrm{N}_{\mathrm{poisson}}$ and KS metrics in Table 9.2.4 the lower these values are, the better is the agreement between the true $N(z)$ and the photo-z-reconstructed one. As for the advantage of using $P(z)$ codes, one can observe in Table 9.2.4 how the $\mathrm{N}_{\mathrm{poisson}}$ values for TPZ and BPZ are significantly smaller in their $P(z)$ versions than in their single-estimate photo-z versions.

On the other hand, although this full redshift distribution is interesting for photo-z analyses, most of the cosmological studies split the galaxy sample into multiple photo-z bins, therefore there is a need to know the true redshift distribution inside each of those photo-z bins. Figure 9.8 shows the redshift distributions, both spectroscopic and photometric, for six photo-z bins of width 0.2 from $z = 0.1$ to $z = 1.3$, and for the four photo-z codes selected. The limited number of spectroscopic galaxies available makes the distributions shown in the figure somewhat noisy, especially in the last photo-z bin, where a very small number of galaxies is available. The third and fourth bins in photo-z are the ones presenting the narrowest spectroscopic redshift distributions, which agrees with the fact that the photo-z precision is the highest in this redshift range.

In Fig. 9.8, we observe how single-estimate photo-z codes produce a top-hat photo-z distribution for each (photo-z selected) redshift bin. In this case, depicted in the left column of Figure 9.8, the photometric and spectroscopic redshift distributions of each bin are very different and therefore a spectroscopic sample is needed to calibrate the broadening of the redshift bin due to photo-z errors. On the other hand, when using $P(z)$ codes to bin a sample in photometric redshift, one selects a galaxy to be inside a given redshift bin by looking at the position of the median of the PDF (other choices are also possible, e.g. the mode), checking whether it is within the boundaries of the the bin and summing the full PDF of the galaxies inside, including probabilities beyond the bin limits. That makes the photo-z distribution broader than the bin limits and

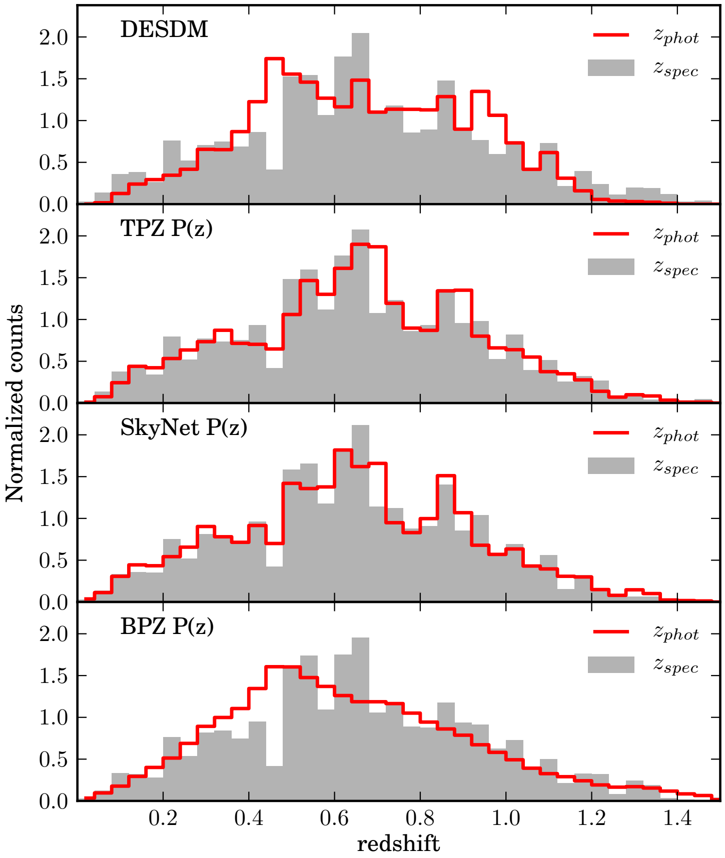


Figure 9.7: Full weighted spectroscopic redshift distribution and its photo-z reconstruction using the four selected codes for Test 1. TPZ , SkyNet and BPZ produce redshift PDFs for each galaxy, thus yielding smoother photo-z distributions.

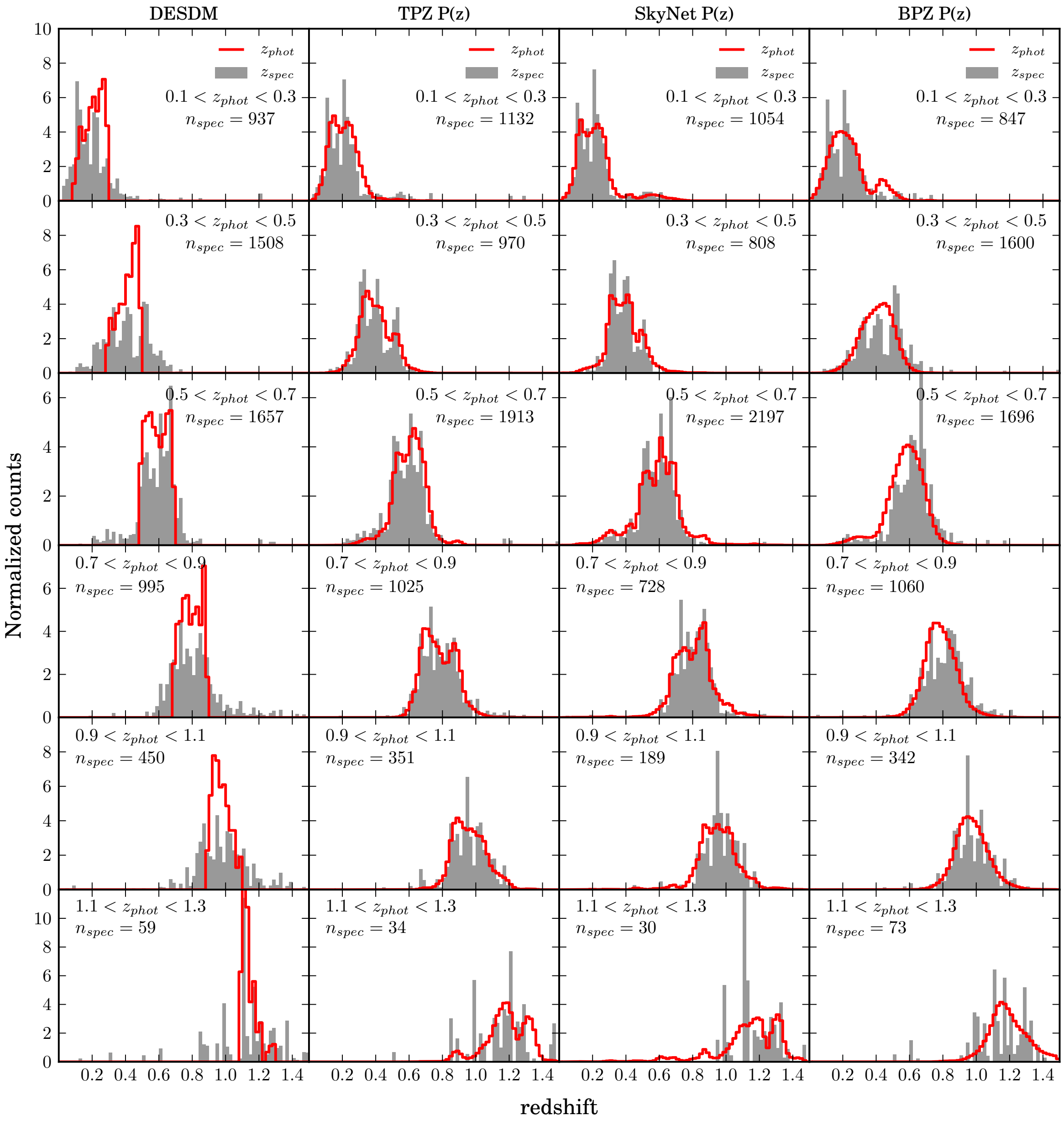


Figure 9.8: Weighted spectroscopic redshift distributions and their photo-z reconstruction using the four selected codes, for photo-z bins of width 0.2. The number of spectroscopic galaxies inside each photo-z bin is shown. The DESDM is a single-estimate photo-z code, while TPZ , SkyNet and BPZ are $P(z)$ codes. This is the reason why the photo-z distributions returned by the latter codes can reconstruct the tails of the spectroscopic distributions beyond the photo-z bins. The photo-z bins are defined using the best estimate z_{phot} for each code, while, for TPZ , SkyNet and BPZ the reconstructed redshift distributions are obtained by stacking the probability density functions for each galaxy.

closer to the spectroscopic redshift distribution of the bin, as can be seen in the three rightmost columns in Figure 9.8. We can see on those panels how the tails of the spectroscopic distributions are well represented by the photo-z distributions. This is an important point in favor of $P(z)$ codes since their ability to reproduce the spectroscopic redshift distribution of a photo-z selected bin by stacking their redshift PDFs makes them less dependent on a spectroscopic calibration sample.

