Dissertation

submitted to the

Combined Faculties of the Natural Sciences and Mathematics

of the Ruperto-Carola-University of Heidelberg. Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Antonio, D'Isanto

born in: Naples, Italy

Oral examination: January 31st, 2019

Probabilistic photometric redshift estimation in massive digital sky surveys via machine learning

Referees: Prof. Dr. Joachim Wambsganß
Dr. Coryn Bailer-Jones

To my wife Simona,
for sharing with me this big adventure.

Zusammenfassung

Das Problem der photometrischen Abschätzung von Rotverschiebungen ist heutzutage ein Schwerpunktthema der Astronomie. Dies ist auf die Notwendigkeit der Berechnung von Entfernungen für eine Vielzahl von Quellen zurückzuführen, wie es die Datenflut der letzten Jahre erfordert. Die Möglichkeit, Rotverschiebungen durch Spektroskopie zu schätzen, ist bei einer solchen Datenlawine allerdings nicht machbar. Photometrische Rotverschiebungen stellt die Antwort auf dieses Problem dar, ist aber nur auf Kosten einer gewissen Genauigkeit möglich. Der Erfolg bevorstehender Vorhaben hängt von der Verfügbarkeit photometrischer Rotverschiebungen ab. Das Ziel dieser Arbeit ist es, innovative Methoden für die photometrische Rotverschiebung vorzuschlagen. Hierzu werden zwei Modelle vorgestellt. Das erste ist ein vollautomatisiertes Modell, das auf der Kombination eines faltenden neuronalen Netzwerks mit einem Gemischdichte-Netzwerk basiert, um die Rotverschiebungen probabilistisch direkt aus den Bildern zu ermitteln. Das zweite Modell basiert auf Merkmalen, indem es eine Kombination von photometrischen Parametern durchführt, um ein Vorauswahlmodell in einem riesigen Merkmalsraum anzuwenden. Die hier vorgeschlagenen Modelle erwiesen sich im Vergleich zu den gängigsten Modellen aus der Literatur als sehr effizient. Ein Teil der Arbeit geht auf die Fehlerabschätzung und die Qualität der Vorhersagen ein. Die vorgeschlagenen Modelle sind sehr allgemein gehalten und können auf verschiedene Themen in der Astronomie und darüber hinaus angewendet werden.

Abstract

The problem of photometric redshift estimation is a major subject in astronomy, since the need of estimating distances for a huge number of sources, as required by the data deluge of the recent years. The ability to estimate redshifts through spectroscopy does not scale with this avalanche of data. Photometric redshifts provide the required redshift estimates at the cost of some precision. The success of several forthcoming missions is highly dependent on the availability of photometric redshifts.

The purpose of this thesis is to provide innovative methods for photometric redshift estimation. Two models are proposed. The first is fully-automatized, based on the combination of a convolutional neural network with a mixture density network, to predict probabilistic multimodal redshifts directly from images. The second model is features-based, performing a massive combination of photometric parameters to apply a forward selection in a huge feature space. The proposed models perform very efficiently compared to some of the most common models used in the literature. An important part of the work is dedicated to the correct estimation of the errors and prediction quality.

The proposed models are very general and can be applied to different topics in astronomy and beyond.

Contents

Chapter 1

Introduction

In the recent years astronomy is experiencing great changes and an era of new discoveries seems ready to start. Much has changed in the last twenty years both in the knowledge and the perception that we have of the Universe, and in the methods with which astronomical research is conducted. This is due, on one hand, to a new generation of instruments, which are exploring the whole electromagnetic spectrum in a detail never reached before. Furthermore, neutrino and astroparticle astronomy, and more recently gravitational waves observations, are opening new windows to the Universe, giving the possibility to observe known phenomenona in a new fashion and to discover ones previously unobserved. Such a revolution is generating a true explosion of the available information, which is bringing astronomy in the regime of Big Data [Szalay and Gray, 2006, Estévez, 2016]. On the other hand, the availability of more powerful hardware solutions, and the realization of novel and clever software applications, are opening the possibility to mine, analyze and make discoveries in this huge amount of data, following Jim Gray's *fourth paradigm* [Hey et al., 2009]. The information explosion experienced by astronomy [Brunner et al., 2002] requires the implementation of new techniques and methods, in order to treat and analyze such an amount of data.

The interest of the astronomical community in machine learning has constantly grown in the last ten years, while pursuing solutions in the fields of data analysis and management, automation, visualization and knowledge discovery [Biehl, 2018]. Traditional techniques are no longer sufficient for these tasks. Instead machine learning (and deep learning) have proved to be very useful in astronomy and several models have been successfully applied. The problem is not just in the amount of data collected (already in the Petabyte domain and looking forward to the Exabyte domain), but also in their complexity and heterogeneity. Therefore, machine learning based techniques are nowadays being introduced at each level of the data processing, from acquisition to storage, analysis and visualization. This is independent from the particular astronomical problem or field of application in which machine learning is used. To give some examples, machine learning models have been successfully adopted for: classification or selection problems, like star-galaxy separation [Ball et al., 2006] and quasar detection [Richards et al., 2009]; morphological classification [Dieleman et al., 2015]; galaxy clustering [Polsterer et al., 2016]; regression problems, e.g. photometric redshift estimation [D'Abrusco et al., 2007]; time series analysis [Mahabal et al., 2017]. It should be pointed out that the enormous growth in popularity of machine learning techniques is also due to the parallel growth of computational power, and the consequent explosion of deep learning applications [Bengio et al., 2012] in several fields of scientific research, and in industrial applications too. An important role is also acquired by the availability of several software packages explicitly devoted to build machine learning models. Just to mention a few, the Python packages scikit-learn [Pedregosa et al., 2011], astroML [Vanderplas et al., 2012], or Theano [specific for deep learning Theano Development Team, 2016] are widely used in the astronomical community.

Machine learning is the automated induction of a model from observed data. The induction is guided by the optimization of a loss function which quantifies how well the data support a model under scrutiny. This translates into tuning the free parameters of a model so that the loss function is minimized. This optimization is known as learning. Typically machine learning is distinguished in supervised and unsupervised learning. Supervised learning seeks a mapping between inputs (covariate/independent variables) and desired outputs (response/outcome). Regression and classification are typical such examples. On the contrary, in unsupervised learning no targets are available and the algorithm is expected to find patterns and structures in the data. For instance, such structures may be clusters, overdensities or outliers.

An important class of machine learning models is constituted by deep learning methods [Dechter, 1986]. This is a class of neural networks which is characterized by many layers of non linear calculation units which are generally able to automatically learn data representation and abstractions, to do dimensionality reduction and feature extraction, in several kind of tasks. For example, these models can be used for classification or for regression problems, for speech recognition, computer vision, natural language processing. Frequently used models are convolutional neural networks [LeCun et al., 2015], recurrent neural networks [Hopfield, 1982] and long short-term memory [LSTM Hochreiter and Schmidhuber, 1997] networks.

In this context, the problem of photometric redshift estimation assumed great importance in the recent years. Redshift is fundamental in astronomy, and in particular in cosmology, for its role in the measurement of distances and other cosmological quantities. In fact, redshift constitutes the last step, in combination with the Hubble's Law, of the cosmic distance ladder. Therefore, the availability of precise redshift estimates for a large amount of sources is mandatory for building models about the structure, the dynamics and the evolution of the Universe. Redshift is typically measured via spectroscopy [Yip et al., 2004], being, by definition, the shift in the spectral lines of galaxies and quasars due to the expansion of the Universe. Unfortunately, spectroscopical analysis is a complex and time consuming task, and it is not possible to obtain spectroscopic redshifts for all sources needed which are currently observed by modern digital surveys [Le Fèvre et al., 2005, Newman et al., 2015]. Photometric redshifts are meant to solve this problem, allowing astronomers to obtain redshift estimates for a huge number of sources, by using only photometric information, at the cost of lower precision. The success of several forthcoming missions and projects is highly based on the availability of precise photometric redshifts.

In literature, the estimation of photometric redshifts is mainly based on two different approaches: template fitting and training set based techniques. The former, also defined as spectral energy distribution (SED) fitting method [Bolzonella et al., 2000], is based on the estimation of a photometric SED, obtained from the known photometry of a certain class of sources. The observed SEDs are then compared to a set of known templates, based on reference spectra, in order to find the best fit through a χ^2 minimization. The estimated redshift is given by the best fitting template spectrum. This method has the advantages of being simple and precise, but it is penalized by the requirement of a high number of templates.

Generally, photometric redshift estimation is mainly based on the application of supervised models [Laurino et al., 2011], with some exceptions [Carrasco Kind and Brunner, 2014]. For this reason, a representative sample of spectrospic redshifts is necessary in order to train the selected model and make predictions. In literature, a wide variety of different machine learning models have been used to solve this task. Decision trees [DT Breiman et al., 1984], k-Nearest Neighbours [kNN Fix and Hodges Jr, 1951], random forest [RF Breiman, 2001] and neural networks are just some of the most used models, due to their efficiency and good general performance. All these methods are based on the use of a vector of input data called features. These features can be defined as a parametrization of the photometric input space. Traditionally, the features used for photometric redshift estimation are plain magnitudes and colours [D'Abrusco et al., 2007]. In order to improve the performance, and the quality of the predictions, one should aim, on one hand, for a better performing model, and on the other hand, on a better representation of the available information as given by synoptic surveys.

This thesis is a cumulative work based on three publications that I developed, together with colleagues, during my PhD studies. The common background of these works is devoted to satisfy the expressed need of improving the global performance of the photometric redshift estimation process by developing novel and affordable methodologies. In the first of the three papers (Publication I), a new model will be presented, based on the application of a modified convolutional neural network [CNN LeCun and Bengio, 1995] to perform the redshift estimation directly from images. Such a model is able to automatically extract features from the original multiband images, therefore optimizing the usage of the available information. It will be demonstrated that the proposed model gives a substantial improvement in the performance of the predictions. This paper, having been published in the proceedings of the *European Symposium on Artificial Neural Networks 2017*, is more dedicated to the technical aspects of the model, which is based on combination of a convolutional neural network and a mixture density network [MDN Bishop, 1994]. The second paper (Publication II), published in the journal Astronomy & Astrophysics, extends the previous work, adopting the proposed model, called *deep convolutional mixture density network* (DCMDN), for different experiments, namely the redshift estimation for galaxies, quasars and a mixed and contamined catalog. Automatic models have several advantages, because the task of feature extraction and selection is completely addressed by the machine. Unfortunately, the risk is that they can become black boxes, in which it is hard to understand the reason for the machine behaviour or to physically interpret the automatically estimated features [Knight, 2018]. Moreover, convolutional neural networks based models are highly computationally expensive and require time to reach the convergence point. Therefore, the possibility to have a fixed set of well defined, interpretable and high performing features is still useful for real applications. In this view, the third and final publication composing this thesis (Publication III), also published in Astronomy & Astrophysics, is dedicated to establishing a method for finding the best performing set of features within a huge set built by the massive combination of all the available photometric and shape informations delivered by the Sloan Digital Sky Survey [Abolfathi et al., 2018] database. This is done by adopting a forward selection model [Pahikkala et al., 2010], based on repeated kNN experiments on random subsamples of the available dataset, in order to generate a tree of features, from which to select the best performing branch. Those features have been then physically interpreted with respect to the typical spectral emission lines for quasars [Charlton and Churchill, 2000]. The final purpose of establishing these two different and, from a certain point of view, alternative methods, is to give to the community a good and affordable way to estimate photometric redshifts for a huge number of sources.

An important part of the work is dedicated to the correct estimation of the errors in the prediction process and to the statistical tools that should be used to deal with this problem. The true nature of the problem has to be found in the way in which photometric redshifts are estimated. Redshifts can be predicted as point estimates or as probability density functions (PDFs). It goes without saying that the latter is a much more correct way to proceed. In fact, a probabilistic description associates an error to the redshift measure, adding much more information with respect to a point estimate. Furthermore, in this thesis the estimated PDFs are treated as multimodal functions, generated by a Gaussian Mixture Model (GMM). Such a description is necessary to take into account the degeneracy introduced by the broadband photometric system, which does not allow the characterization of the redshift with a single peak distribution. In order to correctly deal with the evaluation of an error between a density distribution and a point estimate (the spectroscopic redshift), a novel statistical tool for astronomy has been introduced from the weather forecast field: the continuous rank probability score [CRPS Hersbach, 2000a]. This score, as an integral function, is specifically meant for this task and has been used also as a loss function on which the proposed neural network model has been trained. The choice of the loss function is an important aspect for building a solid machine learning model, and in particular, when dealing with PDFs. Training the model using a proper score is fundamental in order to predict better calibrated and sharper distributions. The concepts of calibration and sharpness, their optimization and a discussion about the correct estimation of errors also constitute an important part of the thesis. Moreover, another tool has been introduced to visually inspect these characteristics of the PDFs, namely the probability integral transform [PIT Gneiting et al., 2005].

This introduction constitutes the starting point for the construction of a workflow for the correct estimation of probabilistic multimodal photometric redshifts. From a technological point of view, the use of such methodologies requires an intensive application of graphics processing units (GPUs), in order to parallelize and speed up the calculations necessary to train the machine learning models. Deep learning models were conceptually known since the '80s, but their implementation for the solution of astronomical problems, and in many other fields, became possible in the recent years, as already stated, due to the big improvements in the computational speed and parallelization techniques. This strong collaboration is driving astronomy toward a new era, and the new field of astroinformatics [Ball and Brunner, 2010] is rising. In the near future, the interoperability between astronomy and computer science will become essential in order to face the challenges brought up by the new generation of instruments and the data explosion that will completely reverse the way in which research is performed. The amount of available data will be so huge, that instead of acquiring new data for the confirmation of hypothesis and theories, scientists will mine into databases, searching for knowledge, patterns and trends, following the paradigm defined as *knowledge discovery in databases* [KDD Zhang et al., 2002]. Photometric redshift estimation is just one important example of this new way to do astronomical research, but the methodologies presented here are very general. The final purpose is to develop methods and techniques that could be applied to several problems and that can allow the community to be ready for the revolution that is to come.

Chapter 2

Photometric redshift

This chapter of the thesis is meant to present a general overview about photometric redshift, as the publications constituting the core of my work are focused on this topic. The chapter will start with a general definition of redshift as a fundamental cosmological parameter. Some historical notes will be given, from the introduction of the photometric redshift concept, following its development through the publications and the methods that constitute the milestones in the field. Therefore, the modern developments and the two main categories of methods for phometric redshift estimation will be presented, namely SED template fitting methods and empirical training set/machine learning methods. Finally, the last section will be dedicated to probabilistic redshift estimation and its importance for the field.

2.1 Cosmological redshift

Cosmological redshift is defined as the physical phenomenon of the shift toward redder wavelengths of the spectral lines of galaxies, due to the expansion of the Universe in the Hubble flow [Hubble, 1929]. Redshift is defined by the general formula:

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}}$$

where λ_{obs} and λ_{emit} are, respectively, the wavelength measured for a particular spectral line from a certain receeding galaxy, and the wavelength of the same spectral line as measured in the laboratory. The Universe expansion generates a stretch of the spectrum of a factor $1 + z$. The concept of redshift is directly related to the scale factor of the Universe, as it is expressed in the Einstein's equations, assuming a Friedmann-Lemaitre-Robertson-Walker (FLRW) metric. In fact, redshift is related to the scale factor by the relation:

$$1 + z = \frac{a(t_0)}{a(t)} \quad \text{that gives:} \quad a(t) = \frac{1}{1 + z}$$

Here $a(t_0)$ is the scale factor at the present age of the Universe, which by definition is equal to 1. Therefore, redshift can be used to calculate the scale factor of the Universe at the age of object which is emitting the light. If the distance between two objects in an expanding Universe described by the FLRW metric is expressed by $d(t) = a(t)d_0$, with d_0 defined as the distance at t_0, then it is possible to define the so called Hubble parameter as:

$$H = \frac{\dot{a(t)}}{a(t)} \quad \text{and the Hubble's law is consequently given by:} \quad \dot{d(t)} = Hd(t)$$

Redshift can be connected, at least for low redshifts, to the expansion velocity, by the relation $v = cz$, where c is the speed of light. Therefore, taking into account Hubble's Law in its form involving the Hubble constant H_0, $v = H_0 d$, it is clear that redshift constitutes a distance measure which can be used to estimate the distances for galaxies and quasars. At higher redshifts, the relation between velocity and redshift does not hold, due to the high velocity involved, and it becomes model dependent, but this does not change the nature of the problem.

In particular, through Hubble's Law, redshift becomes fundamental as a final step in the *cosmic distance ladder*. Up to now, the cosmic distance ladder constitutes the only way to obtain correct distances at increasing scale in the Universe, by calibrating every step through the previous one using the so called *standard candles*, i.e. objects for which the absolute magnitude is known. Therefore their distance can be derived from the distance modulus $m - M = 5 \log d - 5$. In the framework of the cosmic distance ladder, the Hubble's law constitutes the only way to estimate the distance of quasars and distant galaxies, for which other methods cannot be applied. It should be noted that this aspect could change in the near future, as there are indications that the recently observed gravitational waves [Abbott et al., 2016] could be used as *standard sirens* in order to retrieve the Hubble constant without any form of distance ladder [Abbott et al., 2017]. Moreover, redshift is a fundamental quantity in order to estimate the age of the Universe and other important cosmological parameters, upon which the standard model is based.

2.2 Historical overview

Fundamentally, redshift is measured with spectroscopy. Unfortunately, the process of taking spectra, reducing and analyzing them in order to estimate redshift is long and complicated, and can be performed only for a limited number of brighter sources. The spectroscopic redshift obtained in this way can reach a very high precision, but in order to obtain a robust redshift measurement, at least two well identified spectral features have to be detected. On the other hand, the need for redshift measurements (and therefore distances) is nowadays growing in the community. As already sketched in the introduction, the success of several projects and missions will be mainly based on the possibility to access affordable and precise measurements of redshift for a huge number of sources. Photometric redshifts represent the solution to this problem. In fact, an estimate of redshift can be derived by using multiple filters, which are affected by the shift toward the red of the most prominent spectroscopic features through the broadband filters themselves. Relying on the availability of photometric measurements (e.g. magnitudes and colors), the sample of sources for which it is possible to obtain such measurements is much larger than the spectroscopic one. Furthermore, broadband images can be obtained for sources which are several magnitudes fainter compared to spectroscopic observations. The price that has to be paid is the lower precision characterizing photometric redshifts with respect to their spectroscopic counterpart, but this is still enough for many applications for which this kind of information is required. For this reason, nowadays photometric redshifts are becoming fundamental in a number of different astrophysical applications. For example, they are used to study the formation and evolution of galaxies and the relation between their properties and the dark halo [Fontana et al., 2000, Coupon et al., 2015]. Cluster identifications and evolution of Active Galactic Nuclei (AGN) are other important fields of application [Finoguenov et al., 2007, Miyaji et al., 2015]. Finally, dark energy studies are highly dependent on the availability of photometric redshifts, in particular for the weak lensing tomography approach [Hu, 1999], which is/will be used in major projects like the Dark Energy Survey [DES Abbott et al., 2018], the Large Synoptic Survey Telescope [LSST LSST Science Collaborations et al., 2017] or Euclid [Laureijs et al., 2011].

The definition of photometric redshift is older than its spectroscopic counterpart. It was introduced by Baum [1962] who first proposed the idea of comparing spectral energy distributions of different ellipticals in clusters, derived from nine bands, with those of galaxies belonging to the Virgo cluster, in order to estimate their redshift. The concept behind this work was quite simple: multiband photometry can be interpreted as a low resolution spectrum, and hence this could be used to derive an estimate of the redshift. Despite Baum's work, his idea was abandoned for the next 20 years, as redshift was mainly calculated from spectroscopy. The interest in photometric redshift started to grow again in the last 20 years, thanks to the availability of CCD photometry and the new digital surveys. Koo [1985] proposed the use of photometric redshifts as a *poor person's redshift machine*, using linear combinations of four bands photometry to derive the redshift. In this approach, two-color plots were defined, from the considered four bands, in order to separate galaxies by type and redshift and to retrieve an estimate of the redshift by properly calibrating [Bruzual A., 1983] the relation between those colors and the redshift itself. The author found in this case a good agreement with the corresponding spectroscopic detection. Koo's work was based on data taken with photographic plates.

The first attempt to calculate photometric redshifts by mean of CCD data was performed by Loh and Spillar [1986], which used six-band photometry to fit the spectral energy distribution of galaxies. Both these first attempts were biased by the poor photometry quality of the available data, obtained from photographic plates or from the first CCDs ever used in astronomy, which were small and affected by limited quantum efficiency, compared to the detectors available nowadays. The advent of digital imaging and of the big synoptic survey, however, changed this perspective and with these new data available, the interest in photometric redshifts increased. One of the reasons for this renewed interest is surely due to the better quality of the available photometric data. Moreover there is the need to obtain redshift measurements for the huge, and always increasing, number of sources which are observed by the recent digital surveys. The integration time required for spectroscopical analysis simply does not scale with the amount of data collected by the SDSS, for example, and this situation will become increasingly worse when instruments like the LSST will become operational. In Connolly et al. [1995] magnitude and color information were used in combination for the first time to retrieve the redshift by means of a quadratic function fit. All the methods presented until now are mainly based on SED fitting techniques.

A different approach was proposed by Steidel et al. [1996, 1998], where it was demonstrated that redshift can be estimated from photometric data by detecting prominent breaks in the spectral energy distribution of the galaxy, like the $4,000$ Å Balmer break and the $1,216$ Å Lyman break. Following their work, this method, called *dropout* technique, relies on the large break in the continuum flux of the sources occurring at the 912 Å Lyman limit, due to neutral hydrogen absorption around star-forming regions. The breaks correspond to a sudden increase of the flux continuum from lower to higher wavelenghts. The radiation absorption by neutral gas around star forming regions of galaxies causes the spectrum below the Lyman break to become faint and, reversely, it becomea very bright at longer wavelengths. For redshifts around $z = 3$, the Lyman break is shifted to ultraviolet/optical wavelengths and can be used to detect galaxies at that redshift. However, the detection of gradients in the fluxes in contiguous filters could reveal a break and act as a feature to estimate the redshift.

In general, it is demonstrated that most of the information used to generate photo-z comes from the $4,000$ Å Balmer break and the $1,216$ Å Lyman break in galaxy spectra. Following Salvato et al. [2018], the former can be explained by the absorption of photons, which are more energetic with respect to the Balmer limit at $3,646$ Å, and the combination of absorption lines from ionised metals in stellar atmospheres, in particular A-type stars. The latter, as already said, is due to absorption of light below the Lyman limit at 912 Å and the absorption due to the intergalactic medium. For this reason, one expects to find the best photometric redshifts estimates at redshifts where such breaks fall between the used bands. In general, for optical surveys, the best estimates are around $z \sim 0.7$, corresponding to the $4,000$ Å break falling in the i band. For the same reasons, infrared measurements are helpful for galaxies at higher redshifts. Therefore, multi-band images of a field containing high redshift galaxies can be used to identify those objects that disappear in the bluest filters, due to the redshifted Lyman limit. The advantages of this method are its simplicity and speed, the fact that it is based on extrinsic properties and the absence of biases or selection effects which are typically introduced by templates and different fitting methods. On the other hand, the application of SED fitting also brings a redshift probability distribution, further information about galaxy properties and utilizes all the available photometry, not being limited by the band in which the break is prominent. Generally speaking, any photometric redshift estimation method should be designed in order to detect, directly or indirectly, the key features represented by the breaks. This is true for methods based on the assumption of a theoretical/physical model, like SED fitting, but also for empirical methods, like those based on machine learning. However it has to be considered that breaks are broad features. This is the reason for which it is important to plan a multiwavelength approach. As multiple redshifts can correspond to the same color [Benítez et al., 2009], the use of several filters can help to get rid of degeneracies.

2.3 Modern developments and methods for photometric redshift estimation

As already sketched in the introduction, there are basically two main categories of methods for photometric redshift estimation: template fitting and training-set/machine learning based methods. The former was used in almost all the early works on photometric redshifts and it is based on the assumption of a theoretical physically-motivated model, given by the templates. The latter increased in popularity in the last years, as more training sets of spectroscopic redshifts became available for deriving empirically the relationship between photometric data and redshift. These methods are characterized by the absence of an imposed theoretical model, which is instead determined empirically by the algorithm in a data-driven way. In the following subsections the two methods will be analyzed in detail, with the focus on the main milestones in both of them.

2.3.1 Theoretical model-based methods: spectral energy distribution template fitting

The SED fitting method, as previously introduced, is based on the fit of the global shape of spectra and on the detection of prominent spectral features. The spectral energy distributions obtained from the observed photometry are then compared to those obtained from a set of reference spectra, or templates. The photometric redshift is given by the best fit between the photometric SED and the template spectra. Following Bolzonella et al. [2000], the fit is performed through a χ^2 minimization procedure between the observed SED of a given galaxy and the set of templates, by the relation:

$$\chi^2(z) = \sum_{i=1}^{N_{filters}} \left[\frac{F_{obs,i} - b \times F_{temp,i}(z)}{\sigma_i}\right]^2$$

where $F_{obs,i}$, $F_{temp,i}$ and σ_i are the observed and template spectral distributions with the relative uncertainties in the i filter, while b is a normalization constant. Alternatively, in Lanzetta et al. [1996] the χ^2 minimization was substituted by a likelihood function maximization:

$$L(z,T) = \prod_i \exp\left\{ -\frac{1}{2}\left[\frac{f_i - AF_i(z,T)}{\sigma_i}\right]^2\right\}$$

where f_i are the measured fluxes with uncertainties σ_i given the modelled fluxes $F_i(z,T)$ at an assumed redshift z for spectral type T and normalization factor A over four filters $i = 1 \dots 4$. This last work constitutes a good example of the importance of photometric redshift estimation, as it is used to study high redshift galaxy candidates in relation to their environment and photometric and physical properties to understand their star formation history with respect to the evolution of the hosting galaxy.

The templates used for the fit fix the theoretical/physical model that is assumed on top of the method. In general, the first templates commonly used for SED fitting were determined by Coleman et al. [1980] for different spectral type of low redshift galaxies. Such templates are empirical spectro-photometric SEDs. Alternatively, SEDs can be built theoretically from stellar population synthesis models, like those given by Bruzual A. and Charlot [1993], Vazdekis et al. [1996]. Stellar population based models are very important in order to understand the stellar content of sources and are used to derive several properties of galaxies, not just photometric redshifts. For example, the models from Maraston [2005], Maraston et al. [2009] are used in the SDSS to perform the best fit between the observed *ugriz* magnitudes of the SDSS - Baryon Oscillation Spectroscopic Survey [BOSS Dawson et al., 2013] with the spectroscopic redshifts determined by the BOSS pipeline itself. Describing how stellar population models are built is beyond the scope of this thesis, but I will briefly provide some indications for sakeness of completeness. In general, if not considering the case of active galactic nuclei (AGN), galaxies emission is mainly dominated by the stellar component. This light could be detected directly or reprocessed by the gas and dust which constitutes the interstellar medium (ISM). Therefore, the synthetic model has to take into account the spectra of the single stars, or stellar populations, together with effects due to gas and dust extinction, evolution and morphological properties. All these charecteristics are the building blocks of generating a reliable model. This can be done basically by applying two different methods. The first, established by Charlot and Bruzual [1991], is defined as *isochrone synthesis* and uses star isochrones in the Hertzsprung-Russel diagram. The spectra of all the stars are integrated along the isochrones to compute the total flux. The second method [Buzzoni, 1989] is based on the *fuel consumption* approach, in which the fuel is integrated along the evolutionary track. The main idea behind this approach is that the luminosity of post-main sequence stars, which are the most luminous, is directly linked to the available fuel for stars at the turnoff mass. However, it has to be stated that the quality of photometric redshifts predicted by SED fitting does not depend only on the type of available templates, but also on their coverage of the color-redshift space [Chevallard and Charlot, 2016].

SED fitting (and photometric redshift estimation in a more general sense) became very popular with the release of the Hubble Deep Field [HDF Williams et al., 1996], containing multi-band data of galaxies up to 29 magnitude. At that epoch, conventional spectroscopy was not able to observe such faint objects, therefore photometric redshifts became a necessity. HDF represented a big challenge due to the lack of suitable high quality empirical SED templates at short wavelengths, as the UV light is shifted in the optical. Furthermore, one also had to take into account the possibility that high redshift galaxies SEDs were consistently different from those obtained for galaxies in the local Universe, and the effects due to intergalactic hydrogen extinction. For these reasons, in Sawicki et al. [1997], hybrid templates were derived, combining local empirical SEDs with model-based SEDs.

Nowadays, several codes are publicly available to perform photometric redshift estimation based on SED fitting. A very popular library is the GISSEL'98 from Bruzual A. and Charlot [1993], a synthetic collection of galaxies spectra including 200 tracks of ages from 200 million up to 16 billion years for elliptical galaxies, constructed with evolutionary models from the same authors. The HyperZ code from Bolzonella et al. [2000] is based on this library, and it is used to perform photometric redshift estimation through χ^2 minimization. To compare the observed with the fiducial photometry, the absorption from the Lyman forest [Madau, 1995] and the reddening to the redshifted SEDs are applied. Another popular code is Le Phare [PHotometric Analysis for Redshift Estimations Arnouts and Ilbert, 2011], based on PEGASE [Fioc and Rocca-Volmerange, 1999] and GISSEL population synthesis models. Even in this case the program is based on a simple χ^2 fitting method between the theoretical and observed photometric catalog. A simulation program is also available in order to generate realistic multi-color catalogs, taking into account additional observational effects. Le Phare has been used, for example, to derive photometric redshift catalogs for the Canada-France Hawaii Telescope Legacy Survey (CFHTLS) [Ilbert et al., 2006] and for the COSMOS2015 photo-z catalog [Laigle et al., 2016]. In particular, the latter catalog is obtained by combining the visible photometry with near-ultraviolet, near- and mid-infrared information from different instruments. The final catalog contains more than half million sources, over 2 deg^2, with $1 < z < 6$. Taking this last work as example, it is clear that the SED fitting method requires a high number and different types of templates. In particular, this is true if the model is not tuned on a restricted redshift range. In the case of a wide redshift range instead, and many object types, the model requires a configuration that should be as general as possible. In the derivation of the COSMOS2015 catalog, templates from spiral, ellipticals, blue star-forming galaxies were included, together with several extinction laws. In Salvato et al. [2009, 2011], photometric redshifts are estimated for Chandra and XMM-Newton selected sources counterparts via adopting ad hoc built hybrid templates. Additional information like optical variability, morphology and X-ray flux has been used to build the model. Such an approach serves as a baseline for future missions like eROSITA [Cappelluti et al., 2011].

A Bayesian version of the SED fitting has been developed by Benítez [2000] in order to overcome some of the intrinsic weaknesses of the method. In fact, the main sources of error of template-based models, as described by Sawicki et al. [1997] and Fernández-Soto et al. [1999], can be divided in two broad categories: color/redshift degeneracies and templates incompleteness. The former arises when the line corresponding to a certain template self-intersects, or when two different templates interesect at points which correspond to different redshifts. Such degeneracy will clearly increase with increasing redshift range and number of templates included. Random photometric errors have an effect too, causing a blurring and/or thickening of the color-redshift relation. The global effect is an increase in the error between predicted photometric redshifts and in the number of catastrophic outliers (objects which sensibly deviate from the ideal diagonal line in photometric versus spectroscopic redshift plot). Using additional filters does not necessarily improve performance because the information contained in colors can be redundant, and therefore not helpful in breaking the degeneracy. The addition of more templates is also not able to solve the latter source of errors because one should, in principle, include a template for every type of galaxy. Furthermore, the addition of too many templates could worsen the degeneracy. Therefore, the Bayesian treatment is useful because it can allow to include additional information, apart from colors and magnitudes, weighting the templates by its prior probability, helping to break the degeneracy. Apart the BPZ code from Benítez [2000], several other Bayesian-based methods have been developed, like GOODZ [Tanaka, 2015] and BEAGLE [Chevallard and Charlot, 2016]. An important part of these works is focused on finding proper and efficient priors to improve the redshift estimation. In fact, priors have to be used with great care, as there is always the risk of including inaccurate information. The main problem concerning an approach based on Bayesian methods is the high demand of computational resources and the slowness of calculation which is typical of such models.

2.3.2 Empirical model-based methods: training set and machine learning

Empirical training set and machine learning methods are based on the availability of a training set, composed of photometric data and spectroscopic redshifts. The method searches for an optimal fit between photometry and redshifts, by means of several machine learning algorithms, that can be used to predict photometric redshifts from photometric data only. The main advantage of the method is that it is basically empirical, so it does not depend on the knowledge of galaxies SEDs, but it requires a sufficiently large dataset in order to properly optimize the model. For this reason, such methods acquired popularity in the recent years thanks to the data explosion due to the synoptic digital surveys. Empirical methods are not directly based on physically motivated models. Instead, the model is found empirically from the algorithm, namely training the machine to optimize the internal weights which characterize a function in the parameter space. In other words, the model learns, through a training sample, the mapping between photometry and redshift. The training phase is exactly the optimization process of the model by means of the training set, in order to choose the best set of parameters. Once this phase is over, the model can be fixed and used for predictions with an unknown set of data. In general, machine learning methods often demonstrated to produce more accurate and robust results, taking into account complicated correlations existing between the inputs and targets. Machine learning methods have several advantages, like the possibility to incorporate additional observables, and to avoid systematic effects associated with the photometry by means of the adopted training set. However it is very important to use a well representative training sample in order to avoid biases [Hoyle et al., 2015a].

A first attempt was done by Connolly et al. [1995], where two samples of 254 and $2,025$ galaxies, respectively, based on four bands photometry, obtained from digitized photographic plates, were fitted using a linear and a quadratic function. This work is at the basis of all the successive developments obtained using empirical methods. In fact, they used the four filters as *input features*, measuring the root mean square error dispersion (σ_z) in the two cases of the linear fit and the quadratic fit. The latter model gave superior performance with respect to the former and the fit could be further improved reducing the redshift range of the fit, performing it in bins. Considering the broadband photometry as a low-resolution spectrum, it is not possible to identify the spectral signature in a particular emission or absorption feature through a particular bandpass. Instead what the model detects is just the effect of passing the overall shape of the continuum through the different filters as the redshift increases, with the break at $4,000$ Å acting as a prominent spectral feature that is typically identified. Such a method, as specified by the authors, had the goal of being applied at the forthcoming SDSS, which in the following years became a gold mine for this field of research. Further developments were obtained by Brunner et al. [1997], which used CCD photometry instead of photographic plates data, and from Wang et al. [1998], which used a modification of the Connolly et al. [1995] method to determine photometric redshifts for the HDF.

Neural networks have been applied in astronomy since the second half of the '90s for several purposes, from morphological classification of galaxies [Lahav et al., 1996] to star/galaxy separation and object detection [Andreon et al., 2000]. An approach based on neural networks, for photometric redshift estimation, was proposed by Firth et al. [2003], Vanzella et al. [2004], Collister and Lahav [2004]. In these works, the authors applied a multilayer perceptron model [MLP Rosenblatt, 1958, Bishop, 1995], taken from the computer science field, to determine photometric redshifts using magnitudes and colors from the first data release of the SDSS or the HDF as input.

Although neural networks increased their popularity in the following years, many other machine learning models have been used to perform the task. This is the case, for example, of Gerdes et al. [2010], who applied decision trees [DT Quinlan, 1986] to predict photometric redshifts in the form of PDFs. The popular random forest model [RF Breiman, 2001] was instead adopted by Carliles et al. [2010]. Other approaches involve the application of methods based on nearest neighbours [Abazajian et al., 2009] or support vector machines [SVMs Wadadekar, 2005]. A MLP model has been used in D'Abrusco et al. [2007], where the neural network is adopted to estimate photometric redshift for two samples taken from the SDSS, namely general galaxies and luminous red galaxies [LRG Eisenstein et al., 2001]. In this work, objects are pre-classified as nearby ($z < 0.25$) and distant ($0.25 < z < 0.50$). Then, a feature selection is performed, from which magnitudes and colors are identified as the most efficient parameters. Several models from the literature are compared with the method proposed by the authors. This expresses the need of the community to test different methods in order to build an efficient and well performing pipeline for photometric redshift estimation.

All the models presented until now are fully supervised. However, the availability of a spectroscopic sample can constitute a hard issue for the realization of a representative dataset to train the model and achieve the best performance. For this reason, unsupervised methods have been investigated too. They differ from neural networks or random forests because they do not require a sample of targets (e.g. spectroscopic redshifts). Typically, these models perform a clustering in order to find groups and similarities in the photometric space.

However, in order to compare the performance on many different methods, the PHoto-z Accuracy Testing programme [PHAT Hildebrandt et al., 2010] was developed, as an international effort to create a standardized common platform composed by two parts: PHAT0, based on simulations to test the basic functionalities of the different models, and PHAT1, which is based on data from the Great Observatory Origins Deep Survey [GOODS Dickinson et al., 2003]. The PHAT1 catalog has been used with a modified version of the MLP, namely the MLP with a quasi Newton algorithm [MLPQNA Cavuoti et al., 2015] and the platform DAMEWARE [Brescia et al., 2014], by Cavuoti et al. [2012]. The model adopted is based on an approximated calculation of the Hessian matrix, in place of the typically used Jacobian, used for the minimization of the loss function in the MLP. This permits, as claimed by the authors, to improve the convergence of the neural network, and then the global performance, by avoiding local minima. In [Laurino et al., 2011] the authors propose a method called *Weak Gated Experts* (WGE), which is characterized by the application of different data mining techniques on samples of optical galaxies and quasars from the SDSS. The WGE is based on features, basically magnitudes and colors, and can be described by three main steps. First, the feature space is properly partitioned. For each partion, a prediction model, or expert, is determined, which maps each pattern of the feature space to the target space. In this way, a new feature space is defined from the outputs of the predictors. Finally, a new gate predictor is trained, mapping the patterns of the new feature space with respect to the targets. This method combines efficient clustering and regression techniques, as the first partitioning of the features space is done by using k-means clustering [MacQueen, 1967] models. The subsequent steps instead are based on MLPs.

A method based on feature selection and a different way to explore the parameter space has been proposed by Polsterer et al. [2014]. In this work the authors demonstrate how an efficient feature selection can substantially improve the performance of photometric redshift estimation. The features obtained are somewhat unusual with respect to the commonly used features adopted in the literature, namely magnitudes and colors. In order to find the best set of features to be used for the task, the authors apply a simple but striking approach: huge amounts of features combinations are tested via a massive parallel feature selection based on the intensive use of GPUs. The regression model used in this case is a simple k-nearest neighbour. Such model proved to be particularly efficient for the task, having to deal with large amount of patterns in a low-dimensional features space [Polsterer et al., 2013]. The method presented in this work will constitute the starting point for the work presented in the Publication III included in the thesis.

2.3.3 Probabilistic redshift estimation

From a general point of view, most of the works presented in the previous overview were based on finding photometric redshifts in the form of point estimates, independently from the method chosen. Anyhow, in the last years, with the refinement of the techniques and the improvement of the technologies used for the task, the attention of the community shifted to determining probability density functions. The reason for this interest is straightforward: PDFs contain much more information which is important for the estimation of several cosmological measurements, like galaxy clustering, weak lensing, baryon acoustic oscillations and mass functions of galaxy clusters [Reid et al., 2010, Ho et al., 2012, Jee et al., 2013]. The quality of such measures is highly dependent on the availability of huge samples of galaxies with precise distances estimates, as demonstrated by Martí et al. [2014]. Furthermore, a description based on PDFs allows to better describe the whole redshift phenomenon, and in particular the presence of multimodality effects, treatment of extreme outliers, and so on. On the other hand, one has to take into account that the growth of photometric-only surveys simply does not scale with their spectroscopic equivalent. Therefore, the given cosmological measurements will depend on the availability of photometric redshifts. The effectiveness of the methods presented will then tremendously increase if they will be able to provide not just a point estimate but a PDF. This is demonstrated, for example, in Myers et al. [2009], where the additional information given by redshift PDFs is used to measure quasi-stellar objects clustering in the SDSS, improving substantially the performance with respect to the use of point estimates; or in Mandelbaum et al. [2008], where the information contained in photo-z PDFs is used for weak-lensing calibration. Further noticeble works in the field are those from Sheth [2007], van Breukelen and Clewley [2009], where it is demonstrated how the use of PDFs can improve cosmological measurements, namely luminosity function and cluster detection. The two works from the same authors, Carrasco Kind and Brunner [2013, 2014], present two methods, the former being supervised and based on random forests, the latter instead unsupervised with the application of *self-organizing maps* [SOMs Kohonen, 1982], to derive photometric redshift PDFs. In particular, the first work proposes a method called Trees for Photo-Z (TPZ), in which the outputs of the decision trees composing the forest are used to generate PDFs.

In the recent years an increasing number of publications have been released proposing methods to estimate photometric redshifts in the form of PDFs. An example is the one given by Cavuoti et al. [2017], in which the authors establish a method, called Machine-learning Estimation Tool for Accurate PHOtometric Redshifts (METAPHOR), to predict photo-z PDFs by building a modular workflow based on the MLPQNA neural network. Here, the density distributions are derived by properly perturbing the parameter space, in order to obtain multiple test-sets from which deriving different estimates of photometric redshifts. Therefore, the photometric redshifts are binned and for each bin the probability that a given photo-z value belongs to each bin is calculated. The results are compared to popular methods like RF, kNN and Le Phare, with respect to several statistical indicators (bias, standard deviation, median absolute deviation, fraction of outliers, skewness), obtaining a clear improvement. Another remarkable approach is the one given by Sadeh et al. [2016]. In this work, the authors utilize ensembles of machine learning models, like neural networks and decision trees, to estimate PDFs by considering the different sources of uncertainties where the prediction deviates from the ideal result. In fact, in general, the uncertainty on the photometric redshift can be due to: inputs, machine learning model, unrepresentative training set or incomplete training set. Following the cited paper, the first three sources of uncertainty can be incorporated in a meaningful PDF, while the latter causes the arising of degeneracy. The use of ensembles of machine learning models is particularly helpful in the exploitation of uncertainties and differences in performance due to the choice of the hyperparameters and/or architectures. Every machine learning model, in fact, is characterized by a certain number of hyperparameters. Their choice and the configuration of the model is always exposed to a certain degree of arbitrariness. This effect can be controlled and limited by using multiple models with different choices of the hyperparameters, which are commonly known as ensembles.