In summary, we have characterized for each code the true redshift distribution inside each photo-z bin. Regarding the performance in such task, the four codes studied in this section show similar spectroscopic redshift distributions for each photo-z bin, but $P(z)$ codes are able to yield a better reconstruction of these distributions by adding up the redshift PDFs for each galaxy which makes them somewhat less reliant in the precise photo-z calibration.

9.2.6 Discussion

The photo-z codes showing the best performance in the analysis are all training-based methods. Among them, there are various codes using Artificial Neural Networks (ANNs) in different ways and configurations (see Section 9.2.3), and the similarities and differences between them go beyond the network architecture. Aside from ANNs, TPZ, which is a state-of-the-art photo-z code using Prediction Trees and Random Forests, performs remarkably well in all the tests in this work. The prediction trees and random forest techniques used by TPZ have the advantage that they have fewer hyper parameters to be chosen compared to neural networks. Neural networks have, amongst others, to choose the amount of hidden layers, the amount of nodes per hidden layer, the learning rate and at least one regularization parameter if present. Random forests used in TPZ have only 2 hyper parameters to choose: the amount of trees used and the size of the subsample set of features used at each split. This leaves out the choice of activation function in neural networks and the choice of the measure of information gain at each split in random forests, maximizing its performance.

Furthermore, training-based photo-z codes show lower bias compared to that of template-based codes, which indicates possible systematic inaccuracies in the template sets. This can be solved by using adaptive recalibration procedures, which adjust the zero-point offsets in each band using the training sample. Such technique has been successfully applied by LePhare in this work, as was also the case in Hildebrandt et al. (2010).

9.3 Angular Power Spectrum

The analysis of the statistical distribution of fluctuations in the Universe is a potent method for constraining theories or components within Cosmology. The 3D power spectrum of galaxies fully describes these variations, which can be modeled by theory, if these fluctuations are given by a Gaussian random field. Even if they are not, non-gaussianity in the early Universe can be constrained from large-scales of the power spectrum (Dalal et al., 2008). In this thesis, we instead use the angular power spectrum (APS) which is a two dimensional projection of its 3D counterpart. A reconstruction of the full power spectrum is also possible given a well determined galaxy distribution (Dodelson et al., 2002; Nicola et al., 2014).

Given the wealth of information encoded in the APS, this has been used to constrain cosmological parameters (e.g., Tegmark et al., 2002; Blake et al., 2007; Thomas et al., 2011; Ho et al., 2012; Hayes et al., 2012; Leistedt et al., 2013). If the BAO signal is resolved, or by using other features, such as the shape of the APS in different redshift shells, is possible to obtain the angular diameter distance and therefore constrain dark energy models (Cooray et al., 2001; Seo et al., 2012). To compute the APS, we use a similar procedure described in Hayes et al. (2012) and Hayes & Brunner (2013) where C_ℓ is calculated using a quadratic estimation method (e.g., Bond et al., 1998) with Karhunen-Loéve (KL) compression (e.g., Vogeley & Szalay, 1996). This technique fits a quadratic function to the shape of the likelihood function for some initial angular power spectrum, finds the C_ℓ that maximize this quadratic and uses these C_ℓ for a new quadratic fit to iteratively converge to the true maximum of the likelihood function.

9.3.1 Galaxy Overdensities

We compute the overdensities of galaxies by pixelating the area of the survey to be studied by using HEALPIX (Górski et al., 2005). Since we are incorporating photo-z PDFs, we compute the galaxy overdensity field as follows: We consider only galaxy PDFs ($P(z)$) that contribute to that pixel in that redshift bin (this could be the whole range) having part of its integrated area inside that bin above some threshold A_z. This is work in progress and a more carefully analysis remains to be done. For now we select A_z to be 0.2, i.e., only galaxies with areas inside the given bin larger than 20% are considered. Within each pixel the overdensity is computed :

$$\delta_i = \frac{\Omega_{survey} \sum_j^{N_{in}} \int_{z_1}^{z_2} P_{ij}(z)dz}{\Omega_i \sum_j^{N_{tot}} \int_{z_1}^{z_2} P_j(z)dz} - 1 \tag{9.17}$$

where i is the pixel number, $P_{ij}(z)$ is the PDF for galaxy j inside pixel i that contributes to the density, N_{in} is the number of galaxies inside the pixel with area inside the redshift bin larger than A_z, z_1 and z_2 are the boundaries of the redshift bin, Ω_i is the area of pixel i and Ω_{survey} of all pixels analyzed ans N_{tot} is the total number of galaxies. As it happens with $N(z)$, this would produce a better overdensity map in comparison to use the mode or the mean of the PDF. We are currently studying how to quantify this fact and the implications of this approach to obtain overdensities and also to incorporate our sparse representation into this framework to speed the calculations. The overdensity field will also depend on the width of the redshift bin, in large bins this difference is less obvious as most galaxies will be considered. In narrower bins, however, on average we will have the PDFs better resembles the spectroscopic distribution, being less affected by outliers, but more affected by projection effects when considering these fractional values. We will continue studying these effects by using simulations in order to understand the systematics and possible biases introduced by each method.

9.3.2 Theoretical model and cosmological parameters

After the overdensities fields are computed on a pixelated and previously masked maps, we obtain the APS using a quadratic estimation. This is a very computationally expensive task as the time scales as n_{pix}^3 and the required memory scales as n_{pix}^2 (Hayes et al., 2012), where n_{pix} is the number of pixels. We must, therefore, be careful when using large areas or when increasing the resolution of the pixelization. Since we have the overdensities for a subarea of the whole sky, we can't compute the APS C_ℓ at all multipoles; instead we are limited to compute them in multipole bands where the bandwidth is approximately limited to $\Delta_\ell \sim \frac{180}{\phi}$, where ϕ is the smallest dimension of the survey geometry. The pixel resolution will also limits the extend of the multipoles as all the information inside a pixel is lost. These two limits on the computation of C_ℓ are the constrains imposed when fitting cosmological model to the observed data.

In order to model the APS, we need to compute the 3D power spectrum $P(k)$ and project it down to two dimensions using the galaxy distribution $N(z)$ which also plays an important role in the modeling. To compute a linear $P(k)$ we use the prescriptions of Eisenstein & Hu (1998), and for nonlinear $P(k)$ we use CAMB (Lewis et al., 2000) and HALOFIT (Smith et al., 2003) at the desired redshift and scale. Here we use the Limber approximation (Limber, 1953; Crocce et al., 2011), which has been shown to be a good approximation to large scales where $\ell > 30$ (Blake et al., 2007). Without taking into account redshift space distortions at the moment, this 2D projection of the power spectrum can be written as:

$$C_\ell = \frac{\ell(\ell+1)}{2\pi} b^2 \int dz\, n^2(z) \frac{H(z)}{r^2(z)} P\left(\frac{\ell+1/2}{r(z)}, z\right) \tag{9.18}$$

where we assume a scale independent bias b for a given redshift bin, $r(z)$ is the comoving distance, $H(z)$ is the Hubble parameter, $P\left(\frac{\ell+1/2}{r(z)}, z\right)$ is the 3D power spectrum at a given redshift z. In the linear case we can simply use $P(k,z) = D^2(z)P(k,0)$ where $D(z)$ is the growth function of density fluctuations. In equation 9.18, $n(z)$ is the galaxy distribution or $N(z)$ normalized within the redshift range. This is the most important function, as it acts as a window function for the projection and accounts for the mass inside the redshift bin. It is very important to determine this accurately and as we discussed above and show in Figure 9.1, using photo-z PDFs is essential in this process. We are currently explore alternatives to implement our sparse representation in this framework where $N(z)$ can be quickly computed and using a parametrization for its error we can add uncertainty to the mass distribution which can be later be marginalized over for better fits.