Therefore, from all the previous discussion we can state that there are two main methods for the estimation of photometric redshifts in the form of PDFs. The first is based on the so called binned classification [Gerdes et al., 2010]. The full redshift range is divided into small bins and, by using a spectroscopic training set, a set of classifiers runs for every bin. Basically, each classifier examines every pattern, evaluating the probability that its redshift falls within the given bin. The PDF is then reconstructed by examining the distribution of such probabilities with redshift. The second method is the one used in the previously cited Sadeh et al. [2016], based on the construction of the PDF from the results obtained by ensembles of machine learning models.

A good overview of the topics focused in this chapter can be found in Salvato et al. [2018]. Whatever the method or the model adopted, the correct estimation of photometric redshifts, particularly in the case of PDFs, demands the use of proper statistical tools to evaluate and minimize the errors. In the next chapter I will analyze this aspect in detail, in connection with the tools introduced in the publications presented in this thesis.

Chapter 3

Statistical tools and methodologies

Obtaining high quality photometric redshift estimates will be a major issue for a number of projects and missions in the near future. One is the Euclid mission [Laureijs et al., 2011]. Euclid's main goal will be the investigation of dark energy, dark matter and gravity, by means of studying two independent cosmological probes: weak gravitational lensing and baryonic acoustic oscillations. Such measures will be highly dependent on the availability of affordable photometric redshifts, for which the European Space Agency (ESA) and the Euclid Consortium have established specific quality requirements. In particular, following the analysis of Abdalla et al. [2008], Bordoloi et al. [2010], the Euclid Definition Study Report [Laureijs et al., 2011] states that the standard deviation with respect to the true redshifts should be $\sigma_z/(1+z) \leq 0.05$(required)-0.03(goal). Moreover, the percentage of catastrophic outliers should be below 10%(required)-5%(goal) and the error in mean redshift per bin below 0.002. Typically, the performance of the estimates are evaluated by checking:

- bias: $z_{phot} - z_{spec}$;
- precision: defined as the normalized standard deviation $(z_{phot} - z_{spec})/(1 + z_{spec})$ or the normalized median absolute deviation $1.48 \times \text{median}(|z_{phot} - z_{spec}|/(1 + z_{spec}))$;
- fraction of outliers: fraction of sources having $|z_{phot} - z_{spec}| > N\sigma$ or $|z_{phot} - z_{spec}|/(1 + z_{spec}) > 0.15$.

The main problem is the fact that these kind of measurements and error functions work very well when dealing with point estimates, but can lead to misleading results when handling PDFs. This problem will be addressed in Publication II presented in the Chapter 4 of this thesis and further discussed in Sec. 5.4. In the following sections I will just present the new statistical tools that had to be applied in order to correctly estimate the errors and to predict affordable photometric redshift PDFs, namely the continuous rank probability score and the probability integral transform. The third section, in particular, will be dedicated to the concepts of calibration and sharpness. Then I will introduce the concept of proper scoring rules. The last four sections of the chapter will be focused, respectively, on multimodalities, feature selection, artificial neural networks and parallel computing.

3.1 Continuous rank probability score

The *continuous rank probability score* [CRPS Hersbach, 2000b] is a verification tool specifically meant for probabilistic forecast systems. In particular, the CRPS is an ideal tool for estimating an error between a predicted density distribution and a point-like target value. The general definition of the CRPS, following Hersbach [2000b], is given by the formula:

$$CRPS = CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx$$

where $P(x)$ and $P_a(x)$ are the cumulative distribution functions (CDFs) for a corresponding density distribution $\rho(x)$ and an occurence value x_a. Therefore:

$$P(x) = \int_{-\infty}^{x} \rho(y)dy \quad , \quad P_a(x) = H(x - x_a)$$

with $H(x)$ being the well-known Heaviside step function:

$$H(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1, & \text{for } x \geq 0 \end{cases}$$

So the CRPS measures the difference between predicted and occurred cumulative distributions, as it can be seen in Fig. 3.1

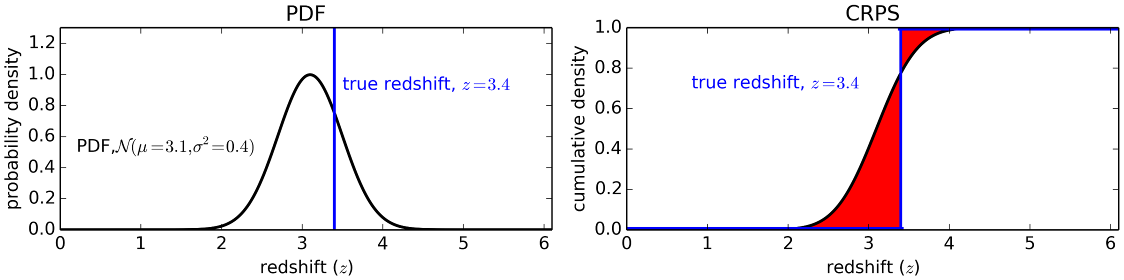


Figure 3.1: Example plot of the CRPS calculation between a Gaussian probability density distribution and a redshift target value (left plot). The cumulative distribution function and the Heaviside function are shown in the right plot and the CRPS corresponds to the integral between these two functions, given by the red area.

Usually, the CRPS is averaged over an area and a number of cases, as:

$$\overline{CRPS} = \sum_k w_k CRPS(P^k, x_a^k)$$

3.2 Probability integral transform

The probability integral transform [PIT Gneiting et al., 2007] is defined as the value that a predictive CDF attains at the target. Practically, it is a visual tool to verify how a set of observations or predictions can be modelled as coming from a particular distribution. It is based on the statement that if a random variable x_t has a continuous distribution, with its cumulative distribution function being $F_t(x_t)$, the PIT is defined as:

$$p_t = F_t(x_t)$$

and, in the ideal case, has a uniform distribution. Therefore the uniformity of the PIT is a necessary condition to achieve a perfect prediction, and this can be practically verified calculating the value p_t for every predicted PDF and plotting a histogram, as shown in Fig. 3.2.

The three different cases shown in Fig. 3.2 depict the reasons for prediction deficiency. Case *a* corresponds to the optimal case of a uniform distribution. Tecnically speaking, we refer to this case as *well calibrated*. The U-shaped histogram shown in figure *b* corresponds to distributions that, on average, are too narrow and it is affected by *underdispersion*. Figure *c* instead refers to the case of *overdispersion*, in which the predicted distributions are too wide. Clearly one could also find triangle-shaped histograms, which means that the distributions are biased.

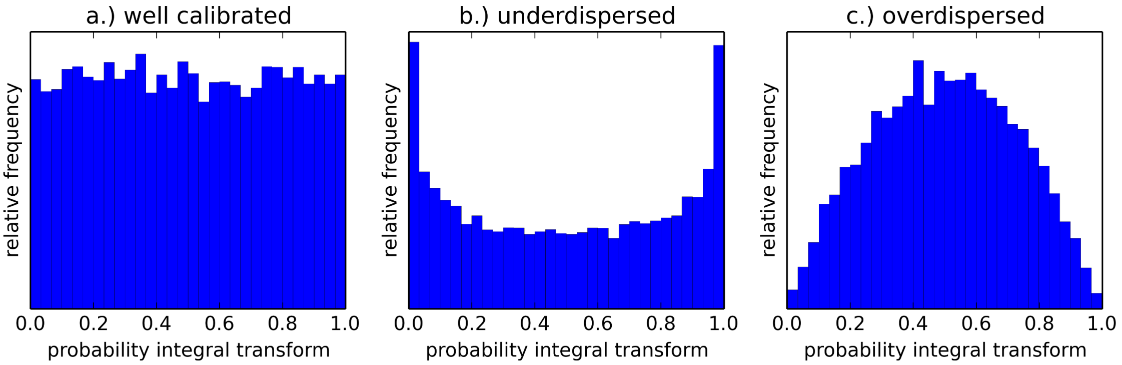


Figure 3.2: Three examples of PIT histograms. Case *a* refers to a well calibrated PIT. Case *b* shows a underdispersed prediction, in which the distributions are on average too narrow. Case *c* shows distributions affected by overdispersion, or too broad.

3.3 Calibration and sharpness

Following Gneiting et al. [2007], the basic idea behind the use of the CRPS and the PIT is to evaluate the performance of the predictions based on the paradigm of "maximizing the sharpness of the predictive distributions subject to calibration". Calibration refers to the statistical consistency between the predicted density distribution and the target value. Sharpness instead expresses how well the predicted distribution is concentrated. Therefore, calibration is a joint property of predictions and target, while sharpness is a property of the predictions only. Clearly, the more concentrated the predictions are, the sharper, and therefore better, they are, subject to a good calibration.

In general, there are several types of calibration. In this case, I will refer to the so called *probabilistic calibration*. Let's consider a CDF distribution G_t as the natural probability distribution of a certain phenomenon, from which x_t is a randomly drawn observed value (target). F_t is instead the CDF of the probabilistic predictive distribution. Therefore the sequence $(F_t)_{t=1,2,...}$ is probabilistically calibrated with respect to the sequence $(G_t)_{t=1,2,...}$ if:

$$\frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \longrightarrow p \quad \text{for all} \quad p \in (0,1)$$

Probabilistic calibration is then equivalent to having uniformity in the PIT values. Hence, when doing predictions, whatever method we choose, the best we can hope is that $F_t = G_t$. From a practical point of view, aiming for this goal means, as said before, maximizing the sharpness of the predictive distributions subject to calibration. This paradigm is defined as the *sharpness principle*, as expressed in Gneiting et al. [2007]. In this publication the authors deeply discuss the importance of using proper scores with respect to the sharpness principle and the optimization of the prediction performance. This is also related to particular conditions which can arise when dealing with real applications, like *conditional heteroscedasticity*. This concept expresses the existence of sub-populations characterized by different variances or statistical dispersion. In this cases, the width of prediction intervals can exhibit strong variability, and therefore the average width is not sufficient to characterize the sharpness. The combined use of proper scores, like the CRPS, and of visual diagnostic tools like the PIT, can help to solve such issues, in particular when dealing with probabilistic predictions of continues variables. The advantages given by the combined use of these tools will be further discussed in Sec. 5.4.

3.4 Proper scoring rules

As stated by Gneiting et al. [2007], the CRPS is a proper scoring rule. In general, a scoring rule assigns a numerical score to a certain probabilistic prediction, giving a summary measure of the performance of the prediction itself. Ideally, a scoring rule should address calibration and sharpness simultaneously. Defining $s(F, x)$ as the score given by a certain predicted distribution F with respect to the target x, the scoring rule is proper if $s(F, x)$ for the target x drawn from the natural distribution G is minimized if $F = G$. If this minimum is unique, then the scoring rule is defined as strictly proper. The definition can also be expressed by stating [Gneiting and Raftery, 2004] that:

$$s(G, G) \leq s(F, G) \tag{3.1}$$

where $s(F, G)$ is the expected value of s under G. Here it is assumed that s is a penalty scoring rule (e.g. the root mean square error), which has to be minimized in order to achieve the best result. In case of a reward (like for the likelihood), s has to be maximized, and therefore the definition becomes $s(G, G) \geq s(F, G)$. However, if this property is valid for every F and G, the scoring rule is defined as proper.

The CRPS is a proper score and with respect to other commonly used scoring rules proved to be sensibly more robust [Gneiting and Raftery, 2004]. In principle, as it will be demonstrated in the appendix of Publication II, the minimization of the negative log-likelihood is equivalently efficient and considered a proper score too. On the other hand, the log-likelihood is much focused on the spatial positioning of the predictions, while the CRPS is better related to sharpness.

Other examples of proper scoring rules are the already cited root mean square error (derived from the Brier score [Brier, 1950]) and the logarithmic score [Good, 1952]. A detailed analysis on proper scoring rules and, in particular, on the advantages of adopting the CRPS in the case of continuous variables and predictive densities is given by Gneiting and Raftery [2004].

3.5 Multimodalities and Gaussian mixtures

We have already seen how the estimation of photometric redshifts takes substantial benefits from a density distribution representation. We can assume such a density distribution to be Gaussian. On the other hand, estimating the redshift as a point estimate or as a single peak distribution would imply that from a given photometric set of features a single photometric redshift can be derived, i.e. there is a single redshift solution. Unfortunately, as showed in Polsterer [2017], this is not true. In fact, in many cases multimodality effects arise, allowing the association of multiple redshifts to the same source. In other words, for a given photometry, the solution is not unique and multiple redshifts are admissible, with varying probability. There are multiple reasons for this behavior. Different physical mechanisms can be at work and modify how the source looks with redshift. Ambiguities can be due also to a limited number of photometric measurements available and to the noise affecting them. Moreover, the broadband filter system used to derive the input features can itself cause degeneracies. For example, in the case of the SDSS, the *ugriz* system [Gunn et al., 1998], shown in Fig. 3.3, is subject to degeneracies coming from the width of the filters. In fact, several narrow features can fall in the same broad filter curve, making it hard to disentangle them. This effect can be immediately verified when looking at Fig. 3.4. Here three spectra of quasars at different redshifts are shown, on top of the same broadband filter curves depicted in Fig. 3.3. It can be noticed how the spectral lines move through the filter system due to the redshift effect. Moreover, it is evident that the degeneracy introduced by the width of the filter curves can be easily removed only in the narrow regions at the intersections between the filters. This explains the many attempts in the literature to get rid of degeneracies by using different features such as several magnitudes and colors. Theoretically this allows us to detect distinctive spectral characteristics (e.g. the Lyman or Balmer break) or to capture narrow features when passing in intersections between the filters.

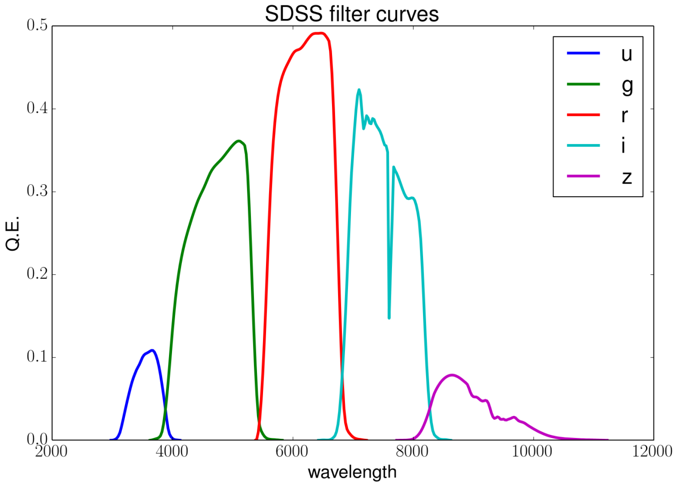


Figure 3.3: Filter curves defining the *ugriz* system of the SDSS

For all of these reasons, it is evident that when predicting redshift PDFs, one should take into account multimodalities. The solution proposed in this thesis is the adoption of a Gaussian mixture model [GMM Mclachlan and Basford, 1988] to fit the input photometry to the spectroscopic target. A GMM is a probability distribution defined as:

$$p(x) = \sum_{m=1}^{M} w_m \mathcal{N}(\mu_m, \sigma_m^2) \tag{3.2}$$

where $\mathcal{N}(\mu_m, \sigma_m^2)$ is a normal distribution and w_m is its weight. The use of the GMM is particularly convenient with respect to the adoption of the CRPS as score function. In fact, in the case of a Gaussian distribution, the CRPS can be easily calculated with the formula:

$$CRPS\Big(\mathcal{N}(\mu, \sigma^2), x\Big) = \frac{\sigma}{\sqrt{\pi}}\left(1 - \sqrt{\pi}\frac{x-\mu}{\sigma}\text{erf}\Big(\frac{x-\mu}{\sqrt{2\sigma^2}}\Big) - \sqrt{2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)$$

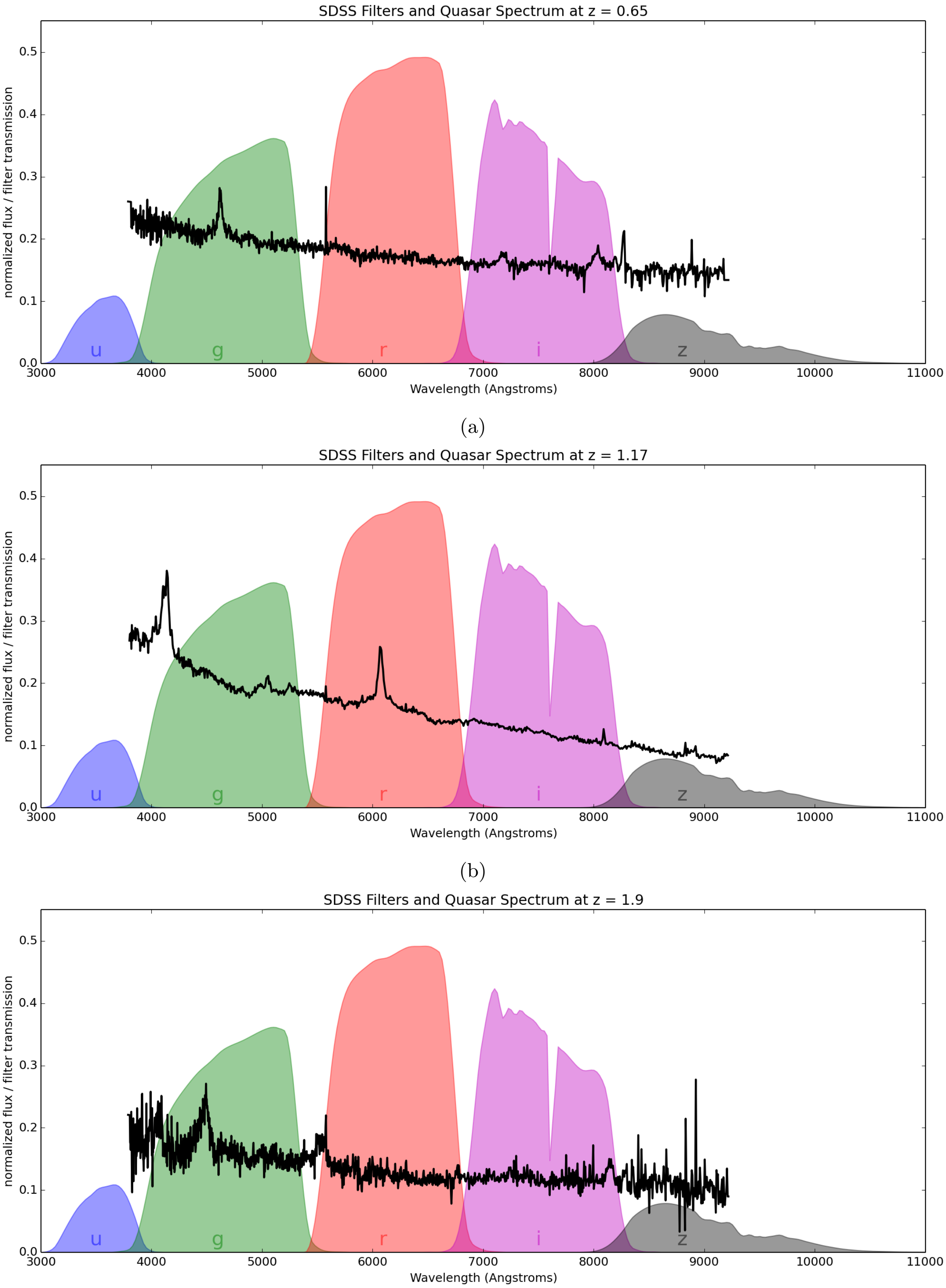
as in Gneiting and Raftery [2004], and with erf indicating the error function. Expressing the Gaussian probability density and its CDF as ϕ and Φ respectively, the previous relation becomes:

$$CRPS\Big(\mathcal{N}(\mu, \sigma^2), x\Big) = \sigma\left(\frac{x-\mu}{\sigma}\left(2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1\right) + 2\phi\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)$$

Similarly it is possible to calculate the CRPS for a GMM, as showed in P. Grimit et al. [2006], obtaining:

$$CRPS\left(\sum_{m=1}^{M} w_m \mathcal{N}(\mu_m, \sigma_m^2), x\right) = \sum_{m=1}^{M} w_m A(x-\mu_m, \sigma_m^2) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} w_m w_n A(\mu_m - \mu_n, \sigma_m^2 + \sigma_n^2)$$

where w_i is the weight of the i-th member of the mixture and:



(a)

(b)

(c)

Figure 3.4: The plots show the spectra of three quasars at different redshifts on top of the SDSS filter curves. It can be noticed how the shift of the spectrum moves the different spectral features through the broadband filters and the degeneracy caused by the width of the filters. The only regions in which the degeneracy is removed are the intersections between neighbouring filter curves.

$$A(\mu,\sigma^2) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right)$$

expresses the expectation for the absolute value of a normal random variable having mean μ and variance σ. With this background, the CRPS can not only be implemented as a score function to estimate the performance of the adopted machine learning models, but also as a loss function, to perform the training of the neural networks, which will be extensively shown in the publications presented in the next chapter.

3.6 Feature selection

Machine learning models require input data in which the original information is condensed to a restricted number of representative parameters. This has to be done in order to make the information itself and the models more interpretable for humans, to reduce the training time and to avoid the *curse of dimensionality* [Bellman, 1961]. This expression refers to the fact that, intuitively, in high-dimensional spaces the increase of dimensions generates a faster increase of the space volume, causing the data to become sparse. Therefore, the data are no longer representative and the results can become poor or unreliable. On the other hand, increasing the number of input features is necessary to increase the generalization capability of the model, avoiding overfitting or, for example, to break degeneracies, as in the case of photometric redshift estimation. For this reason, one should always try to find a decent balance between increasing the dimensionality of the input space and limiting the number of parameters to avoid the curse of dimensionality.

In this sense, feature selection is a fundamental step to build a reliable and efficient model. Typically it follows the related step of feature extraction, which is used to express the original information in a number of parameters, like, for example, magnitudes and colors extraction from astronomical images. Feature selection is meant to select only the necessary parameters, removing those which are shown to not improve the result, are redundant or highly correlated. There are plenty of methods to perform this task, from a manual selection to algorithms based on selection criteria and completely automated methods. Feature selection has the advantage to not transform the input data, in contrast to other techniques, e.g. applying principal component analysis. This can be preferable when the meaning of the features is relevant in order to find relationships between the parameters for a better comprehension of the physics of the problem. Feature selection methods can be roughly divided in three categories. Filter methods are completely independent from the model used to solve the problem or perform the predictions and are based on correlation coefficients to rank the features. That is to say, they do not involve specific learning methods, relying only on general characteristics of the data to evaluate and select the most reliable feature subsets. Wrapper methods instead are based on the performance of the same machine learning model selected for the task. Therefore, they select the features which best fit with that specific model. For example, the greedy forward selection [Pahikkala et al., 2010] used for the feature selection presented in Publication III is a wrapper method. Finally, embedded methods are a group of techniques which perform the feature selection as part of the model building process. The feature selection automatically performed by the DCMDN, as showed in Publication II, is an embedded method. A good overview about the field can be found in Guyon and Elisseeff [2003]. A detailed description of these methods goes beyond the scope of this thesis but, as the problem also involves machine learning models applied to astronomical problems, I want to give a brief overview of the major milestones in this field.

An important work is that by Donalek et al. [2013], in which the authors apply several feature selection techniques on data from the Catalina Real-Time Transient Survey [CRTS Djorgovski et al., 2011] and Kepler [Borucki et al., 2010] to classify transient events. In this work five different methods are adopted to select the best performing features for the classification with several machine learning models. The results are then compared with those obtained using all the available features, generally showing an improvement in the performance when finding the best feature selection. The problem of feature selection has been treated in detail in the literature, not just in relation to the solution of a specific problem, but also with the aim of defining a specific set of features which could be used in general by the community. For example, in the case of time series analysis, the work from Richards et al. [2011] is particularly relevant, as the authors define a set of periodic and non-periodic features from light curves, which have become widely used in the community. The use of features is fundamental when dealing with light curves, or time series in general, as the original data are typically not regularly sampled nor observed with the same number of epochs and signal-to-noise ratio. Therefore features represent a homogeneization of the data, which are transformed to a vector of real numbers by using statistical and model fitting procedures. The same features have been used, for example, in D'Isanto et al. [2016] in order to develop a novel method of feature selection, based on the random forest feature importance, to classify transient sources from the CRTS with neural networks. In particular, this work is focused on the identification of cataclismic variables, of supernovae and the separation between galactic and extra-galactic sources.

With regard to photometric redshift estimation, many works, even some reported in Sec. 2.2, focus on the choice of parameters used in order to improve the performance. The features proposed in Laurino et al. [2011], which are point spread function (psf) magnitudes, model magnitudes and the related colors, have become almost a standard in machine learning based models used to derive photo-z. Instead, in Polsterer et al. [2014], a different approach is adopted in order to find the best features for photometric redshift estimation. The authors perform a brute force combination of several photometric parameters, and select the most efficient with a greedy forward selection method. This constitutes the basis for the work presented in Publication III, as it permits us to make a better use of the photometry contained in the SDSS catalog, improving the general performance, and to find new features, which can be interpreted in a physical sense. The new features are based not only on magnitudes and colors but on different combinations of them and they include further photometric information too, like radii and ellipticities. The improvement obtained by considering additional photometric information in the feature selection is further demonstrated by Hoyle et al. [2015b]. In this work the selection is based on the feature importance estimated by the Gini index [Gini, 1912] as it is calculated in Decision Trees from the scikit-learn [Pedregosa et al., 2011] implementation. In Publication III a comparison between the feature ranking obtained with the proposed method and the importance calculated with the Gini index as used in the RF model is also performed.

3.7 Artificial neural networks

A detailed description of machine learning models would be beyond the scope of this thesis, but in this section I want to give the rudiments, without any claim of completeness, about artificial neural networks (ANN), which are useful to understand the models presented in Chapter 4.

ANNs are a class of machine learning models inspired by the structure of the biological brain [McCulloch and Pitts, 1943]. In fact, the main calculation unit is called *neuron*, and neurons are interconnected, in order to transmit signals, like it happens for synapses in the biological brain. Typically, neural networks are based on supervised learning and can be used for classification and regression tasks. One of the simplest ANN models is the multilayer perceptron (see Fig. 3.5), composed of an input layer, one or more hidden layers and an output layer. Every layer is composed of several neurons. All the neurons are fully-connected to all the neurons of the sub-sequent layer and every connection includes a weight.

ANNs in general, and the MLP more specifically, are basically complex functions of the inputs $\mathbf{x}$, parametrized by weights $\mathbf{w}$, i.e. the parameters optimized during the training, and a particular activation function f, which is the function applied to the input of every neuron. Following Bishop [1995], the relation between input and output, for a MLP like that shown in Fig. 3.5, is given by:

$$y_k(\mathbf{x}, \mathbf{w}) = g\Big(\sum_j w_{kj}^{(2)} f\Big(\sum_i w_{ji}^{(1)} x_i + b_j^{(1)}\Big) + b_k^{(2)}\Big)$$

Here b indicates a bias added to the sum and the function g used for the output typically is the identity for regression problems and a logistic or softmax function for classification problems. In general, the activation functions most commonly used are

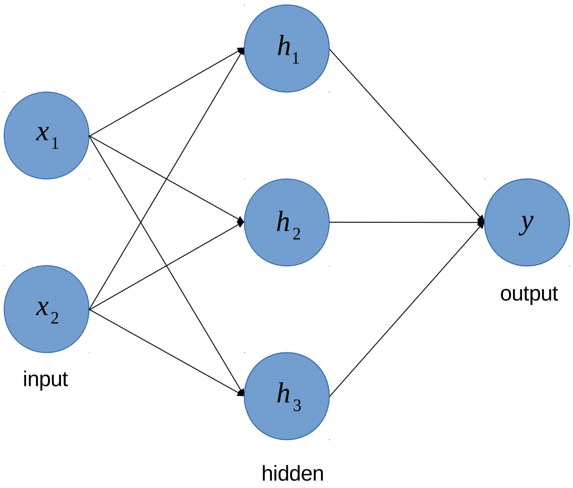


Figure 3.5: Example of a multilayer perceptron characterized by two input neurons, one hidden layer with three neurons and a single output.

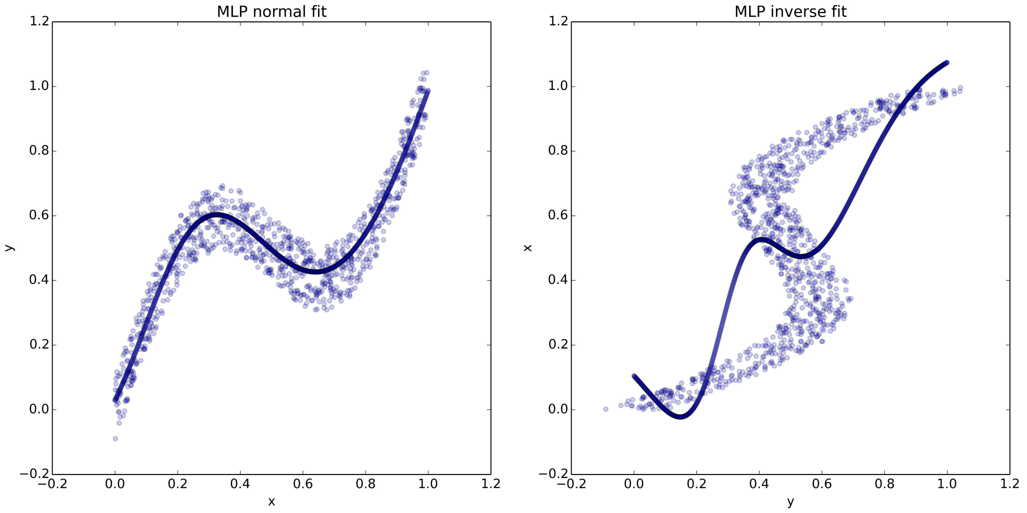
$$\tanh = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad , \quad sigmoid = \frac{1}{1 + e^{-z}} \quad , \quad ReLu = max(0, z)$$

namely the hyperbolic tangent, the sigmoid and the rectified linear unit. The process of evaluating y_k is commonly referred as *forward propagation*. Therefore the network is trained to optimize the weights, minimizing the loss function between the predictions and the targets, in our case the spectroscopic redshifts. Once the network is trained, it is frozen, so not updating anymore the parameters, and used to perform the prediction on an unknown dataset. In other words, the neural network performs a high dimensional interpolation from which the model learns to generalize the task by means of the training sample. For this reason, it is very important that the training set used is well representative with respect to the specific problem. For example, in the case of photometric redshift estimation the feature space needs to be representative of the the photometric and physical properties of the sources. If this should not be the case the model could have problems generalizing correctly, losing accuracy. Typically, the minimization of the loss function is achieved by adopting the *backpropagation algorithm* [Rumelhart et al., 1988], during which the weights are updated using the so called gradient descent. If $LF = LF(y(x), x_t)$ is the loss function between the predictions y and the targets x_t, then the weights are updated following the rule:

$$W_i' = W_i - \eta \nabla (LF)$$

where the parameter η, called *learning rate*, represents, intuitively, how quickly or slowly the update of the weights is done. The choice of the learning rate is particularly important, as it is related with the capability of the network to converge efficiently and to avoid local minima.

In the papers cited in Sec. 2.3.2, the MLPs adopted for the experiments have a quite simple structure, based on few hidden layers and a limited number of neurons. The root mean square error (RMSE) is typically used as error function. This is a common choice when predicting point estimates, but I will deeply analyze in Publication II and in Sec. 5.4 the reasons why it is not ideal when dealing with probability density functions. The use of *deep* architectures, with many fully connected layers, became popular only in the recent years, due to the increase of the data complexity and hence the need to apply more elaborate models that can deal such complexity. However, a fundamental role was played by the increase of computational power and the improvements in the field of parallel computing, as it will be discussed in Sec. 3.8.



(a)

(b)

Figure 3.6: Example representing the reconstruction of a noisy sinusoidal function performed with a MLP, from x to y (a), and from y to x (b). In case (b) the prediction is poor as the relation we are trying to approximate is not a function, as the same input value can correspond to multiple output values.

3.7.1 Mixture density network

An important extension of the MLP is the so called mixture density network [MDN Bishop, 1994, 1995]. Typically the output of a MLP is modelled as a single Gaussian density (regression task) or as a Bernoulli distribution (classification task). In constrast, the MDN outputs are input dependent density distributions, modelled as a Gaussian mixture model, given by Eq. 3.2. Following Bishop [1995], if $\mathbf{y}$ are the outputs of the neural network, then the MDN will define the means, variances, and mixing coefficients of a GMM with K components, respectively, as:

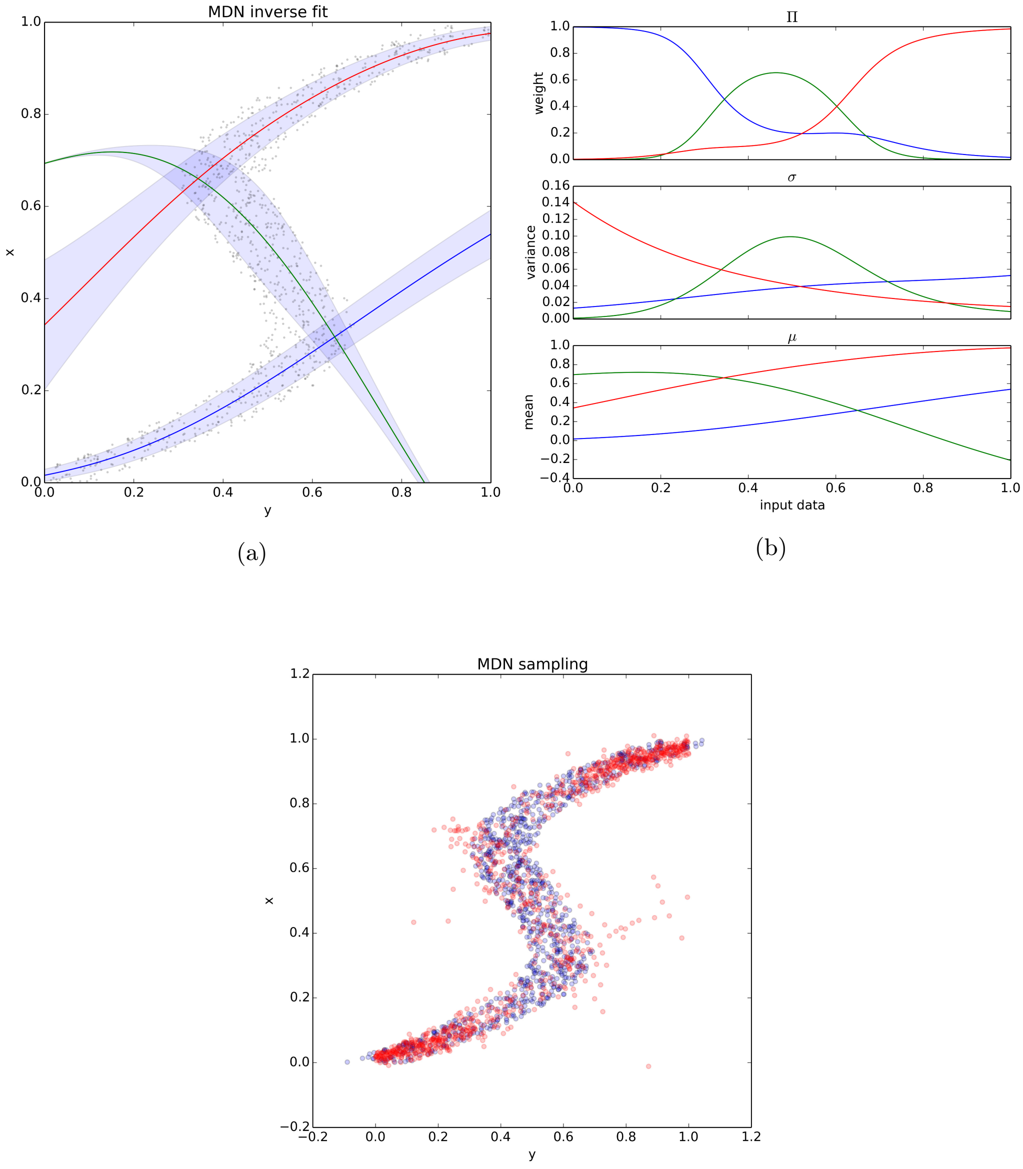
$$\mu_k = y_k^{\mu} \quad , \quad \sigma_k = \exp(y_k^{\sigma}) \quad , \quad \pi_k = \frac{\exp(y_k^{\pi})}{\sum_{l=1}^{K} \exp(y_l^{\pi})}$$

where the mixing coefficients must satisfy:

$$\sum_{k=1}^{K} \pi_k = 1 \quad , \quad 0 \leq \pi_k \leq 1$$

The MDN is particularly useful when dealing with data characterized by multimodalities, and for this reason in Publication I I will present how to build a model composed by combining a convolutional neural network and a mixture density network for photometric redshift estimation.

In Fig. 3.6 and Fig. 3.7, an example taken from Bishop [1994] is given, in order to clarify the utility of MDNs in such a case. In Fig. 3.6a I plot noisy data generated from a sinusoidal function. The task is a regression problem in which we would like to reconstruct the underlying function. The function is approximated with a MLP (with a single output neuron), which learns to predict the values y given the input x. The prediction, as shown by the plot, is good enough. Now we consider the inverse problem, where the y axis is swapped with the x axis. In other words, y are now the inputs and x are the outputs. As shown in Fig. 3.6b, training the same MLP on this task, the prediction gives poor results. The reason why the MLP does not perform well is the fact that the relation we are trying to approximate is no longer a function, as the same input value can correspond to multiple output values. Therefore, to perform this prediction correctly, a



(a)

(b)

(c)

Figure 3.7: Example representing the reconstruction of the same noisy sinusoidal function from Fig. 3.6, performed with a MDN based on a GMM with $K = 3$ Gaussian components. In (a) the results are shown, while in (b) the behavior of the parameters of the GMM with respect to the input data is depicted. In (c) a sample from the MDN predictions is shown with respect to the original data.

model able to predict multiple outputs for each input is required. The standard MLP is not able to fullfil such a requirement. For this reason, a mixture density network has to be used. Here I use a MDN characterized by a GMM with $K = 3$ Gaussian components and the likelihood as loss function. In Fig. 3.7a and Fig. 3.7b the results are given. It can be noticed that for every input data point, the output can be given by any of the k predicted means μ_k (bottom plot in Fig. 3.7b), each characterized by a probability represented by the mixing coefficients. The mixing coefficiens indicate the regions of input space for which each mean μ_k is responsible for modelling the data. In particular, Fig. 3.7b shows how the GMM parameters vary with respect to the input data. Finally, in Fig. 3.7c a sample from the MDN predictions is shown with respect to the original data. It can be noticed that the shape is captured very well by the sampled data. This example clarifies the necessity to use a mixture density network, based on a probabilistic model, in order to deal with multimodalities.

3.8 Parallel computing

Most of the work presented in this thesis is based on an intensive application of parallel computing, performed by adopting graphics processing units (GPUs). This type of computation is carried out by executing many calculations or processes concurrently. Typically this is done by breaking a computational task in several similar subtasks, which can be processed independently and whose results can be combined afterward. It can be performed on a single machine equipped with multi-core or multi-processor hardware, or on multiple machines, arranged in a cluster or grid structure, which work on the same task. Until the first years of the 2000s, frequency scaling was the principal way to increase a computer performance, as the runtime of a program is trivially given by the number of instructions multiplied by the average time per instruction. The problem in this kind of approach lies in the fact that power consumption P is directly related to the processor frequency F by the relation:

$$P = C \times V^2 \times F$$

where C is the capacitance per clock cycle and V is the voltage. Therefore, increases in frequency imply increases in power consumption, and this brought about a drastic change in the way improvement of performances in computing is obtained. Producers started to develop (also in the desktop sector) central processing units (CPUs) characterized by multiple cores in order to deal with the problems of power consumption and overheating.

Ideally, one would expect that the speed-up obtained from parallelization should be linear. This means that doubling the number of processing units should half the runtime. Unfortunately, this is only partially true and the speed-up is almost linear only for a small number of units, flattening out as they increase. The relation expressing the potential speed-up of an algorithm on a parallel platform is given by Amdahl's law [Amdahl, 1967, Rodgers, 1985]:

$$S_{latency}(s) = \frac{1}{1 - p + \frac{p}{s}}$$

where $S_{latency}(s)$ is the potential speed-up in latency to execute the whole task, s is the speed-up related only to the parallelizable part of the task and p is the percentage of execution time that the part of the task affected by the parallelization originally had. Amdahl's law is shown graphically in Fig. 3.8. The two relations:

$$S_{latency}(s) \leq \frac{1}{1-p} \quad , \quad \lim_{s \to \infty} S_{latency}(s) = \frac{1}{1-p}$$

show, on one hand, that the theoretical speed-up in the execution of the task increases by improving the resources and, on the other hand, that the part of the task which cannot be improved by the parallelization constitutes always a limit for the speed-up itself.

For our purposes, I would like to focus on parallelization performed with GPUs. This is also addressed as *general purpose computing on graphics processing units* (GPGPU). Typically GPUs are applied for computers graphics and their use for different applications only recently became a trend in computer science and engineering. There is to say that the availability of GPU-based solutions contributed enormously to the growth in popularity of deep learning applications in every field. The reason lies in the fact that computer graphics is mainly based on algebric matrix operations, which is used in most deep learning models, for example in convolutional neural networks [Lecun et al., 1998]. In the case of neural networks, the speed up is mainly due to the parallelization of the stochastic gradient descent [Kiefer and Wolfowitz, 1952], based on the calculation of the gradient in minibatches, during the backpropagation phase. The gradient, in fact, is typically not calculated on the whole data sample, as this operation would be too much computationally demanding. The data are divided in minibatches and the gradient is calculated for each object in the minibatch,

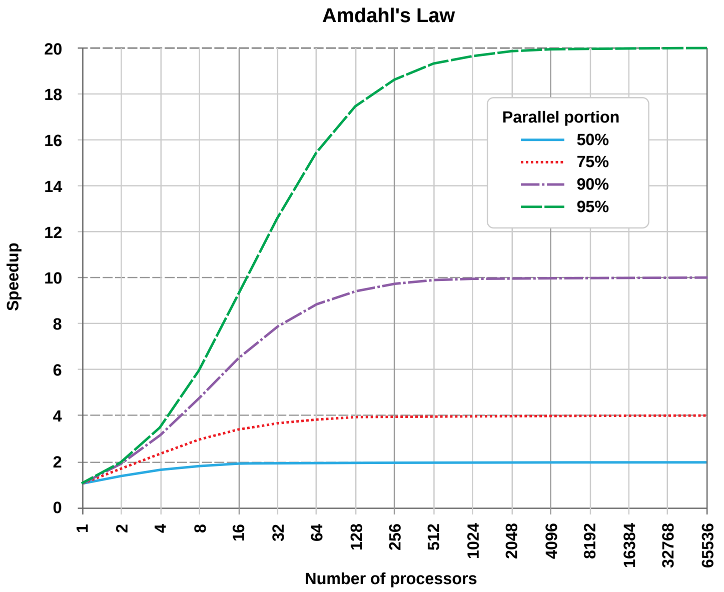


Figure 3.8: Evolution of the theoretical speed-up in latency, according to Amdahl's law, of a certain task with respect to the number of processors used and for different percentages of p. Image from Wikipedia.

taking the mean of the values obtained. When the calculation in a minibatch is completed, the weights are updated, and the procedure is repeated for the subsquent minibatch. As the weights are fixed for the calculation in every minibatch, this operation can be parallelized, performing it simultaneously for every object contained in the minibatch.