In order to fit the observed angular power spectrum and estimate different cosmological parameters, we use a Markov Chain Morte Carlo approach which is much faster than the standard χ^2 minimization to explore the parameter space. To sample from the parameter distribution to obtain the set that maximizes the likelihood (or minimizes the χ^2) we use the following χ^2 (Tegmark et al., 2002):

$$\chi^2(a_p) = \sum_{bb'} (\ln \mathcal{C}_b - \ln \mathcal{C}_b^T)\mathcal{C}_b F_{bb'} \mathcal{C}_{b'} (\ln \mathcal{C}_{b'} - \ln \mathcal{C}_{b'}^T) \tag{9.19}$$

where a_p are set of the cosmological parameters of interest, $F_{bb'}$ is the Fisher matrix, and $\mathcal{C}_b$ is the angular power spectrum within a bandpower. Using this definition, we then compute the likelihood function and by using a Metropolis-Hasting sampler we estimate the parameter distributions by marginalizing over the nuisance parameters which account for uncertainties and systematics in the model fitting. The main goal for this machinery is to be applied on data from deep and large photometric surveys like the Dark Energy Survey or from the LSST in the future as these surveys will cover a large area of the sky at deeper magnitudes than current surveys like SDSS. This will allow to have enough statistics to compute the APS on several redshift bins which allows to put constrains not only for Dark Matter but also for Dark Energy.

9.3.3 Example application on simulated data

The DES simulation working group have carried out a series of simulations by performing dark matter particles simulations and adding galaxies to the halos and using the appropriate masking to match the DES footprint. To try out our methods on calculating the APS and to estimate cosmological parameters, we used one of the simulated catalogs where they provide observables as well as the true redshifts of galaxies. We then run our TPZ code described in Chapter 3 on the simulated data to obtain photo-z PDF for all the galaxies in a selected patch of the sky. The area of the catalog we used consists in 833 squares degrees with $\sim$ 43 million

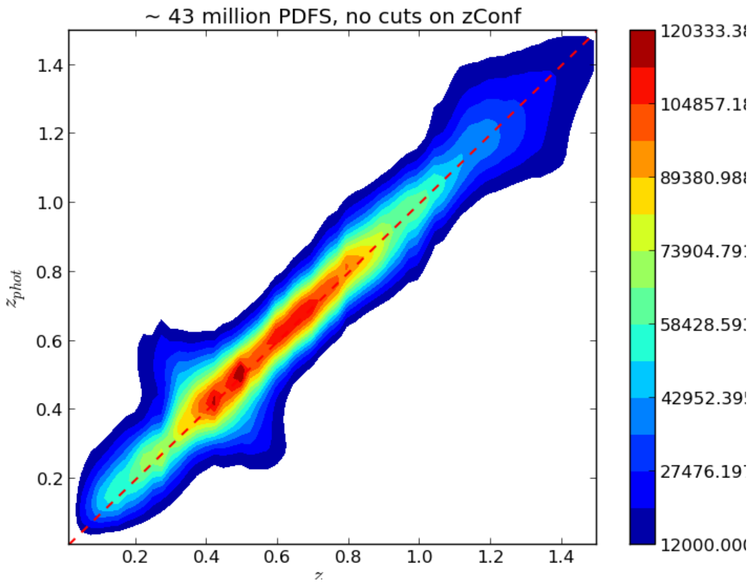


Figure 9.9: TPZ photo-z vs spectroscopic redshifts using PDFs, no quality flags cuts applied for the $\sim$ 43 million simulated galaxies.

galaxies between redshift 0 and 1.5. We use the observable quantities only (except for the true redshift for comparison), i.e., the RA,DEC and magnitudes corrected by lensing effects and photometric errors models. We randomly select 100,000 galaxies for training TPZ , which correspond to 0.2 % of all galaxies, and use the remaining 43 million galaxies for testing.

We compute photo-z PDFs for the full sample with a resolution of 0.04, and the main results are shown in Figure 9.9 which shows the 43 million PDFs combined to produce a photo-z vs spec-z to see how TPZ provides a very good solution to these data. The metrics also are within the DES original requirements discussed briefly in §9.2 and in more detailed in Sánchez et al. (2014). The feature or degeneracy at low z is a small issue with the data that is also showing for other codes, which is related to the filter curves used during the generation of the synthetic magnitudes. Figure 9.10 shows the computed $N(z)$ for these galaxies which, as discussed, is better recovered when using the full PDF as oppose of the other single estimates, here the difference is much stronger than previously discussed due to the large number of objects in the catalog. This plays an important role for the APS analysis as discussed in before, and the differences in the shape will also strongly affect the overdensity maps as well as the projection of the 3D power spectrum to the angular version.

We also calculated $N(z)$ for different redshift shells which will be used in the calculation of the APS. These distributions are shown in the right panel of Figure 9.11 and were calculated with photo-z PDF that had more than 20% of their area inside the shell. As we can observed, the shapes of the distributions are similar to a Gaussian distribution and usually a standard approach is to use the photo-z estimation and assume a

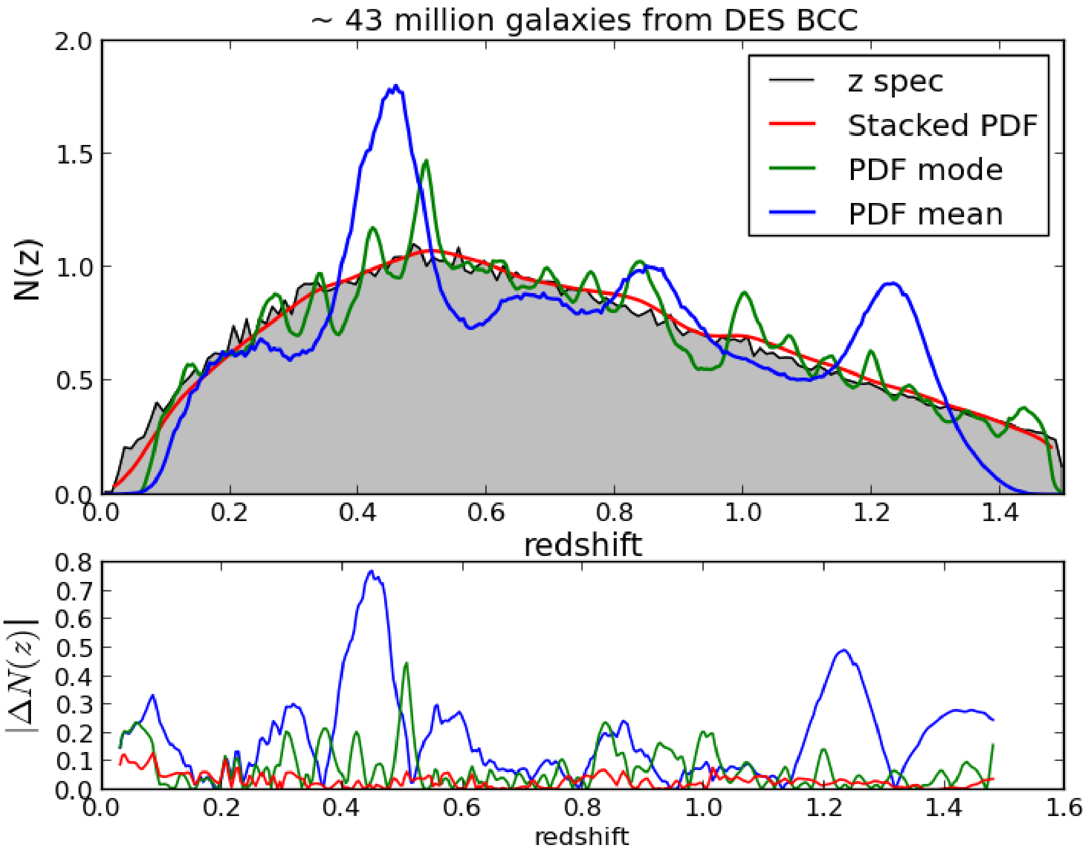


Figure 9.10: Normalize $N(z)$ reconstruction for all the galaxies ion our sample using the mean (green solid line), the mean (blue solid line) and the stacked PDF (red line). The tru $N(z)$ for the same galaxies is shown in gray. We observe that the stacking of the PDFs agrees very well with the underlying distribution.

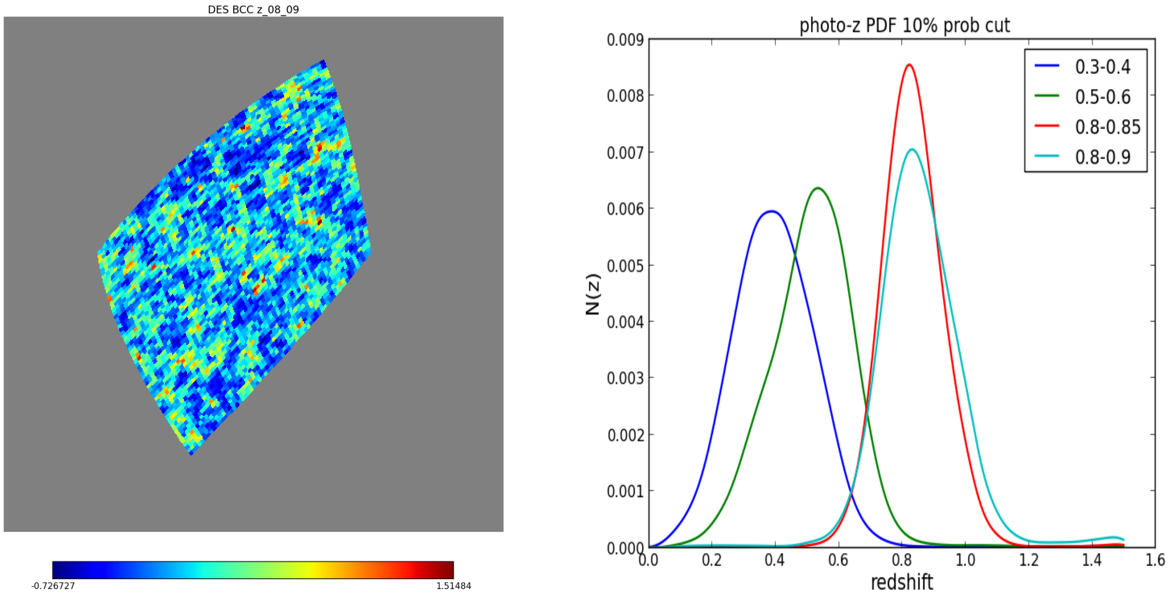


Figure 9.11: (*Left*): Overdensity maps using equation 9.17 for the 833 sq. degrees used in the analysis for 4096 pixels. (*Right*): $N(z)$ for 4 redshift bins using the BCC simulation data and photo-z PDF produced by TPZ . As discussed in the text only galaxies PDF with more than 20% are considered on the calculation.

Gaussian distribution for a given estimated sigma. Using a full PDF overcomes this problem and is a more accurate way to describe the photo-z and its error as well as the distribution. To compute the overdensity fields we pixelated the area being analyzed by using HEALPIX. In total we have 4096 pixels, each pixel has the same area, which is approximately 0.2 square degrees for the resolution used. The reason we don't use a larger area or a higher resolution is due to the computational cost that would result in computing the APS. The overdensity map for one of the redshift shells (0.8 – 0.9) is shown in the left panel of Figure 9.11 for reference.

Preliminary results for the APS are presented in Figure 9.12. Due to the area geometry we used a bandwidth of $\Delta\ell = 11$. The left panel shows the APS calculated for almost the whole redshift distribution by using the spectroscopic redshift (blue circles), the mean of the PDF (red triangles), and the full photo-z PDF (green squares). These three methods do not show much difference since the redshift bin is wide enough to account for all projection effects, in which basically all the galaxies considered are projected to the same plane and they all have their PDF within the redshift bin. In this same panel we show the best fitting model where the data in gray was not included due to the pixel window function. This window function removes power from the APS which becomes significantly at around half of the computed multipoles. We can model the loss of power and this can be corrected to fit higher multipoles. On the other hand, the maximum multipole used here is $\ell \sim 300$, after that there are other effects that make the APS less reliable at higher ℓ

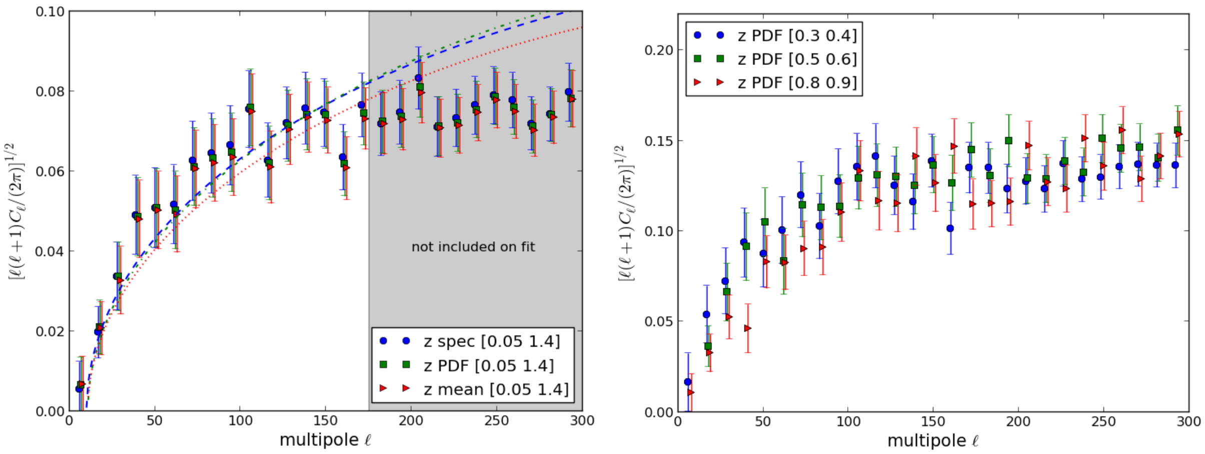


Figure 9.12: (*Left*): Computed APS for the redshift range [0.05-1.4] (almost all galaxies) using the spectroscopic redshift (blue), the mean of the PDF (red) and the full PDF (green). The lines shows the fitting to the data using the same colors where the shaded area was not included in the fit due to the pixel window function limitation and the ”red” leak problem. (*Right*): The APS using full photo-z PDFs for three redshift shell as indicated, no fit has been performed on this data so far.

(red leak effect; Tegmark et al., 2002).

We see a big difference between the fitted models (especially between the mean of the PDF and the other two) which can be explained in terms of the differences seen in the $N(z)$ from Figure 9.10, where $N(z)$ computed using the mean of the PDF shows a large disagreement between the ones computed using stacked PDFs and true redshifts. This re-validates our discussion about including photo-z PDF on cosmological analysis. Small differences in the model due to these effects can lead to a wrong estimation of cosmological parameters, therefore the fitting process must be done very carefully considering all possible sources of uncertainties in which the our use of PDF can greatly contribute.

The right panel of Figure 9.12 shows the APS computed by using full PDF on three different redshift shells, namely, 0.3 – 0.4 (blue), 0.5 – 0.6 (green) and 0.8 – 0.9 (blue), the $N(z)$ for these bins are shown in the right panel of Figure 9.11. No fitting to these parameters has been performed so far, however this is an illustration and the APS in different bins can be used to compute the angular diameter distance and constrains dark energy models.