A GPU is a programmable heterogeneous multi-processor chip, highly tuned for graphics. Their architecture based on multiple cores is meant to deal with vertex, primitive and fragment generation and processing, in order to create 3D graphics. In particular, fragment operations are based on shaders [Pixar, 1989, Upstill, 1989]. These are a class of algorithms which were originally meant to deal with the appropriate levels of light, darkness and color within an image. Nowadays they are widely used in computer graphics for a number of specialized functions, and even for tasks not related to graphics at all. By using shaders, the elements which characterize an image (position, hue, saturation, brightness, contrast, vertices, or textures) can be customized on the fly. This can be done using algorithms specifically defined in the shader and characterized by a high degree of flexibility. All these parameters can therefore be modified by external variables or textures introduced by the program calling the shader. However, each shader needs to be processed independently, and cannot be explicitly parallelized. In other words, every fragment is characterized by an independent logical sequence of control. From this comes the necessity of having multiple cores which each control a single fragment process, allowing parallelization. Older cards needed to utilize separate processing units for each different shader type (geometry, vertex, pixel, etc.). Modern GPUs are instead based on the *Unified Shader Model*, characterized by the fact that all the shader stages in the rendering pipeline possess the same capabilities. Therefore GPUs are designed to work on textures mapping and polygons rendering, rotation and translation of vertices, manipulation of the same vertices and textures using shaders, oversampling and interpolation techniques to reduce aliasing. In other words, they handle all the visual elements that characterize gaming, videos, graphic softwares and so on. Basically, all these tasks are based on matrix and vector operations, and this makes GPUs so suitable for other kind of applications which demand a high level of parallelization for such calculations. A modern dedicated GPU typically interfaces with the motherboard by the PCI Express (PCIe) slot. The term *dedicated* refers to the fact that the GPU has its own RAM dedicated to the card's use. Furthermore, with the increase in the employment of GPUs for different purposes, a dedicated cache has also been included in the new generation cards.

The increase in terms of performance with respect to CPU based applications is dramatic. This

Type	Hardware	Running time/epoch
CPU	Intel Core i7 - 1 core	4h 8m
CPU	Intel Core i7 - 8 cores	82m
GPU	Nvidia Titan X	90 s
GPU	Nvidia Pascal P100	83 s
GPU	Nvidia Pascal P40	81 s

Table 3.1: Running time per epoch in a photometric redshift estimation experiment using the deep learning model presented in the publications of this thesis. The performance of different architectures, based on CPU and GPU, are compared.

can be seen from Tab. 3.1, where the execution time of a single epoch for the deep learning model presented in the publications of this thesis in a photometric redshift estimation experiment is compared with respect to different hardware architectures. The running time per epoch goes from the order of hours, for CPU based architectures, to seconds, in the presence of dedicated GPUs. Considering that the convergence of the model requires at least some hundreds or thousands of iterations, it is clear that such technologies can be successfully adopted only in combination with an intensive use of GPU computing. Clearly, one has always to deal with possible bottlenecking effects which can limit the performance and take into account that Amdahl's law is valid also in the case of parallelization with GPUs. Therefore the availability on the market of powerful GPUs at reasonable prices is one of the main reasons behind the explosion in the interest in deep learning solutions. Several implementations have been developed in order to make use of such technologies in many fields, from image to speech recognition, using common smartphone or powerful clusters. In the last few years, astronomy has started to take advantage of these techniques too, and GPU computing and deep learning solutions are constantly increasing in popularity in the community, in several fields of research, from the experimental and observational side to the simulation one. To use Prof. Alex Szalay's words, "modern science is approaching the point where novel computational algorithms and tools, combined with computational thinking, will become as indispensable as mathematics" [Szalay, 2011].

Chapter 4

Publications

4.1 Overview

The present thesis is written in cumulative form, as allowed from the 'Appendix 2, to §7 (2) Regulations concerning the conferral of doctoral degrees / *Promotionsordnung*' of the University of Heidelberg. Three publications are required for the submission of a cumulative thesis, and during my PhD studies I have been involved in the articles hereby reported. In all these publications I am first author and they have been published in well-known and refereed scientific journals or conference proceedings. The articles are included here in the same form they have been published in the indicated journals. None of these publications has been or will be used in a cumulative thesis of another co-author.

Publication I

Title: 'Uncertain photometric redshifts via combining deep convolutional and mixture density networks.'

Authors: Antonio D'Isanto, Kai Lars Polsterer

This article has been published in ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. It has been reviewed by three independent referees.

All rights reserved.

The publication is accessible via `https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2017-56.pdf`

Publication II

Title: 'Photometric redshift estimation via deep learning - Generalized and pre-classification-less, image based, fully probabilistic redshifts'

Authors: Antonio D'Isanto, Kai Lars Polsterer

This article has been published in Astronomy & Astrophysics (A&A).

Credit: A. D'Isanto & K.L. Polsterer, A&A, 609, A111, 2018, reproduced with permission ©ESO.

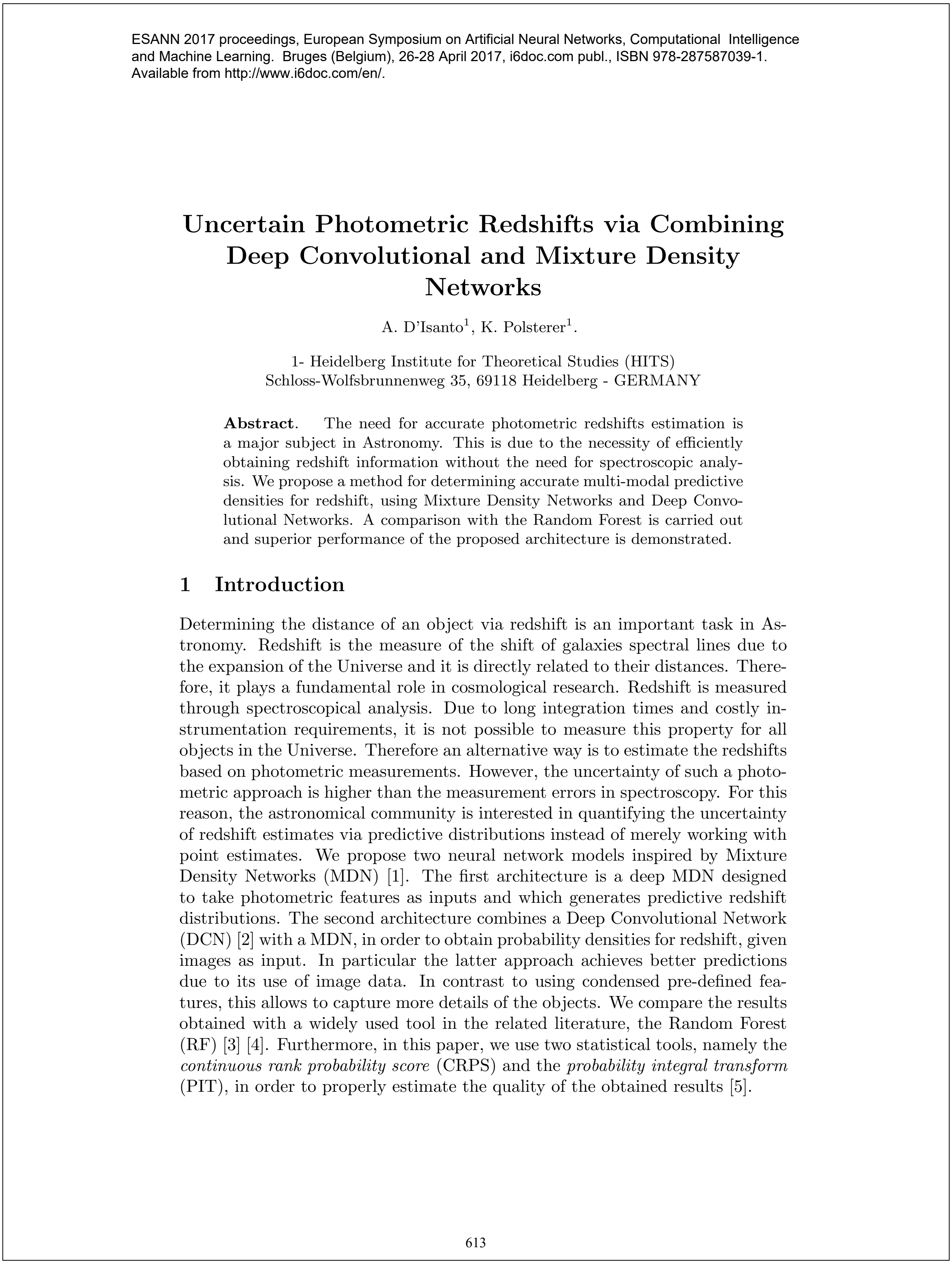
Publication III

Title: 'Return of the features - Efficient feature selection and interpretation for photometric redshifts'

Authors: Antonio D'Isanto, Stefano Cavuoti, Fabian Gieseke, Kai Lars Polsterer

This article has been published in Astronomy & Astrophysics (A&A).

Credit: A. D'Isanto, S. Cavuoti, F. Gieseke and K. L. Polsterer, 2018, A&A, 616, A97, reproduced with permission ©ESO.



ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Uncertain Photometric Redshifts via Combining Deep Convolutional and Mixture Density Networks

A. D'Isanto[1], K. Polsterer[1].

1- Heidelberg Institute for Theoretical Studies (HITS)
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg - GERMANY

Abstract. The need for accurate photometric redshifts estimation is a major subject in Astronomy. This is due to the necessity of efficiently obtaining redshift information without the need for spectroscopic analysis. We propose a method for determining accurate multi-modal predictive densities for redshift, using Mixture Density Networks and Deep Convolutional Networks. A comparison with the Random Forest is carried out and superior performance of the proposed architecture is demonstrated.

1 Introduction

Determining the distance of an object via redshift is an important task in Astronomy. Redshift is the measure of the shift of galaxies spectral lines due to the expansion of the Universe and it is directly related to their distances. Therefore, it plays a fundamental role in cosmological research. Redshift is measured through spectroscopical analysis. Due to long integration times and costly instrumentation requirements, it is not possible to measure this property for all objects in the Universe. Therefore an alternative way is to estimate the redshifts based on photometric measurements. However, the uncertainty of such a photometric approach is higher than the measurement errors in spectroscopy. For this reason, the astronomical community is interested in quantifying the uncertainty of redshift estimates via predictive distributions instead of merely working with point estimates. We propose two neural network models inspired by Mixture Density Networks (MDN) [1]. The first architecture is a deep MDN designed to take photometric features as inputs and which generates predictive redshift distributions. The second architecture combines a Deep Convolutional Network (DCN) [2] with a MDN, in order to obtain probability densities for redshift, given images as input. In particular the latter approach achieves better predictions due to its use of image data. In contrast to using condensed pre-defined features, this allows to capture more details of the objects. We compare the results obtained with a widely used tool in the related literature, the Random Forest (RF) [3] [4]. Furthermore, in this paper, we use two statistical tools, namely the *continuous rank probability score* (CRPS) and the *probability integral transform* (PIT), in order to properly estimate the quality of the obtained results [5].

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

2 Statistical tools: CRPS and PIT

In this section we briefly describe the statistical tools used to evaluate the predictions of the proposed models. As discussed in [6], a predictive distribution explains well an observation if it is well calibrated and sharp; as stated in [6], *calibration expresses the consistency between predictions and observations, while sharpness refers to the concentration of the predictions in the probability distribution.* CRPS quantifies both desired properties, while the PIT provides a visual appreciation of them. The CRPS [7] is meant to compare a distribution with an observation (see Fig.1):

$$CRPS = CRPS(F, x_a) = \int_{-\infty}^{+\infty} [F(x) - F_a(x)]^2 dx \tag{1}$$

where $F(x)$ and $F_a(x)$ represent respectively the cumulative density functions (CDFs) of the probability density function (PDF) and of the observation, namely: $F(x) = \int_{-\infty}^{x} f(t)dt$ and $F_x = H(x - x_a)$, with $H(x)$ being the Heaviside step-function. We use the CRPS as a score function to express the results of the predictions and as a loss function for the proposed neural networks.

The PIT is defined by the value given by the CDF of the predictions F_t at the observation x_t, that is to say: $p_t = F_t(x_t)$. If the predictions are ideal, then

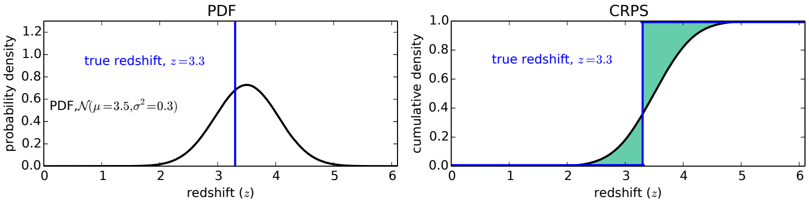


Fig. 1: Meaning of probability density function (*PDF*) and continuous ranked probability score (*CRPS*).

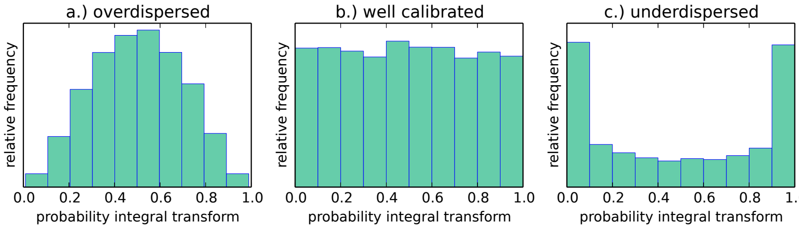
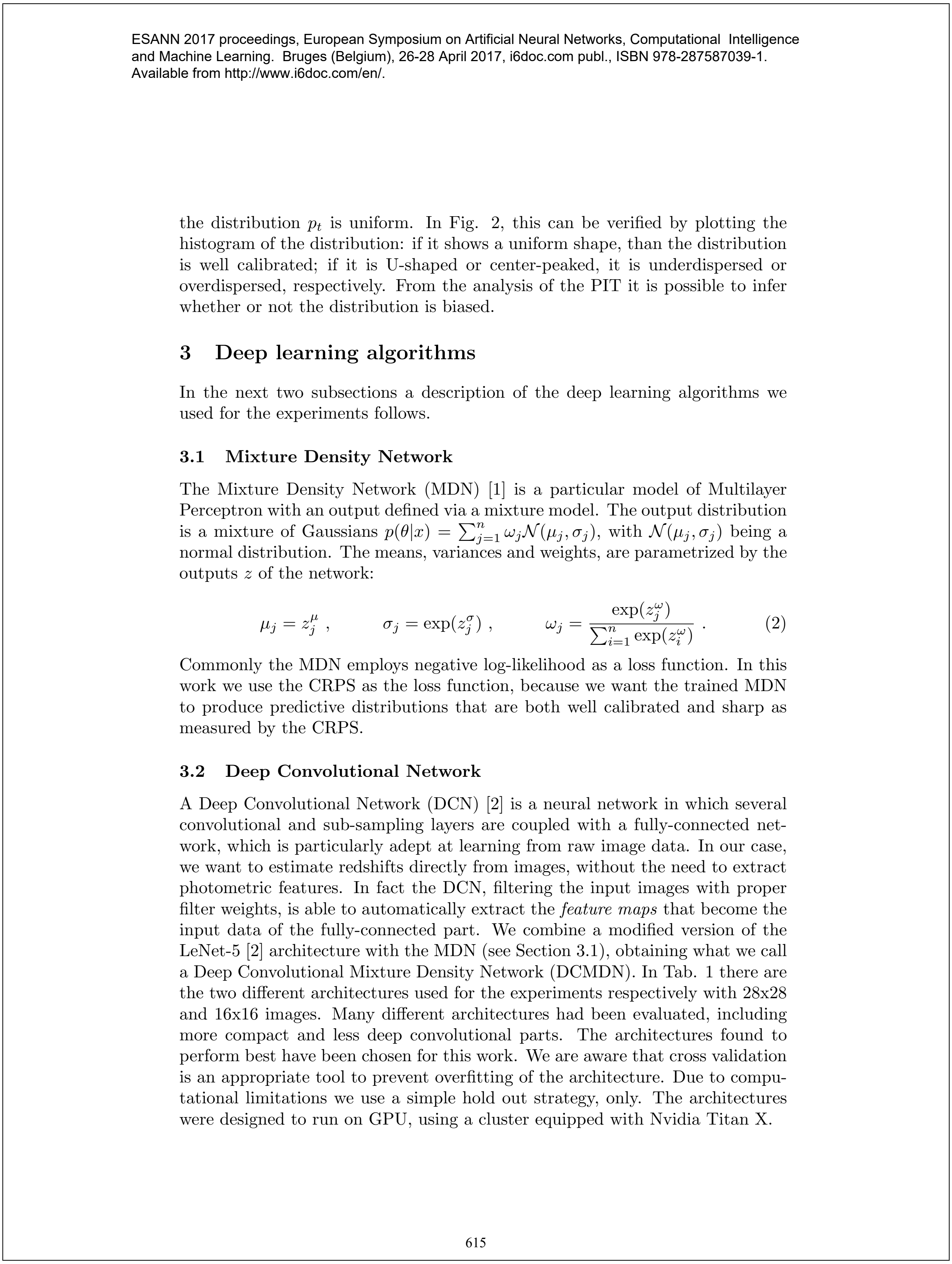


Fig. 2: Three different examples of probability integral transforms (*PIT*s) for overdispersed, well calibrated and underdispersed distributions.



ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

the distribution p_t is uniform. In Fig. 2, this can be verified by plotting the histogram of the distribution: if it shows a uniform shape, than the distribution is well calibrated; if it is U-shaped or center-peaked, it is underdispersed or overdispersed, respectively. From the analysis of the PIT it is possible to infer whether or not the distribution is biased.

3 Deep learning algorithms

In the next two subsections a description of the deep learning algorithms we used for the experiments follows.

3.1 Mixture Density Network

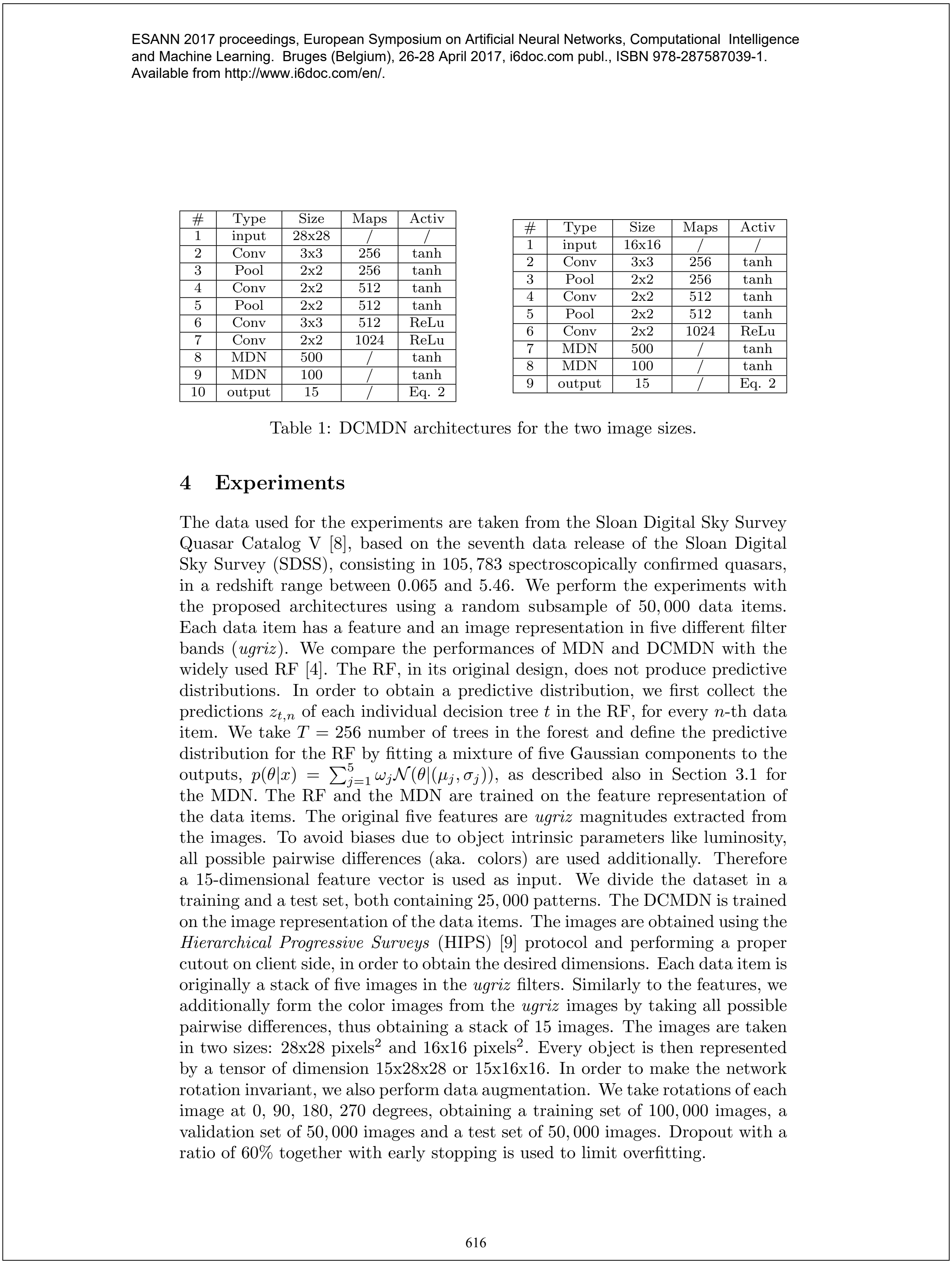
The Mixture Density Network (MDN) [1] is a particular model of Multilayer Perceptron with an output defined via a mixture model. The output distribution is a mixture of Gaussians $p(\theta|x) = \sum_{j=1}^{n} \omega_j \mathcal{N}(\mu_j, \sigma_j)$, with $\mathcal{N}(\mu_j, \sigma_j)$ being a normal distribution. The means, variances and weights, are parametrized by the outputs z of the network:

$$\mu_j = z_j^{\mu} \ , \qquad \sigma_j = \exp(z_j^{\sigma}) \ , \qquad \omega_j = \frac{\exp(z_j^{\omega})}{\sum_{i=1}^{n} \exp(z_i^{\omega})} \ . \tag{2}$$

Commonly the MDN employs negative log-likelihood as a loss function. In this work we use the CRPS as the loss function, because we want the trained MDN to produce predictive distributions that are both well calibrated and sharp as measured by the CRPS.

3.2 Deep Convolutional Network

A Deep Convolutional Network (DCN) [2] is a neural network in which several convolutional and sub-sampling layers are coupled with a fully-connected network, which is particularly adept at learning from raw image data. In our case, we want to estimate redshifts directly from images, without the need to extract photometric features. In fact the DCN, filtering the input images with proper filter weights, is able to automatically extract the *feature maps* that become the input data of the fully-connected part. We combine a modified version of the LeNet-5 [2] architecture with the MDN (see Section 3.1), obtaining what we call a Deep Convolutional Mixture Density Network (DCMDN). In Tab. 1 there are the two different architectures used for the experiments respectively with 28x28 and 16x16 images. Many different architectures had been evaluated, including more compact and less deep convolutional parts. The architectures found to perform best have been chosen for this work. We are aware that cross validation is an appropriate tool to prevent overfitting of the architecture. Due to computational limitations we use a simple hold out strategy, only. The architectures were designed to run on GPU, using a cluster equipped with Nvidia Titan X.



#	Type	Size	Maps	Activ
1	input	28x28	/	/
2	Conv	3x3	256	tanh
3	Pool	2x2	256	tanh
4	Conv	2x2	512	tanh
5	Pool	2x2	512	tanh
6	Conv	3x3	512	ReLu
7	Conv	2x2	1024	ReLu
8	MDN	500	/	tanh
9	MDN	100	/	tanh
10	output	15	/	Eq. 2

#	Type	Size	Maps	Activ
1	input	16x16	/	/
2	Conv	3x3	256	tanh
3	Pool	2x2	256	tanh
4	Conv	2x2	512	tanh
5	Pool	2x2	512	tanh
6	Conv	2x2	1024	ReLu
7	MDN	500	/	tanh
8	MDN	100	/	tanh
9	output	15	/	Eq. 2

Table 1: DCMDN architectures for the two image sizes.

4 Experiments

The data used for the experiments are taken from the Sloan Digital Sky Survey Quasar Catalog V [8], based on the seventh data release of the Sloan Digital Sky Survey (SDSS), consisting in $105,783$ spectroscopically confirmed quasars, in a redshift range between 0.065 and 5.46. We perform the experiments with the proposed architectures using a random subsample of $50,000$ data items. Each data item has a feature and an image representation in five different filter bands (*ugriz*). We compare the performances of MDN and DCMDN with the widely used RF [4]. The RF, in its original design, does not produce predictive distributions. In order to obtain a predictive distribution, we first collect the predictions $z_{t,n}$ of each individual decision tree t in the RF, for every n-th data item. We take $T = 256$ number of trees in the forest and define the predictive distribution for the RF by fitting a mixture of five Gaussian components to the outputs, $p(\theta|x) = \sum_{j=1}^{5} \omega_j \mathcal{N}(\theta|(\mu_j, \sigma_j))$, as described also in Section 3.1 for the MDN. The RF and the MDN are trained on the feature representation of the data items. The original five features are *ugriz* magnitudes extracted from the images. To avoid biases due to object intrinsic parameters like luminosity, all possible pairwise differences (aka. colors) are used additionally. Therefore a 15-dimensional feature vector is used as input. We divide the dataset in a training and a test set, both containing $25,000$ patterns. The DCMDN is trained on the image representation of the data items. The images are obtained using the *Hierarchical Progressive Surveys* (HIPS) [9] protocol and performing a proper cutout on client side, in order to obtain the desired dimensions. Each data item is originally a stack of five images in the *ugriz* filters. Similarly to the features, we additionally form the color images from the *ugriz* images by taking all possible pairwise differences, thus obtaining a stack of 15 images. The images are taken in two sizes: 28x28 pixels^2 and 16x16 pixels^2. Every object is then represented by a tensor of dimension 15x28x28 or 15x16x16. In order to make the network rotation invariant, we also perform data augmentation. We take rotations of each image at 0, 90, 180, 270 degrees, obtaining a training set of $100,000$ images, a validation set of $50,000$ images and a test set of $50,000$ images. Dropout with a ratio of 60% together with early stopping is used to limit overfitting.

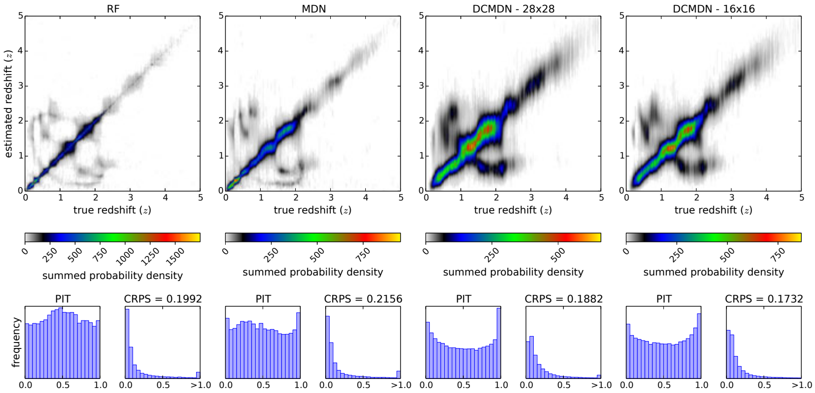
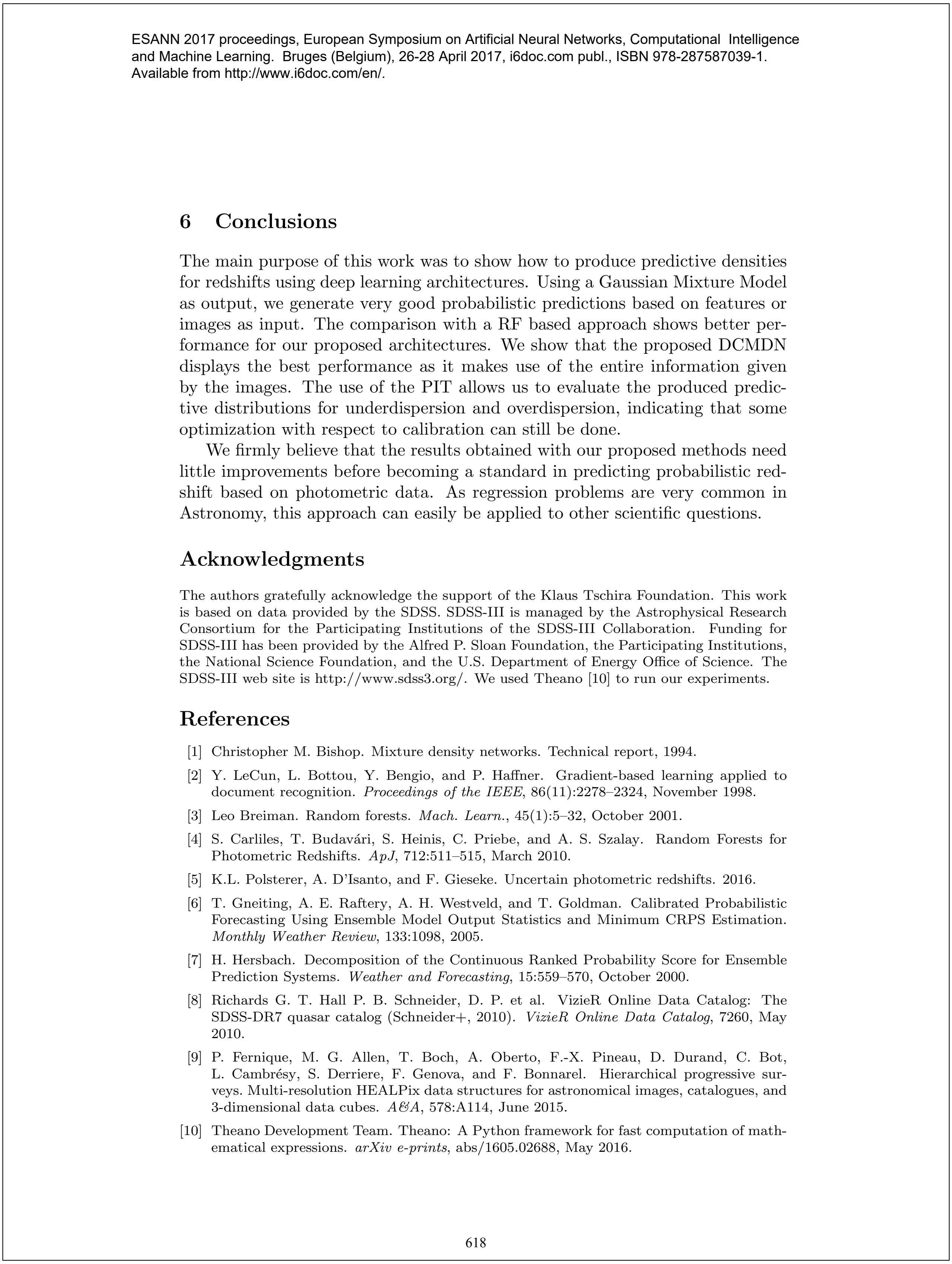


Fig. 3: Results of the prediction obtained with the MDN and the DCMDN (two different input sizes), compared with the RF results. For each experiment, three plots are present. The upper plots compare the spectroscopic redshift with the predictive density produced by the model, with the color indicating the summed probability density of the distributions. In the two lower plots, the histogram of the PIT values and the histogram of the individual CRPS values, are shown. The mean CRPS value is also reported.

5 Analysis of results

The results of the experiments are reported in Fig. 3. In the RF experiment, the model reaches a score of 0.20 and the PIT histogram shows overdispersion. Feeding a RF with the plain pixel information, as done for the DCMDN, results in a CRPS of 0.195 with high overdispersion. The performance of the MDN is a bit worse in terms of the CRPS (score of 0.21) compared to the RF, with a better calibrated PIT. With the DCMDN architecture a significantly better result in terms of the CRPS (0.19 for 28x28, 0.17 for 16x16 images) is achieved. The usual deviation of experiments with other data folds is 0.005 in CRPS. The resulting PIT is acceptable, although it is still underdispersed, which slightly improves for the 16x16 images experiment. The reduced size of the images is more focusing on the central region and ignores neighboring objects and hence improves the result in both, CRPS and PIT. The reason for the better overall performance of the DCMDN is the fact, that the features described in Section 4 use only a fraction of the available information. In fact, the process of extracting historically motivated features is common in Astronomy. In this process a lot of information gets lost. Instead, using images, the DCMDN is able to automatically determine thousands of features, leading to a better prediction of the redshifts.



ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

6 Conclusions

The main purpose of this work was to show how to produce predictive densities for redshifts using deep learning architectures. Using a Gaussian Mixture Model as output, we generate very good probabilistic predictions based on features or images as input. The comparison with a RF based approach shows better performance for our proposed architectures. We show that the proposed DCMDN displays the best performance as it makes use of the entire information given by the images. The use of the PIT allows us to evaluate the produced predictive distributions for underdispersion and overdispersion, indicating that some optimization with respect to calibration can still be done.

We firmly believe that the results obtained with our proposed methods need little improvements before becoming a standard in predicting probabilistic redshift based on photometric data. As regression problems are very common in Astronomy, this approach can easily be applied to other scientific questions.

Acknowledgments

The authors gratefully acknowledge the support of the Klaus Tschira Foundation. This work is based on data provided by the SDSS. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. We used Theano [10] to run our experiments.

References

[1] Christopher M. Bishop. Mixture density networks. Technical report, 1994.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[3] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[4] S. Carliles, T. Budavári, S. Heinis, C. Priebe, and A. S. Szalay. Random Forests for Photometric Redshifts. *ApJ*, 712:511–515, March 2010.

[5] K.L. Polsterer, A. D'Isanto, and F. Gieseke. Uncertain photometric redshifts. 2016.

[6] T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133:1098, 2005.

[7] H. Hersbach. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15:559–570, October 2000.

[8] Richards G. T. Hall P. B. Schneider, D. P. et al. VizieR Online Data Catalog: The SDSS-DR7 quasar catalog (Schneider+, 2010). *VizieR Online Data Catalog*, 7260, May 2010.

[9] P. Fernique, M. G. Allen, T. Boch, A. Oberto, F.-X. Pineau, D. Durand, C. Bot, L. Cambrésy, S. Derriere, F. Genova, and F. Bonnarel. Hierarchical progressive surveys. Multi-resolution HEALPix data structures for astronomical images, catalogues, and 3-dimensional data cubes. *A&A*, 578:A114, June 2015.

[10] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

A&A 609, A111 (2018)
DOI: 10.1051/0004-6361/201731326
© ESO 2018

Astronomy & Astrophysics

Photometric redshift estimation via deep learning

Generalized and pre-classification-less, image based, fully probabilistic redshifts

A. D'Isanto and K. L. Polsterer

Astroinformatics, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: [antonio.disanto;kai.polsterer]@h-its.org

Received 7 June 2017 / Accepted 20 July 2017

ABSTRACT

Context. The need to analyze the available large synoptic multi-band surveys drives the development of new data-analysis methods. Photometric redshift estimation is one field of application where such new methods improved the results, substantially. Up to now, the vast majority of applied redshift estimation methods have utilized photometric features.
Aims. We aim to develop a method to derive probabilistic photometric redshift directly from multi-band imaging data, rendering pre-classification of objects and feature extraction obsolete.
Methods. A modified version of a deep convolutional network was combined with a mixture density network. The estimates are expressed as Gaussian mixture models representing the probability density functions (PDFs) in the redshift space. In addition to the traditional scores, the continuous ranked probability score (CRPS) and the probability integral transform (PIT) were applied as performance criteria. We have adopted a feature based random forest and a plain mixture density network to compare performances on experiments with data from SDSS (DR9).
Results. We show that the proposed method is able to predict redshift PDFs independently from the type of source, for example galaxies, quasars or stars. Thereby the prediction performance is better than both presented reference methods and is comparable to results from the literature.
Conclusions. The presented method is extremely general and allows us to solve of any kind of probabilistic regression problems based on imaging data, for example estimating metallicity or star formation rate of galaxies. This kind of methodology is tremendously important for the next generation of surveys.

Key words. methods: data analysis – methods: statistical – galaxies: distances and redshifts

1. Introduction

In recent years, the availability of large synoptic multi-band surveys increased the need of new and more efficient data analysis methods. The astronomical community is currently experiencing a data deluge. Machine learning and, in particular, deep-learning technologies are increasing in popularity and can deliver a solution to automate complex tasks on large data sets. Similar trends can be observed in the business sector for companies like Google and Facebook. In astronomy, machine learning techniques have been applied to many different uses (Ball & Brunner 2010). Redshift estimation is just one relevant field of application for these statistical methods. Constant improvements in performances had been achieved by adopting and modifying machine learning approaches. The need for precise redshift estimation is increasing due to its importance in cosmology (Blake & Bridle 2005). For example, the *Euclid* mission (Laureijs et al. 2012) is highly dependent on accurate redshift values. Unfortunately, measuring the redshift directly is a time consuming and expensive task as strong spectral features have to be clearly recognized. Therefore, redshifts extracted via photometry based models provide a good alternative (Beck et al. 2016). At the price of a lower accuracy compared to the spectroscopic measurements, photometric redshift estimates enable the processing of huge numbers of sources (Abdalla et al. 2011). Moreover, by combining photometric and spectroscopic techniques, low signal-to-noise spectra of faint objects can be better calibrated and processed (Fernández-Soto et al. 2001).

Photometric redshift estimation methods found in the literature can be divided in two categories: template based spectral energy distribution (SED) fitting (e.g. Bolzonella et al. 2000; Salvato et al. 2009) and statistical and/or machine learning algorithms (e.g. Benítez 2000; Laurino et al. 2011). In this work we will focus on the latter ones and in particular on the application of deep-learning models. Most machine learning approaches demand a large knowledge base to train the model. Once trained, such models allow us to process huge amounts of data and to automatically generate catalogs with millions of sources (as done in Brescia et al. 2014). Instead of generating point estimates only, extracting density distributions that grant access to the uncertainty in the prediction is gaining more focus in the community (Carrasco Kind & Brunner 2013). This scheme is increasingly important for the success of the *Euclid* mission, which depends on highly precise and affordable probabilistic photometric redshifts (Dubath et al. 2017).

Due to the advent of faster and more specialized compute architectures as well as improvements in the design and optimization of very deep and large artificial neural networks, more complex tasks could be solved. In recent years, the so called field of deep-learning was hyped together with big-data-analytics, both in the commercial sector as well as in science. Astronomy has always been a data-intense science but the next generation of survey missions will generate a data-tsunami. Projects such as the Large Synoptic Survey Telescope (LSST) as well as the Square Kilometre Array (SKA) are just some examples that demand processing and storage of data in the peta- and exabyte regime. Deep-learning models could provide a possible solution to analyze those data-sets, even though those large networks are lacking the ability to be completely interpretable by humans.

In this work, the challenge of deriving redshift values from photometric data is addressed. Besides template fitting approaches, several machine-learning models have been used in the past to deal with this problem (Collister & Lahav 2004; Laurino et al. 2011; Polsterer et al. 2013; Cavuoti et al. 2015; Hoyle 2016). Up to now, the estimation of photometric redshifts was mostly based on features that had been extracted in advance. A Le-Net deep convolutional network (DCN; LeCun et al. 1998) is able to automatically derive a set of features based on imaging data and thereby make the extraction of tailored features obsolete. Most machine-learning based photometric redshift estimation approaches found in the literature just generate single value estimates. Template fitting approaches typically deliver a probability density function (PDF) in the redshift space. Therefore currently the machine learning based approaches are either modified or redesigned to deliver PDFs as well (Sadeh et al. 2016). We have proposed a combined model of a mixture density network (MDN; Bishop 1994) and a DCN. The MDN hereby replaces the fully-connected feed-forward part of the DCN to directly generate density distributions as output. This enables the model to use images directly as input and thereby utilize more information in contrast to using a condensed and restricted feature-based input set. The convolutional part of the DCN automatically extracts useful information from the images which are used as inputs for the MDN. The conceptual idea of combining a DCN with a MDN together with details on the implementation have been presented to the computer science community in D'Isanto & Polsterer (2017) while this publication addresses the challenges of photometric redshift estimation on real world data by using the proposed method. A broad range of network structures have been evaluated with respect to the performed experiments. The layout of the best performing one is presented here.

The performance of the proposed image-based model is compared to two feature-based reference methods: a modified version of the widely used random forest (RF; Breiman 2001) which is able to produce PDFs and a plain MDN. The predicted photometric redshifts are expressed as PDFs using a Gaussian mixture model (GMM). This allows the capture of the uncertainty of the prediction in a compact format. Due to degeneracies in the physical problem of identifying a complex spectral energy distribution with a few broadband filters only, multi-modal results are expected. A single Gaussian is not enough to represent the photometric redshift PDF. When using PDFs instead of point estimates, proper statistical analysis tools must be used, taking into account the probabilistic representation. The continuous ranked probability score (CRPS; Hersbach 2000) is a proper score that is used in the field of weather forecasting and expresses how well a predicted PDF represents the true value. In this work, the CRPS reflects how well the photometrically estimate PDF represents the spectroscopically measured redshift value. The CRPS is calculated as the integral taken over the squared difference between the cumulative distribution function (CDF) and the Heaviside step function of the true redshift value. In contrast to the likelihood, the CRPS is more focused on the location and not on the sharpness of the prediction. We have adopted the probability integral transform (PIT; Gneiting et al. 2005) to check the sharpness and calibration of the predicted PDFs. By plotting a histogram of the CDF values at the true redshift over all predictions, overdispersion, underdispersion and biases can be visually detected. We demonstrate that the CRPS and the PIT are proper tools with respect to the traditional scores used in astronomy. A detailed description of how to calculate and evaluate the CRPS and the PIT are given in Appendix A.

The experiments were performed using data from the Sloan Digital Sky Survey (SDSS-DR9; Ahn et al. 2012). A combination of the SDSS-DR7 Quasar Catalog (Schneider et al. 2010) and the SDSS-DR9 Quasar Catalog (Pâris et al. 2012) as well as two subsamples of 200 000 galaxies and 200 000 stars from SDSS-DR9 are used in Sect. 3.

In Sect. 2 the machine learning models used in this work are described. The layout of the experiments and the used data are described in Sect. 3. Next, in Sect. 4 the results of the experiments are presented and analyzed. Finally, in Sect. 5 a summary of the whole work is given. In the appendix, an introduction on CRPS and PIT is given, alongside with the discussion of the applied loss function. Furthermore, we motivate the choice of the number of components in the GMM used to describe the predicted PDFs. Finally, the SQL queries used to download the training and test data as well as references to the developed code are listed.

2. Models

In this section the different models used for the experiments are described. The RF is used as a basic reference method while the MDN and the DCN are used to compare the difference between feature based and image based approaches.

2.1. Random forest

The RF is a widely used supervized model in astronomy to solve regression and classification problems. By partitioning the high dimensional features space, predictions can be efficiently generated. Bagging as well as bootstrapping make those ensembles of decision trees statistically very robust. Therefore RF is often used as a reference method to compare performances of new approaches. The RF is intended to be used on a set of input features which could be plain magnitudes or colors in the case of photometric redshift estimation (Carliles et al. 2010). In its original design, the RF does not generate density distributions. To produce PDFs, the results of every decision tree in the forest are collected and a Gaussian mixture model (GMM) is fitted. The PDF is then presented by a mixture of n components:

$$p(x) = \sum_{j=1}^{n} \omega_j \mathcal{N}(x|\mu_j, \sigma_j), \quad (1)$$

where $\mathcal{N}(x|\mu_j, \sigma_j)$ is a normal distribution with a given mean μ_j and standard deviation μ_j at a given value x. Each component is weighted by ω_j and all weights sum to one. To calculate the CRPS for a GMM, the equations by Grimit et al. (2006) can be used. For the proposed experiments a model composed of 256 trees is used. As input the five *ugriz* magnitudes and the pairwise color combinations are used, obtaining a feature vector of 15 dimensions per data item.

2.2. Mixture density network

An MDN (Bishop 1994) is a modified version of the widely known multilayer perceptron (MLP; Rosenblatt 1962), producing distributions instead of point estimates. The MLP is a supervized feed-forward neural network, whose main calculation unit

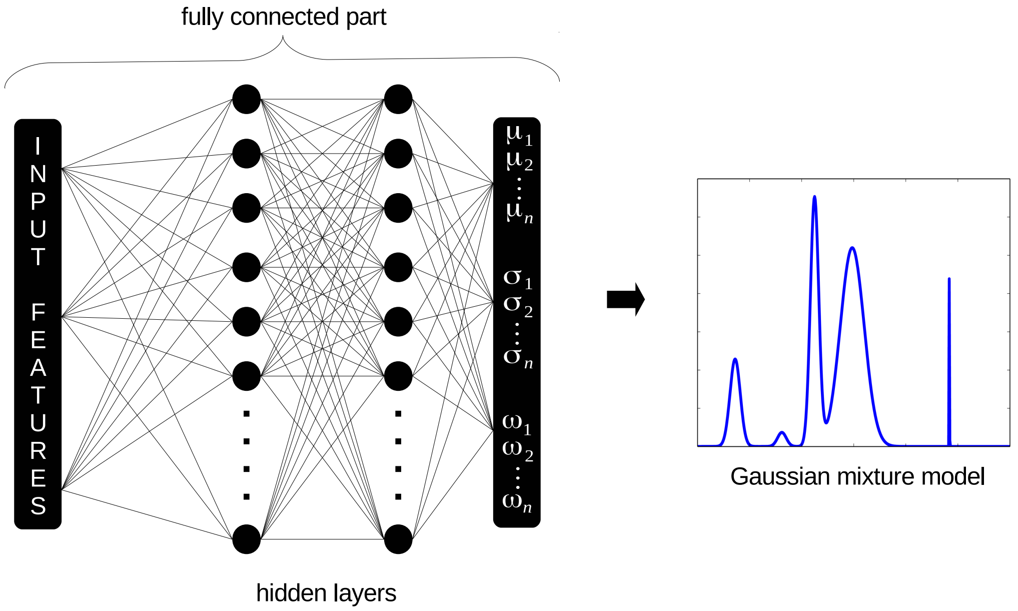


Fig. 1. Architecture of the mixture density network. Next to the input layer, two hidden and fully interconnected layers are depicted. As output a vector for each parameter of the GMM is predicted (μ, σ, ω). Based on this compact description, a density distribution can be generated.

is called neuron. The neurons are arranged in layers, an input layer, one or more hidden layers, and an output layer. Several hyperparameters characterize the architecture of an MLP. The activation function is a non-linear function applied to the data as they pass through the neurons. Commonly a sigmoidal function or hyperbolic tangent (tan$h$) are utilized. In recent years, the MLP has been widely used in astronomy, for example to estimate redshifts based on photometric features (Brescia et al. 2013).

The MDN interprets the output vector o of the network as the means μ, standard deviations σ and weights ω of a mixture model, using n Gaussians as basis functions:

$$\begin{aligned} \mu_j &= o_j^{\mu}, \\ \sigma_j &= \exp(o_j^{\sigma}), \\ \omega_j &= \frac{\exp(o_j^{\omega})}{\sum_{i=1}^{n} \exp(o_i^{\omega})} \\ &\text{with } j \in 1 ... n \text{ and } o = \{o^{\mu}, o^{\sigma}, o^{\omega}\}. \end{aligned} \tag{2}$$

Commonly, MDNs are trained using the log-likelihood as loss function. In this application the focus is more on the distribution and shape than on the location of the prediction; hence the CRPS is adopted as loss function. A detailed analysis of the performances of both loss functions is provided in Appendix B. The CRPS increases linearly with the distance from the reference value while the log-likelihood increases with the square of the distance. Like the RF, the MDN is a feature-based model and therefore can use exactly the same input features: plain magnitudes and colors. A generalized architecture of an MDN is shown in Fig. 1.

2.3. Deep convolutional mixture density network

A DCN is a feed-forward neural network in which several convolutional and sub-sampling layers are coupled with a fully-connected part. In some sense, it is a specialization of the fully interconnected MLP model. By locally limiting the used interconnections in the first part, a spatial sensitivity is realized. Thereby the dimensionality of the inputs is reduced step-wise in every layer. The second part makes use of the so-called feature maps in a flattened representation, using them as input for an ordinary MLP.

This kind of network architecture finds wide application in the fields of image, video and speech recognition due to its capability of performing some kind of dimensionality reduction and automatic feature extraction. A DCN model was chosen because of its ability to deal directly with images, without the need of preprocessing and an explicit features extraction process. The network is trained to capture the important aspects of the input data. By optimizing the condensed representation of the input data in the feature maps, the performance of the fully connected part is improved. Every input data item is a tensor of multi-band images. In the experiments the five *ugriz* filters from SDSS as well as the pixel-wise differences were used. Those image gradients can be compared to the colors in feature based approaches that minimize the effects of object intrinsic variations in luminosity. In the convolutional layers, the input images are spatially convolved with proper filters of fixed dimensions. Hereby "proper" denotes a size that corresponds to the expected spatial extension of distinctive characteristics (filter of the dimension 3 × 3 and 2 × 2 have been selected for the application at hand). The filters constitute the receptive field of the model.

After the convolution with the filters, a non-linear function is applied. The tan$h$ was used in this work as expressed by the following relation:

$$H_k = \tan h(W_k * X + b_k). \tag{3}$$

In general, for every filter k the hidden activation H_k is calculated by using the filter weight matrix W_k and the filter bias b_k. The outputs of the previous layer H are used as inputs X for the succeeding convolutional layer. The first layer is directly fed with the imaging data that should be processed. This filtering system

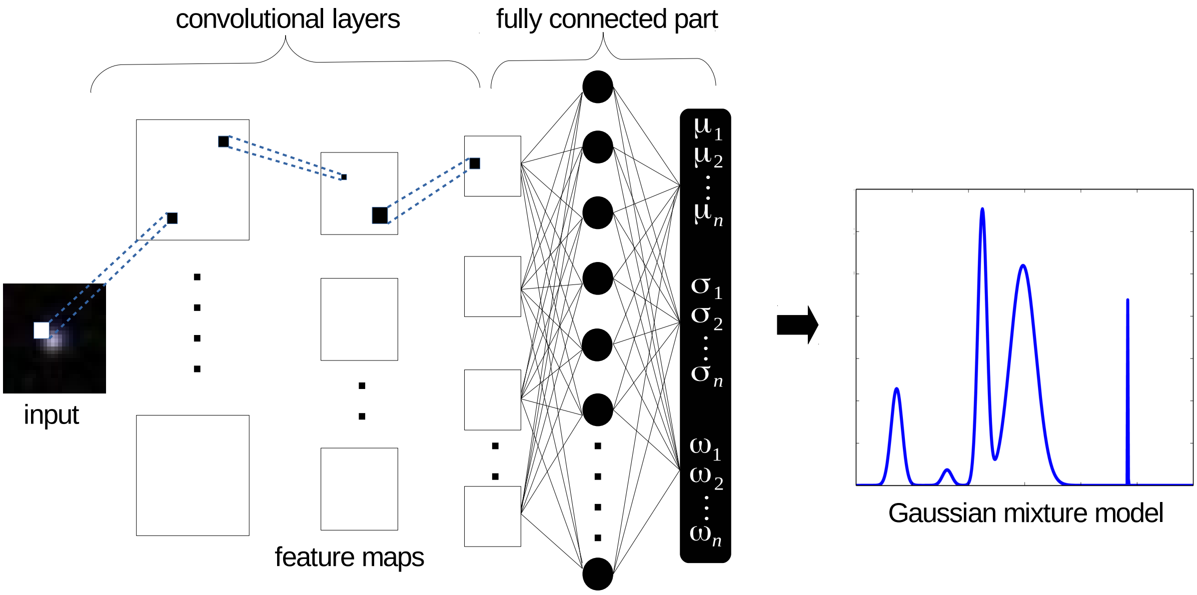


Fig. 2. Architecture of the deep convolutional mixture density network. This architecture directly uses image matrices as input. In this figure, three convolutional layers are drawn, showing the local connection between the different layers. In the fully connected part, a MDN with one hidden layer is used and a vector for each parameter of the GMM is given as result (μ, σ, ω). A sample of a predicted PDF is shown exemplarily.

with local connection make the architecture sensitive to spatial input patterns. The filters are tiled through the entire visual field, with a fixed stride, generating a new representation of the input, that is, the feature maps. A useful tool to pursue the down sampling is to apply max-pooling between the convolutional layers. Those pooling filter typically select only the maximum value in confined region, typically of dimension 2×2 or 3×3. Convolutional and pooling layers are alternated until the desired complexity in terms of feature maps is obtained. In addition, those layers are alternated with *rectified linear unit* (*ReLu*) layers, in which the non-linear activation function is substituted by a non saturating function:

$$H_k = \max(0, W_k * X + b_k). \tag{4}$$

This function has many advantages; in particular it increases the efficiency of the gradient descent process, avoiding vanishing gradient problems. Furthermore, using only comparison, addition and multiplication operations it is computationally more efficient. The choice of the activation functions, namely the tan*h* and the *rectified linear unit*, influences the convergence and performance of the neural network. Therefore, the activation function has to be considered as a free parameter when designing the network architecture. In our case, switching to the *rectified linear unit* improved the performance of the predictions notably. Many different combinations have been tested, choosing the best performing one. The feature maps constitute the input for the fully connected part, which has in general the same behavior as a MLP. In the proposed model, we substitute the usual fully connected part with a MDN in order to generate PDFs instead of single point estimates. For this reason, this combined architecture is denoted by us as deep convolutional mixture density network (DCMDN). The structure of the DCMDN is sketched in Fig 2. Furthermore, as for the MDN, the CRPS is used as loss function.

Several hyperparameters influence the layout of the network architecture as well as the training phase. Multiple combinations have been tested extensively. Due to the immense amount of possible parameter combinations the currently used solution was obtained by clever manual engineering. The most influencing parameters are listed below:

- *global architecture*: this includes the number and types of layers in the local-connected part and the number of hidden layers and neurons characterizing the fully-connected part.
- *activation function*: defines the non-linear function to process the input values of a neuron.
- *number of filters*: influences the number of generated feature maps and therefore the amount of extracted features.
- *filter shape*: characterizes the dimensions of the filters used; it can vary from layer to layer.
- *max-pooling shape*: as for the filters, it specifies the dimension of the area to which the max-pooling is applied.
- *learning rate*: influences the step-size during the gradient descent optimization. This value can decrease with the number of trained epochs.
- *number of epochs*: defines how often the whole training data set is used to optimize the weights and biases.
- *batch size*: as a stochastic gradient descent optimization strategy was chosen, this number defines the amount of training patterns to be used in one training step.

The presented model exhibits many advantages for the application to photometric redshift estimation tasks. It can natively handle images as input and extract feature maps in a fully-automatized way. The DCMDN does not need any sort of pre-processing and pre-classification on the input data and as shown in the experiments section, the performances are far better with respect to the reference models. The reason for an improvement with respect to the estimation performance is the better and extensive use of the available information content of the images. Besides automatically extracting features, their importance with respect to the redshift estimation task is automatically determined too. Feature based approaches make use of the condensed information that is provided through previously extracted features only. Those features are good for a broad set of applications but not optimized for machine learning methods applied to very specific tasks.

Depending on the size of the network architecture, an extremely high number of weights and biases has to be optimized. This allows the network to adopt to not significant correlations in the data and therefore overfit. While good performances are achieved on the training data, the application to another data-sets would exhibit poor results. To limit the effect of overfitting, the dropout technique can be applied; randomly setting a given percentage of the weights in both parts of the network to zero. As deep-learning methods highly benefit from a huge amount of training objects, data augmentation is a common technique. By simply rotating the images by 90°, 180° and 270° and flipping the images, the size of the training data-set can be increased by a factor of four and eight, respectively. This reduces the negative effects of background sources and artifacts on the prediction performance. Moreover, the early stopping technique can be applied to limit the chance of overfitting. As soon as the performance that is evaluated on a separate validation set starts to degrade while the training error is still improving, the training is stopped even before reaching the anticipated number of epochs. The DCMDN is based on the LeNet-5 architecture (LeCun et al. 1998) and realized in Python, making use of the Theano library (Theano Development Team 2016).

The architecture implemented by us is meant to run on graphics processing units (GPUs) as the training on simple central processing units (CPUs) is by far too time consuming. In our experiments a speedup of factor of ≈40× between an eight-core CPU (i7 4710MQ 2.50 GHz × 8) and the GPU hardware was observed. During the experiments a cluster equipped with Nvidia Titan X was intensively used, allowing us to evaluate a larger combination of network architectures and hyperparameters.

3. Experiments

In the following sections the experiments performed with the presented models are described. Those experiments are intended to compare the probabilistic redshift prediction performances of different models on different data-sets. The data-sets used for training the models as well as evaluating the performances are described in the following.

3.1. Data

All data-sets have been compiled using data from SDSS. To generate a set of objects that cover the whole range of redshifts, separate data-sets for quasars and galaxies have been created. The SDSS Quasar Catalog Seventh Data Release (Schneider et al. 2010), containing 105, 783 spectroscopically confirmed quasars and the SDSS Quasar Catalog Ninth Data Release (Pâris et al. 2012), containing 87, 822 spectroscopically confirmed quasars, are used as basis for the quasar data-set. The two catalogs had to be combined because the former contains confirmed sources from SDSS II only, while the latter is composed of 91% of new objects observed in SDSS III-BOSS (Dawson et al. 2013). In this way a catalog with a much better coverage of the redshift space has been composed (see Fig. 3). Furthermore, two samples composed of 200 000 randomly chosen galaxies and 200 000 randomly picked stars from DR9 (Ahn et al. 2012) have been selected (queries are stated in Appendix F). The objects that have been classified by the spectroscopic pipeline as stars are assigned a redshift of zero. These objects are mandatory to crosscheck the performance on objects that have not been pre-classified and therefore might be contaminated with stellar sources.

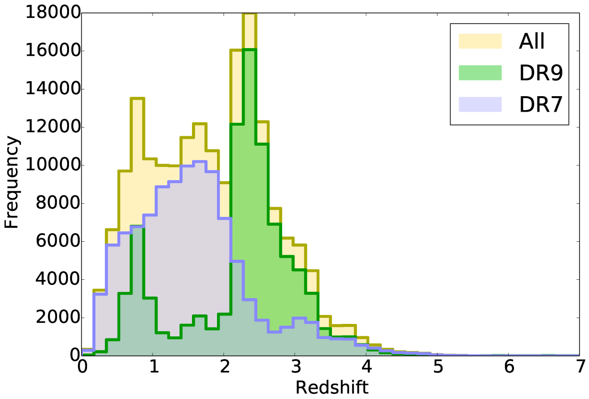


Fig. 3. Redshift distribution of the quasar data from DR7, DR9 and the combined catalog, respectively.

For every catalog the corresponding images from SDSS DR9 in the five *ugriz* filters have been downloaded. The images have been downloaded using the Hierarchical Progressive Surveys (HiPS) data partitioning format (Fernique et al. 2015), in a 28 × 28 square-pixels format, and subsequently pairwise subtracted, in order to obtain color images corresponding to the colors used in the feature-based experiments. As described above, those pixel wise gradients minimize the effects of object intrinsic variations in luminosity.

3.2. Setup and general configuration

Three different categories of experiments have been performed, each using the three models of Sect. 2 to compare the performances. Before stating the details of the single experiments, a brief description of the general parameters adopted for the models is given. The experiments with the RF and the MDN are feature-based, while the DCMDN is trained on the image representations of the data items. As input features for the two reference models, magnitudes in the *ugriz* filter are used together with all pair-wise color combinations. For the RF models a fixed structure with 256 decision trees was chosen. Bootstrapping was used for creating the individual trees. Neither the depth nor the number of used features have been limited. The MDN architectures use 15 inputs neurons (corresponding to the 15 input features) followed by two hidden layers containing 50 and 100 neurons, respectively. 15 output neurons are used to characterize the parameters of a GMM with $n = 5$. Photometrically estimated redshift distributions are of complex and multimodal nature (Kügler et al. 2016). The choice of 5 Gaussian components is based on experiments where the Bayesian information criterion (BIC) was used as a metric. Depending on the redshift region calculated for, the BIC is indicating values between one and five components. The results presented in Appendix C indicate that five components are, on average, a good choice. For the DCMDN, many different architectures have been tested, comprising less deep and compact convolutional parts. The architecture that gave the best performances was finally created via a clever manual and empirical engineering. The learning rate influences the size of the steps when applying gradient descent during the training with the backpropagation algorithm. For the experiments presented in this work a tan*h* was chosen with the learning

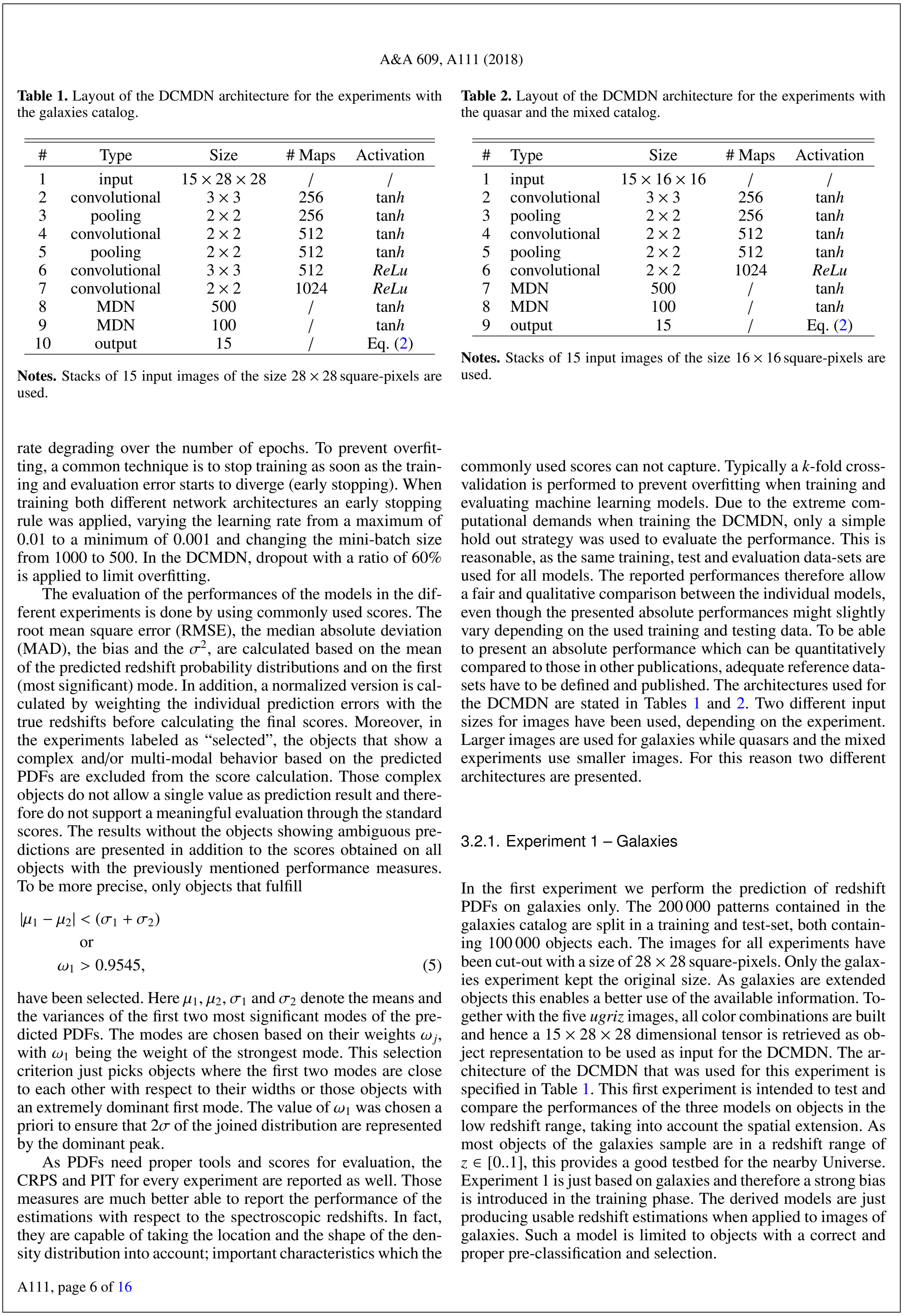


Table 1. Layout of the DCMDN architecture for the experiments with the galaxies catalog.

#	Type	Size	# Maps	Activation
1	input	$15 \times 28 \times 28$	/	/
2	convolutional	3×3	256	tan*h*
3	pooling	2×2	256	tan*h*
4	convolutional	2×2	512	tan*h*
5	pooling	2×2	512	tan*h*
6	convolutional	3×3	512	*ReLu*
7	convolutional	2×2	1024	*ReLu*
8	MDN	500	/	tan*h*
9	MDN	100	/	tan*h*
10	output	15	/	Eq. (2)

Notes. Stacks of 15 input images of the size 28×28 square-pixels are used.

Table 2. Layout of the DCMDN architecture for the experiments with the quasar and the mixed catalog.

#	Type	Size	# Maps	Activation
1	input	$15 \times 16 \times 16$	/	/
2	convolutional	3×3	256	tan*h*
3	pooling	2×2	256	tan*h*
4	convolutional	2×2	512	tan*h*
5	pooling	2×2	512	tan*h*
6	convolutional	2×2	1024	*ReLu*
7	MDN	500	/	tan*h*
8	MDN	100	/	tan*h*
9	output	15	/	Eq. (2)

Notes. Stacks of 15 input images of the size 16×16 square-pixels are used.

rate degrading over the number of epochs. To prevent overfitting, a common technique is to stop training as soon as the training and evaluation error starts to diverge (early stopping). When training both different network architectures an early stopping rule was applied, varying the learning rate from a maximum of 0.01 to a minimum of 0.001 and changing the mini-batch size from 1000 to 500. In the DCMDN, dropout with a ratio of 60% is applied to limit overfitting.

The evaluation of the performances of the models in the different experiments is done by using commonly used scores. The root mean square error (RMSE), the median absolute deviation (MAD), the bias and the σ^2, are calculated based on the mean of the predicted redshift probability distributions and on the first (most significant) mode. In addition, a normalized version is calculated by weighting the individual prediction errors with the true redshifts before calculating the final scores. Moreover, in the experiments labeled as "selected", the objects that show a complex and/or multi-modal behavior based on the predicted PDFs are excluded from the score calculation. Those complex objects do not allow a single value as prediction result and therefore do not support a meaningful evaluation through the standard scores. The results without the objects showing ambiguous predictions are presented in addition to the scores obtained on all objects with the previously mentioned performance measures. To be more precise, only objects that fulfill

$$|\mu_1 - \mu_2| < (\sigma_1 + \sigma_2) \quad \text{or} \quad \omega_1 > 0.9545, \tag{5}$$

have been selected. Here μ_1, μ_2, σ_1 and σ_2 denote the means and the variances of the first two most significant modes of the predicted PDFs. The modes are chosen based on their weights ω_j, with ω_1 being the weight of the strongest mode. This selection criterion just picks objects where the first two modes are close to each other with respect to their widths or those objects with an extremely dominant first mode. The value of ω_1 was chosen a priori to ensure that 2σ of the joined distribution are represented by the dominant peak.

As PDFs need proper tools and scores for evaluation, the CRPS and PIT for every experiment are reported as well. Those measures are much better able to report the performance of the estimations with respect to the spectroscopic redshifts. In fact, they are capable of taking the location and the shape of the density distribution into account; important characteristics which the commonly used scores can not capture. Typically a *k*-fold cross-validation is performed to prevent overfitting when training and evaluating machine learning models. Due to the extreme computational demands when training the DCMDN, only a simple hold out strategy was used to evaluate the performance. This is reasonable, as the same training, test and evaluation data-sets are used for all models. The reported performances therefore allow a fair and qualitative comparison between the individual models, even though the presented absolute performances might slightly vary depending on the used training and testing data. To be able to present an absolute performance which can be quantitatively compared to those in other publications, adequate reference data-sets have to be defined and published. The architectures used for the DCMDN are stated in Tables 1 and 2. Two different input sizes for images have been used, depending on the experiment. Larger images are used for galaxies while quasars and the mixed experiments use smaller images. For this reason two different architectures are presented.

3.2.1. Experiment 1 – Galaxies

In the first experiment we perform the prediction of redshift PDFs on galaxies only. The 200 000 patterns contained in the galaxies catalog are split in a training and test-set, both containing 100 000 objects each. The images for all experiments have been cut-out with a size of 28×28 square-pixels. Only the galaxies experiment kept the original size. As galaxies are extended objects this enables a better use of the available information. Together with the five *ugriz* images, all color combinations are built and hence a $15 \times 28 \times 28$ dimensional tensor is retrieved as object representation to be used as input for the DCMDN. The architecture of the DCMDN that was used for this experiment is specified in Table 1. This first experiment is intended to test and compare the performances of the three models on objects in the low redshift range, taking into account the spatial extension. As most objects of the galaxies sample are in a redshift range of $z \in [0..1]$, this provides a good testbed for the nearby Universe. Experiment 1 is just based on galaxies and therefore a strong bias is introduced in the training phase. The derived models are just producing usable redshift estimations when applied to images of galaxies. Such a model is limited to objects with a correct and proper pre-classification and selection.

3.2.2. Experiment 2 – Quasars

In the second experiment PDFs are estimated for quasars. The quasar catalog is composed of 185 000 objects; the DR7 and DR9 catalogs are combined selecting all the objects and removing double ones. The quasar experiments have been performed using 100 000 objects in the training-set and 85 000 in the test-set. This makes the size of the data used for training comparable to the previous experiment. In this experiment the size of the input images for the DCMDN is reduced to 16 × 16 square-pixels in order to save computational resources and speed up the training. As quasars are more compact sources, a smaller cut-out should be sufficient to capture the details of the spatial distribution and still include information from the hosting galaxy. The same color combinations used in the first experiment were created and a $15 \times 16 \times 16$ tensor is used as input for the DCMDN. The architecture of the DCMDN that was used for this experiment is specified in Table 2. The quasar experiment tests the performance for less extended objects that cover a wider range of redshift $z \in [0...6]$. Similarly to the first experiment, the models of the second experiment were heavily dependent on a correct pre-classification of objects.

3.2.3. Experiment 3 – Mixed

Finally, in the third experiment a mixed catalog was used. By combining quasars, galaxies and stars we are able to test and evaluate the performances of the three models independently from the nature of the sources. The step of pre-classifying objects is hereby made obsolete. The stars that have been added can be considered as contamination; as faint cool stars can be easily confused with quasars. This makes the use-case of photometric redshift estimation more realistic to the challenges of processing yet unseen objects with an uncertain classification. To be able to derive a proper PDF for all objects, stars have been assigned a redshift of $z = 0$. As stated above, the whole catalog is composed of 585 000 objects. In this experiment, 300 000 patterns were used for training and 285 000 for testing and the dimension of the input images is reduced to 16 × 16 square-pixels. The DCMDN has therefore the same architecture as used in the previous experiment (see Table 2). This experiment is intended to evaluate the performances of the models in a realistic use-case. Hence the results of this experiments are the most notable, as no biases through a pre-classification phase are introduced. Such an experiment should be part of every publication, introducing a new data-driven method for photometric redshift estimation.

4. Results

The experiments of Sect. 3 have been performed on a GPU cluster. The detailed results are presented in the following.

4.1. Experiment 1 – Galaxies

The results of the first experiment are depicted in Figs. 4 and 5. In both figures the estimated redshifts are plotted against the spectroscopic redshifts. To be comparable to results from other publications in the field of photometric redshift estimation, in Fig. 4 the complex estimated PDFs are compressed into a single point estimate. This compression is realized by either taking the plain mean, the first and most dominant mode of the mixture model or by taking the mean of objects that do not exhibit an ambiguous PDF (see Eq. (5)). Based on these three simplified representations of the estimated PDFs, the traditional scores have been calculated. Those values are reported in Table 3 and show a similar performance as other publications (e.g. Laurino et al. 2011) that used a comparable data-set. All three models used for testing show a very similar performance. It is notable, that due to the nature of the used different models, the MDN and the DCMDN show a slightly better generalization performance especially in those regions where the transition of characteristic spectroscopic features through the broadband filters do not allow a distinct separation. In those redshift regions where the degeneracy of the reverse determination of the spectral energy distribution and distinct spectral features through the low spectral resolution image data is dominating, selecting the first mode instead of using the mean value shows a poorer performance. This is especially the case for $z \approx 0.35$ and $z \approx 0.45$. When selecting only those objects for evaluation that do not have an ambiguous behavior, the mentioned redshift regions become underpopulated in the diagnostic plots. For the DCMDN this effect is not as prominent as for the other two reference models due to the ability to make use of a larger base of information. As presented in Fig. 4, compressing the PDFs into single values does not cover the full complexity of the redshift estimations. In particular, the selection of outliers having no unique single dominant redshift component in the PDF demonstrates the multi-modal nature of the photometric redshift estimation task. Therefore, a visible improvement of the performance can be noticed when selecting only the subset of patterns which show no multi-modal behavior. For these reasons, a proper probabilistic evaluation of the PDFs has to be performed. Thus, in Fig. 5 a diagnostic plot is introduced which preserves the overall information of the density distributions. Alongside this probabilistic comparison between the estimated redshift distributions and the spectroscopic values, the PIT and the CRPS are used as proper tools. In the upper plot, the spectroscopic redshift is compared with the generated predictive density of every data item. Hereby the logarithm of the summed probability density for each redshift bin is plotted. In the two lower plots, the histograms for the PIT and the individual CRPS are given, together with the value of the mean CRPS. Analyzing the plots, it is notable that the RF performs slightly better with respect to the PIT. In all cases the PIT shows a more or less well calibrated uniform distribution of the evaluated CDFs at the corresponding spectroscopic redshifts. The MDN and the DCMDN in particular exhibit a more uniform and cleaner alignment toward the diagonal line which indicates the ideal performance.

4.2. Experiment 2 – Quasars

In Fig. 6 the results of the second experiment are shown in the standard plot. All estimated PDFs have been transferred into point estimates as it was done for the results of the first experiment. In Table 4 the corresponding scores are presented. As expected, the performances achieved with the mean values are comparable to point estimates obtained by other methods (Laurino et al. 2011). When using the first modes of the predicted density distributions, the degeneracies caused by the poor spectral resolution of the broadband photometry become visible. Especially in the redshift areas around $z \approx [0.5...0.9]$ and $z \approx [1.5...2.5]$ a strong multi-modal behavior can be observed. In those areas, picking the first mode does not necessary provide the best estimate, as both redshifts are equally likely. The task of uniquely assigning a redshift based on poorly resolved broadband observations is in some regions a degenerated problem. The better description of the probability distributions results therefore in a symmetry along the line of ideal

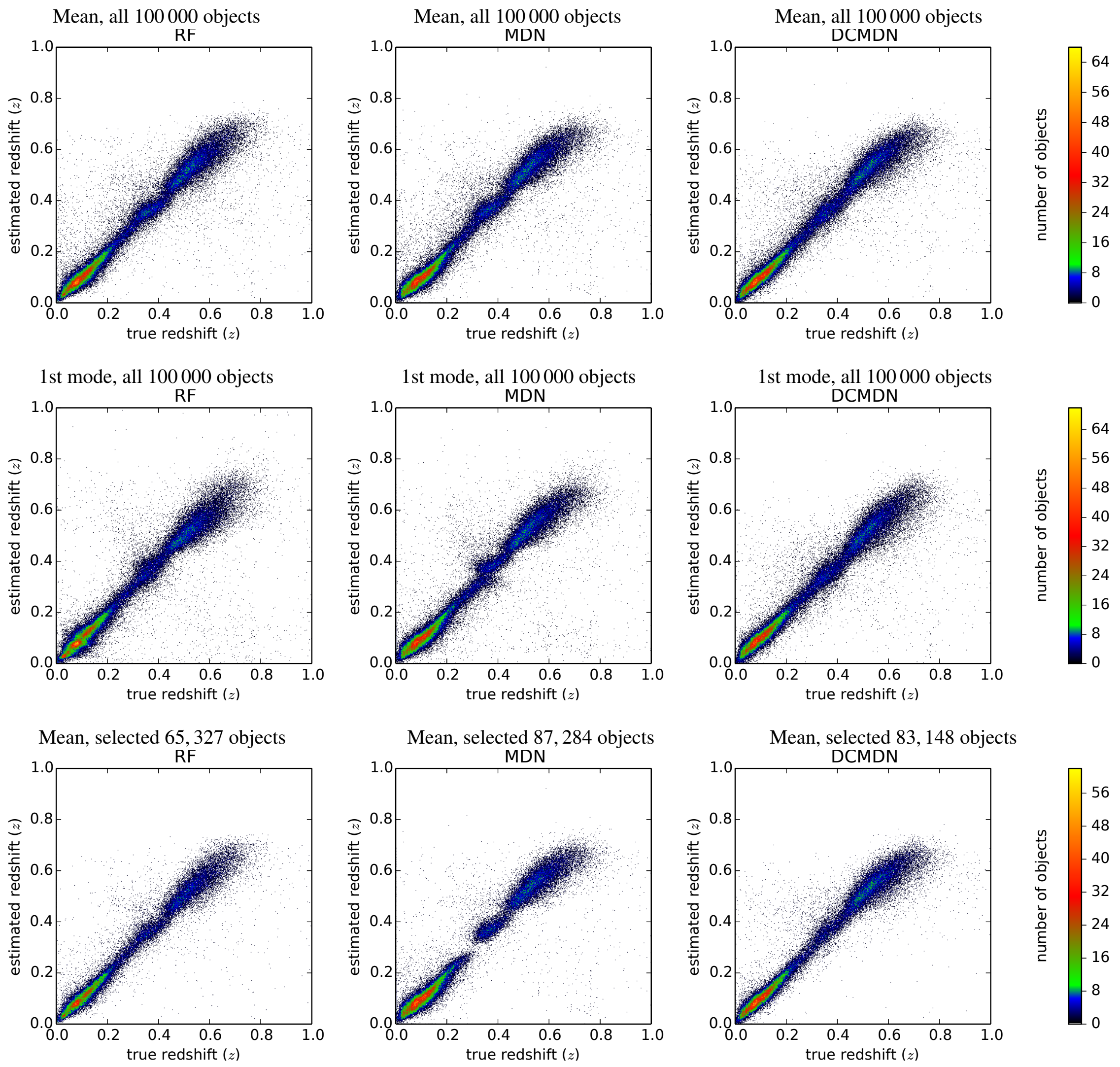


Fig. 4. Results of Experiment 1 (galaxies). The estimated redshifts are plotted against the spectroscopic redshifts, which are considered ground trues. As in Table 3 the plots are done for all three considered models (*left to right*) as well as with the mean, dominant mode and unambiguous objects (*top to bottom*). In addition, the number of unambiguous objects is reported.

Table 3. Results of Experiment 1 (galaxies).

Criterion	Model	Bias(Δz)	$\sigma^2(\Delta z)$	Mad(Δz)	rmse(Δz)	Bias(Δz_{norm})	$\sigma^2(\Delta z_{\text{norm}})$	Mad(Δz_{norm})	rmse(Δz_{norm})
Mean	RF	0.0001	0.0033	0.0164	0.0575	–0.0017	0.0018	0.0133	0.0431
	MDN	0.0016	0.0034	0.0174	0.0589	–0.0003	0.0019	0.0141	0.0442
	DCMDN	0.0018	0.0030	0.0157	0.0548	–0.0003	0.0017	0.0128	0.0409
First mode	RF	0.0031	0.0042	0.0171	0.0652	0.0010	0.0022	0.0139	0.0471
	MDN	0.0029	0.0039	0.0172	0.0628	0.0010	0.0021	0.0140	0.0459
	DCMDN	0.0060	0.0031	0.0167	0.0561	0.0029	0.0016	0.0135	0.0407
Selected	RF	0.0001	0.0023	0.0147	0.0484	–0.0011	0.0013	0.0121	0.0365
	MDN	0.0021	0.0027	0.0165	0.0523	0.0006	0.0015	0.0136	0.0391
	DCMDN	0.0017	0.0023	0.0146	0.0485	–0.0001	0.0013	0.0120	0.0366

Notes. Based on the estimated PDFs the traditional scores have been calculated. This was done by using the mean, the most dominant mode and the mean of the selected unambiguous objects, respectively.