Figure 9.13 shows preliminary results using an MCMC approach to fit cosmological parameters to the APS shown in Figure 16 (top), as mentioned above we use CAMB and HALOFIT to generate $P(k)$ for different models which are projected to the C_ℓ using the galaxy distribution $N(z)$. These plots show the cosmological parameters histogram and the 2D contour plots when marginalizing over the rest. For illustration and testing purposes only we choose to fit Ω_m, Ω_b , $bias$, n_s (spectral index) and the Hubble parameter. The cosmology used in this simulation is unknown, but it is known that it corresponds to a ΛCDM with parameters close

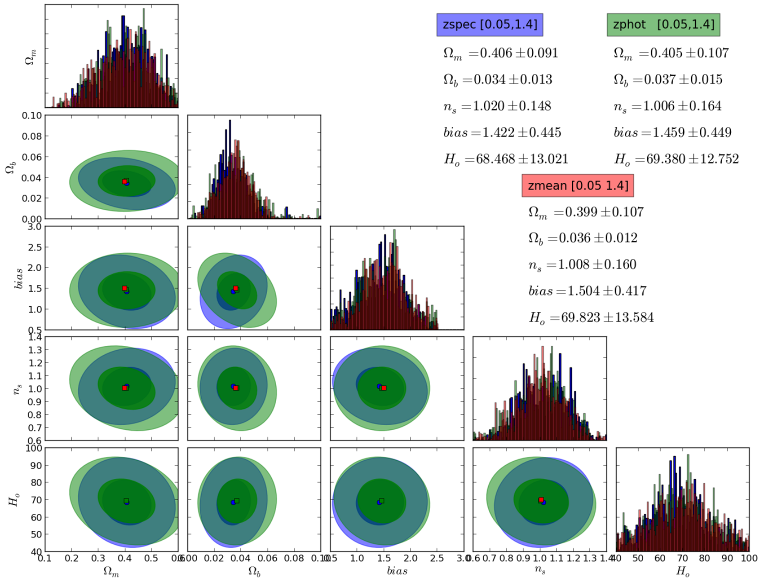


Figure 9.13: Preliminary results from cosmological parameters estimation using MCMC mehods on the APS.

to those computed by WMAP7 (Komatsu et al., 2011; Larson et al., 2011). We still need to select the best combination for the fitting (like $\Omega_m h^2$), determine the most useful nuisance parameters, the burn-in phase, and apply some convergence tests to the process. We also want to study the difference between the linear and nonlinear models for $P(k)$, and the possibility to observe the BAO signal (included in the simulations) by for example increasing the survey area to reduce the bandwidth.

9.4 Summary

By calculating the normalized distribution of galaxies as a function of redshift, we were able to demonstrate the advantages of using a full photo-z PDF as opposed to using one single estimator of the PDF or any other point metric. Specifically, by simply stacking each individual PDF, we recover the underlying galaxy redshift distribution to a much higher precision than by simplifying using the mean of each individual photo-z PDF. This is true for all the samples we have used in this thesis, from real observations as with the data introduced in Chapter 2 to simulated data shown here.

In this chapter, we also demonstrated that, as a simple cosmological application of our photo-z PDF reconstruction using sparse representation, we could accurately recover the underlying $N(z)$ distribution to great accuracy. In particular, we recovered the $N(z)$ of our CFHTLenS test sample to an accuracy of 99.87% by using only ten points per photo-z PDF. Given their compact nature and the fact that they are predetermined, we showed that we could obtain the sample $N(z)$ by integrating the bases over the sample redshift range and later multiplying by the basis coefficients, which can also be prefactored, thereby significantly reducing the number of required integrations. This same principle can be applied to other linear combinations of photo-z PDFs or to more complex analyses if they can be expressed in terms of the underlying bases.

Also in this chapter, we present the photometric redshift performance of the DES survey in the SV period. Most of the relevant photo-z codes have been used in the analysis. Since spectroscopic galaxy samples are generally shallower, a weighting technique is used to make the calibration sample of galaxies to mimic the DES full sample in magnitude and color space in order to properly estimate the photo-z performance in the DES galaxy sample. Calibration and testing samples have been produced with two different depths: Main is the default depth in the DES survey, and Deep corresponds to the depth in SNe fields. Test 1, which uses the Main training and testing samples, represents the default case for photo-z estimation in DES. Results from 13 different codes are analyzed in this case, showing fluctuations in photo-z performance but a general agreement in codes of the same type (machine-learning or template fitting algorithms). In particular, most of the codes analyzed comfortably meet the DES science requirements in terms of photo-z precision and several also meet the requirements on the fractions of outliers.

Generally speaking, training-based photo-z codes show the best performance in the tests in terms of photo-z precision and bias. Among them, TPZ , using Prediction Trees and Random Forest seem to yield the most accurate results, achieving a core photometric redshift resolution below $\sigma_{68} = 0.08$ and is today one of the best, if not the best one, among the training based codes available in the literature. The photo-z analyses carried out as part of this chapter using these early stage DES data will serve as a benchmark for future DES data releases, and as the survey area grows during the observation period, more spectroscopic data will be available allowing a better calibration and a better sampling for training algorithms. Therefore these promising early results will do nothing but improve in the near future, which will allow putting tighter constrains on several cosmological parameters.

Regarding to the latter point and as an ongoing application, we have develop a complete framework to compute the angular power spectrum from generating overdensity maps to fit cosmological parameters by computing the Angular Power Spectrum that uses the full photo-z PDF which allows, at the end, a better estimation of the cosmological fitting. We showed, by using simulated data, that this is ready to be apply on real observation coming from future surveys.

Chapter 10

Conclusions and Future work

A primary output of this thesis has been the development of a fully parallelized software to compute and represent the photo-z PDFs, which is publicly available at http://lcdm.astro.illinois.edu/static/code/mlz/MLZ-1.1/doc/html/index.html. We plan to continue making improvements and to develop this software to provide the most accurate photo-z PDFs possible, with additional tools to simplify their use in cosmological applications.

As part of this framework, we have developed two new, state-of-the-art machine learning photo-z estimation codes; the first is `TPZ` , which is a supervised technique that uses prediction trees and a random forest. Hundreds of different trees are built recursively by asking a series of questions about the properties of the data guided by the spectroscopic redshift information, the prediction results from these trees is then combined to provide not only photo-z PDFs but also ancillary information that can be used to calibrate spectroscopic observations, to rank variables regarding their importance, and to provide targeting areas to improve the photo-z solution. `TPZ` has been extensively applied on several datasets including SDSS, DEEP2, PHAT, CFHTLens and DES among other with excellent results. It is currently one of the most accurate photo-z codes available. The development of `TPZ` was published in Carrasco Kind & Brunner (2013a).

The second method is an unsupervised machine learning algorithm that uses Self-Organizing maps whose goal is to project the multidimensional color/magnitude space into a two dimensional space where the topology is closely conserved. Following the `TPZ` framework, we introduced the concept of a *random atlas* where multiple SOMs are generated and their results are aggregated to produce a photo-z PDF. We explored different 2D topologies and configurations, ultimately reaching similar performance to `TPZ` but in a quasi-independent manner. The advantage of this unsupervised technique is that we use the redshift information only at the final step. This reduces possible bias in the data, and also allows a different characterization of the redshift distribution or any other variable unused under the same deprojection, as we have subsequently shown when we combine multiple techniques by comparing their performances via a SOM map. The development of this method and its applications was published in Carrasco Kind & Brunner (2014a).

During the development of this thesis, we realized that despite there are several template fitting and machine learning methods available to compute photo-z PDFs, most previous studies ignore the fact that the estimation from independent codes can provide extra information about the performance of individual methods. For example, not only can you use machine learning to improve template methods, but you can use the independent nature of both approaches to our advantage. To address the lack of analysis in this aspect, we explored and introduced new Bayesian methods to efficiently combine PDFs from multiple techniques. We used our own methods that are different in their formulation, and we modified and adopted an SED fitting approach for the benefit of this work. We demonstrated that by using a sophisticated framework it was possible to extract even more information than is available to a single method, which translated to a better photo-z PDF estimation.

We employ techniques of cross validation, in which each algorithm is evaluated carefully on the available training data, to identify the strengths and weaknesses of each individual method. By using a SOM, we were able to best identify those areas in the multidimensional color space in which each method work best, thus providing important insights not only into the methods but also within the data itself. We concluded that not only do we need to fully understand a photo-z code and its performance, but also the structure of the data. Both aspects provide powerful information when combining multiple techniques and they must be considered together.

We explored these techniques under different situations, and we found that it is often possible to get improvements by efficiently using the available information. Within the same framework, we used information from these multiple techniques and a Bayesian Classification scheme called a Naive Bayesian Classifier to build a better method to identify catastrophic outliers withing the galaxy sample than standard approaches. This reinforces the idea that multiple methods contain different properties about the multidimensional space that in combination provides a powerful method to exhaust information contained within the dataset. Our early work on combination models was published in a conference proceeding in Carrasco Kind & Brunner (2013b). The detailed study of different Bayesian combination models and the identification of outliers was published in Carrasco Kind & Brunner (2014c).