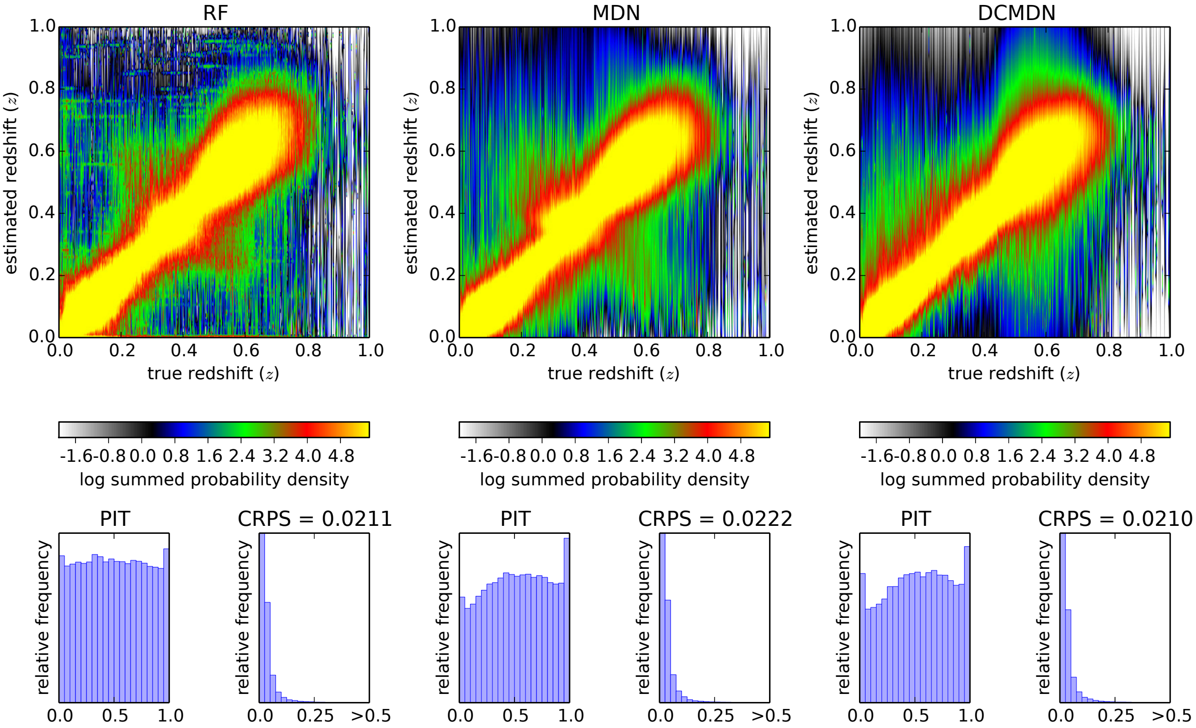
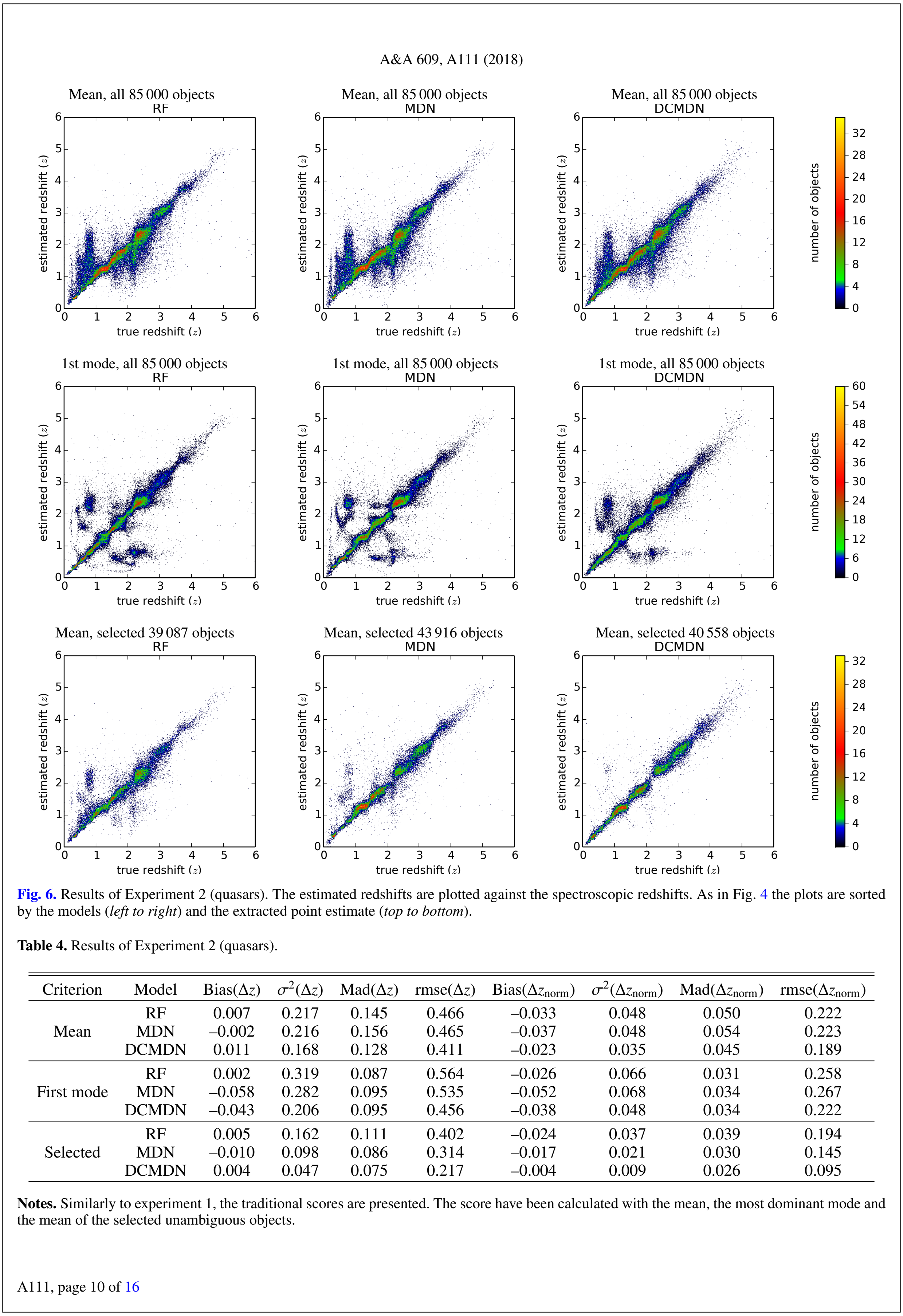


Fig. 5. Results of Experiment 1 (galaxies) with a fully probabilistic evaluation and representation of the estimated PDFs. For each model, three plots are present. In the upper one, the predicted density distribution for each individual object is plotted at the corresponding spectroscopic redshift. The colors hereby indicate the logarithm of the summed probability densities using 500 redshift bins per axis. In the two lower plots, the histogram of the PIT values and the histogram of the CRPS values are shown, respectively. The mean CRPS values are reported, too.

performance. Due to selection effects by observing a limited and cone-shaped volume, the distribution of objects with respect to cosmological scales results in not perfect symmetric behavior. The necessity of a multi-modal description becomes even more obvious when objects with an ambiguous density distribution are not taken into consideration. This way of excluding sources with estimated ambiguous redshift distributions can be considered a correct method to filter possible outliers, hence the selection of objects that show an uni-modal behavior is purely done based on the probabilistic description. In all three different representation of the complex PDFs through a single point estimate, the DCMDN exhibits a better performance. This is also reflected by the scores that are presented in Table 4. For the RF, the partitioning of the high-dimensional feature space orthogonal to the dimension axis does not provide a generalized representation of the regression problem. When covering a wider redshift range the predictions exhibit more differences performance wise. Therefore the performance of the RF drops in this experiment with respect to the other models as the MDN and DCMDN produce much smoother predictions along the ideal diagonal of the used diagnostic plot. With respect to catastrophic outliers, the DCMDN has a superior performance when compared with the other two models. This is consistent with the results from the first experiment. When analyzing the probabilistic representation of the prediction results in Fig. 7, the superior performance of the DCMDN is striking. The CRPS especially indicates the quality of the predictions made by the DCMDN. In fact, by using images the DCMDN can utilize all the contained information and automatically extract the best usable features.

4.3. *Experiment 3 – Mixed*

The experiment with the mixed catalog is the most challenging one. It tests how well the advantages of a deep convolutional network architecture can be used to render the step of pre-processing and pre-classifying objects obsolete. This experiment is much closer to the real application compared to the previous cases. In Fig. 8 and Table 5 the results of Experiment 3 are reported and evaluated as single point estimates. As in the previous experiments, the PDFs are therefore compressed. When using the mean or the first mode as representation, the DCMDN significantly outperforms the RF and the MDN. In the case of selecting unambiguous PDFs only, the result of the DCMDN is close to the ideal performance. The obtained results confirm the indications of the previous experiments. The fully probabilistic representation of the results of the third experiment are presented in Fig. 9. When using proper tools and scores for evaluation, the DCMDN is the best model with respect to the CRPS. Both, the RF and the MDN exhibit similar CRPS results, as they did in all the three experiments. In the representation of the PITs the differences in calibration can be further analyzed. The RF shows a nearly uniform distribution with an extreme peak in the center. Due to the partitioning of the high-dimensional feature space performed by the RF, most of the stars are perfectly recovered. As stars are fixedly assigned a redshift of $z = 0$, the chosen way of fitting a GMM to the individual decision tree results produces the central peak. This can be seen as an extreme overdispersion of the predictions in relation to the true values. With respect to the estimated PDFs, the CDFs at the truncated true value $z = 0$ is always very close to 0.5. For a stellar object, only a very few



Mean, all 85 000 objects
RF
Mean, all 85 000 objects
MDN
Mean, all 85 000 objects
DCMDN
1st mode, all 85 000 objects
RF
1st mode, all 85 000 objects
MDN
1st mode, all 85 000 objects
DCMDN
Mean, selected 39 087 objects
RF
Mean, selected 43 916 objects
MDN
Mean, selected 40 558 objects
DCMDN
estimated redshift (z)
true redshift (z)
number of objects
0
1
2
3
4
5
6
0
4
8
12
16
20
24
28
32
0
6
12
18
24
30
36
42
48
54
60

Fig. 6. Results of Experiment 2 (quasars). The estimated redshifts are plotted against the spectroscopic redshifts. As in Fig. 4 the plots are sorted by the models (*left to right*) and the extracted point estimate (*top to bottom*).

Table 4. Results of Experiment 2 (quasars).

Criterion	Model	Bias(Δz)	$\sigma^2(\Delta z)$	Mad(Δz)	rmse(Δz)	Bias(Δz_{norm})	$\sigma^2(\Delta z_{\text{norm}})$	Mad(Δz_{norm})	rmse(Δz_{norm})
Mean	RF	0.007	0.217	0.145	0.466	–0.033	0.048	0.050	0.222
	MDN	–0.002	0.216	0.156	0.465	–0.037	0.048	0.054	0.223
	DCMDN	0.011	0.168	0.128	0.411	–0.023	0.035	0.045	0.189
First mode	RF	0.002	0.319	0.087	0.564	–0.026	0.066	0.031	0.258
	MDN	–0.058	0.282	0.095	0.535	–0.052	0.068	0.034	0.267
	DCMDN	–0.043	0.206	0.095	0.456	–0.038	0.048	0.034	0.222
Selected	RF	0.005	0.162	0.111	0.402	–0.024	0.037	0.039	0.194
	MDN	–0.010	0.098	0.086	0.314	–0.017	0.021	0.030	0.145
	DCMDN	0.004	0.047	0.075	0.217	–0.004	0.009	0.026	0.095

Notes. Similarly to experiment 1, the traditional scores are presented. The score have been calculated with the mean, the most dominant mode and the mean of the selected unambiguous objects.

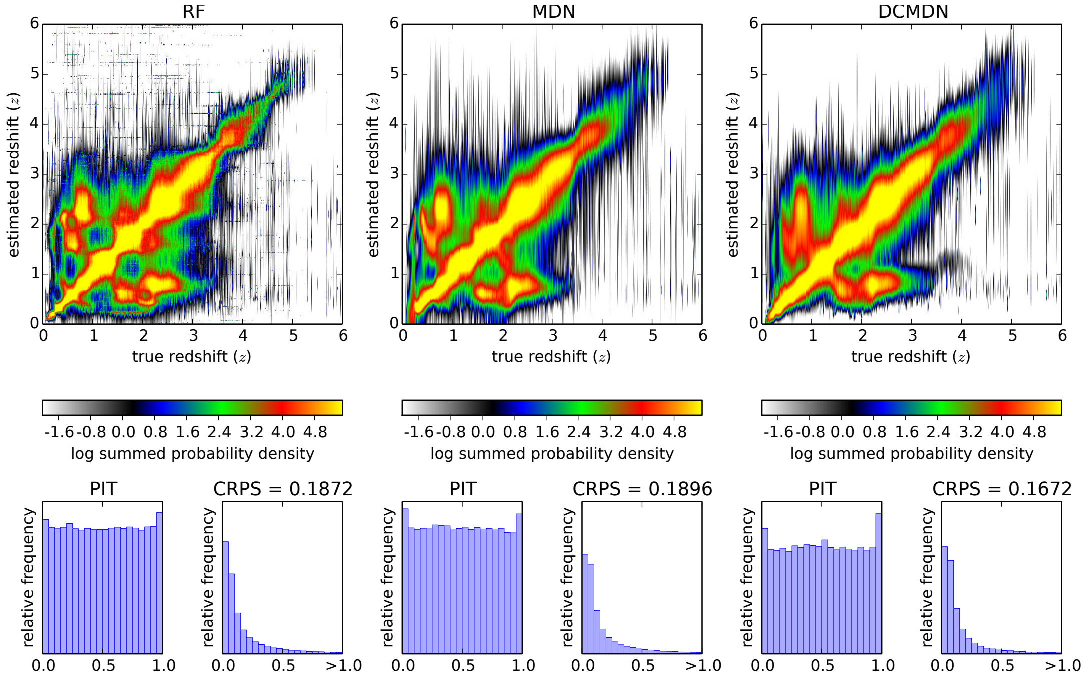


Fig. 7. Results of Experiment 2 (quasars) with a fully probabilistic evaluation. As in Fig. 5 three plots per model are used to visualize the probabilistic performance of the estimated PDFs.

decision trees might return a redshift $z > 0$, given the large number of ≈100 000 stars used in the training sample. During training the MDN got biased by the large number of stellar components with redshift set to $z = 0$. Therefore the corresponding PIT indicates a tendency to underestimate the redshift. A complex, asymmetric behavior in the shape of the estimated PDFs is not captured when using the classical scores and diagnostic plots only, as the bias on the mean of the PDFs is close to zero for this experiment. For the DCMDN we can observe two short comings. Overall, the generated PDFs are overdispersed. This effect is caused by the large fraction of stellar objects with a fixed value and the other objects covering the full redshift range. The model tries to account for this highly unbalanced distribution in the redshift space when being optimized. Generating broader distributions for one part of the objects while creating very narrow estimates for the stellar population is hard to be achieved by a generalizing model. The second effect observed is the presence of outliers where the CDF indicates an extreme value. This is caused by objects in the high redshift range being highly underrepresented in the training sample. The DCMDN architecture has a superior capability of generalization and makes better use of the information contained in the original images, giving a boost in the final performance, together with a good separation of the stars from quasars.

5. Conclusions

The aim of this work is to present and test a new method for photometric redshift estimation. The novelty of the presented approach is to estimate probability density functions for redshifts based on imaging data. The final goal is to make the additional steps of feature-extraction and feature-selection obsolete. To achieve this, a deep convolutional network architecture was combined with a mixture density network. Essential for a proper training is the use of the CRPS as loss function, taking into account not only the location but also the shape and layout of the estimated density distributions. The new architecture is described in a general and conceptional way that allows using the concept for many other regression tasks in astronomy.

In order to perform a fair evaluation of the performance of the proposed method, three experiments have been performed. Three different catalogs have been utilized for evaluation, containing galaxies, quasars and a mix of the previous catalogs plus a sample of stars. The experiments were chosen to test the performance on different redshift ranges and different sources. The last experiment was designed to test the model in a more realistic scenario, where a contamination with stellar sources and therefore a confusion between no and high redshift objects is synthetically introduced. This experiment is intended to test whether the pre-classification of objects can be omitted, too.

In all three experiments a modified version of the random forest and a mixture density network are used as feature-based reference models. To be comparable to the literature, the traditional scores and diagnostic plots have been used. As we demonstrate, the usual way of expressing the results of the prediction quality through the traditional scores is not able to capture the complexity of often multi-modal and asymmetric distributions that are required to correctly describe the redshift estimates. Therefore, proper scores and tools have been applied in addition to analyzing the results in a probabilistic way, that is, the CRPS and PIT. These indicators can be considered proper tools to estimate the quality of photometric redshift density distributions. As summarized in Table 6, the DCMDN architecture outperforms the two reference models. The same relative performances can be observed with the traditional diagnostics, too. When using the PDFs to find and exclude ambiguous objects, the DCMDN

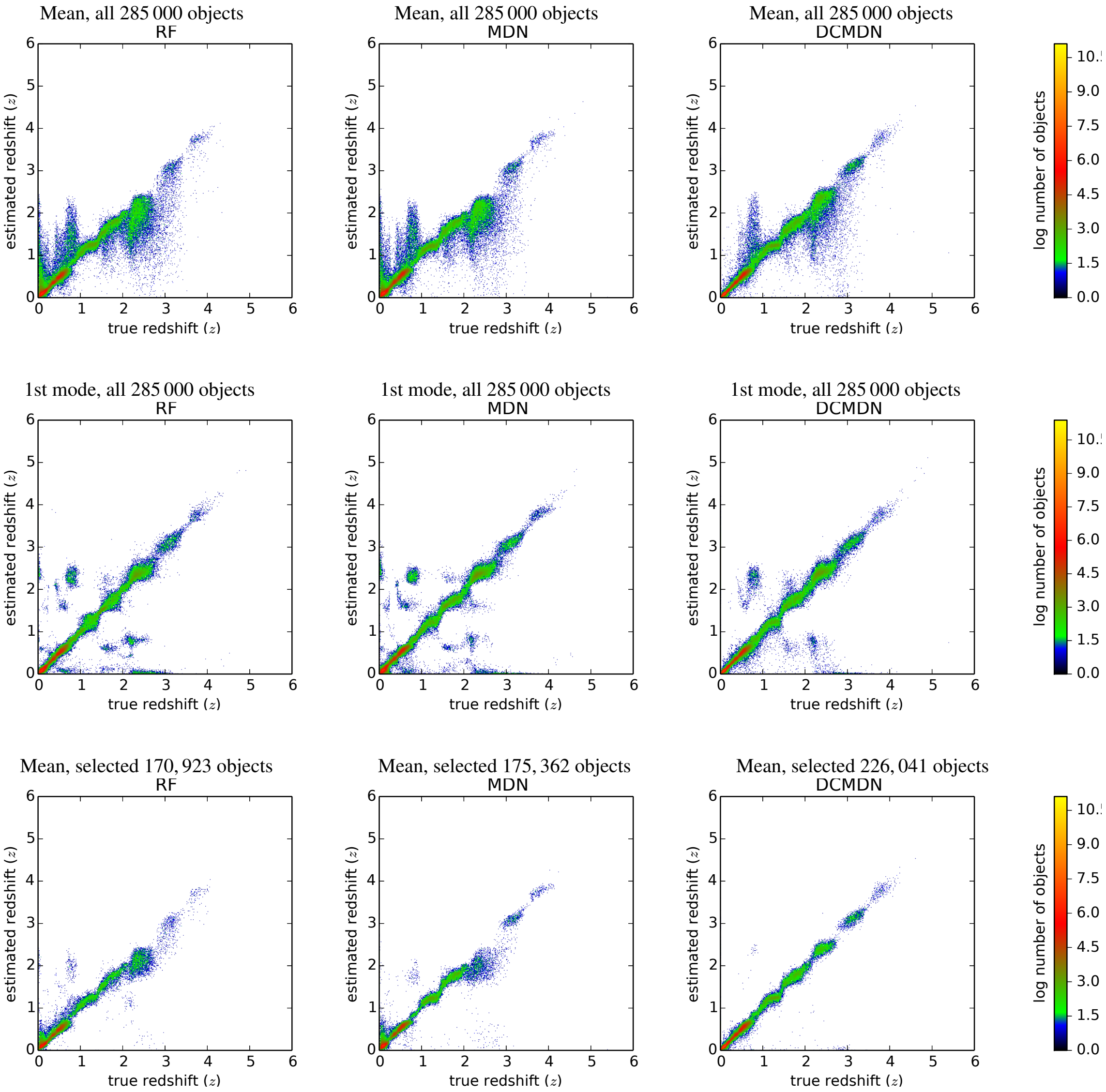


Fig. 8. Results of Experiment 3 (mixed). The estimated redshifts are plotted against the spectroscopic redshifts. As in the figures of the two other experiments, the plots are sorted by the models (*left to right*) and the extracted point estimate (*top to bottom*).

Table 5. Results of Experiment 3 (mixed).

Criterion	Model	Bias(Δz)	$\sigma^2(\Delta z)$	Mad(Δz)	rmse(Δz)	Bias($\Delta z_{\rm norm}$)	$\sigma^2(\Delta z_{\rm norm})$	Mad($\Delta z_{\rm norm}$)	rmse($\Delta z_{\rm norm}$)
Mean	RF	–0.001	0.288	0.043	0.536	–0.072	0.126	0.028	0.363
	MDN	–0.006	0.279	0.043	0.528	–0.073	0.125	0.031	0.362
	DCMDN	0.007	0.210	0.022	0.458	–0.041	0.089	0.016	0.301
First mode	RF	0.040	0.435	0.020	0.660	–0.029	0.150	0.013	0.388
	MDN	–0.027	0.393	0.024	0.627	–0.067	0.191	0.016	0.442
	DCMDN	–0.001	0.287	0.018	0.536	–0.036	0.124	0.013	0.355
Selected	RF	–0.001	0.118	0.016	0.343	–0.029	0.059	0.012	0.245
	MDN	–0.006	0.114	0.023	0.337	–0.034	0.052	0.017	0.230
	DCMDN	0.002	0.043	0.012	0.206	–0.007	0.014	0.010	0.120

Notes. Similarly to experiment 1 and 2, the traditional scores have been calculated.

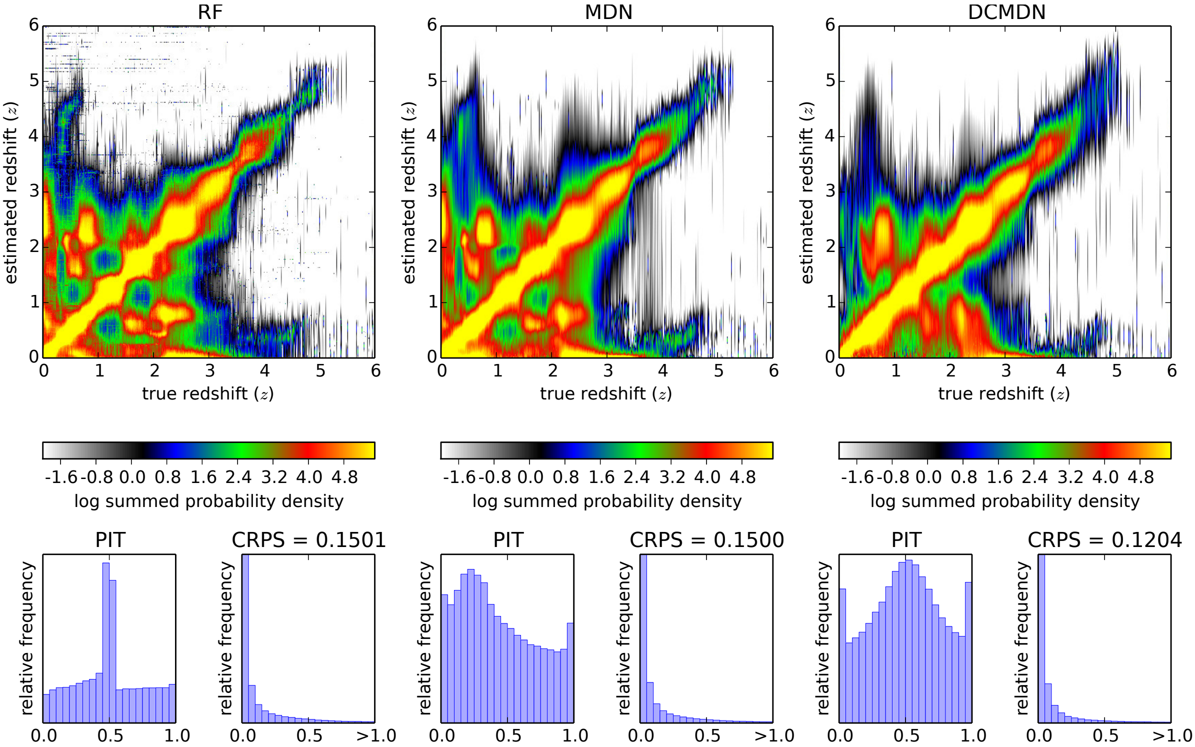


Fig. 9. Results of Experiment 3 (mixed) with a fully probabilistic evaluation. As in Fig. 5 three plots per model are used to visualize the probabilistic performance of the estimated PDFs.

Table 6. Summary table of all the experiments.

Exp.	Model	CRPS	PIT
Galaxies	RF	0.021	well calibrated
	MDN	0.022	biased
	DCMDN	0.021	slightly overdispersed
Quasars	RF	0.187	well calibrated
	MDN	0.190	well calibrated, few outliers
	DCMDN	0.167	well calibrated, few outliers
Mixed	RF	0.150	extremely overdispersed
	MDN	0.150	overdispersed and biased
	DCMDN	0.120	overdispersed, some outliers

always produces the best results. Its ability to generalize from the training data, together with the larger amount of used available information permits a better probabilistic estimate of the redshift distribution and therefore a better selection of spurious predictions.

As shown in the last experiment, the DCMDN performs two very important tasks automatically. During the training of the network, a set of thousands of features is automatically extracted and selected from the imaging data. This minimizes the biases that are introduced when manually extracting and selecting features and increases the amount of utilized information. The second and probably most important capability of the DCMDN is to solve the regression problem without the necessity of pre-classifying the objects. The estimated PDFs reflect in the distribution of the probabilities the uncertainties of the classification as well as the uncertainties of the redshift estimation. This is extremely important when dealing with data from larger surveys. The errors introduced through a hard classification into distinct classes limits the ability of finding rare but interesting objects, like high redshifted quasars that are easily mistaken with cool faint stars. A wrong initial classification would mark those quasars as stellar components even though the probability of being a high redshifted object is not negligible. A fully probabilistic approach including feature extraction, feature selection and source classification is less affected by selection biases.

The performance of the two feature-based reference methods is in accordance with results from the literature. The boost in performance observed for the DCMDN is related to the better use of information and the automatic selection of the best performing features. As the DCMDN model was trained with data from the SDSS it could in principle be applied to every source in the SDSS database, without concern for the nature of the source, by directly using the images. Only the selection biases that have been introduced when selecting the targets for spectroscopic follow-up observations in SDSS have to be considered, as those are preserved by the trained model. Our approach provides the machinery to deal with the avalanche of data we are facing with the new generation of survey instruments already today. Such fully-automatized and probabilistic approaches, based on deep learning architectures, are therefore necessary.

Acknowledgements. The authors gratefully acknowledge the support of the Klaus Tschira Foundation. We gratefully thank the anonymous referee for the constructive and precise comments, leading to a more detailed discussion of the loss functions and the number of Gaussian components used. We would like to thank Nikos Gianniotis and Erica Hopkins for proofreading and commenting on this work. This work is based on the usage of Theano (Theano Development Team 2016) to run our experiments. Topcat (Taylor 2005) was used to access and visualize data from the VO and to create the used catalogs. This work is based on data provided by SDSS. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

References

Abdalla, F. B., Banerji, M., Lahav, O., & Rashkov, V. 2011, MNRAS, 417, 1891
Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, ApJS, 203, 21
Ball, N. M., & Brunner, R. J. 2010, Int. J. Mod. Phys. D, 19, 1049
Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371
Benítez, N. 2000, ApJ, 536, 571
Bishop, C. M. 1994, Mixture density networks, Tech. Rep., Aston University
Blake, C., & Bridle, S. 2005, MNRAS, 363, 1329
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Breiman, L. 2001, Mach. Learn., 45, 5
Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772
Brescia, M., Cavuoti, S., Longo, G., & De Stefano, V. 2014, A&A, 568, A126
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Carrasco Kind, M., & Brunner, R. J. 2013, MNRAS, 432, 1483
Cavuoti, S., Brescia, M., Tortora, C., et al. 2015, MNRAS, 452, 3100
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
D'Isanto, A., & Polsterer, K. L. 2017, in ESANN 2017, 25th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 26–28, 2017, Proceedings
Dubath, P., Apostolakos, N., Bonchi, A., et al. 2017, IAU Symp., 325, 73
Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Pascarelle, S. M., & Yahata, N. 2001, ApJS, 135, 41
Fernique, P., Allen, M. G., Boch, T., et al. 2015, A&A, 578, A114
Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. 2005, Mon. Weather Rev., 133, 1098
Grimit, E. P., Gneiting, T., Berrocal, V. J., & Johnson, N. A. 2006, Quarterly J. Roy. Meteorol. Soc., 132, 2925
Hersbach, H. 2000, Weather and Forecasting, 15, 559
Hoyle, B. 2016, Astron. Comput., 16, 34
Kügler, S. D., Gianniotis, N., & Polsterer, K. L. 2016, in 2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016, Athens, Greece, December 6–9, 2016, 1
Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, in Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave, Proc. SPIE, 8442, 84420
Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, MNRAS, 418, 2165
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Pâris, I., Petitjean, P., Aubourg, É., et al. 2012, A&A, 548, A66
Polsterer, K. L., Zinn, P.-C., & Gieseke, F. 2013, MNRAS, 428, 226
Rosenblatt, F. 1962, Principles of neurodynamics: perceptrons and the theory of brain mechanisms, Report (Cornell Aeronautical Laboratory) (Spartan Books)
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360
Schwarz, G. 1978, Ann. Stat., 6, 461
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Theano Development Team. 2016, ArXiv e-prints [arXiv:1605.02688]

Appendix A: CRPS and PIT

The CRPS is defined by the relation:

$$\mathrm{CRPS} = \mathrm{CRPS}(F, x_a) = \int_{-\infty}^{+\infty} [F(x) - F_a(x)]^2 \mathrm{d}x \tag{A.1}$$

where $F(x)$ and $F_a(x)$ are the CDFs relative to the predicted PDF $f(t)$ and the observation x_a, respectively. Namely: $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ and $F_x = H(x - x_a)$, with $H(x)$ being the Heaviside step-function. In case the PDF is given as a GMM, the CRPS can be calculated through the following formula in closed form:

$$\mathrm{CRPS}\Big(\sum_{m=1}^{M} \omega_m \mathcal{N}(\mu_m, \sigma_m^2)\Big) = \sum_{m=1}^{M} \omega_m A\left(x - \mu_m, \sigma_m^2\right)$$

$$- \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} \omega_m \omega_n A\left(\mu_m - \mu_n, \sigma_m^2 + \sigma_n^2\right) \tag{A.2}$$

where

$$A(\mu, \sigma^2) = 2\sigma\phi\Big(\frac{\mu}{\sigma}\Big) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right) \tag{A.3}$$

and $\phi\Big(\frac{y-\mu}{\sigma}\Big)$, $\Phi\Big(\frac{y-\mu}{\sigma}\Big)$ respectively represent the PDF and the CDF of a normal distribution with a mean of zero and a variance of one evaluated through the normalized prediction error $\frac{y-\mu}{\sigma}$.

The probability integral transform (PIT) is generated with the histogram of the values:

$$p_t = F_t(x_t) \tag{A.4}$$

being F_t the CDF of the predicted PDF evaluated at the observation x_t. In Fig. A.1 some example PITs are given.

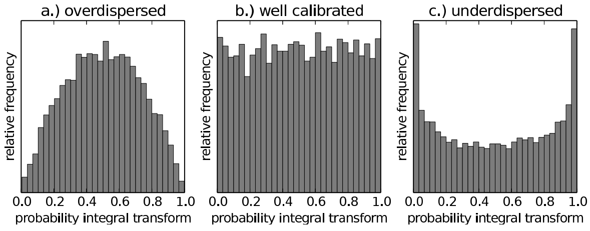


Fig. A.1. Visual guide to the usage of a PIT. In case the estimated PDFs are to broad with respect to the position of the true value, a convex histogram with a peak in the center can be observed (*a*). As soon as the predicted densities are too narrow, the evaluation of the CDF at the true redshift exhibits in most cases just very low and very high values. Therefore a concave, U-shaped histogram will be produced (*c*). Only in the case where the widths of the predicted densities is in accordance with the deviations from the true measurements, a uniformly distributed histogram is generated (*b*). This indicates sharp and well calibrated predictions.

Appendix B: CRPS vs. log-likelihood as loss-function

We compared the performance of the MDN when trained with the log-likelihood and the CRPS as loss functions, in order to investigate the differences in the behavior of the model. The experiment was performed on the data used for the quasar experiment, split in independent subsets for training (100 000 objects) and testing (85 000 objects). The results are shown in Fig. B.1. In this plot we report the performance expressed by using both the log-likelihood and the CRPS as score functions, at different epochs of the training phase. Moreover, the correspondent PIT histograms for these epochs are shown. Training the network with the log-likelihood improves the performance in terms of the log-likelihood itself, respect to the same architecture trained using the CRPS. As expected, when training with the CRPS, the observed results are opposite to the previous case. The PIT histograms indicate a better performance when the neural network is trained using the CRPS, leading to a well calibrated PIT already after 500 epochs. Moreover, the PIT of the model trained using the log-likelihood starts to degrade again at 10 000 epochs. This does not happen when using the CRPS for training, as the CRPS accounts for calibration and sharpness of the predictions. For this reason, the choice of the CRPS as loss function is reasonable, in order to obtain sharper and better calibrated PDFs.

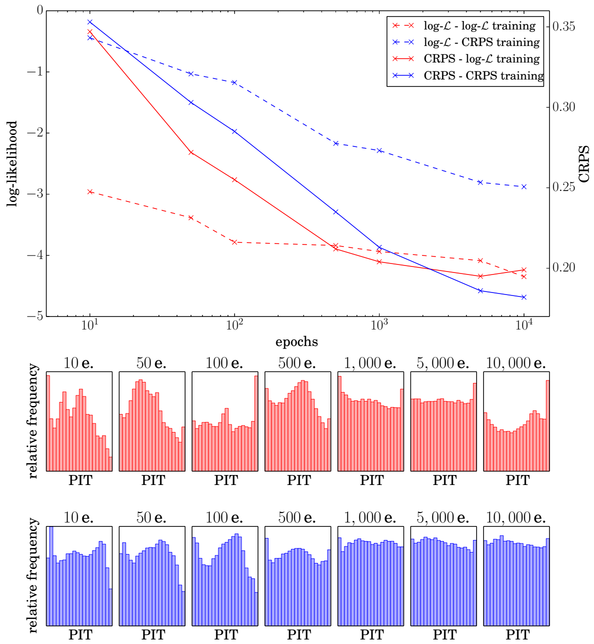


Fig. B.1. Comparison of the performance when using the log-likelihood (red) and the CRPS (blue) as loss-functions for training the MDN. For different epochs the CRPS and log-likelihood scores as well as the PIT histograms are provided.

Appendix C: Number of Gaussian components

In order to choose an appropriate number of Gaussian components, an experiment has been performed using the RF model for the quasar data-set and calculating the BIC following Schwarz (1978). The mean score has been calculated over different redshift bins, for different numbers of Gaussian components. The plot shows that extreme values, like $n = 1$ or $n = 7$ tend to exhibit bad performance in multiple regions. Instead, the results using $n = 3$ and $n = 5$ are comparably good. The BIC score is based on the log-likelihood calculation, therefore we consider a choice of $n = 5$ to be reasonable for this work, keeping in mind that our model is trained using the CRPS.

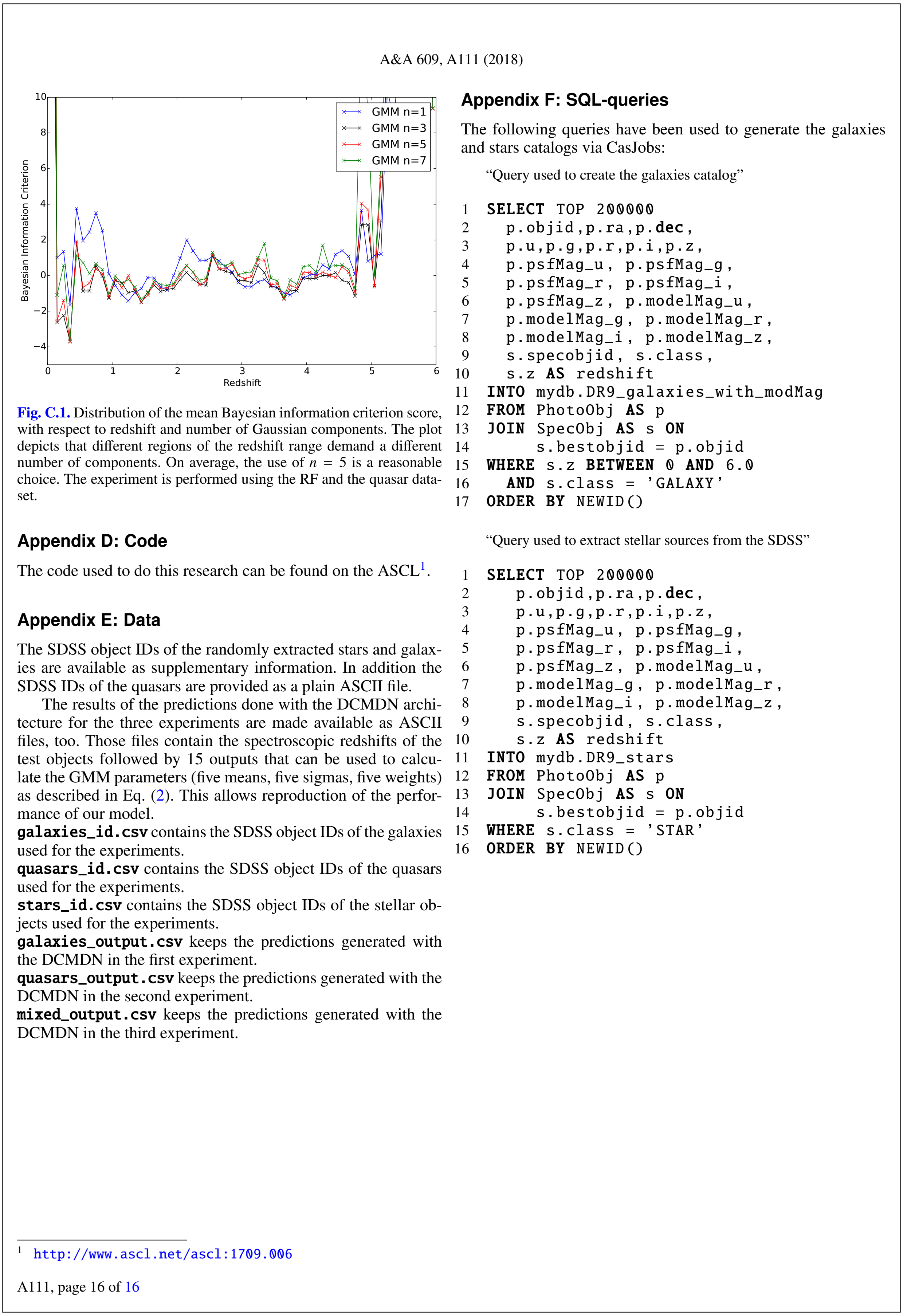


Fig. C.1. Distribution of the mean Bayesian information criterion score, with respect to redshift and number of Gaussian components. The plot depicts that different regions of the redshift range demand a different number of components. On average, the use of $n = 5$ is a reasonable choice. The experiment is performed using the RF and the quasar dataset.

Appendix D: Code

The code used to do this research can be found on the ASCL[1].

Appendix E: Data

The SDSS object IDs of the randomly extracted stars and galaxies are available as supplementary information. In addition the SDSS IDs of the quasars are provided as a plain ASCII file.

The results of the predictions done with the DCMDN architecture for the three experiments are made available as ASCII files, too. Those files contain the spectroscopic redshifts of the test objects followed by 15 outputs that can be used to calculate the GMM parameters (five means, five sigmas, five weights) as described in Eq. (2). This allows reproduction of the performance of our model.

galaxies_id.csv contains the SDSS object IDs of the galaxies used for the experiments.
quasars_id.csv contains the SDSS object IDs of the quasars used for the experiments.
stars_id.csv contains the SDSS object IDs of the stellar objects used for the experiments.
galaxies_output.csv keeps the predictions generated with the DCMDN in the first experiment.
quasars_output.csv keeps the predictions generated with the DCMDN in the second experiment.
mixed_output.csv keeps the predictions generated with the DCMDN in the third experiment.

Appendix F: SQL-queries

The following queries have been used to generate the galaxies and stars catalogs via CasJobs:

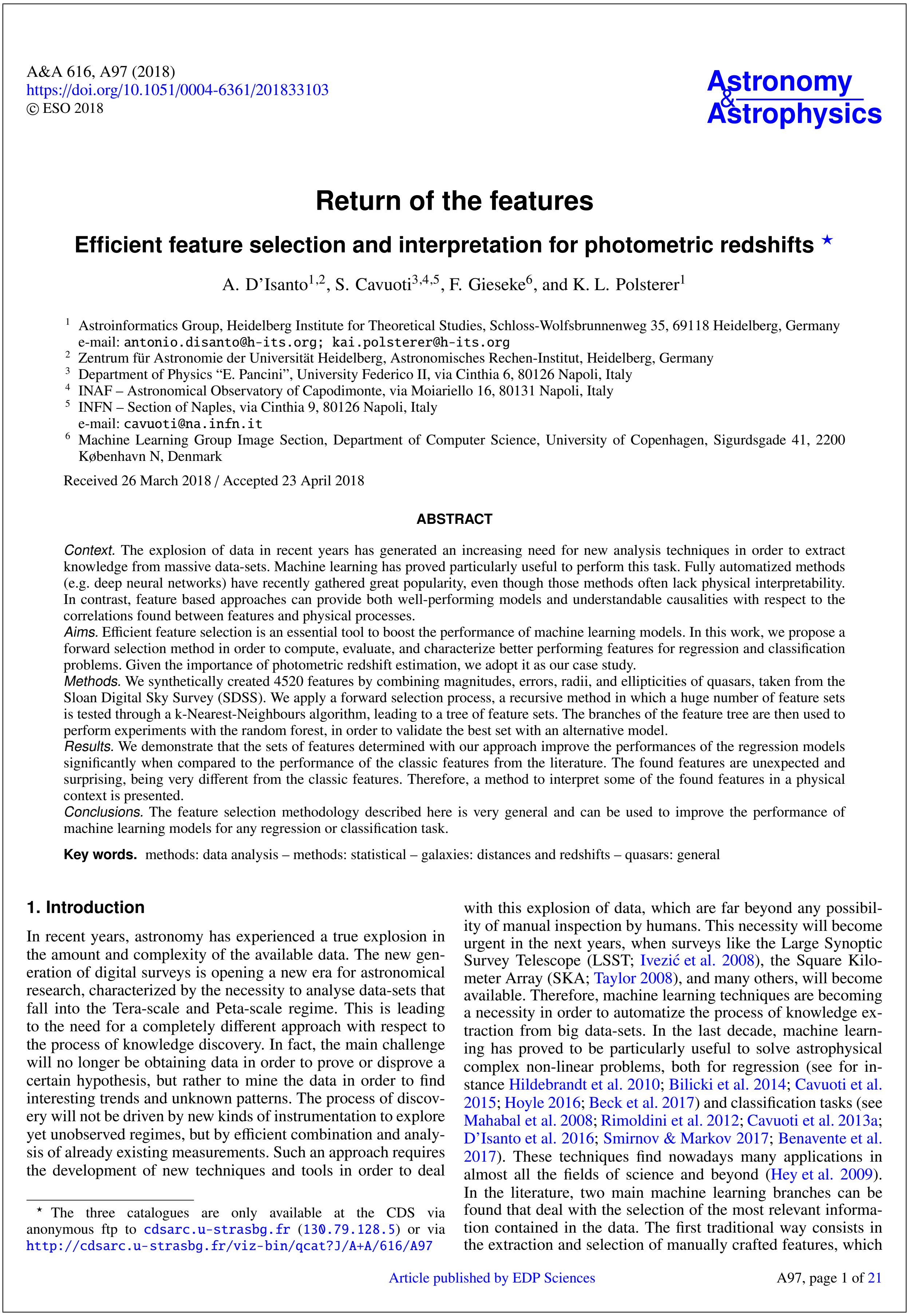
"Query used to create the galaxies catalog"

```
SELECT TOP 200000
  p.objid,p.ra,p.dec,
  p.u,p.g,p.r,p.i,p.z,
  p.psfMag_u, p.psfMag_g,
  p.psfMag_r, p.psfMag_i,
  p.psfMag_z, p.modelMag_u,
  p.modelMag_g, p.modelMag_r,
  p.modelMag_i, p.modelMag_z,
  s.specobjid, s.class,
  s.z AS redshift
INTO mydb.DR9_galaxies_with_modMag
FROM PhotoObj AS p
JOIN SpecObj AS s ON
     s.bestobjid = p.objid
WHERE s.z BETWEEN 0 AND 6.0
  AND s.class = 'GALAXY'
ORDER BY NEWID()
```

"Query used to extract stellar sources from the SDSS"

```
SELECT TOP 200000
   p.objid,p.ra,p.dec,
   p.u,p.g,p.r,p.i,p.z,
   p.psfMag_u, p.psfMag_g,
   p.psfMag_r, p.psfMag_i,
   p.psfMag_z, p.modelMag_u,
   p.modelMag_g, p.modelMag_r,
   p.modelMag_i, p.modelMag_z,
   s.specobjid, s.class,
   s.z AS redshift
INTO mydb.DR9_stars
FROM PhotoObj AS p
JOIN SpecObj AS s ON
     s.bestobjid = p.objid
WHERE s.class = 'STAR'
ORDER BY NEWID()
```

[1] http://www.ascl.net/ascl:1709.006



A&A 616, A97 (2018)
https://doi.org/10.1051/0004-6361/201833103
© ESO 2018

Astronomy & Astrophysics

Return of the features

Efficient feature selection and interpretation for photometric redshifts *

A. D'Isanto[1,2], S. Cavuoti[3,4,5], F. Gieseke[6], and K. L. Polsterer[1]

[1] Astroinformatics Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: antonio.disanto@h-its.org; kai.polsterer@h-its.org
[2] Zentrum für Astronomie der Universität Heidelberg, Astronomisches Rechen-Institut, Heidelberg, Germany
[3] Department of Physics "E. Pancini", University Federico II, via Cinthia 6, 80126 Napoli, Italy
[4] INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, 80131 Napoli, Italy
[5] INFN – Section of Naples, via Cinthia 9, 80126 Napoli, Italy
e-mail: cavuoti@na.infn.it
[6] Machine Learning Group Image Section, Department of Computer Science, University of Copenhagen, Sigurdsgade 41, 2200 København N, Denmark

Received 26 March 2018 / Accepted 23 April 2018

ABSTRACT

Context. The explosion of data in recent years has generated an increasing need for new analysis techniques in order to extract knowledge from massive data-sets. Machine learning has proved particularly useful to perform this task. Fully automatized methods (e.g. deep neural networks) have recently gathered great popularity, even though those methods often lack physical interpretability. In contrast, feature based approaches can provide both well-performing models and understandable causalities with respect to the correlations found between features and physical processes.

Aims. Efficient feature selection is an essential tool to boost the performance of machine learning models. In this work, we propose a forward selection method in order to compute, evaluate, and characterize better performing features for regression and classification problems. Given the importance of photometric redshift estimation, we adopt it as our case study.

Methods. We synthetically created 4520 features by combining magnitudes, errors, radii, and ellipticities of quasars, taken from the Sloan Digital Sky Survey (SDSS). We apply a forward selection process, a recursive method in which a huge number of feature sets is tested through a k-Nearest-Neighbours algorithm, leading to a tree of feature sets. The branches of the feature tree are then used to perform experiments with the random forest, in order to validate the best set with an alternative model.

Results. We demonstrate that the sets of features determined with our approach improve the performances of the regression models significantly when compared to the performance of the classic features from the literature. The found features are unexpected and surprising, being very different from the classic features. Therefore, a method to interpret some of the found features in a physical context is presented.

Conclusions. The feature selection methodology described here is very general and can be used to improve the performance of machine learning models for any regression or classification task.

Key words. methods: data analysis – methods: statistical – galaxies: distances and redshifts – quasars: general

1. Introduction

In recent years, astronomy has experienced a true explosion in the amount and complexity of the available data. The new generation of digital surveys is opening a new era for astronomical research, characterized by the necessity to analyse data-sets that fall into the Tera-scale and Peta-scale regime. This is leading to the need for a completely different approach with respect to the process of knowledge discovery. In fact, the main challenge will no longer be obtaining data in order to prove or disprove a certain hypothesis, but rather to mine the data in order to find interesting trends and unknown patterns. The process of discovery will not be driven by new kinds of instrumentation to explore yet unobserved regimes, but by efficient combination and analysis of already existing measurements. Such an approach requires the development of new techniques and tools in order to deal with this explosion of data, which are far beyond any possibility of manual inspection by humans. This necessity will become urgent in the next years, when surveys like the Large Synoptic Survey Telescope (LSST; Ivezić et al. 2008), the Square Kilometer Array (SKA; Taylor 2008), and many others, will become available. Therefore, machine learning techniques are becoming a necessity in order to automatize the process of knowledge extraction from big data-sets. In the last decade, machine learning has proved to be particularly useful to solve astrophysical complex non-linear problems, both for regression (see for instance Hildebrandt et al. 2010; Bilicki et al. 2014; Cavuoti et al. 2015; Hoyle 2016; Beck et al. 2017) and classification tasks (see Mahabal et al. 2008; Rimoldini et al. 2012; Cavuoti et al. 2013a; D'Isanto et al. 2016; Smirnov & Markov 2017; Benavente et al. 2017). These techniques find nowadays many applications in almost all the fields of science and beyond (Hey et al. 2009). In the literature, two main machine learning branches can be found that deal with the selection of the most relevant information contained in the data. The first traditional way consists in the extraction and selection of manually crafted features, which

* The three catalogues are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/616/A97

are theoretically more suitable to optimize the performance. In Donalek et al. (2013) feature selection strategies are compared in an astrophysical context.

The second option is using automatic feature selection models and became more popular in more recent years. For example, Athiwaratkun & Kang (2015) delegate this task to the machine by analysing the automatically extracted feature representations of convolutional neural networks. In convolutional neural networks, during the training phase the model itself determines and optimizes the extraction of available information in order to obtain the best performance. The challenge of feature selection is fundamental for machine learning applications, due to the necessity of balancing between overfitting and the curse of dimensionality (Bishop 2006), which arises when dealing with very high-dimensional spaces. Therefore a clever process of feature selection is needed to overcome this issue. In this setting, a different strategy was chosen for this work, in which a forward selection algorithm (Guyon & Elisseeff 2003) is adopted to identify the best performing features out of thousands of them. We decided to apply this procedure in a very important field: photometric redshift estimation. Due to the enormous importance that this measurement has in cosmology, great efforts have been lavished by the astronomical community on building efficient methods for the determination of affordable and precise photometric redshifts (Richards et al. 2001; Hildebrandt et al. 2008, 2010; Ball et al. 2008). Photometric redshifts are of extreme importance with respect to upcoming missions, for example the forthcoming Euclid mission (Laureijs et al. 2011), which will be based on the availability of photometric redshift measures, and the Kilo Degree Survey (KiDS; de Jong et al. 2017), which aims to map the large-scale matter distribution in the Universe, using weak lensing shear and photometric redshift measurements (Hildebrandt et al. 2016; Tortora et al. 2016; Harnois-Déraps et al. 2017; Joudaki et al. 2017; Köhlinger et al. 2017). Furthermore, photometric redshifts estimation is crucial for several other projects, the most important being the Evolutionary Map of the Universe (EMU; Norris et al. 2011), the Low Frequency Array (LOFAR; van Haarlem et al. 2013), Dark Energy Survey (Bonnett et al. 2016), the Panoramic Survey Telescope and Rapid Response System (PANSTARRS; Chambers et al. 2016), and the VST Optical Imaging of the CDFS and ES1 Fields (VST-VOICE; Vaccari et al. 2016). In light of this, we propose to invert the task of photometric redshift estimation. That is to say, having stated the possibility to determine the redshift of a galaxy based on its photometry, we want to build a method that allows us to investigate the parameter space and to extract the features to be used to achieve the best performance. As thoroughly analysed in D'Isanto & Polsterer (2018), the implementation of deep learning techniques is providing an alternative to feature based methods, allowing the estimation of photometric redshifts directly from images. The main concerns when adopting deep learning models are related to the amount of data needed to efficiently perform the training of the networks, the cost in terms of resources and computation time, and the lack of interpretability related to the features automatically extracted. In fact, deep learning models can easily become like magic boxes and it is really hard to assign any kind of physical meaning to the features estimated by the model itself. Therefore, a catalogue-based approach still has great importance, due to the gains in time, resources, and interpretability. In particular, this is true if a set of significant features is provided, in order to concentrate the important information with respect to the problem in a reduced number of parameters. Both methods, based on automatically extracted features or on selected features, constitute the starting point to build an efficient and performing model for redshift estimation, respectively. The topic of feature selection is a well-treated subject in the literature (see for example Rimoldini et al. 2012; Tangaro et al. 2015; Hoyle et al. 2015; D'Isanto et al. 2016). The forward selection approach we used (Gieseke et al. 2014) is meant to select between thousands of features generated by combining plain photometric features as they are given in the original catalogue. No matter what selection strategy is applied, the final results have to be compared to those obtained with the traditional features from the literature (D'Abrusco et al. 2007; Richards et al. 2009; Laurino et al. 2011) and with automatically extracted features. The aim is to find the subsets that give a better performance for the proposed experiments, mining into this new, huge feature space and to build a method useful to find the best features for any kind of problem. Moreover, we propose to analyse the obtained features, in order to give them a physical explanation and a connection with the processes occurring in the specific category of sources. Such an approach also demands a huge effort in terms of computational time and resources. Therefore, we need an extreme parallelization to deal with this task. This has been done through the intensive use of graphics processing units (GPU), a technology that is opening new doors for Astroinformatics (Cavuoti et al. 2013b; Polsterer et al. 2015; D'Isanto & Polsterer 2018), allowing the adoption of deep learning and/or massive feature selection strategies. In particular, in this work, the feature combinations are computed following Gieseke et al. (2014) and Polsterer et al. (2014), using a GPU cluster equipped with four Nvidia Pascal P40 graphic cards[1]. Likewise for Zhang et al. (2013), the k-Nearest-Neighbours (kNN; Fix & Hodges 1951) model is used, running recursive experiments in order to estimate the best features through the forward selection process. This choice has been done because the kNN model scales very well with the use of GPU, with respect to performance and quality of the prediction, as shown in Heinermann et al. (2013). In this way, for each run of the experiment, the most contributing features are identified and added to previous subsets. Thereby, a tree of feature groups is created that afterwards can be compared with the traditional ones. The validation experiments are performed using a random forest (RF) model (application in astronomy Carliles et al. 2010). We will show that this approach can strongly improve performance for the task of redshift estimation. The improvement is due to the identification of specific feature subsets containing more information and capable of better characterizing the physics of the sources. In the present work, we perform the experiments on quasar data samples extracted from the Sloan Digital Sky Survey Data Release 7 (SDSS DR7; Abazajian et al. 2009) and Data Release 9 (SDSS DR9; Ahn et al. 2012). The proposed approach is very general and could be also used to solve many other tasks in astronomy, including both regression and classification problems.

In Sect. 2 the methodology and models used to perform the experiments are described together with the statistical estimators used to evaluate the performance. The strategy adopted for the feature selection is also explained. Section 3 is dedicated to the data used and the feature extraction process. In Sect. 4 the experiments performed and the results obtained are described. Finally, in Sect. 5 the results are discussed in detail and in Sect. 6 some conclusions are drawn.

[1] https://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf

2. Methods

The main purpose of this work is to build an efficient method capable of generating, handling and selecting the best features for photometric redshift estimation, even though the proposed method is also able to deal with any other task of regression or even classification. We calculate thousands of feature combinations of photometric data taken from quasars. Then, a forward selection process is applied, as will be explained in more detail in the next sections. This is done to build a tree of best performing feature subsets. This method has to be considered as an alternative to the automatic features extraction used in D'Isanto & Polsterer (2018). Both methods can be useful and efficient, depending on the nature of the problem, and on the availability of data and resources. For this reason, the results obtained with both methods will be compared. The experimental strategy is based on the application of two different machine learning models and evaluated on the basis of several statistical tools. In the following these models, kNN and RF, are presented. The strategy used to perform the feature selection is then depicted in detail and we give a description of the statistical framework used for the experiments' evaluation and of the cross validation algorithm.

2.1. Regression models

As mentioned above, our method makes use of kNN and RF models, which are described in detail in the following subsections, while the details regarding the deep convolutional mixture density network (DCMDN) used to compare the results with an automatic features extraction based model can be found in D'Isanto & Polsterer (2018).

2.1.1. kNN

The kNN (Fix & Hodges 1951) is a machine learning model used both for regression and classification tasks (Zhang et al. 2013). This model explores the feature space by estimating the k nearest points (or neighbours) belonging to the training sample with respect to each test item. In our case the distance involved is calculated through a Euclidean metric. In the case of a regression problem (like redshift estimation), the kNN algorithm is used to find a continuous variable averaging the distances of the k selected neighbours. The efficiency of the algorithm is strongly related to the choice of the parameter k, which represents the number of neighbours to be selected from the training set. The best choice of this parameter is directly related to the input data, their complexity, and the way in which the input space is sampled. Clearly, the most simple case is a model with $k = 1$. In this case, a prediction equal to the target of the closest pattern in the training set is associated to each pattern. Increasing the k parameter could improve the precision of the model (this is due to the increasing generalization capability), but can also generate overfitting (Duda et al. 2000). In our experiments, the choice of the k parameter was part of the learning task by evaluating a set of possible values. The kNN is one of the simplest machine learning algorithms, but even if it could be outperformed by more complex models, it has the advantage of being very fast and in any case quite efficient. Another possible problem concerning the use of the kNN model is given by possible differences in the range of the input features. This could generate problems and misleading results in the estimation of distances in the parameter space. For this reason, all the features used in this work have been normalized using the min-max normalization technique (Aksoy & Haralick 2000).

2.1.2. Random forest

The RF (Breiman et al. 1984) is one of the most popular ensemble-based machine learning models, and could be used for regression and classification tasks (see Carliles et al. 2010, for an application to photometric redshift estimation). It is an ensemble of decision trees, where each tree is meant to partition the feature space in order to find the best split that minimizes the variance. Each decision tree is built by adding leaf nodes where the input data are partitioned with respect to a different chosen feature, repeating the process for all the possible choices of variables to be split. In case of a regression problem, the root mean square error (RMSE) is computed for each possible partition, and the partition which minimizes the RMSE is chosen. The RF averages the results provided by many decision trees, each trained on a different part of the training set through the bagging technique (Breiman 1996). This avoids overfitting due to single decision trees growing too deep. Moreover, the decision tree makes use of the bootstrapping technique (Breiman 1996) in order to increase the performance and stability of the method and reduce overfitting at the same time. This consists in giving, as input, a different random sub-sample of the training data to each decision tree. The RF uses the feature bagging during the training phase. This consists in selecting a random subset of features at each split. Bootstrapping and bagging help to avoid correlations between single decision trees, which could appear when training them on the same training set and in the presence of strong features selected multiple times.

2.2. Features selection strategy

The huge number of features evaluated, as described in Sect. 3, imposes the need to establish an efficient feature selection process. In fact, in order to estimate a subset of the best $f = 10$ features[2], starting with $r = 4520$ features, would imply, if we want to test all the possible combinations, the following number of experiments:

$$n = \frac{r!}{f! * (r - f)!} = 9.7 \times 10^{29}. \qquad (1)$$

Assuming that a nonillion experiments are too many to be performed, a more efficient approach had to be chosen. Therefore, we decided to apply a forward selection process (Mao 2004) as described in the following. The number of features used for the experiment was iteratively increased. In other words, to select the first best feature a kNN model for each of the $r = 4520$ features was trained in a one-dimensional feature space. Due to the memory limitations of the hardware architecture used, the feature selection was done by performing 100 kNN experiments, selecting for each of them a random subset of 20 000 data points and using a five-fold cross validation (see Sect. 2.4 for more details). The repeated experiments on different training samples were meant to generate statistics of the features in order to identify the most frequently selected ones. This was done to minimize the biases introduced by the random extraction of the training data. Since 100 runs were performed, sometimes more than one feature was selected. The basic idea behind the proposed strategy is to select a limited number of best-performing features per step. The number of features which were actually selected were chosen by evaluating the occurrence of each of them as the best feature in all of the 100 runs. Therefore, for

[2] The reason for selecting 10 features is discussed in Sect. 4.3 and Fig. 6.

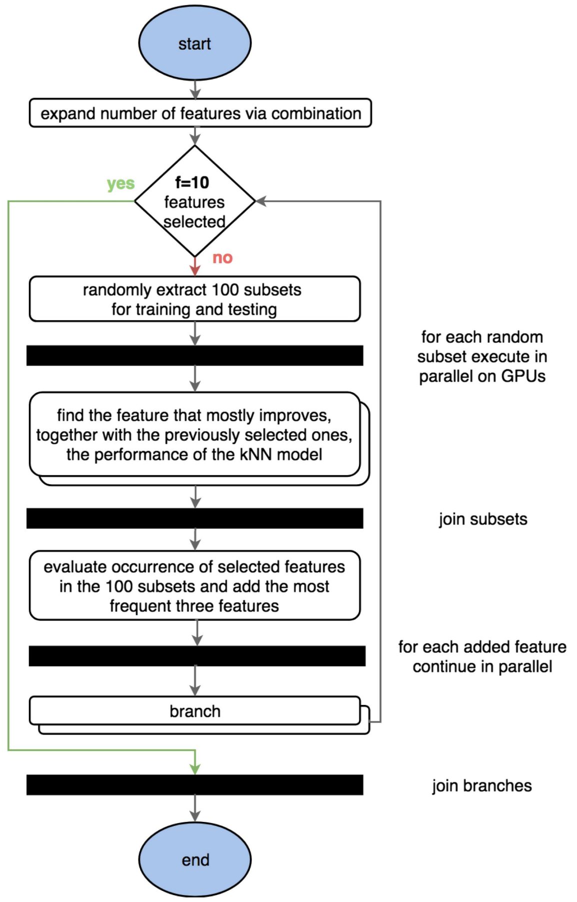

start
expand number of features via combination
yes
f=10 features selected
no
randomly extract 100 subsets for training and testing
for each random subset execute in parallel on GPUs
find the feature that mostly improves, together with the previously selected ones, the performance of the kNN model
join subsets
evaluate occurrence of selected features in the 100 subsets and add the most frequent three features
for each added feature continue in parallel
branch
join branches
end

Fig. 1. Workflow used to generate tree structure. The black boxes represent states where multiple operations are started in parallel or parallel operations are joined. The iteration is stopped when each branch of the tree has a depth of 10. A five-fold cross validation is applied for every model evaluation step.

each iteration a minimum of one and a maximum of three features were selected. After choosing the best features, they were fixed and the next run was performed in order to choose the subsequent features. This method was iterated until the tenth feature was selected. A tree with a maximum branching number of three was derived, because in every step a maximum number of three features that best improve the model were chosen. Each branch can be seen as a set of best-performing-feature combinations. The necessity of performing a high number of experiments on different data subsets is caused by the slightly varying behaviour of the kNN model with respect to different input patterns. The whole workflow is summarized in Fig. 1. The cross validation, moreover, was used in order to further reduce any risk of overfitting.

2.3. GPU parallelization for kNN

The feature selection is done by parallelizing the experiments on a GPU cluster. The massive use of GPUs proved to be mandatory in order to deal with such an amount of data, features, k values, and runs on randomly sampled data-sets. Following Heinermann et al. (2013) and Gieseke et al. (2014), the kNN algorithm has been parallelized by using GPUs. Typically, GPU-based programs are composed by a host program running on central processing unit (CPU) and a kernel program running on the GPU itself, which is parallelized on the GPU cores in several threads or kernel instances. This scheme is particularly adapted to kNN models, due to the advantages obtained by parallelizing matrix multiplications. In the code used for this work (Gieseke et al. 2014) the calculation is performed by generating matrices containing the distances of the selected features from the query object. This calculation is entirely performed on the GPU, while the CPU is mainly used for synchronization and for updating a vector containing the selected features at every step. The approach based on this method proved to speed up the calculation by a factor of ~150. We modified the given code to start the selection process with a given set of already selected features. This was done to enable the generation of the feature trees based on 100 random subsets.

2.4. Statistical estimators and cross validation

The results have been evaluated using the following set of statistical scores for the quantity $\Delta z = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$ expressing the estimation error[3] on the objects in the blind validation set:

- bias: defined as the mean value of the normalized residuals Δz;
- RMSE: root mean square error;
- NMAD: normalized median absolute deviation of the normalized residuals, defined as NMAD(Δz) = 1.48 × median $(|\Delta z_i - \text{median}(\Delta z)|)$;
- CRPS: the continuous rank probability score (Hersbach 2000) is a proper score to estimate how well a single value is represented by a distribution. It is used following D'Isanto & Polsterer (2018).

The prediction of redshifts in a probabilistic framework has many advantages. The ability of reporting the uncertainty is the most important one to mention. In order to correctly evaluate the performance of the features in a probabilistic setting, the CRPS was added to the set of scores. By using the RF as a quantile regression forest and fitting a mixture of Gaussians to the predictions of the ensemble members, a probability distribution can be generated and the CRPS can be calculated. The DCMDN, by definition, predicts density distributions that are represented by their mean when calculating the scores used for point estimates.

As stated before, all the indicators are then averaged on the k folds of the cross validation. Through this approach, the standard deviation is also obtained as a measure of the error on each statistical estimator. We do not report those values as the errors were small enough to be considered negligible. Cross validation (Kohavi 1995) is a statistical tool used to estimate the generalization error. The phenomenon of overfitting arises when the model is too well adapted to the training data. In this case, the performance on the test set will be poor as the model is not general enough. A validation set is defined, in order to test this generalization of the model, with respect to the training data, on an unseen and omitted set of data. In particular, cross validation becomes necessary when dealing with small training sets or high-dimensional feature spaces.

[3] We note that Δz denotes the normalized error in redshift estimation and not the usually used plain error.

In this kind of approach, the data-set is divided into k subsets and each of them is used for the prediction phase, while all the $k-1$ subsets constitute the training set. The training is then repeated k times, using all the subsets. The final performance is obtained by averaging the results of the single folds and the error on the performance is obtained by evaluating the standard deviation of the results coming from the different folds. In this work, we adopt a k-fold cross validation approach, with $k = 5$ for the kNN experiments and $k = 10$ for the RF experiments.

3. Data

In the following subsections the details about the data-set used and the feature combinations performed for the experiments are outlined.

3.1. Data-sets

The experiments are based on quasar data extracted from the SDSS DR7 (Abazajian et al. 2009) and SDSS DR9 (Ahn et al. 2012). Three catalogues have been retrieved for the experiments. Moreover, images for the DCMDN experiments have been downloaded making use of Hierarchical Progressive Survey (HiPS; Fernique et al. 2015).

Catalogue DR7a. Catalogue DR7a is the most conservative with respect to the presence of bad data or problematic objects. It is based on DR7 only, with clean photometry and no missing data; the query used is reported in Appendix D. Furthermore, to be more conservative, we checked the spectroscopic redshifts in two different data releases (9 and 12) and we decided to cut all the objects with a discrepancy in $z_{\rm spec}$ not fulfilling the given criteria

$$|z_{\rm DR7} - z_{\rm DR9}| < 0.01, \text{ and}$$
$$|z_{\rm DR7} - z_{\rm DR12}| < 0.01, \text{ and}$$
$$|z_{\rm DR12} - z_{\rm DR9}| < 0.01.$$

The final catalogue contains 83 982 objects with a spectroscopically determined redshift.

Catalogue DR7b. Catalogue DR7b has been obtained using the same query used for Catalogue DR7a, but removing the image processing flags. This has been done in order to verify if the presence of objects previously discarded by the use of these flags could affect the feature selection process. The catalogue has been cleaned by removing all the objects with NaNs and errors bigger than a value of one, ending with a catalogue containing 97 041 objects.

Catalogue DR7+9. Catalogue DR7+9 has been prepared mixing quasars from DR7 and DR9 in order to perform the feature selection with a different and more complete redshift distribution. The difference in the redshift distribution of the two catalogues can be seen from the histogram in Fig. 2. The catalogue has been cleaned with the same procedure adopted for Catalogue DR7b and the common objects between DR7 and DR9 have been used only once. This produced a catalogue of 152 137 objects. In the following sections, the results obtained with this catalogue are discussed in depth.

3.2. Classic features

In classic redshift estimation experiments for quasars and galaxies, as can be found in the literature (e.g. D'Abrusco et al. 2007), for SDSS data colours are mainly used as features. To be comparable, we decided to use a set of ten features as our benchmark feature set. Colours of the adjacent filterbands for the point spread function (PSF) and model magnitudes are used together with the plain PSF and model magnitudes. In SDSS, the model magnitudes are the best fitting result of an exponential or de Vaucouleurs model. All Classic$_{10}$ features can be found in the first column of Table 2.

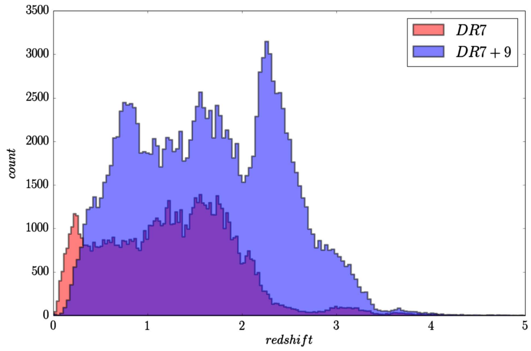


Fig. 2. Histogram showing the redshift distribution of the catalogues with objects from DR7 only and DR7 plus DR9. The distribution for the catalogue DR7b is not reported here because the difference with respect to catalogue DR7a is practically negligible.

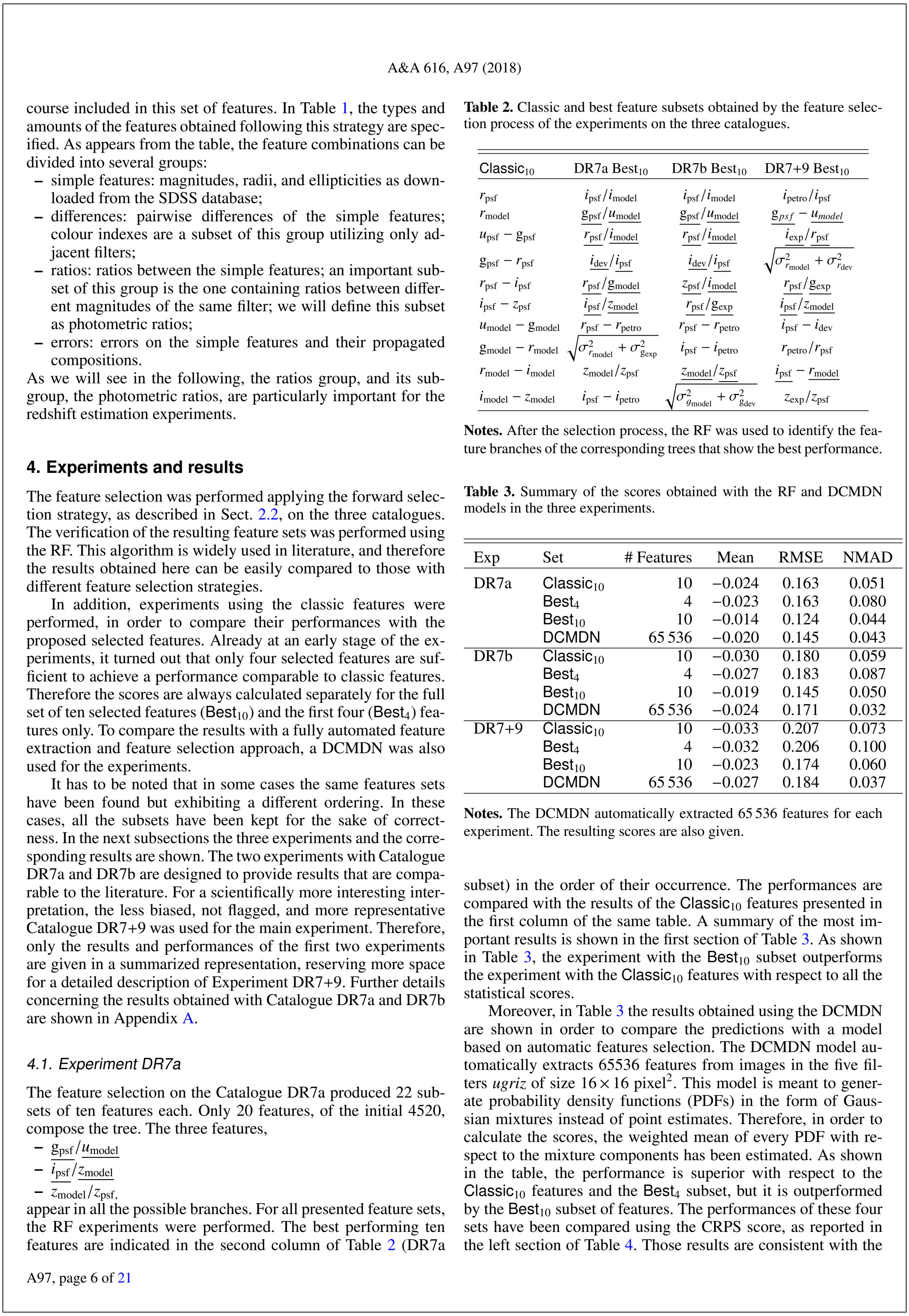
Table 1. Types of features downloaded from SDSS and their combinations in order to obtain the final catalogue used for the experiments.

	Magnitudes	σ	Radii	Ellipticities
	modelMag / Extinction	✓	devRad	devAB
	petroMag / Extinction	✓	expRad	expAB
	psfMag / Extinction	✓	petroRad	
	devMag / Extinction	✓	petroR50	
	expMag / Extinction	✓	petroR90	
Plain	25 + 25 dereddened	25	25	10
Combined	1 225 Differences 2450 Ratios	300 Pairs	300 Differences	45 Differences 90 Ratios
Total 4520	3725	325	325	145

Notes. The number of each feature type is given alongside with the final number of synthetically derived features.

3.3. Combined features

For each of the three catalogues, the features concerning magnitudes and their errors, radii, ellipticities, and extinction are retrieved. An overview of the features is shown in Table 1. Magnitudes that have been corrected for extinction are denoted with an underline indicating that, for example, $\underline{u_{\rm model}}$ is equivalent to $u_{\rm model} - u_{\rm extinction}$. The parameter space has been enriched by performing several combinations of the original features (Gieseke et al. 2014). A similar feature generation approach was applied also in Polsterer et al. (2014) but with a limited set of plain features and combination rules. In other words, the magnitude features were combined obtaining all the pairwise differences and ratios, both in the normal and dereddened version. The errors on the magnitudes have been composed taking their quadratic sums. Finally, radii and ellipticities have been composed through pairwise differences with ratios only for the ellipticities. The final catalogue consists of 4520 features for each data item. It has to be noted that the Classic$_{10}$ features are of



course included in this set of features. In Table 1, the types and amounts of the features obtained following this strategy are specified. As appears from the table, the feature combinations can be divided into several groups:

- simple features: magnitudes, radii, and ellipticities as downloaded from the SDSS database;
- differences: pairwise differences of the simple features; colour indexes are a subset of this group utilizing only adjacent filters;
- ratios: ratios between the simple features; an important subset of this group is the one containing ratios between different magnitudes of the same filter; we will define this subset as photometric ratios;
- errors: errors on the simple features and their propagated compositions.

As we will see in the following, the ratios group, and its subgroup, the photometric ratios, are particularly important for the redshift estimation experiments.

4. Experiments and results

The feature selection was performed applying the forward selection strategy, as described in Sect. 2.2, on the three catalogues. The verification of the resulting feature sets was performed using the RF. This algorithm is widely used in literature, and therefore the results obtained here can be easily compared to those with different feature selection strategies.

In addition, experiments using the classic features were performed, in order to compare their performances with the proposed selected features. Already at an early stage of the experiments, it turned out that only four selected features are sufficient to achieve a performance comparable to classic features. Therefore the scores are always calculated separately for the full set of ten selected features (Best_{10}) and the first four (Best_4) features only. To compare the results with a fully automated feature extraction and feature selection approach, a DCMDN was also used for the experiments.

It has to be noted that in some cases the same features sets have been found but exhibiting a different ordering. In these cases, all the subsets have been kept for the sake of correctness. In the next subsections the three experiments and the corresponding results are shown. The two experiments with Catalogue DR7a and DR7b are designed to provide results that are comparable to the literature. For a scientifically more interesting interpretation, the less biased, not flagged, and more representative Catalogue DR7+9 was used for the main experiment. Therefore, only the results and performances of the first two experiments are given in a summarized representation, reserving more space for a detailed description of Experiment DR7+9. Further details concerning the results obtained with Catalogue DR7a and DR7b are shown in Appendix A.

4.1. Experiment DR7a

The feature selection on the Catalogue DR7a produced 22 subsets of ten features each. Only 20 features, of the initial 4520, compose the tree. The three features,

- $\underline{g_{\mathrm{psf}}}/\underline{u_{\mathrm{model}}}$
- $\underline{i_{\mathrm{psf}}}/\underline{z_{\mathrm{model}}}$
- $z_{\mathrm{model}}/z_{\mathrm{psf}}$,

appear in all the possible branches. For all presented feature sets, the RF experiments were performed. The best performing ten features are indicated in the second column of Table 2 (DR7a subset) in the order of their occurrence. The performances are compared with the results of the $\mathsf{Classic}_{10}$ features presented in the first column of the same table. A summary of the most important results is shown in the first section of Table 3. As shown in Table 3, the experiment with the Best_{10} subset outperforms the experiment with the $\mathsf{Classic}_{10}$ features with respect to all the statistical scores.

Table 2. Classic and best feature subsets obtained by the feature selection process of the experiments on the three catalogues.

$\mathsf{Classic}_{10}$	DR7a Best_{10}	DR7b Best_{10}	DR7+9 Best_{10}
r_{psf}	$i_{\mathrm{psf}}/i_{\mathrm{model}}$	$i_{\mathrm{psf}}/i_{\mathrm{model}}$	$i_{\mathrm{petro}}/i_{\mathrm{psf}}$
r_{model}	$\underline{g_{\mathrm{psf}}}/\underline{u_{\mathrm{model}}}$	$\underline{g_{\mathrm{psf}}}/\underline{u_{\mathrm{model}}}$	$\underline{g_{psf}} - \underline{u_{model}}$
$u_{\mathrm{psf}} - g_{\mathrm{psf}}$	$\underline{r_{\mathrm{psf}}}/\underline{i_{\mathrm{model}}}$	$\underline{r_{\mathrm{psf}}}/\underline{i_{\mathrm{model}}}$	$\underline{i_{\mathrm{exp}}}/\underline{r_{\mathrm{psf}}}$
$g_{\mathrm{psf}} - r_{\mathrm{psf}}$	$\underline{i_{\mathrm{dev}}}/\underline{i_{\mathrm{psf}}}$	$\underline{i_{\mathrm{dev}}}/\underline{i_{\mathrm{psf}}}$	$\sqrt{\sigma^2_{r_{\mathrm{model}}} + \sigma^2_{r_{\mathrm{dev}}}}$
$r_{\mathrm{psf}} - i_{\mathrm{psf}}$	$\underline{r_{\mathrm{psf}}}/\underline{g_{\mathrm{model}}}$	$\underline{z_{\mathrm{psf}}}/\underline{i_{\mathrm{model}}}$	$\underline{r_{\mathrm{psf}}}/\underline{g_{\mathrm{exp}}}$
$i_{\mathrm{psf}} - z_{\mathrm{psf}}$	$\underline{i_{\mathrm{psf}}}/\underline{z_{\mathrm{model}}}$	$\underline{r_{\mathrm{psf}}}/\underline{g_{\mathrm{exp}}}$	$\underline{i_{\mathrm{psf}}}/\underline{z_{\mathrm{model}}}$
$u_{\mathrm{model}} - g_{\mathrm{model}}$	$r_{\mathrm{psf}} - r_{\mathrm{petro}}$	$r_{\mathrm{psf}} - r_{\mathrm{petro}}$	$i_{\mathrm{psf}} - i_{\mathrm{dev}}$
$g_{\mathrm{model}} - r_{\mathrm{model}}$	$\sqrt{\sigma^2_{r_{\mathrm{model}}} + \sigma^2_{g_{\mathrm{exp}}}}$	$i_{\mathrm{psf}} - i_{\mathrm{petro}}$	$r_{\mathrm{petro}}/r_{\mathrm{psf}}$
$r_{\mathrm{model}} - i_{\mathrm{model}}$	$z_{\mathrm{model}}/z_{\mathrm{psf}}$	$\underline{z_{\mathrm{model}}}/\underline{z_{\mathrm{psf}}}$	$\underline{i_{\mathrm{psf}}} - \underline{r_{\mathrm{model}}}$
$i_{\mathrm{model}} - z_{\mathrm{model}}$	$i_{\mathrm{psf}} - i_{\mathrm{petro}}$	$\sqrt{\sigma^2_{g_{\mathrm{model}}} + \sigma^2_{g_{\mathrm{dev}}}}$	$z_{\mathrm{exp}}/z_{\mathrm{psf}}$

Notes. After the selection process, the RF was used to identify the feature branches of the corresponding trees that show the best performance.

Table 3. Summary of the scores obtained with the RF and DCMDN models in the three experiments.

Exp	Set	# Features	Mean	RMSE	NMAD
DR7a	$\mathsf{Classic}_{10}$	10	−0.024	0.163	0.051
	Best_4	4	−0.023	0.163	0.080
	Best_{10}	10	−0.014	0.124	0.044
	DCMDN	65 536	−0.020	0.145	0.043
DR7b	$\mathsf{Classic}_{10}$	10	−0.030	0.180	0.059
	Best_4	4	−0.027	0.183	0.087
	Best_{10}	10	−0.019	0.145	0.050
	DCMDN	65 536	−0.024	0.171	0.032
DR7+9	$\mathsf{Classic}_{10}$	10	−0.033	0.207	0.073
	Best_4	4	−0.032	0.206	0.100
	Best_{10}	10	−0.023	0.174	0.060
	DCMDN	65 536	−0.027	0.184	0.037

Notes. The DCMDN automatically extracted 65 536 features for each experiment. The resulting scores are also given.

Moreover, in Table 3 the results obtained using the DCMDN are shown in order to compare the predictions with a model based on automatic features selection. The DCMDN model automatically extracts 65536 features from images in the five filters *ugriz* of size 16×16 pixel^2. This model is meant to generate probability density functions (PDFs) in the form of Gaussian mixtures instead of point estimates. Therefore, in order to calculate the scores, the weighted mean of every PDF with respect to the mixture components has been estimated. As shown in the table, the performance is superior with respect to the $\mathsf{Classic}_{10}$ features and the Best_4 subset, but it is outperformed by the Best_{10} subset of features. The performances of these four sets have been compared using the CRPS score, as reported in the left section of Table 4. Those results are consistent with the

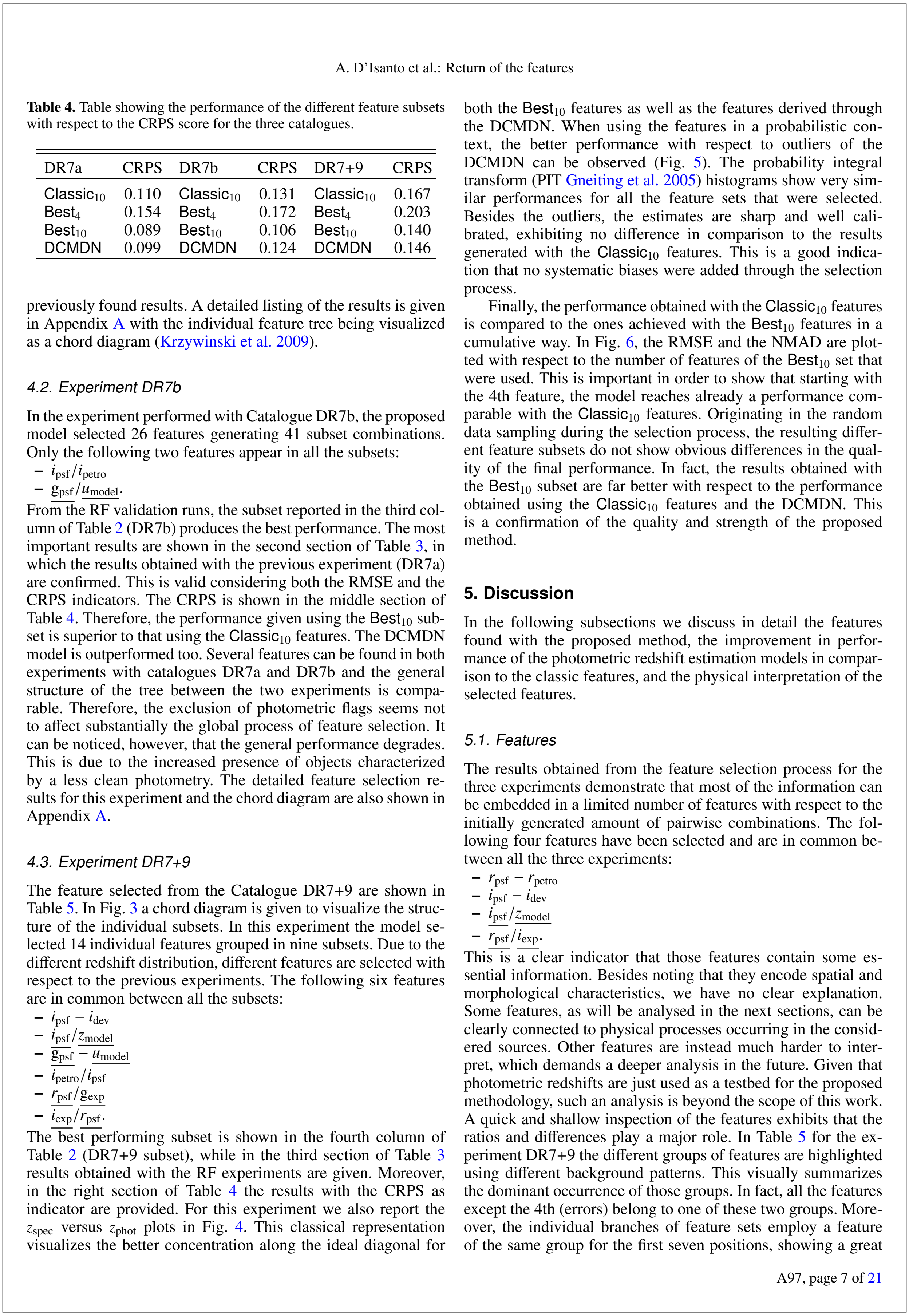


Table 4. Table showing the performance of the different feature subsets with respect to the CRPS score for the three catalogues.

DR7a	CRPS	DR7b	CRPS	DR7+9	CRPS
Classic$_{10}$	0.110	Classic$_{10}$	0.131	Classic$_{10}$	0.167
Best$_{4}$	0.154	Best$_{4}$	0.172	Best$_{4}$	0.203
Best$_{10}$	0.089	Best$_{10}$	0.106	Best$_{10}$	0.140
DCMDN	0.099	DCMDN	0.124	DCMDN	0.146

previously found results. A detailed listing of the results is given in Appendix A with the individual feature tree being visualized as a chord diagram (Krzywinski et al. 2009).