While working on the combination model, we faced the problem of dealing with multiple photo-z PDFs from multiple algorithms and realized that this will quickly become a problem when the number of galaxies and the number of methods increase significantly. This will clearly happens with future photometric surveys, and to handle these PDF and to store them in a database will be even more challenging. We therefore focused our efforts on new techniques to represent and reconstruct these photo-z PDFs by minimizing the number of points required. We demonstrated that by using a sparse representation, we can write each PDF as a finite sum of these pre-defined bases (using Gaussian and Voigt profiles) where the number of bases used determines the reconstruction accuracy. We showed that this approach outperforms current techniques like using multi-gaussian fitting, and allows us, for a typical dataset, to represent a full PDF with high resolution by using a very compressed format of less than 20 4-bytes integers per galaxy. This compressed format retains an unprecedented 99.9% of accuracy while at the same time facilitating its storage and reconstruction. We also developed a new mathematical framework where we can write different cosmological measurements, like the galaxy distribution, in this basis system that also allows us to reduce the computational cost over the standard approach. The work on this new approach to represent and to reconstruct PDFs in a very compressed format was published in Carrasco Kind & Brunner (2014b).

We demonstrated by using simple examples the applicability of our methods in different contexts. We showed that by using the full PDF we obtain a better reconstruction of the distribution of galaxies,,$N(z)$, and also that this can be computed by using our sparse basis framework. We applied our TPZ code on early Dark Energy Survey data with outstanding results, we carried out a photo-z code comparison to asses the quality of the data from DES, and we found that most of today's codes meet the established scientific requirements when compared to similar surveys. As a results of this photometric redshift study and photo-z code comparison, we found that TPZ is among the best codes available. Our analysis and findings on this DES data was submitted to a refereed journal and published in the public arXiv repository in Sánchez et al. (2014). We finalized our example applications by showing how to compute the angular power spectrum of the galaxy distribution for photometric surveys, and how to incorporate the use of photo-z PDFs in this analysis to constrain cosmological models. Our preliminary results using simulated data show promising results and indicate that our full machinery will be ready when more data from DES or data from LSST become available.

One of the goals of this thesis was to provide new tools to compute, combine, and use accurate and robust photo-z PDFs for cosmological applications, which will enable better and tighter constrains on models of galaxy formation and evolution. We believe our work on probabilistic photometric redshifts has contributed enormously to this field, and will enable important scientific discoveries in the understanding of our Universe as we enter the era of Petascale astronomy.

Future work

Despite that our work has covered several old and new areas regarding probabilistic photometric redshifts, we still have new ideas that promote ongoing work. In the future we plan to continue investigating new alternatives to improve upon our photo-z approach. In this regard, we have already started to study the applicability of machine learning algorithms to predict photo-z at the image pixel level instead of by using pre-built catalogs. This means that we are working directly on calibrated images, which, after being astrometrically aligned, are used to identify objects and the photo-z analysis (i.e., , training and validating) is performed on every pixel. This is computational challenging, but we believe that by using this new approach we can improve photo-z estimation, especially for faint and blended objects and will avoid systematics introduced by using different apertures and other related issues in the photometric measurements made by the surveys from the images. Our preliminary results show that this new approach has great potential and could open an entirely new line of investigation.

Both of the developed machine learning methods in this thesis also have the capability to be used for classification problems, like when we need to separate or classify objects into pre-defined classes. We have successfully applied both of these algorithms to the star-galaxy classification problem which arises from the fact that faint galaxies are very hard to separate from faint stars as their morphological profiles are very close. This will become a big issue with the development of deep photometric surveys, and new methodologies to separate both populations will be very important when constraining cosmological models, as any possible contamination of stars within the galaxy catalog will bias and degrade measurements. While we did not discuss this application within this thesis, this is an ongoing investigation where we use our developed codes as a starting point to tackle these important issues.

We are also continuing to work on the cosmological applications of photo-z PDFs. We are currently optimizing our APS code to be applicable to large numbers of pixels, and we are also planning to incorporate our sparse representation framework into this computation. By using a fast method to compute $N(z)$ with the fact that we can use multiple photo-z PDF techniques to do so, we can incorporate this into the cosmological parameter estimation in a way that has never been done, where we can parametrize and later marginalize over the errors associated in the computation of $N(z)$, which usually is considered static and a known variable even if its computation might not be very accurate. In this regard we also plan to explore cross correlations techniques as introduced by Newman (2008) in combination with our approach to improve the computations of $N(z)$ when representative data is not available.

Finally, and related to our last point, we will continue our work on spectroscopic calibrations by studying different techniques to maximize the available telescope time when observing new spectroscopic galaxies which will form a training set. As mentioned in this thesis, all of our methods rely on the availability of representative and high quality spectroscopic training data. We have worked extensively on new techniques to retrieve all possible information from the available training data. In the future, we will focus our research into methods to obtain the best possible training data.

Bibliography

Abazajian K. N. et al., 2009, ApJS, 182, 543 (Cited on page 10.)

Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891 (Cited on page 3.)

Abrahamse A., Knox L., Schmidt S., Thorman P., Tyson J. A., Zhan H., 2011, ApJ, 734, 36 (Cited on pages 6 and 138.)

Abramowitz M., Stegun I., 1972, Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, Applied mathematics series. Dover Publications (Cited on page 126.)

Adelman-McCarthy J. K. et al., 2008, ApJS, 175, 297 (Cited on page 24.)

Ahn C. P. et al., 2013, ArXiv e-prints (Cited on pages 11 and 14.)

Aihara H. et al., 2011, ApJS, 193, 29 (Cited on page 2.)

Arnouts S. et al., 2002, MNRAS, 329, 355 (Cited on pages 3 and 149.)

Assef R. J. et al., 2010, ApJ, 713, 970 (Cited on page 3.)

Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12 (Cited on page 4.)

Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., Llorà X., 2007, ApJ, 663, 774 (Cited on page 4.)

Banerji M. et al., 2014, ArXiv e-prints (Cited on page 15.)

Baum W. A., 1962, in IAU Symposium, Vol. 15, Problems of Extra-Galactic Research, McVittie G. C., ed., p. 390 (Cited on pages 3 and 74.)

Bender R. et al., 2001, in Deep Fields, Cristiani S., Renzini A., Williams R. E., eds., p. 96 (Cited on page 149.)

Benítez N., 2000, ApJ, 536, 571 (Cited on pages 3, 74, 75, 79, 80, 85, and 149.)

Bernstein G., Jain B., 2004, ApJ, 600, 17 (Cited on page 2.)

Blake C., Collister A., Bridle S., Lahav O., 2007, MNRAS, 374, 1527 (Cited on pages 138, 158, and 159.)

Blake C. et al., 2011, MNRAS, 418, 1707 (Cited on page 2.)

Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476 (Cited on page 3.)

Bond J. R., Jaffe A. H., Knox L., 1998, PhRvD, 57, 2117 (Cited on page 158.)

Bonfield D. G., Sun Y., Davey N., Jarvis M. J., Abdalla F. B., Banerji M., Adams R. G., 2010, MNRAS, 405, 987 (Cited on page 4.)

Bonnett C., 2013, ArXiv e-prints (Cited on pages 4 and 149.)

Bordoloi R., Lilly S. J., Amara A., 2010, MNRAS, 406, 881 (Cited on pages 6 and 138.)

Bovy J. et al., 2011, ApJ, 729, 141 (Cited on page 123.)

Bovy J. et al., 2012, ApJ, 749, 41 (Cited on page 123.)

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503 (Cited on pages 3 and 149.)

Breiman L., 1996, Machine Learning, 24, 123, 10.1007/BF00058655 (Cited on pages 21 and 22.)

Breiman L., 2001, Machine Learning, 45, 5 (Cited on pages 7, 21, 37, 81, and 120.)

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A. (Cited on pages 7, 16, and 17.)

Brescia M., Cavuoti S., Longo G., De Stefano V., 2014, ArXiv e-prints (Cited on page 4.)

Brett D. R., West R. G., Wheatley P. J., 2004, MNRAS, 353, 369 (Cited on page 5.)

Brunner R. J., Connolly A. J., Szalay A. S., Bershady M. A., 1997, ApJL, 482, L21 (Cited on page 4.)

Bruzual G., Charlot S., 2003, MNRAS, 344, 1000 (Cited on page 3.)

Bundy K., Ellis R. S., Conselice C. J., 2005, ApJ, 625, 621 (Cited on page 12.)

Capak P. et al., 2004, AJ, 127, 180 (Cited on page 12.)