4.2. Experiment DR7b

In the experiment performed with Catalogue DR7b, the proposed model selected 26 features generating 41 subset combinations. Only the following two features appear in all the subsets:

- $i_{\rm psf}/i_{\rm petro}$
- $\underline{g_{\rm psf}}/\underline{u_{\rm model}}$.

From the RF validation runs, the subset reported in the third column of Table 2 (DR7b) produces the best performance. The most important results are shown in the second section of Table 3, in which the results obtained with the previous experiment (DR7a) are confirmed. This is valid considering both the RMSE and the CRPS indicators. The CRPS is shown in the middle section of Table 4. Therefore, the performance given using the Best$_{10}$ subset is superior to that using the Classic$_{10}$ features. The DCMDN model is outperformed too. Several features can be found in both experiments with catalogues DR7a and DR7b and the general structure of the tree between the two experiments is comparable. Therefore, the exclusion of photometric flags seems not to affect substantially the global process of feature selection. It can be noticed, however, that the general performance degrades. This is due to the increased presence of objects characterized by a less clean photometry. The detailed feature selection results for this experiment and the chord diagram are also shown in Appendix A.

4.3. Experiment DR7+9

The feature selected from the Catalogue DR7+9 are shown in Table 5. In Fig. 3 a chord diagram is given to visualize the structure of the individual subsets. In this experiment the model selected 14 individual features grouped in nine subsets. Due to the different redshift distribution, different features are selected with respect to the previous experiments. The following six features are in common between all the subsets:

- $i_{\rm psf} - i_{\rm dev}$
- $\underline{i_{\rm psf}}/\underline{z_{\rm model}}$
- $\underline{g_{\rm psf}} - \underline{u_{\rm model}}$
- $i_{\rm petro}/i_{\rm psf}$
- $\underline{r_{\rm psf}}/\underline{g_{\rm exp}}$
- $\underline{i_{\rm exp}}/\underline{r_{\rm psf}}$.

The best performing subset is shown in the fourth column of Table 2 (DR7+9 subset), while in the third section of Table 3 results obtained with the RF experiments are given. Moreover, in the right section of Table 4 the results with the CRPS as indicator are provided. For this experiment we also report the $z_{\rm spec}$ versus $z_{\rm phot}$ plots in Fig. 4. This classical representation visualizes the better concentration along the ideal diagonal for both the Best$_{10}$ features as well as the features derived through the DCMDN. When using the features in a probabilistic context, the better performance with respect to outliers of the DCMDN can be observed (Fig. 5). The probability integral transform (PIT Gneiting et al. 2005) histograms show very similar performances for all the feature sets that were selected. Besides the outliers, the estimates are sharp and well calibrated, exhibiting no difference in comparison to the results generated with the Classic$_{10}$ features. This is a good indication that no systematic biases were added through the selection process.

Finally, the performance obtained with the Classic$_{10}$ features is compared to the ones achieved with the Best$_{10}$ features in a cumulative way. In Fig. 6, the RMSE and the NMAD are plotted with respect to the number of features of the Best$_{10}$ set that were used. This is important in order to show that starting with the 4th feature, the model reaches already a performance comparable with the Classic$_{10}$ features. Originating in the random data sampling during the selection process, the resulting different feature subsets do not show obvious differences in the quality of the final performance. In fact, the results obtained with the Best$_{10}$ subset are far better with respect to the performance obtained using the Classic$_{10}$ features and the DCMDN. This is a confirmation of the quality and strength of the proposed method.

5. Discussion

In the following subsections we discuss in detail the features found with the proposed method, the improvement in performance of the photometric redshift estimation models in comparison to the classic features, and the physical interpretation of the selected features.

5.1. Features

The results obtained from the feature selection process for the three experiments demonstrate that most of the information can be embedded in a limited number of features with respect to the initially generated amount of pairwise combinations. The following four features have been selected and are in common between all the three experiments:

- $r_{\rm psf} - r_{\rm petro}$
- $i_{\rm psf} - i_{\rm dev}$
- $\underline{i_{\rm psf}}/\underline{z_{\rm model}}$
- $\underline{r_{\rm psf}}/\underline{i_{\rm exp}}$.

This is a clear indicator that those features contain some essential information. Besides noting that they encode spatial and morphological characteristics, we have no clear explanation. Some features, as will be analysed in the next sections, can be clearly connected to physical processes occurring in the considered sources. Other features are instead much harder to interpret, which demands a deeper analysis in the future. Given that photometric redshifts are just used as a testbed for the proposed methodology, such an analysis is beyond the scope of this work. A quick and shallow inspection of the features exhibits that the ratios and differences play a major role. In Table 5 for the experiment DR7+9 the different groups of features are highlighted using different background patterns. This visually summarizes the dominant occurrence of those groups. In fact, all the features except the 4th (errors) belong to one of these two groups. Moreover, the individual branches of feature sets employ a feature of the same group for the first seven positions, showing a great

Table 5. Detailed feature branches obtained from the feature selection for the DR7+9 experiment.

id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	i_{petro}/i_{psf}	$\underline{g_{psf}} - \underline{u_{model}}$	$\underline{i_{exp}}/\underline{r_{psf}}$	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{r_{dev}}}$	$\underline{r_{psf}}/\underline{g_{exp}}$	$\underline{i_{psf}}/\underline{z_{model}}$	$i_{psf} - i_{dev}$	r_{petro}/r_{psf}	$\underline{i_{psf}} - \underline{i_{dev}}$	$z_{psf} - z_{model}$
2*										z_{exp}/z_{psf}
3									$\underline{i_{psf}} - \underline{r_{model}}$	
4										$\underline{i_{psf}} - \underline{i_{dev}}$
5								$z_{psf} - z_{model}$	r_{petro}/r_{psf}	
6				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{r_{dev}}}$				r_{petro}/r_{psf}	$\underline{i_{psf}} - \underline{i_{dev}}$	$z_{psf} - z_{model}$
7									$z_{psf} - z_{model}$	$\underline{i_{psf}} - \underline{i_{dev}}$
8								$z_{psf} - z_{model}$	r_{petro}/r_{psf}	
9									$r_{psf} - r_{petro}$	

Notes. The 2nd branch, indicated with the $*$ symbol, is the best performing subset with respect to the experiments using the RF. The ratios and photometric ratios are indicated, respectively, with vertical lines and dots. The differences are with horizontal lines and the errors are with north west lines. The colour code for the features is the same as shown in the chord diagram in Fig. 3.

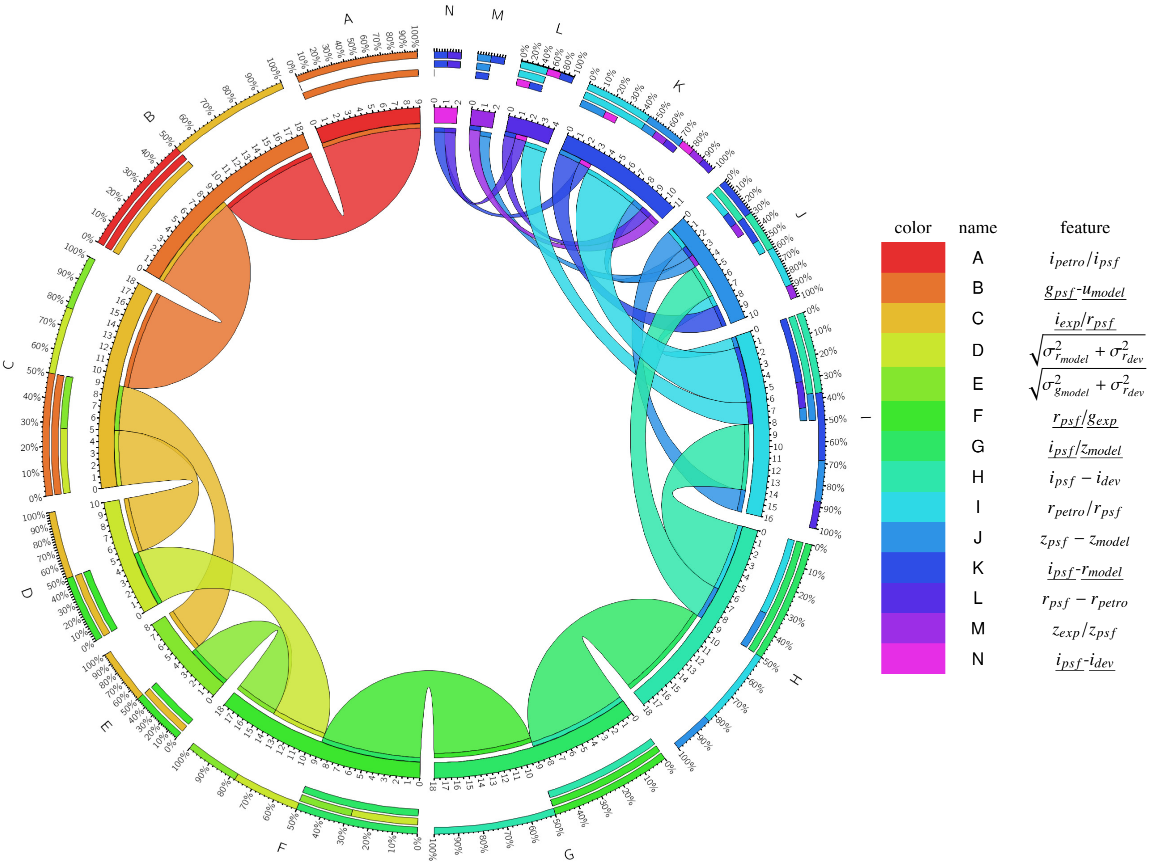


Fig. 3. Chord diagram of the features derived in Experiment DR7+9. Every feature is associated to a specific colour, and starting from the first feature A it is possible to follow all the possible paths of the tree, depicting the different feature subsets. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. (Note the branches splitting in feature C and re-joining in feature F).

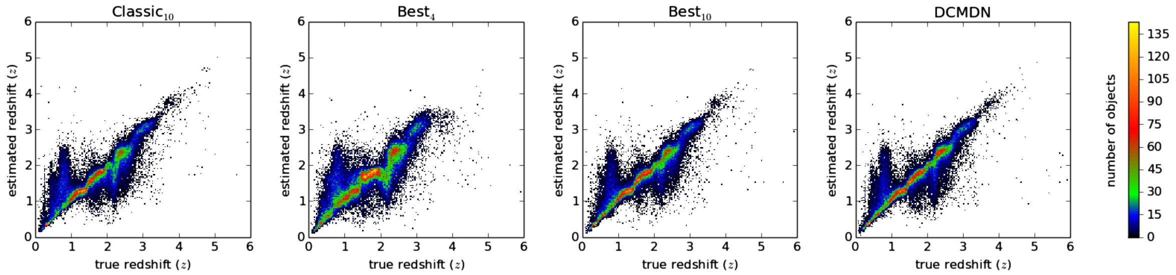


Fig. 4. Comparison of the spectroscopic (true) redshifts (z_{spec}) against the photometrically estimated redshifts (z_{phot}) of the different feature sets in experiment DR7+9.

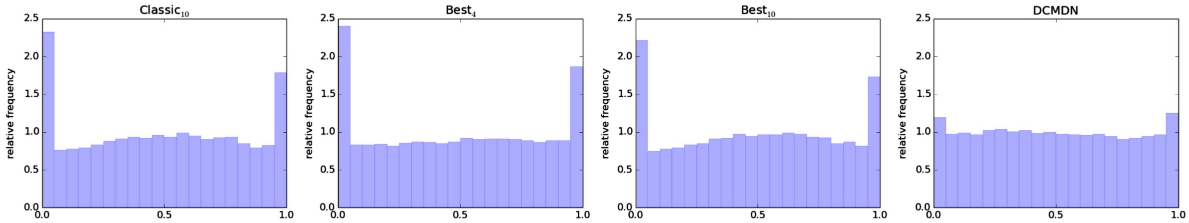


Fig. 5. PIT histograms for experiment DR7+9 for the different features sets, as shown in Table 4. Except the PIT of the DCMDN, all other feature sets generate results with significant outliers at the extrema.

stability in the composition of the branches. The experiment based on the DR7+9 catalogue generates a much less complex structure of the tree of feature sets with respect to experiments DR7a and DR7b. Fewer branches and a reduced number of features are selected. Reasons for this behaviour are the more complete redshift distribution of catalogue DR7+9 with respect to the other two and the improvement in SDSS photometry from DR7 to DR9. This drives the model to find the required information in a reduced number of efficient features. The analysis of the tree composition and features distribution can be done following the chord diagram shown in Fig. 3. The chord diagram is an optimal visualization tool for the description of a complex data structure. In this diagram, every feature is associated to a specific colour, and starting from the first feature (A) it is possible to follow all the possible paths of the tree, depicting the different feature subsets. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. Therefore, the chord diagram, coupled with Table 5, gives a clear description of the structure and composition of the tree of features. In addition, in Table 5 the same colour code as in the chord diagram is adopted, to identify the features and their distribution. The chord diagram clearly visualizes that the feature trees split at feature C and later rejoin at feature F. In comparison to the chord diagram obtained for Experiment DR7+9, the two chord diagrams for experiments DR7a and DR7b (see Appendix A) immediately visualize the higher complexity of those trees. From Fig. 3 and Table 5 it appears that, apart from a few exceptions, the selected features follow a precise scheme. No classic colour indexes or any of the Classic$_{10}$ features have been chosen, while only differences between different magnitudes of the same band or differences between different type of magnitudes play a certain role. The ratios have been all selected in the extinction-corrected version, except for the subcategory of the photometric ratios. This can be understood considering that the latter are ratios between magnitudes of the same filter where the contribution of the extinction correction tends to cancel out.

Another relevant aspect in experiment DR7+9 is that all the 15 features in the tree are exclusively a composition of magnitudes and their errors. Neither radii nor ellipticities have been chosen during the selection process. As only quasars have been used in the experiments, this introduces a bias to the selection process in favour of magnitudes and against shape-based features. This is a clear indication that just the magnitudes are required to describe the objects and explore the parameter space in the setting of photometric redshift estimation. Although photometric ratios are shape-related parameters, they express the ratio between the centred and the extended part of a component that can be interpreted as flux of the hosting galaxy. Therefore, here a bias introduced by using quasars for the experiments cannot be observed.

It is remarkable that photometric errors are selected as features, given that there is no obvious physical relation between the redshift of the considered objects and the measurement errors reported by the photometric pipeline of SDSS. Therefore it is important to consider how errors are derived in the SDSS, based on flux measurements (Lupton et al. 1999). Magnitude errors quantify the discrepancy between the fitted photometric model (psf, model, petrosian, etc.) and the observed pixel-wise distribution of spatially correlated fluxes with respect to the applied noise model. Therefore, it is evident that the errors on the single magnitudes appear to be larger for fainter objects, a physical property that is directly correlated to distance. In addition, the deviation of spatial flux distributions from the applied spatial photometric models are good morphological indicators; for example, the shape and size of the hosting galaxy are correlated with redshift. The workflow adopted is able to capture these dependencies, selecting a composition of errors as an important feature of the best set.

Even though 4520 features were synthetically created by combining base features, only 15 were selected in experiment

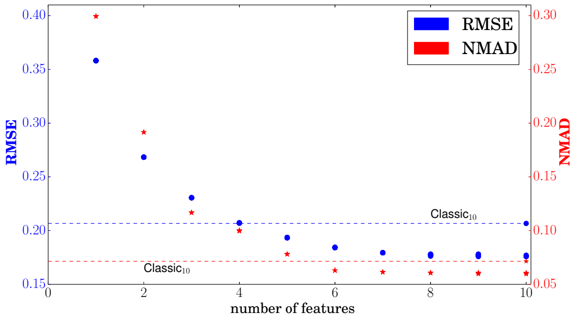


Fig. 6. Comparison of model performance with regard to the number of used features. The root mean square error and normalized median absolute deviation of the results from the DR7+9 RF experiments are presented. As reference line the performance achieved with the Classic$_{10}$ features is shown. As it can be seen, from the fourth feature on, the performance of the subsets outperforms the Classic$_{10}$ features. After the ninth feature, the improvement settles. When adding many more features, the performance will start to degrade.

DR7+9 (19 and 26 for experiments DR7a and DR7b, respectively). Furthermore, some features encode the same type of information with just subtle differences in composition. It is remarkable that every feature that is built on magnitudes incorporates a PSF magnitude. Moreover, the model and exponential magnitude in the SDSS are related[4], with the model magnitude being just the better fitting model when comparing an exponential and a de Vaucouleurs profile. In the first stages of the selection process, the proposed algorithm does not select differing branches but identifies essential features to produce good results when photometrically estimating redshifts. These observations are also valid for the results found in experiments DR7a and DR7b.

5.2. Comparison of performance

Using the RF, the validation experiments were carried out on every feature set. The second subset, indicated as Best$_{10}$, gave a slightly better performance than the others. Even though we would not consider this as a substantial effect, we decided to choose this as our reference set. It can be noticed from Fig. 6 that from the 4th feature on, every subset delivers a performance comparable to the performance of all ten features in the Classic$_{10}$ set, with respect to the RMSE. Consistently, the use of more than four features outperforms the Classic$_{10}$, independently of the subset used. Adding more features improves further the performance and the trend becomes asymptotic around the 9th feature. At a certain point, adding many more features results in a degradation of the redshift estimation performance. After the 8th feature, the contribution is of a minor nature. Just to have a fair comparison to the Classic$_{10}$ features, we decided to pick the same number of ten features, even though a smaller number is sufficient to outperform the Classic$_{10}$ features. The performance improvement is evident seeing the results reported in Table 3 and Fig. 4. It is important to note that the CRPS results (Table 4) confirm the performance shown with respect to the other scores. When predicting PDFs instead of point estimates, the PIT histograms (Fig. 5) indicate the DCMDN as the best calibrated model. This result is reasonable because the DCMDN is the only model trained using the CRPS as loss function, which is focused on the PDFs calibration. The kNN and the RF are instead based on the optimization of point estimates using the RMSE. Therefore, the calibration of the PDFs estimated using the DCMDN is superior. The use of such a probabilistic model is helpful to handle the presence of extreme outliers, since it is not based on the minimization of the RMSE, as discussed in D'Isanto & Polsterer (2018). The usage of PDFs allows us to identify objects with an ambiguous redshift distribution, while in a point estimation scenario, where just the mean of such a distribution would be considered, the estimates of those objects would result in extreme outliers.

Table 6. Cross experiments performed with the RF, using the Best$_{10}$ sets obtained from every experiment with all the three catalogues.

Exp.	Catalogue DR7a	Catalogue DR7b	Catalogue DR7+9
DR7a	0.124	0.146	0.176
DR7b	0.125	0.145	0.176
DR7+9	0.124	0.147	0.174

Notes. The results are expressed using the RMSE. It can be noticed the negligible difference of performance, for every catalogue, independently from the feature set used.

Six features of the best subset are ratios of different magnitudes. Three of them are plain ratios, while three are photometric ratios. Analysing the fourth column of Table 2, it appears that one of the components of these features is always a PSF magnitude, coupled with a model, petro, or exp magnitude. Therefore, from the analysis of the results obtained, we can state that the reason for the performance improvement is not in the choice of some specific features, or in a particular subset of features, but in their type and in the combination of certain groups.

All these aspects are clear indicators to demonstrate the following two conclusions. The proposed method is highly stable, enabling us to derive subsets of features that are equivalently well-performing and similar, based on a common structure. In this sense, the improvement with respect to the use of Classic$_{10}$ features is clear. In order to prove the robustness of the proposed method, we performed some experiments using for each data-set the Best$_{10}$ features obtained with the other two catalogues, as shown in Table 6, and the results were almost as good as in the other cases. The method captures the inherent structure of the physical properties of the sources, which is essential to provide good photometrically estimated redshifts for quasars.

5.3. Physical interpretation

In contrast to deep learning models, feature-based approaches have the advantage of allowing an interpretation in a physical context. Therefore the features selected by our approach are discussed in the following. By analysing the importance of each feature of the Best$_{10}$ set in smaller redshift bins, the contribution of certain spectral features can be understood. In Fig. 7 the importance is presented for sliding bins of $\Delta z = 0.2$ based on the Gini index (Breiman et al. 1984). The Gini index is used in the RF to perform the segmentation of the parameter space orthogonally to its dimensions at every node. As all ten features contribute individually, the total contribution is normalized to one and the individual lines are presented in a cumulative way. The relative importance of each feature clearly does not reflect their ordering, as they have been assembled by a forward feature selection algorithm. In particular, the first feature of the best

[4] http://classic.sdss.org/dr7/algorithms/photometry.html#mag_model

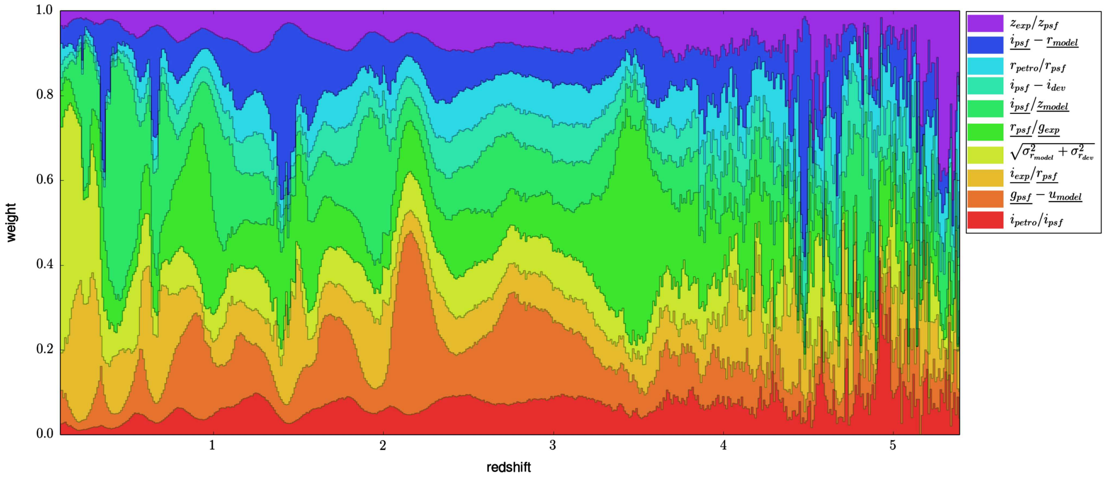


Fig. 7. Importance of every feature of the Best$_{10}$ subset from experiment DR7+9. For a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated in a localized regression model based on the Gini index as utilized by the RF. The colour code used is the same adopted for the chord diagram in Fig. 3.

set does not show a dominant role when using multiple features. When building a photometric regression model based on just a single feature, the concentration index in the i band provides the best tracer for distance. Therefore a concentration index in the i band is consequently chosen in all the three experiments. This selection is of course heavily biased by the distribution of our training objects with respect to redshift and by the fact that objects for training are selected based on the classification of the spectral template fitting of SDSS.

As soon as more photometric features are used, the spectral energy distribution and distinct spectral features are the dominant source of information for estimating the redshifts. Those features are mainly ratios. To use ratios instead of colours is a surprising fact, as in the literature colours are the usual choice for photometric redshift estimation models. In Fig. 7 one can inspect how the different features contribute at different redshift bins, building a well-performing model that covers the full redshift range. Besides some very narrow redshift regions, no clear structure with preference of some photometric features can be observed at higher redshifts ($z > 4$). This is due to the poor coverage of the training and validation data in that range. The ordering of the features in the Best$_{10}$ set and their importance as shown in Fig. 7 can be compared with the global feature importance as obtained from the RF experiment (Table 7). The feature importance calculated on the overall redshift distribution gives different indications with respect to the bin-wise analysis, but it is quite consistent with the original order obtained from the feature selection. This is a further demonstration of the stability and robustness of the proposed method.

The different behaviours and importance found for the features in the individual redshift bins can be partially explained by analysing distinct features in the spectral energy distribution. By carefully inspecting the emission lines of quasars as reported by the SDSS spectral pipeline, a connection between some photometric features and emission lines could be found. Those features that are composed of adjacent filter bands are very sensitive to spectral lines that are in the vicinity of the overlapping area of filter transmission curves. This can be explained by a flipping of the feature, for example positive or negative for colours and above or below one for ratios. Already a little shift of an emission line with respect to the redshift is enough to create a significant change in the feature space that is detected and utilized by the machine learning model. Five features of the Best$_{10}$ share this characteristic. Therefore the discussion with respect to emission lines is focused on selected features that are composed of magnitudes from neighbouring filter bands. Using the well known relation

$$z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}} - 1 = \frac{\lambda_{\text{filter intersection}}}{\lambda_{\text{qso emission line}}} - 1, \tag{2}$$

it is possible to calculate the redshift at which a specific emission line becomes traceable when using a certain filter combination. The proposed features capture many distinct emission lines, showing peaks in the redshift bins where the lines appear. This is shown in Figs. 8 and 9, where the feature importance has been compared with the classic features of the corresponding bands. To understand better the influence of the usage of magnitudes describing extended objects, both the PSF and the model magnitudes of the classic features where used for comparison. In Fig. 8

Table 7. Features of the Best$_{10}$ set from experiment DR7+9, ordered by decreasing importance as expressed by the score of the RF based on the Gini criterion.

Position		Feature	Score
1	▲1	$\underline{g_{\text{psf}}} - \underline{u_{\text{model}}}$	0.424
2	▼1	$i_{\text{petro}}/i_{\text{psf}}$	0.121
3	▲1	$\sqrt{\sigma^2_{r_{\text{model}}} + \sigma^2_{r_{\text{dev}}}}$	0.092
4	▼1	$\underline{i_{\text{exp}}}/\underline{r_{\text{psf}}}$	0.072
5	=	$\underline{r_{\text{psf}}}/\underline{g_{\text{exp}}}$	0.071
6	=	$\underline{i_{\text{psf}}}/\underline{z_{\text{model}}}$	0.064
7	▲2	$\underline{i_{\text{psf}}} - \underline{r_{\text{model}}}$	0.062
8	▼1	$i_{\text{psf}} - i_{\text{dev}}$	0.042
9	▲1	$z_{\text{exp}}/z_{\text{psf}}$	0.026
10	▼2	$r_{\text{petro}}/r_{\text{psf}}$	0.025

Notes. The change with respect to the initially found ordering of the presented approach, and the RF score are reported, too.

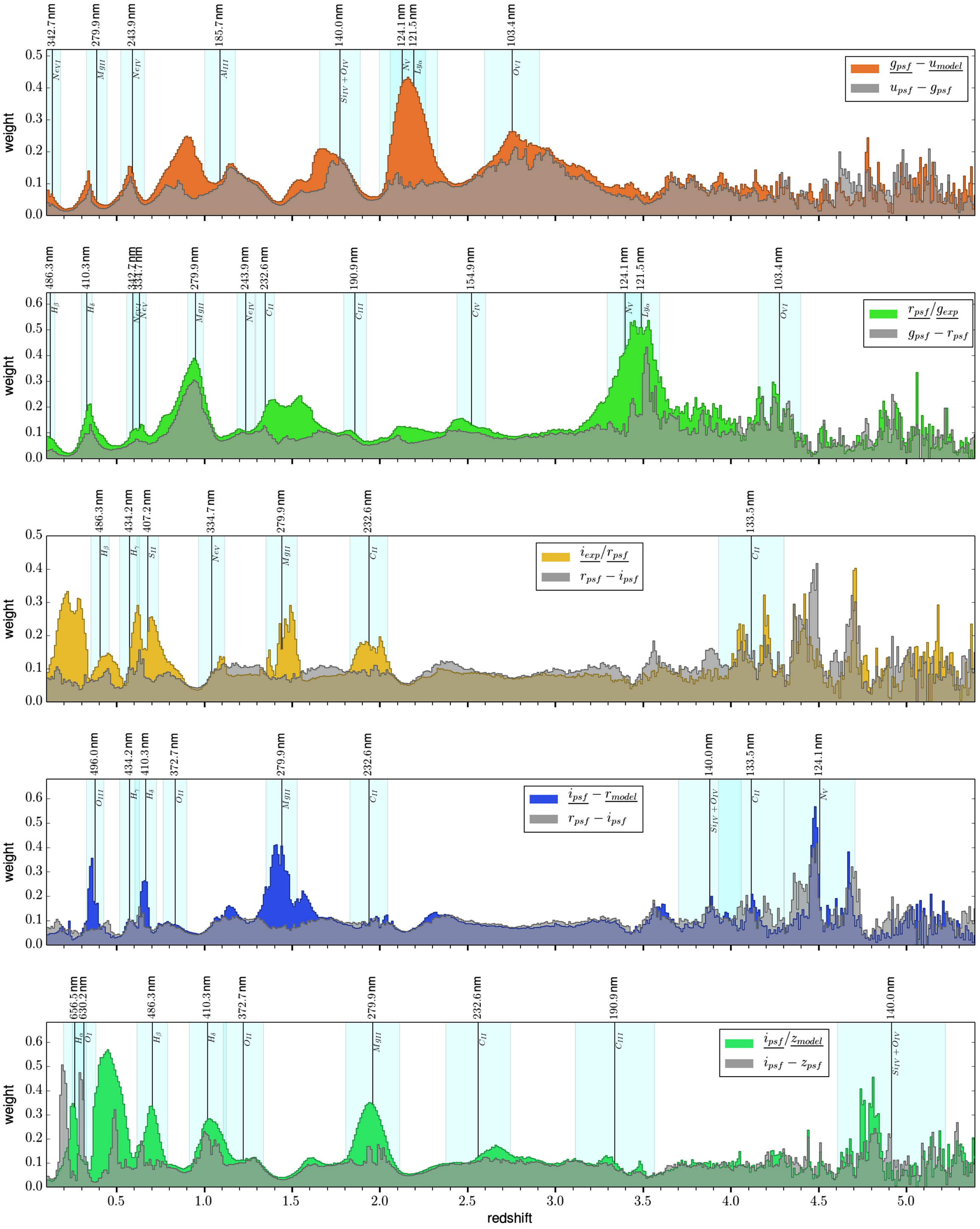


Fig. 8. Feature importance of the five features from the Best$_{10}$ set composed by magnitudes from neighbouring bands. As in Fig. 7, for a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated. The results are compared to the classic features using PSF magnitudes of the same bands. Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2). The used colour code is the same as in Fig. 3, while corresponding features of the Classic$_{10}$ set are always shown in grey.

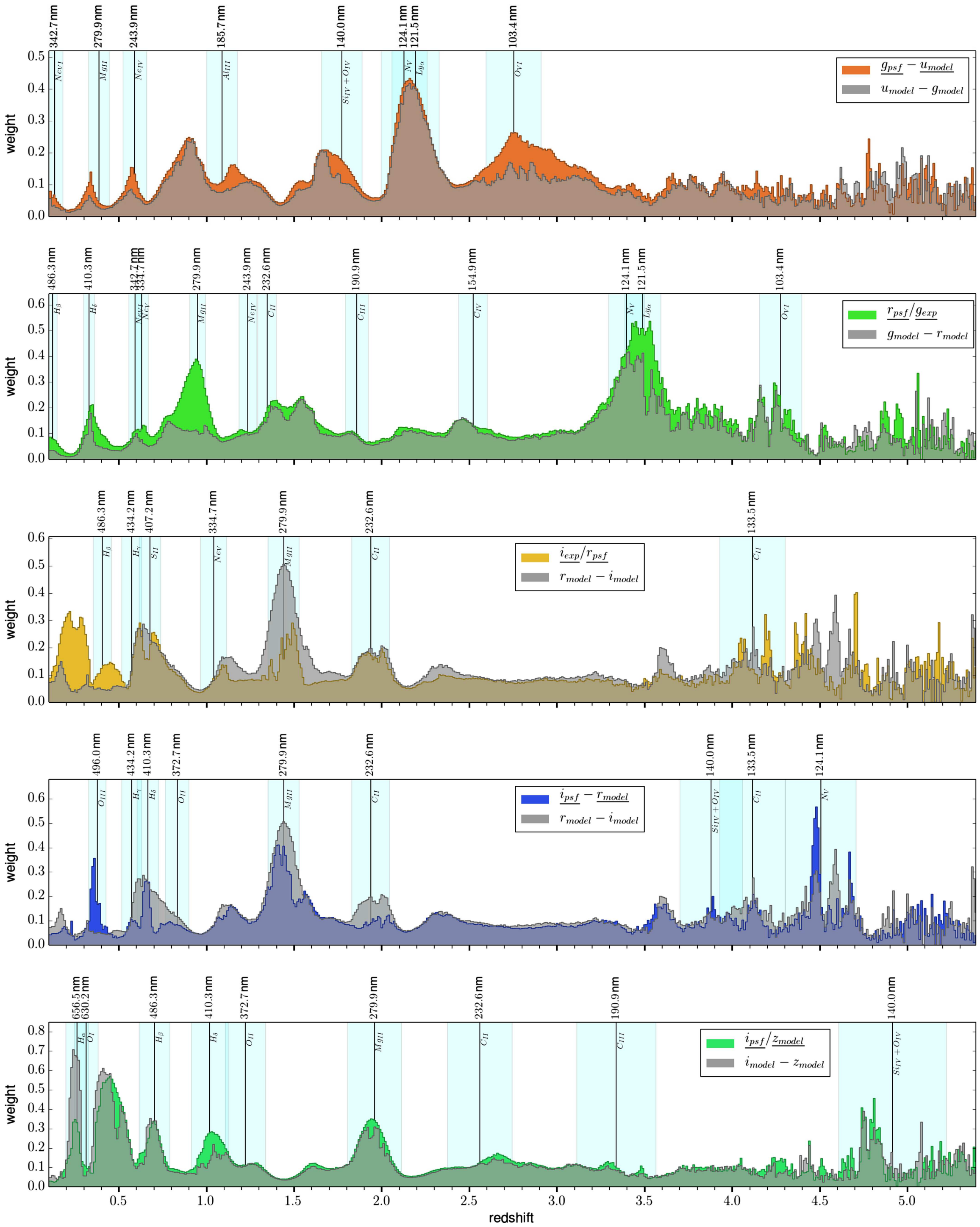


Fig. 9. Feature importance of the five features from the Best_{10} set composed by magnitudes from neighbouring bands. As in Fig. 7, for a sliding redshift bin of $\Delta z = 0.2$, the importance of every feature was calculated. The results are compared to the classic features using model magnitudes of the same bands. Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2). The used colour code is the same as in Fig. 3, while corresponding features of the Classic_{10} set are always shown in grey.

the comparison is performed with respect to PSF colours, while in Fig. 9 the same comparison is done with respect to model colours. By using Eq. 2, a selected set of spectral emission lines of quasars has been convolved with the corresponding filter characteristics to annotate the plots. Besides the maximum of the overlapping region, the start and the end of the intersection are depicted. We defined the upper and lower limits as the points at which the sensitivity of the filter curve is equal to 0.001 in quantum efficiency. It can be seen that many emission lines perfectly correspond to peaks in importance exhibited by the features of the Best_{10} set. This can be observed only partially for the classic features.

In particular, purely PSF or model magnitude-based colours have a different and often complementary contribution for several spectral lines. This is due to the fact that either concentrated or extended characteristics of the analysed objects are considered. The proposed features are more suitable than classic features to describe the peaks at distinct emission lines. Considering the $N_V - Ly_\alpha$ lines for the $\underline{g_{\rm psf}} - \underline{u_{\rm model}}$ feature, the comparison between the extended and concentrated classic features clearly indicates that an extended component of the source is captured via this feature. Keeping in mind that a pixel size of 0.4″ of the SDSS camera corresponds[5] at a redshift of $z \approx 2.2$ to ≈3.4 kpc, this is a clear indicator that the hosting galaxy is significantly contributing to the solution of the photometric redshift estimation model. A similar behaviour can be observed for the $N_V - Ly_\alpha$ lines in the $\underline{r_{\rm psf}}/\underline{g_{\rm exp}}$ feature, while the $Mg_{\rm II}$ emission line mainly appears in the PSF colour. Therefore the $Mg_{\rm II}$ emission line can be considered to be more prominent in the central region of the objects. Between the most notable lines, the Lyman-α and the Balmer series can be identified. Other important lines found are the C_{II}, C_{III}, $\mathrm{C_{IV}}$, $\mathrm{O_I}$, $\mathrm{O_{II}}$, $\mathrm{O_{III}}$, $\mathrm{O_{VI}}$, and the $Mg_{\rm II}$ lines. Besides the identified peaks caused by specific emission lines, some peaks in weight stay unexplained. Even though it is possible to distinguish between mostly spatially extended or concentrated characteristics of the objects, an association of a single emission line fails. In those cases not the transition of a line between two filters but an overall shape relation is captured by the selected parameters. As the selected features combine the strength of identifying line transitions as well as morphological characteristics, the resulting boost in performance of the photometric redshift estimation model can be well explained. To explain the meaning of the selected features that use a combination of features extracted from the same photometric band and thereby describe a morphological structure of the source, further image-based investigations are necessary. This proves that a model using the proposed feature selection approach is better able to exploit the information that represents the underlying physical and morphological structure as well as the processes going on in the sources.

6. Conclusions

In this work a method to select the best features for photometric redshift estimation is proposed. The features are calculated via a greedy forward selection approach, in which the features are selected from a set of 4520 combinations based on the photometric and shape information stored in the SDSS DR7 and DR9 catalogues. By randomly sampling the training data and running multiple kNN experiments, trees in which every branch constitutes a subset of features were generated for all the experiments. The obtained branches were then validated using a RF model and compared to the results obtained using classic sets of features. Moreover, the results were compared with a convolutional neural network based model, meant to automatically perform the feature extraction and selection. Three experiments, based on different catalogues, were carried out. The first catalogue was obtained selecting quasars from SDSS DR7 and applying photometric flags. The second catalogue was composed of quasars from SDSS DR7 too, but without using photometric flags. Finally, the third catalogue was made by mixing SDSS DR7 and DR9 quasars, in order to extend the redshift distribution. We have shown that all the sets obtained in all the experiments outperform the $\mathsf{Classic}_{10}$, and in particular a best-performing branch has been identified for each catalogue. The best sets also gave a better performance with respect to the automatic model (even though the latter typically shows a better calibration and is less affected by outliers when predicting PDFs instead of point estimates). The new best features obtained in the present work are not immediately comprehensible. Further analysis shows a relation between the dominant features of the Best_{10} set and the emission lines of quasars, which correspond to the peaks of importance of the different features along the redshift distribution. The same analysis carried out on the $\mathsf{Classic}_{10}$ features proves that the latter are not able to capture the same physical information as compactly as the selected features. This explains why the results obtained with the proposed method are outstanding with respect to the ones obtained with the $\mathsf{Classic}_{10}$ features. Moreover, we demonstrate that the proposed features fill the redshift space in a complementary way, each adding information that is relevant in different redshift ranges. The proposed method is highly stable, as shown from the distribution of the features and the groups to which they belong. The experiments show that the useful information is concentrated in a reduced number of features, which are typically very different from the $\mathsf{Classic}_{10}$. Furthermore, we verified that the difference in terms of performance with respect to the various sets is almost negligible. This demonstrates that the true advantage with respect to the $\mathsf{Classic}_{10}$ features is not given by the selected features themselves, but from their distribution and type in the specific set. Therefore, the stability shown from the different branches, for example the common distribution scheme of the features, and the ability to better capture the underlying physical processes, explains the superior performance obtained. The method is very general and could be applied to several tasks in astrophysics (and not only in astrophysics). In the future we propose to apply it to different sources (i.e. galaxies with and without an active nuclei) in order to verify if the obtained features are general or if they are only related to the fine structure of the data itself and to this specific population of sources. This includes the question of how much the processes of the active galactic nuclei dominate with respect to the processes in the surrounding galaxy the feature selection approach. It goes without saying that this first step made in the interpretation of the new features could open new doors in the understanding of the physics of quasars with respect to distance and age by providing better and more precise tracers. On the other hand, the method shows a different approach alternative to the application of deep learning, but also employing GPUs intensively. Both approaches are meant to establish an affordable and well-performing method to precisely predict photometric redshifts, in light of the upcoming missions and instruments in the near future.

Acknowledgements. The authors gratefully acknowledge the support of the Klaus Tschira Foundation. SC acknowledges support from the project “Quasars at high redshift: physics and Cosmology” financed by the ASI/INAF agreement 2017-14-H.0. We would like to thank Nikos Gianniotis and Erica

[5] Using Wright (2006) with $H_0 = 69.6, \Omega_{\rm M} = 0.286, \Omega_{\rm DE} = 0.714$.

Hopkins for proofreading and commenting on this work. Topcat has been used for this work (Taylor 2005). The DCMDN model has been developed using Theano (The Theano Development Team et al. 2016). To extract the image for the DCMDN we made use of HIPS (Fernique et al. 2015) and Montage (Berriman et al. 2004). Montage is funded by the National Science Foundation under Grant Number ACI-1440620, and was previously funded by the National Aeronautics and Space Administration's Earth Science Technology Office, Computation Technologies Project, under Cooperative Agreement Number NCC5-626 between NASA and the California Institute of Technology. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. 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 University, 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.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, ApJS, 203, 21

Aksoy, S., & Haralick, R. M. 2000, Pattern Recognit. Lett. 22, 563

Athiwaratkun, B., & Kang, K. 2015, ArXiv e-prints [arXiv:1507.02313]

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12

Beck, R., Lin, Ishida, E., et al. 2017, Mon. Notes Astron. Soc. S. Afr., 468, 4323

Benavente, P., Protopapas, P., & Pichara, K. 2017, ApJ, 845

Berriman, G. B., Good, J. C., Laity, A. C., et al. 2004, ASP Conf. Ser., 314, 593

Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9

Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Information Science and Statistics) (Secaucus, NJ: Springer-Verlag New York, Inc.)