Carliles S., Budavári T., Heinis S., Priebe C., Szalay A., 2008, in Astronomical Society of the Pacific Conference Series, Vol. 394, Astronomical Data Analysis Software and Systems XVII, Argyle R. W., Bunclark P. S., Lewis J. R., eds., p. 521 (Cited on page 24.)

Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511 (Cited on pages 4 and 24.)

Carnero A., Sánchez E., Crocce M., Cabré A., Gaztañaga E., 2012, MNRAS, 419, 1689 (Cited on page 6.)

Carrasco Kind M., Brunner R. J., 2013a, MNRAS, 432, 1483 (Cited on pages 4, 16, 64, 81, 116, 120, 122, 149, and 167.)

Carrasco Kind M., Brunner R. J., 2013b, in Astronomical Society of the Pacific Conference Series, Vol. 475, Astronomical Society of the Pacific Conference Series, Friedel D. N., ed., p. 69 (Cited on pages 7, 71, 80, 84, and 168.)

Carrasco Kind M., Brunner R. J., 2013c, TPZ: Trees for Photo-Z. Astrophysics Source Code Library, record ascl:1304.011 (Cited on page 4.)

Carrasco Kind M., Brunner R. J., 2014a, MNRAS, 438, 3409 (Cited on pages 37, 120, and 167.)

Carrasco Kind M., Brunner R. J., 2014b, MNRAS, 441, 3550 (Cited on pages 131 and 169.)

Carrasco Kind M., Brunner R. J., 2014c, MNRAS, 442, 3380 (Cited on pages 5, 71, 149, 152, and 168.)

Caruana R., Karampatziakis N., Yessenalina A., 2008, in Proceedings of the 25th international conference on Machine learning, ICML '08, ACM, New York, NY, USA, pp. 96–103 (Cited on pages 16, 21, 57, and 80.)

Cavuoti S., Brescia M., Longo G., Mercurio A., 2012, ArXiv e-prints (Cited on page 4.)

Chisari N. E., Mandelbaum R., Strauss M. A., Huff E., Bahcall N., 2014, ArXiv e-prints (Cited on page 138.)

Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, AJ, 132, 926 (Cited on pages 3 and 149.)

Coe D., Moustakas L. A., 2009, ApJ, 706, 45 (Cited on page 2.)

Coil A. L., 2013, The Large-Scale Structure of the Universe, Oswalt T. D., Keel W. C., eds., p. 387 (Cited on page 1.)

Coleman G. D., Wu C.-C., Weedman D. W., 1980, ApJS, 43, 393 (Cited on pages 3, 76, and 78.)

Collister A. A., Lahav O., 2004, PASP, 116, 345 (Cited on pages 4, 5, 24, and 149.)

Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655 (Cited on page 4.)

Cooper M. C. et al., 2012, MNRAS, 425, 2116 (Cited on page 15.)

Cooray A., Hu W., Huterer D., Joffre M., 2001, ApJL, 557, L7 (Cited on page 158.)

Cowie L. L., Barger A. J., Hu E. M., Capak P., Songaila A., 2004, AJ, 127, 3137 (Cited on page 12.)

Crocce M., Cabré A., Gaztañaga E., 2011, MNRAS, 414, 329 (Cited on page 159.)

Csabai I. et al., 2003, AJ, 125, 580 (Cited on page 3.)

Cunha C., Huterer D., Doré O., 2010, PhRvD, 82, 023004 (Cited on page 1.)

Cunha C. E., Huterer D., Busha M. T., Wechsler R. H., 2012a, MNRAS, 423, 909 (Cited on pages 5 and 6.)

Cunha C. E., Huterer D., Lin H., Busha M. T., Wechsler R. H., 2012b, ArXiv e-prints (Cited on pages 5 and 6.)

Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, MNRAS, 396, 2379 (Cited on pages 137 and 145.)

Dahlen T. et al., 2013, ApJ, 775, 93 (Cited on pages 7, 84, 86, 90, and 91.)

Dalal N., Doré O., Huterer D., Shirokov A., 2008, PhRvD, 77, 123514 (Cited on page 158.)

Davis M. et al., 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4834, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Guhathakurta P., ed., pp. 161–172 (Cited on page 13.)

Davis M. et al., 2007, ApJL, 660, L1 (Cited on page 14.)

Dawson K. S. et al., 2013, AJ, 145, 10 (Cited on pages 2, 11, and 14.)

de Putter R. et al., 2012, ApJ, 761, 12 (Cited on page 1.)

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159 (Cited on page 87.)

Dodelson S. et al., 2002, ApJ, 572, 140 (Cited on page 158.)

Domingos P., Pazzani M., 1997, Machine Learning, 29, 103 (Cited on page 112.)

Drinkwater M. J. et al., 2010, MNRAS, 401, 1429 (Cited on page 2.)

Eisenstein D. J., Hu W., 1998, ApJ, 496, 605 (Cited on page 159.)

Erben T. et al., 2013, MNRAS, 433, 2545 (Cited on page 14.)

Faber S. M. et al., 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4841, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Iye M., Moorwood A. F. M., eds., pp. 1657–1669 (Cited on page 13.)

Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15 (Cited on page 90.)

Feldmann R. et al., 2006, MNRAS, 372, 565 (Cited on pages 3 and 149.)

Flaugher B., 2005, International Journal of Modern Physics A, 20, 3121 (Cited on page 14.)

Frank E., Trigg L., Holmes G., Witten I., 2000, Machine Learning, 41, 5 (Cited on page 112.)

Freeman P. E., Newman J. A., Lee A. B., Richards J. W., Schafer C. M., 2009, MNRAS, 398, 2012 (Cited on page 4.)

Fustes D., Manteiga M., Dafonte C., Arcay B., Ulla A., Smith K., Borrachero R., Sordo R., 2013, ArXiv e-prints: 1309.2418 (Cited on page 5.)

Garilli B. et al., 2014, A&A, 562, A23 (Cited on page 14.)

Garilli B. et al., 2008, A&A, 486, 683 (Cited on pages 14 and 15.)

Geach J. E., 2012, MNRAS, 419, 2633 (Cited on pages 5 and 7.)

Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823 (Cited on pages 4, 24, 26, 31, and 149.)

Giavalisco M. et al., 2004, ApJL, 600, L93 (Cited on page 12.)

Gorecki A., Abate A., Ansari R., Barrau A., Baumont S., Moniez M., Ricol J.-S., 2014, A&A, 561, A128 (Cited on pages 6 and 113.)

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759 (Cited on pages 57 and 158.)

Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741 (Cited on page 149.)

Gregory P. C., Loredo T. J., 1992, ApJ, 398, 146 (Cited on page 87.)

Gwyn S. D. J., 2012, AJ, 143, 38 (Cited on pages 13 and 14.)

Hayes B., Brunner R., 2013, MNRAS, 428, 3487 (Cited on page 158.)

Hayes B., Brunner R., Ross A., 2012, MNRAS, 421, 2043 (Cited on pages 138, 158, and 159.)

Heymans C. et al., 2012, MNRAS, 427, 146 (Cited on pages 14 and 154.)

Hildebrandt H. et al., 2010, A&A, 523, A31 (Cited on pages 3, 12, 13, 38, 39, and 157.)

Hildebrandt H. et al., 2012, MNRAS, 421, 2355 (Cited on pages 14 and 154.)

Ho S. et al., 2012, ApJ, 761, 14 (Cited on pages 2 and 158.)

Högbom J. A., 1974, A&AS, 15, 417 (Cited on page 124.)

Hogg D. W. et al., 1998, AJ, 115, 1418 (Cited on page 3.)

Ilbert O. et al., 2006, A&A, 457, 841 (Cited on pages 3 and 149.)

in der Au A., Meusinger H., Schalldach P. F., Newholm M., 2012, A&A, 547, A115 (Cited on page 5.)

Jee M. J., Tyson J. A., Schneider M. D., Wittman D., Schmidt S., Hilbert S., 2013, ApJ, 765, 74 (Cited on pages 6 and 138.)

Kinney A. L., Calzetti D., Bohlin R. C., McQuade K., Storchi-Bergmann T., Schmitt H. R., 1996, ApJ, 467, 38 (Cited on pages 3 and 76.)

Kohonen T., 1982, Biological Cybernetics, 43, 59 (Cited on page 50.)

Kohonen T., 1990, Proceedings of the IEEE, 78, 1464 (Cited on page 4.)

Kohonen T., 2001, Self-Organizing Maps, Physics and astronomy online library. Springer-Verlag GmbH (Cited on pages 4 and 50.)

Komatsu E. et al., 2011, ApJS, 192, 18 (Cited on page 165.)

Koo D. C., 1985, AJ, 90, 418 (Cited on page 3.)

Krone-Martins A., Ishida E. E. O., de Souza R. S., 2014, MNRAS, 443, L34 (Cited on page 4.)

Larson D. et al., 2011, ApJS, 192, 16 (Cited on page 165.)

Laurino O., D'Abrusco R., Longo G., Riccio G., 2011, MNRAS, 418, 2165 (Cited on page 30.)

Lawrence R., Almasi G., Rushmeier H., 1999, Data Mining and Knowledge Discovery, 3, 171 (Cited on page 4.)

Le Fèvre O. et al., 2013, A&A, 559, A14 (Cited on page 15.)

Le Fèvre O. et al., 2005, A&A, 439, 877 (Cited on pages 14 and 15.)

Leistedt B., Peiris H. V., Mortlock D. J., Benoit-Lévy A., Pontzen A., 2013, MNRAS, 435, 1857 (Cited on page 158.)

Lemire D., Boytsov L., 2012, CoRR, abs/1209.2137 (Cited on page 134.)

Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473 (Cited on page 159.)

Lilly S. J. et al., 2009, ApJS, 184, 218 (Cited on page 15.)

Lilly S. J. et al., 2007, ApJS, 172, 70 (Cited on page 15.)

Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118 (Cited on pages 4, 145, and 146.)

Limber D. N., 1953, ApJ, 117, 134 (Cited on page 159.)

Loh E. D., Spillar E. J., 1986, ApJ, 303, 154 (Cited on page 3.)

Mallat S. G., Zhang Z., 1993, IEEE Transactions on Signal Processing, 41, 3397 (Cited on page 124.)

Mandelbaum R. et al., 2008, MNRAS, 386, 781 (Cited on pages 6, 137, and 138.)

Mannucci F., Basile F., Poggianti B. M., Cimatti A., Daddi E., Pozzetti L., Vanzi L., 2001, MNRAS, 326, 745 (Cited on page 3.)

Matthews D. J., Newman J. A., Coil A. L., Cooper M. C., Gwyn S. D. J., 2013, ApJS, 204, 21 (Cited on pages 13, 71, and 78.)

McMahon R. G., Banerji M., Gonzalez E., Koposov S. E., Bejar V. J., Lodieu N., Rebolo R., the VHS Collaboration, 2013, The Messenger, 154, 35 (Cited on page 15.)

Melchior P. et al., 2014, ArXiv e-prints (Cited on page 15.)

Monteith K., Carroll J. L., Seppi K., Martinez T., 2011, The 2011 International Joint Conference on Neural Networks, 2657 (Cited on pages 87 and 88.)

Myers A. D., White M., Ball N. M., 2009, MNRAS, 399, 2279 (Cited on pages 6 and 138.)

Naim A., Ratnatunga K. U., Griffiths R. E., 1997, ApJS, 111, 357 (Cited on page 5.)

Newman J. et al., 2013b, ArXiv e-prints : 1309.5384 (Cited on page 3.)

Newman J. A., 2008, ApJ, 684, 88 (Cited on page 170.)

Newman J. A. et al., 2013a, ApJS, 208, 5 (Cited on pages 13 and 14.)

Nicola A., Refregier A., Amara A., Paranjape A., 2014, ArXiv e-prints (Cited on page 158.)

Oke J. B. et al., 1995, PASP, 107, 375 (Cited on page 13.)

Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008a, ApJ, 689, 709 (Cited on pages 5 and 6.)

Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., Sheldon E. S., 2008b, ApJ, 674, 768 (Cited on pages 4, 24, 31, and 149.)

Parkinson D., Liddle A. R., 2013, Statistical Analysis and Data Mining, 6, 3 (Cited on page 87.)

Percival W. J. et al., 2010, MNRAS, 401, 2148 (Cited on page 2.)

Perlmutter S. et al., 1999, ApJ, 517, 565 (Cited on page 1.)

Planck Collaboration et al., 2013, ArXiv e-prints (Cited on page 1.)

Reddy N. A., Steidel C. C., Erb D. K., Shapley A. E., Pettini M., 2006, ApJ, 653, 1004 (Cited on page 12.)

Reid B. A. et al., 2010, MNRAS, 404, 60 (Cited on page 2.)

Riess A. G. et al., 1998, AJ, 116, 1009 (Cited on page 1.)

Rokach L., 2010, Artificial Intelligence Review, 33, 1 (Cited on page 84.)

Ross A. J., Percival W. J., Brunner R. J., 2010, MNRAS, 407, 420 (Cited on page 1.)

Sánchez A. G. et al., 2013, MNRAS, 433, 1202 (Cited on page 2.)

Sánchez C. et al., 2014, ArXiv e-prints (Cited on pages 3, 15, 144, 145, 149, 150, 161, and 169.)

Sawicki M., 2012, PASP, 124, 1208 (Cited on page 3.)

Schapire R. E., Freund Y., Bartlett P., Lee W. S., 1998, The Annals of Statistics, 26, 1651 (Cited on page 24.)

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525 (Cited on page 13.)

Schmidt M., Lipson H., 2009, Science, 324, 81 (Cited on page 4.)

Seo H.-J. et al., 2012, ApJ, 761, 13 (Cited on page 158.)

Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J., Weaver B. A., 2012, ApJS, 201, 32 (Cited on page 6.)

Sheth R. K., 2007, MNRAS, 378, 709 (Cited on page 6.)

Simet M. et al., 2012, ApJ, 748, 128 (Cited on page 1.)

Smith R. E. et al., 2003, MNRAS, 341, 1311 (Cited on page 159.)

Strateva I. et al., 2001, AJ, 122, 1861 (Cited on page 33.)

Strauss M. A. et al., 2002, AJ, 124, 1810 (Cited on page 10.)

Tegmark M. et al., 2002, ApJ, 571, 191 (Cited on pages 158, 160, and 164.)

Thomas S. A., Abdalla F. B., Lahav O., 2011, MNRAS, 412, 1669 (Cited on page 158.)

Tipping M. E., 2001, J. Mach. Learn. Res., 1, 211 (Cited on page 149.)

Treu T., Ellis R. S., Liao T. X., van Dokkum P. G., 2005, ApJL, 622, L5 (Cited on page 12.)

Trimble V., 1987, ARA&A, 25, 425 (Cited on page 1.)

Trotta R., 2007, MNRAS, 378, 72 (Cited on page 87.)

van Breukelen C., Clewley L., 2009, MNRAS, 395, 1845 (Cited on page 6.)

Vogeley M. S., Szalay A. S., 1996, ApJ, 465, 34 (Cited on page 158.)

Wadadekar Y., 2005, PASP, 117, 79 (Cited on page 4.)

Wang D., Zhang Y.-X., Liu C., Zhao Y.-H., 2008, ChJAA, 8, 119 (Cited on page 3.)

Wang W.-H., Cowie L. L., Barger A. J., 2006, ApJ, 647, 74 (Cited on page 12.)

Wang Y., Brunner R. J., Dolence J. C., 2013, MNRAS, 432, 1961 (Cited on page 138.)

Way M. J., Foster L. V., Gazis P. R., Srivastava A. N., 2009, ApJ, 706, 623 (Cited on page 4.)

Way M. J., Klose C. D., 2012, PASP, 124, 274 (Cited on pages 5, 7, and 68.)

Wirth G. D. et al., 2004, AJ, 127, 3121 (Cited on page 12.)

Wittman D., 2009, ApJL, 700, L174 (Cited on pages 123, 138, and 141.)

Wolf C., 2009, MNRAS, 397, 520 (Cited on page 3.)

Yip C.-W., Szalay A. S., Carliles S., Budavári T., 2011, ApJ, 730, 54 (Cited on page 40.)

York D. G. et al., 2000, AJ, 120, 1579 (Cited on pages 2 and 10.)

Zhan H., Knox L., 2006, ApJ, 644, 663 (Cited on page 2.)

Zhang H., 2004, in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), Barr V., Markov Z., eds., AAAI Press (Cited on pages 80, 112, and 113.)