Bonnett, C., Troxel, M. A., Hartley, W., et al. 2016, Phys. Rev. D, 94, 042005

Breiman, L. 1996, Mach. Learn. 24, 123

Breiman, L., Friedman, J., Olshen, R., & Stone, C. 1984, Classification and Regression Trees (Monterey, CA: Wadsworth and Brooks)

Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2013a, MNRAS, 437, 968

Cavuoti, S., Garofalo, M., Brescia, M., et al. 2013b, Smart Innov. Syst. Technol. 19, 29

Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015, Exp. Astron. 39, 45

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]

de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134

D'Abrusco, R., Staiano, A., Longo, G., et al. 2007, ApJ, 663, 752

D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111

D'Isanto, A., Cavuoti, S., Brescia, M., et al. 2016, MNRAS, 457, 3119

Donalek, C., Arun Kumar, A., Djorgovski, S. G., et al. 2013, ArXiv e-prints [arXiv:1310.1976]

Duda, R. O., Hart, P. E., & Stork, D. G. 2000, Pattern Classification, 2nd Edition (New York: Wiley-Interscience)

Fernique, P., Allen, M. G., Boch, T., et al. 2015, A&A, 578, A114

Fix, E., & Hodges, J. L. 1951, in US Air Force School of Aviation Medicine, Technical Report 4, 477

Gieseke, F., Polsterer, K. L., Oancea, C. E., & Igel, C. 2014, in 22th European Symposium on Artificial Neural Networks, ESANN 2014

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. 2005, Mon. Weather Rev. 133, 1098

Guyon, I., & Elisseeff, A. 2003, J. Mach. Learn. Res., 3, 1157

Harnois-Déraps, J., Tröster, T., Chisari, N., et al. 2017, MNRAS, 471, 1619

Heinermann, J., Kramer, O., Polsterer, K., & Gieseke, F. 2013, Lect. Notes Comput. Sci. Ser., 8077, 86

Hersbach, H. 2000, Weather Forecasting, 15, 559

Hey, T., Tansley, S., & Tolle, K., eds. 2009, The Fourth Paradigm: Data-Intensive Scientific Discovery (Redmond, WA: Microsoft Research)

Hildebrandt, H., Wolf, C., & Benítez, N. 2008, A&A, 480, 703

Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31

Hildebrandt, H., Viola, M., Heymans, C., et al. 2016, MNRAS, 465, 1

Hoyle, B. 2016, Astron. Comput. 16, 34

Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S., & Weller, J. 2015, MNRAS, 449, 1275

Ivezić, v., Tyson, J. A., Acosta, E., et al. 2008, ArXiv e-prints [arXiv:0805.2366v4]

Joudaki, S., Mead, A., Blake, C., et al. 2017, MNRAS, 471, 1259

Kohavi, R. 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI'95 (San Francisco, USA: Morgan Kaufmann Publishers Inc.), 2, 1137

Köhlinger, F., Viola, M., Joachimi, B., et al. 2017, MNRAS, 471, 4412

Krzywinski, M. I., Schein, J. E., Birol, I., et al. 2009, Genome Res., 19, 1639

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]

Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, MNRAS, 418, 2165

Lupton, R. H., Gunn, J. E., & Szalay, A. S. 1999, AJ, 118, 1406

Mahabal, A., Djorgovski, S. G., Turmon, M., et al. 2008, Astron. Nachr., 329, 288

Mao, K. 2004, IEEE Trans. Syst. Man Cybern. Part B Cybern. 34, 629

Norris, R. P., Hopkins, A. M., Afonso, J., et al. 2011, PASA, 28, 215

Polsterer, K. L., Gieseke, F., Igel, C., & Goto, T. 2014, ASP Conf. Ser., 485, 425

Polsterer, K., Gieseke, F., & Igel, C. 2015, ASP Conf. Ser., 495, 81

Richards, G. T., Weinstein, M. A., Schneider, D. P., et al. 2001, AJ, 122, 1151

Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, ApJS, 180, 67

Rimoldini, L., Dubath, P., Süveges, M., et al. 2012, MNRAS, 427, 2917

Smirnov, E., & Markov, A. 2017, MNRAS, 469, 2024

Tangaro, S., Amoroso, N., Brescia, M., et al. 2015, Comput. Math. Methods Med. 2015

Taylor, A. R. 2008, IAU Symp., 248, 164

Taylor, M. B. 2005, ASP Conf. Ser., 347, 29

The Theano Development Team, Al-Rfou, R., Alain, G., et al. 2016, ArXiv e-prints [arXiv:1605.02688]

Tortora, C., La Barbera, F., Napolitano, N., et al. 2016, MNRAS, 457, 2845

Vaccari, M., Covone, G., Radovich, M., et al. 2016, in Proceedings of the 4th Annual Conference on High Energy Astrophysics in Southern Africa (HEASA 2016), online at http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=275, id.26, 26

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2

Wright, E. L. 2006, PASP, 118, 1711

Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. 2013, AJ, 146, 22

Appendix A: Additional tables and figures

In this section, the additional tables for the features selection and the tree structure, together with the related chord diagrams for the experiments DR7a and DR7b are given. A brief explanation of how to read a chord diagram follows.

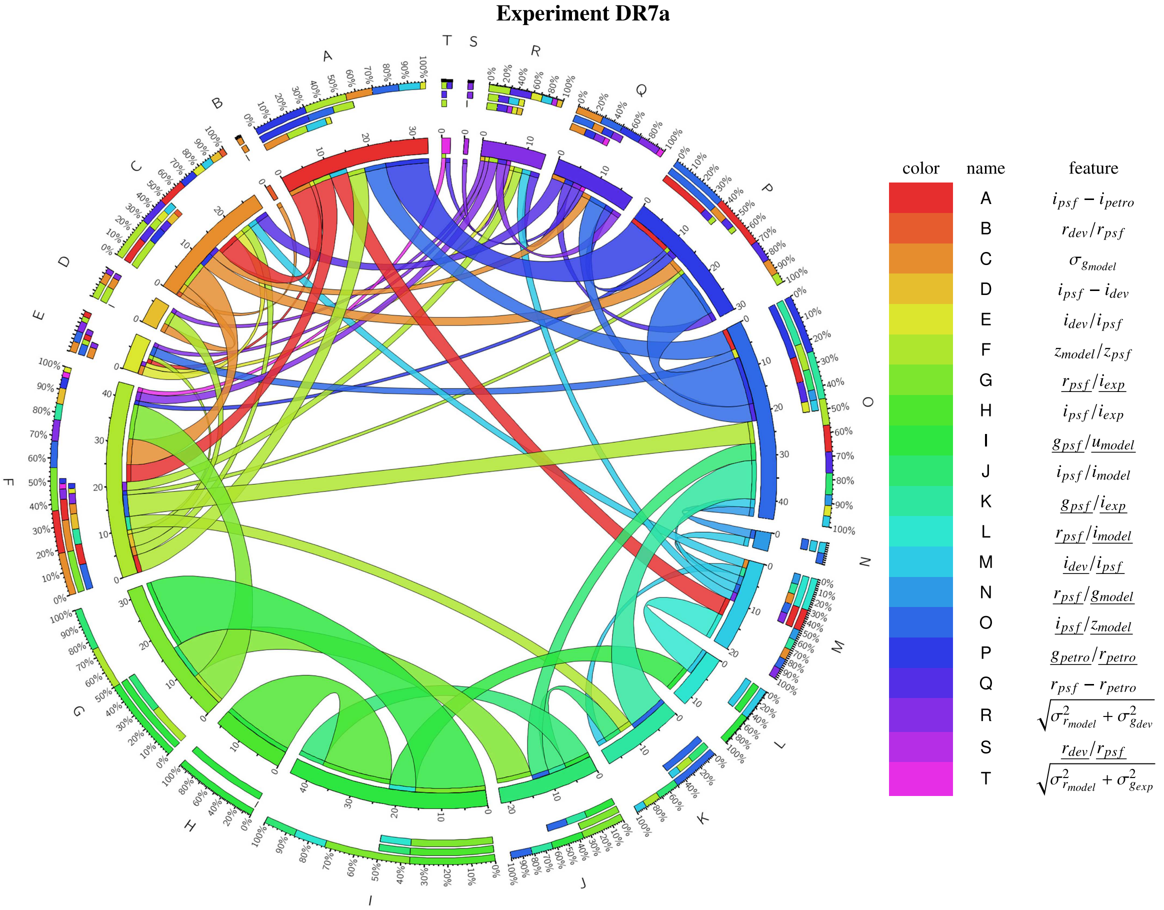
A.1. Chord diagram: how to read

The chord diagram is a tool to visualize complex structures and relations in multidimensional data, which is arranged in a matrix shape. The data are disposed in a circle and each element, in our case the features, is associated with a different colour. The relations between the elements are expressed by ribbons which connect them, with a specific width related to the importance of that specific connection. Therefore, the different ribbons can enter or exit from every arc, representing the features. The chord diagrams utilized for this work are characterized by three external arcs for each feature. Ordered from outside to inside, the external arcs represent the occurrences of a particular feature: the total percentage of the individual connections, the numbers and sources of connections entering, and the numbers and targets of connections exiting. Therefore, starting from the first features indicated in the captions, it is possible to follow all the possible paths of the tree, depicting the different feature subsets and their global scheme. Splitting points, joints, and complex interplay between feature groups can thereby be analyzed intuitively.

Table A.1. Detailed feature branches obtained from the feature selection for the experiment DR7a.

id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	i_{psf}/i_{exp}	g_{psf}/u_{model}	r_{psf}/i_{exp}	i_{psf}/i_{model}	g_{psf}/i_{exp}	i_{psf}/z_{model}	$i_{psf}-i_{petro}$	$\sigma_{g_{model}}$	z_{model}/z_{psf}	$i_{psf}-i_{dev}$
2								z_{model}/z_{psf}	$\sigma_{g_{model}}$	
3								$\sigma_{g_{model}}$	z_{model}/z_{psf}	g_{petro}/r_{petro}
4										i_{dev}/i_{psf}
5					i_{psf}/z_{model}	g_{petro}/r_{petro}				$i_{psf}-i_{dev}$
6								z_{model}/z_{psf}	$\sigma_{g_{model}}$	
7									$\sqrt{\sigma^2_{r_{model}}+\sigma^2_{g_{dev}}}$	
8										i_{dev}/i_{psf}
9				z_{model}/z_{psf}	g_{psf}/i_{exp}	i_{psf}/z_{model}		i_{dev}/i_{psf}	$\sigma_{g_{model}}$	g_{petro}/r_{petro}
10							i_{dev}/i_{psf}	$\sigma_{g_{model}}$	$r_{psf}-r_{petro}$	
11								$\sqrt{\sigma^2_{r_{model}}+\sigma^2_{g_{dev}}}$		
12					i_{psf}/z_{model}	g_{petro}/r_{petro}	$i_{psf}-i_{petro}$	$\sigma_{g_{model}}$	z_{model}/z_{psf}	$i_{psf}-i_{dev}$
13								i_{dev}/i_{psf}	$\sigma_{g_{model}}$	$r_{psf}-r_{petro}$
14										r_{dev}/r_{psf}
15									$\sqrt{\sigma^2_{r_{model}}+\sigma^2_{g_{dev}}}$	$r_{psf}-r_{petro}$
16										r_{dev}/r_{psf}
17	i_{psf}/i_{model}		r_{psf}/i_{model}	i_{dev}/i_{psf}			$r_{psf}-r_{petro}$	$\sigma_{g_{model}}$	z_{model}/z_{psf}	$i_{psf}-i_{petro}$
18								$\sqrt{\sigma^2_{r_{model}}+\sigma^2_{g_{dev}}}$		
19					r_{psf}/g_{model}	i_{psf}/z_{model}				
20*								$\sqrt{\sigma^2_{r_{model}}+\sigma^2_{g_{exp}}}$		
21					g_{psf}/i_{exp}			$\sigma_{g_{model}}$	g_{petro}/r_{petro}	z_{model}/z_{psf}
22										$i_{psf}-i_{petro}$

Notes. The 20th branch, indicated with the ∗ symbol, is the best performing subset with respect to the experiments using the RF. The *ratios* and *photometric ratios* are indicated, respectively, with vertical lines and dots. The *differences* are marked with horizontal lines and the *errors* with north west lines. The color code for the features is the same as shown in the chord diagram in Fig. A.1.



color	name	feature
🟥	A	$i_{psf} - i_{petro}$
🟧	B	r_{dev}/r_{psf}
🟧	C	$\sigma_{g_{model}}$
🟨	D	$i_{psf} - i_{dev}$
🟨	E	i_{dev}/i_{psf}
🟩	F	z_{model}/z_{psf}
🟩	G	$\underline{r_{psf}/i_{exp}}$
🟩	H	i_{psf}/i_{exp}
🟩	I	$\underline{g_{psf}/u_{model}}$
🟩	J	i_{psf}/i_{model}
🟩	K	$\underline{g_{psf}/i_{exp}}$
🟦	L	$\underline{r_{psf}/i_{model}}$
🟦	M	$\underline{i_{dev}/i_{psf}}$
🟦	N	$\underline{r_{psf}/g_{model}}$
🟦	O	$\underline{i_{psf}/z_{model}}$
🟦	P	$\underline{g_{petro}/r_{petro}}$
🟪	Q	$r_{psf} - r_{petro}$
🟪	R	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{dev}}}$
🟪	S	$\underline{r_{dev}/r_{psf}}$
🟪	T	$\sqrt{\sigma^2_{r_{model}} + \sigma^2_{g_{exp}}}$

Fig. A.1. Chord diagram for the experiment DR7a. Every feature is associated to a specific colour, and starting from the first features (H, J) it is possible to follow all the possible paths of the tree, depicting the different feature subsets.

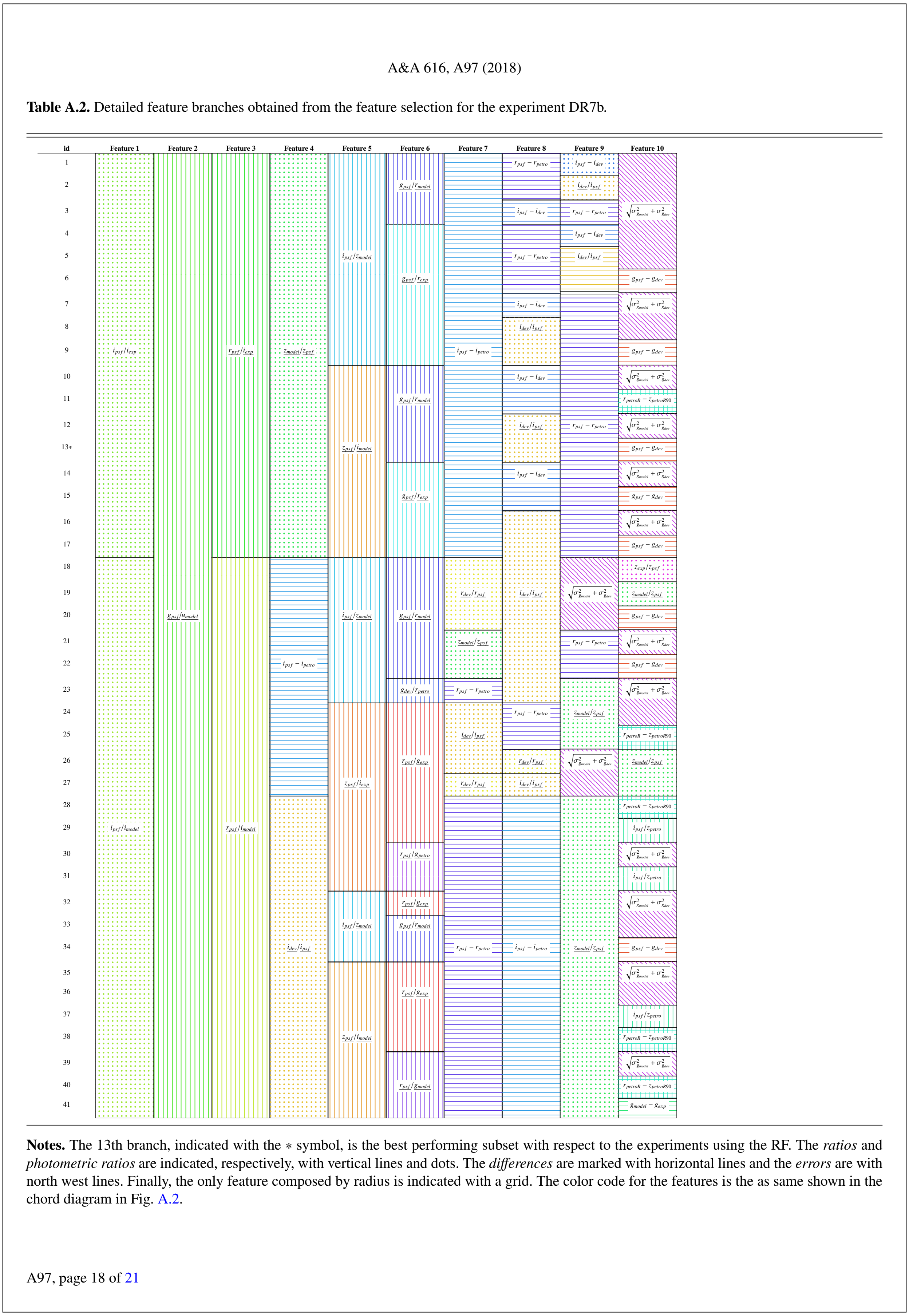


Table A.2. Detailed feature branches obtained from the feature selection for the experiment DR7b.

id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	i_{psf}/i_{exp}	g_{psf}/u_{model}	r_{psf}/i_{exp}	z_{model}/z_{psf}	i_{psf}/z_{model}	g_{psf}/r_{model}	$i_{psf} - i_{petro}$	$r_{psf} - r_{petro}$	$i_{psf} - i_{dev}$	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
2									i_{dev}/i_{psf}	
3								$i_{psf} - i_{dev}$	$r_{psf} - r_{petro}$	
4						g_{psf}/r_{exp}		$r_{psf} - r_{petro}$	$i_{psf} - i_{dev}$	
5									i_{dev}/i_{psf}	
6										$g_{psf} - g_{dev}$
7								$i_{psf} - i_{dev}$	$r_{psf} - r_{petro}$	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
8								i_{dev}/i_{psf}		
9										$g_{psf} - g_{dev}$
10					z_{psf}/i_{model}	g_{psf}/r_{model}		$i_{psf} - i_{dev}$		$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
11										$r_{petroR} - z_{petroR90}$
12								i_{dev}/i_{psf}		$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
13*										$g_{psf} - g_{dev}$
14						g_{psf}/r_{exp}		$i_{psf} - i_{dev}$		$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
15										$g_{psf} - g_{dev}$
16								i_{dev}/i_{psf}		$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
17										$g_{psf} - g_{dev}$
18	i_{psf}/i_{model}		r_{psf}/i_{model}	$i_{psf} - i_{petro}$	i_{psf}/z_{model}	g_{psf}/r_{model}	r_{dev}/r_{psf}		$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$	z_{exp}/z_{psf}
19										z_{model}/z_{psf}
20										$g_{psf} - g_{dev}$
21							z_{model}/z_{psf}		$r_{psf} - r_{petro}$	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
22										$g_{psf} - g_{dev}$
23						g_{dev}/r_{petro}	$r_{psf} - r_{petro}$		z_{model}/z_{psf}	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
24					z_{psf}/i_{exp}	r_{psf}/g_{exp}	i_{dev}/i_{psf}	$r_{psf} - r_{petro}$		
25										$r_{petroR} - z_{petroR90}$
26								r_{dev}/r_{psf}	$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$	z_{model}/z_{psf}
27							r_{dev}/r_{psf}	i_{dev}/i_{psf}		
28				i_{dev}/i_{psf}			$r_{psf} - r_{petro}$	$i_{psf} - i_{petro}$	z_{model}/z_{psf}	$r_{petroR} - z_{petroR90}$
29										i_{psf}/z_{petro}
30						r_{psf}/g_{petro}				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
31										i_{psf}/z_{petro}
32					i_{psf}/z_{model}	r_{psf}/g_{exp}				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
33						g_{psf}/r_{model}				
34										$g_{psf} - g_{dev}$
35					z_{psf}/i_{model}	r_{psf}/g_{exp}				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
36										
37										i_{psf}/z_{petro}
38										$r_{petroR} - z_{petroR90}$
39						r_{psf}/g_{model}				$\sqrt{\sigma^2_{g_{model}} + \sigma^2_{g_{dev}}}$
40										$r_{petroR} - z_{petroR90}$
41										$g_{model} - g_{exp}$

Notes. The 13th branch, indicated with the $*$ symbol, is the best performing subset with respect to the experiments using the RF. The *ratios* and *photometric ratios* are indicated, respectively, with vertical lines and dots. The *differences* are marked with horizontal lines and the *errors* are with north west lines. Finally, the only feature composed by radius is indicated with a grid. The color code for the features is the as same shown in the chord diagram in Fig. A.2.

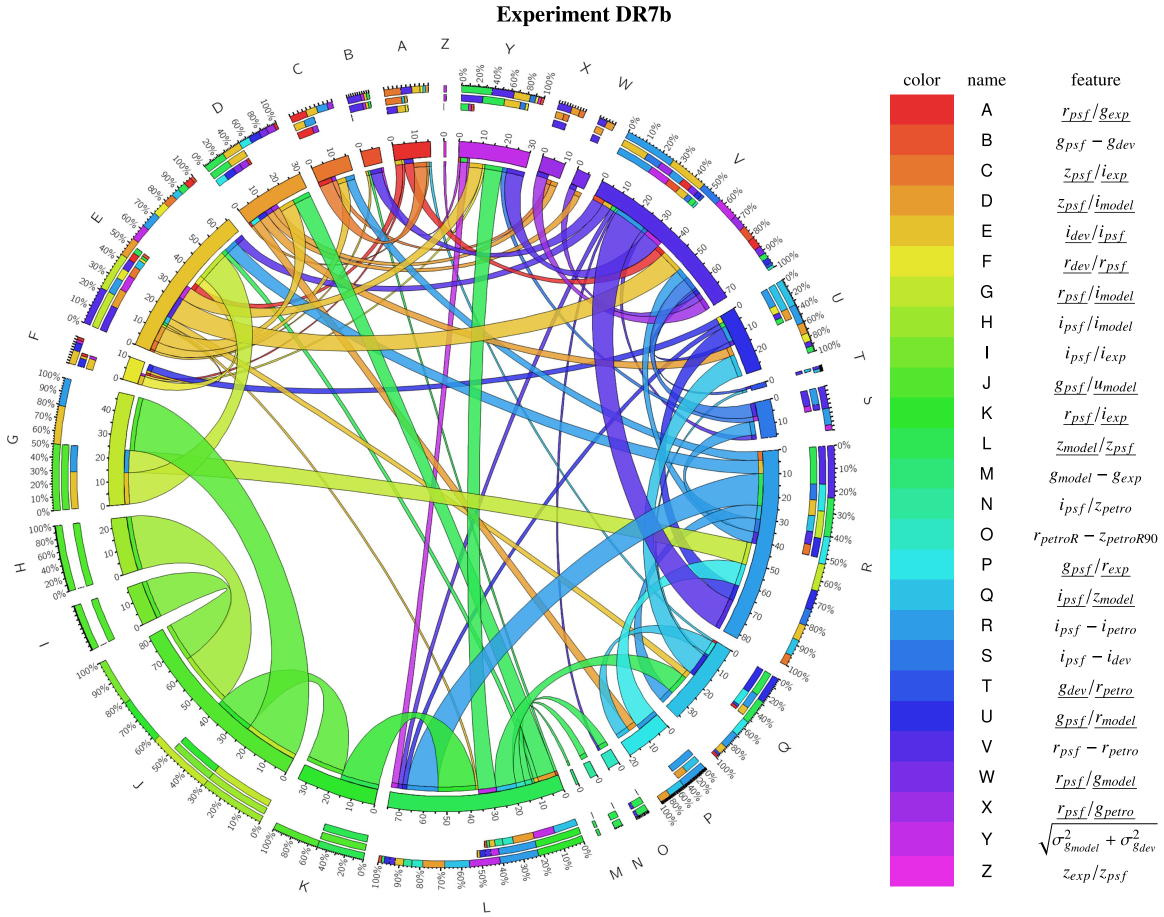


Fig. A.2. Chord diagram for the experiment DR7b. Every feature is associated to a specific colour, and starting from the first features (H, I) it is possible to follow all the possible paths of the tree, depicting the different feature subsets.

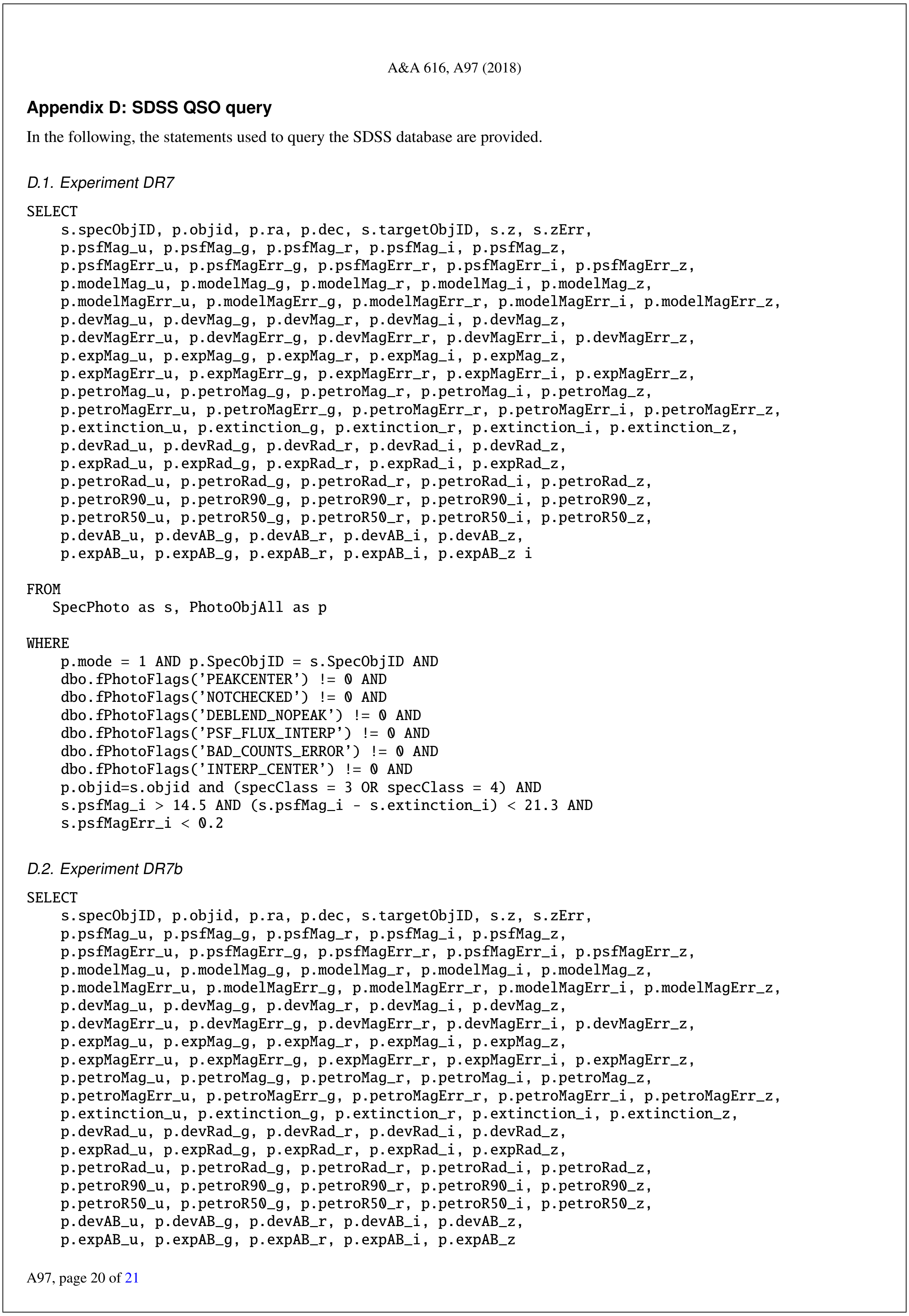
Appendix B: Data

The SDSS object IDs and coordinates of the extracted quasars for the three catalogues are available as supplementary information, as ASCII files.
dr7a.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7a.
dr7b.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7b.
dr7+9.csv contains the SDSS object IDs and coordinates of the quasars for experiment DR7+9.

Appendix C: Code

The code of the DCMDN model is available on the ASCL[6].

[6] http://www.ascl.net/ascl:1709.006



Appendix D: SDSS QSO query

In the following, the statements used to query the SDSS database are provided.

D.1. Experiment DR7

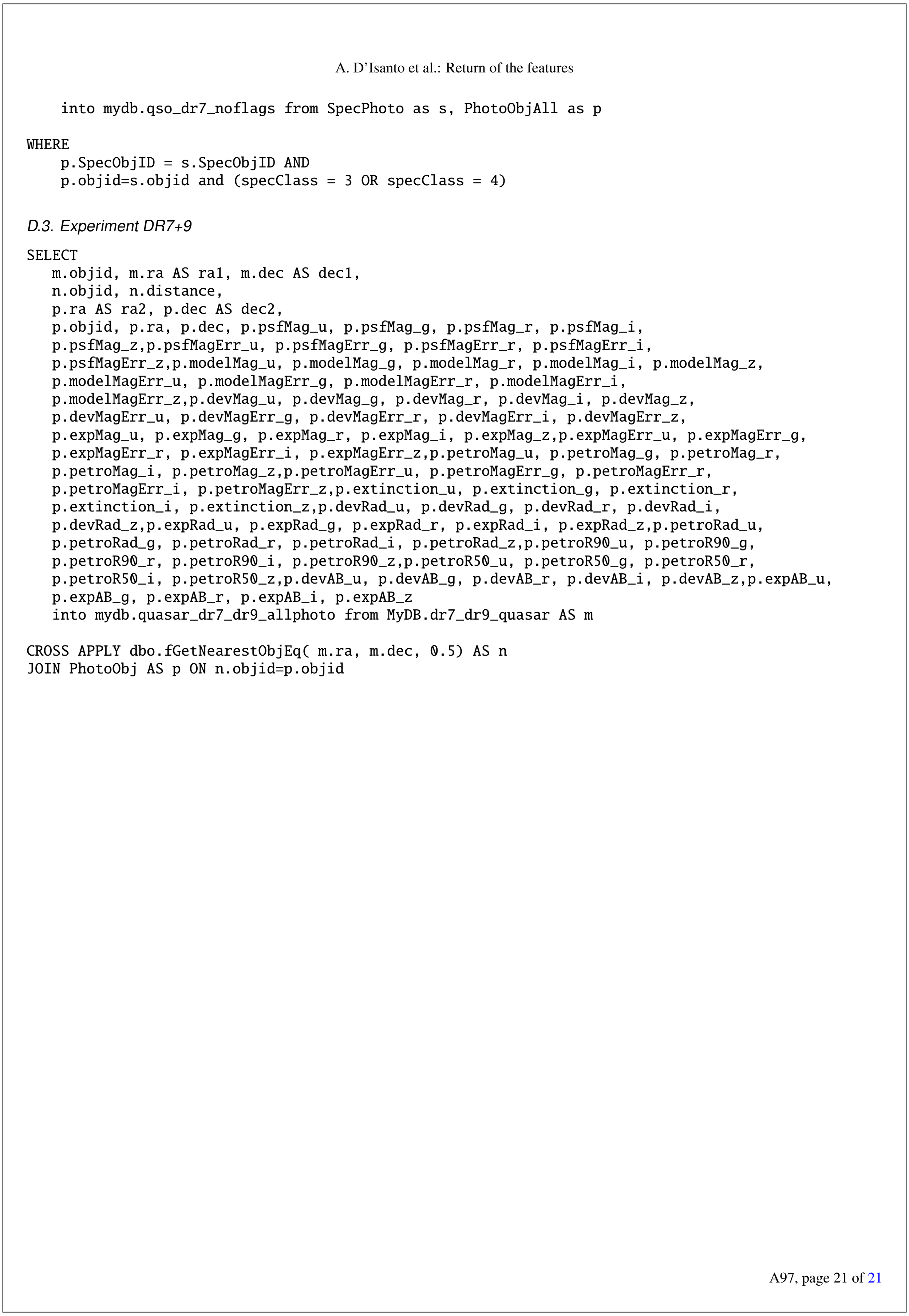
```
SELECT
    s.specObjID, p.objid, p.ra, p.dec, s.targetObjID, s.z, s.zErr,
    p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i, p.psfMag_z,
    p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i, p.psfMagErr_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i, p.modelMagErr_z,
    p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
    p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
    p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,
    p.expMagErr_u, p.expMagErr_g, p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,
    p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
    p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r, p.petroMagErr_i, p.petroMagErr_z,
    p.extinction_u, p.extinction_g, p.extinction_r, p.extinction_i, p.extinction_z,
    p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i, p.devRad_z,
    p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,
    p.petroRad_u, p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,
    p.petroR90_u, p.petroR90_g, p.petroR90_r, p.petroR90_i, p.petroR90_z,
    p.petroR50_u, p.petroR50_g, p.petroR50_r, p.petroR50_i, p.petroR50_z,
    p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,
    p.expAB_u, p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z i

FROM
   SpecPhoto as s, PhotoObjAll as p

WHERE
    p.mode = 1 AND p.SpecObjID = s.SpecObjID AND
    dbo.fPhotoFlags(’PEAKCENTER’) != 0 AND
    dbo.fPhotoFlags(’NOTCHECKED’) != 0 AND
    dbo.fPhotoFlags(’DEBLEND_NOPEAK’) != 0 AND
    dbo.fPhotoFlags(’PSF_FLUX_INTERP’) != 0 AND
    dbo.fPhotoFlags(’BAD_COUNTS_ERROR’) != 0 AND
    dbo.fPhotoFlags(’INTERP_CENTER’) != 0 AND
    p.objid=s.objid and (specClass = 3 OR specClass = 4) AND
    s.psfMag_i > 14.5 AND (s.psfMag_i - s.extinction_i) < 21.3 AND
    s.psfMagErr_i < 0.2
```

D.2. Experiment DR7b

```
SELECT
    s.specObjID, p.objid, p.ra, p.dec, s.targetObjID, s.z, s.zErr,
    p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i, p.psfMag_z,
    p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i, p.psfMagErr_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i, p.modelMagErr_z,
    p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
    p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
    p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,
    p.expMagErr_u, p.expMagErr_g, p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,
    p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
    p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r, p.petroMagErr_i, p.petroMagErr_z,
    p.extinction_u, p.extinction_g, p.extinction_r, p.extinction_i, p.extinction_z,
    p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i, p.devRad_z,
    p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,
    p.petroRad_u, p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,
    p.petroR90_u, p.petroR90_g, p.petroR90_r, p.petroR90_i, p.petroR90_z,
    p.petroR50_u, p.petroR50_g, p.petroR50_r, p.petroR50_i, p.petroR50_z,
    p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,
    p.expAB_u, p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z
```



```
    into mydb.qso_dr7_noflags from SpecPhoto as s, PhotoObjAll as p

WHERE
    p.SpecObjID = s.SpecObjID AND
    p.objid=s.objid and (specClass = 3 OR specClass = 4)
```

D.3. Experiment DR7+9

```
SELECT
   m.objid, m.ra AS ra1, m.dec AS dec1,
   n.objid, n.distance,
   p.ra AS ra2, p.dec AS dec2,
   p.objid, p.ra, p.dec, p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i,
   p.psfMag_z,p.psfMagErr_u, p.psfMagErr_g, p.psfMagErr_r, p.psfMagErr_i,
   p.psfMagErr_z,p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
   p.modelMagErr_u, p.modelMagErr_g, p.modelMagErr_r, p.modelMagErr_i,
   p.modelMagErr_z,p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
   p.devMagErr_u, p.devMagErr_g, p.devMagErr_r, p.devMagErr_i, p.devMagErr_z,
   p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,p.expMagErr_u, p.expMagErr_g,
   p.expMagErr_r, p.expMagErr_i, p.expMagErr_z,p.petroMag_u, p.petroMag_g, p.petroMag_r,
   p.petroMag_i, p.petroMag_z,p.petroMagErr_u, p.petroMagErr_g, p.petroMagErr_r,
   p.petroMagErr_i, p.petroMagErr_z,p.extinction_u, p.extinction_g, p.extinction_r,
   p.extinction_i, p.extinction_z,p.devRad_u, p.devRad_g, p.devRad_r, p.devRad_i,
   p.devRad_z,p.expRad_u, p.expRad_g, p.expRad_r, p.expRad_i, p.expRad_z,p.petroRad_u,
   p.petroRad_g, p.petroRad_r, p.petroRad_i, p.petroRad_z,p.petroR90_u, p.petroR90_g,
   p.petroR90_r, p.petroR90_i, p.petroR90_z,p.petroR50_u, p.petroR50_g, p.petroR50_r,
   p.petroR50_i, p.petroR50_z,p.devAB_u, p.devAB_g, p.devAB_r, p.devAB_i, p.devAB_z,p.expAB_u,
   p.expAB_g, p.expAB_r, p.expAB_i, p.expAB_z
   into mydb.quasar_dr7_dr9_allphoto from MyDB.dr7_dr9_quasar AS m

CROSS APPLY dbo.fGetNearestObjEq( m.ra, m.dec, 0.5) AS n
JOIN PhotoObj AS p ON n.objid=p.objid
```

Chapter 5

Discussion

The work presented in the three publications reported in Chapter 4 has been carried out with the aim of developing well performing methods for photometric redshift estimation. Two main methods have been presented and several novel approaches introduced into the field, both from a methodological and statistical point of view. In the following sections I will discuss these aspects in detail and combine the results obtained in a common overview in order to clarify the most important advances that this work brings to the field. I will also outline the strenghts and weaknesses of the two models and discuss the open questions which demand further analysis.

5.1 Fully automated model

The first method is based on the implementation of a deep learning model, the DCMDN, which is fully automated and provides the feature extraction and redshift estimation directly from images in the form of density distributions. Such a model has been succesfully applied to different catalogs taken from the SDSS. This method presents several novel aspects; while the application of convolutional neural networks for redshift estimation directly from images had been done already by Hoyle [2016], this latter work is based on the prediction of pure point estimates on a catalog composed only by galaxies and with an error estimation based on the classical parameters used in the literature. The model presented in Publication I and Publication II, thanks to the combination of a convolutional neural network with a mixture density network, allows the prediction of multimodal PDFs, with all the advantages already stated in Sec. 2.3.3. Moreover the use of the CRPS and of the PIT allows for the correct estimate of the error for the predicted PDFs, seeking the best calibration and sharpness. The use of the CRPS as a loss function for the training of the neural network constitutes an absolute novelty, both in astronomy and computer science. The correct implementation of this function was a challenge also from a technical point of view. In fact, the code of the DCMDN has been realized in the Theano environment [The Theano Development Team et al., 2016]. This library is specifically meant to realize machine learning models with the Python language and it is characterized by a symbolic structure which generates a *graph* containing the architecture of the model before the runtime phase. The introduction of a loss function based on the calculation of an integral with symbolic variables brought up many technical issues which had to be solved in order to reach a correct convergence of the model. In the field of computer science, much effort has been put by the community into finding proper and efficient loss functions when dealing with machine learning models, as shown in Janocha and Czarnecki [2017]. However, the use of the CRPS for this task is particularly important to predict distributions by means of a model which is already built with a focus on the maximization of the sharpness, subject to calibration. As already stated, the CRPS has been originally adopted in the weather forecast field, but we applied it for the first time in astronomy and as a loss function for a neural network (also related to the computer science field). Therefore, I have to point out that the probabilistic model used to realize the DCMDN, with the CRPS implemented as loss function for a neural network, has been recently applied in the same field, as reported in Rasp and Lerch [2018], where our model is cited. In this work the authors adopt a fully connected neural network model to perform weather forecast predictions in a novel way with respect to statistical post-processing of the errors, which is traditionally used in the field. This approach highly benefits from the use of the CRPS for a Gaussian distribution as a loss function, giving a correct mathematical background for the attempt of sharpness maximization subject to calibration. A modified version of the CRPS has been recently adopted in the medical sector [Avati et al., 2018] as well. In this paper the CRPS is introduced as loss function to optimize a recurrent neural network [Hopfield, 1982] and the performances are compared to those obtained by using the maximum likelihood, finding significant benefits. This is due to the fact that the CRPS is focused, as already said, on the sharpness maximization, while the likelihood is more related to the spatial positioning of the predictions.

Concerning the astronomical field, I want to report that the approach proposed in Publication II constitutes the basis for the work from Pasquet et al. [2018]. In this paper, the authors adopt the CRPS and the PIT as scoring rules to assess the quality of their predictions, namely photometric redshift PDFs determined from images, as in our model, with a convolutional neural network based architecture. The results shown are, theoretically, superior to what we obtain with the DCMDN, but the training is performed in several, smaller, redshift ranges, allowing in this way a higher precision.

The use of the DCMDN brings several advantages, which we verified from the results of the experiments and the extensive discussion given in Publication II. However there are some points that require further analysis to better understand the advantages of such a model and what requires further improvement. A fully automatic model based on a convolutional network is able to automatically generate thousands of features, leaving the machine performing the task of feature extraction and selection. In other words, the network performs a dimensionality reduction, reducing the original images to a (high) number of parameters and focusing on those which prove to be most important to maximizing the performance. This has two main disadvantages. The main risk is losing control of what happens inside the machine, which tends to become a sort of black box [Knight, 2018]. In fact, the process of weights optimization, which produces the features out of the feature maps, is completely managed by the machine without any need of intervention by humans. It goes without saying that the automation of the pipeline which extracts photometric redshift starting from raw images is a clear advantage, but it is important to preserve knowledge and control of every step of the process, to avoid mistakes or misleading results. There are several studies in the field with respect to this problem, and in particular in computer science there are several possible solutions proposed in the literature [Hohman et al., 2018]. This goes deep into the field of visualization problems and visual analytics and essentially the problem is reconducted to the comprehension of the *five W's and How* (Why, Who, What, How, When and Where). In more detail, following Hohman et al. [2018], the following questions should find an answer:

- Why: why would one want to apply visualization in deep learning?
- Who: who would use it and have benefit from visualization?
- What: what data, features and relationships can eventually be visualized?
- How: how can we practically visualize these data, features and relationships?
- When: when, in the learning process, is the visualization applied?
- Where: where, in the architecture of the model, has the visualization been used?

The detailed description of the techniques at study to answer the *five W's and How* goes beyond the scope of this thesis. Unfortunately at the current stage none of the proposed techniques has been applied in the astronomical field. Currently I am discussing the problem with several collegues and working to find solutions.

The second disadvantage given by fully automated models is deeply connected with the first one: the lack of interpretability of the produced features. Traditionally, the features manually generated and used in the literature have a well defined physical meaning which can be connected to the nature and the properties of the problem itself. In the case of photometric redshift estimation, for example, magnitudes represent a low resolution approximation of the spectrum from which the spectroscopic redshift has been extracted. Features can be manually engineered by taking into account the specific problem considered and/or the model adopted, as pointed out in Sec. 3.6, or as a result of a dimensionality reduction. The automatically generated features of the DCMDN are instead nothing more than a huge vector of numbers. They constitute a discretization of the original information contained in the images and even if some physical property or meaning correlated with the characteristics of the sources should be embedded into them, it is very hard, at the current stage, to specifically identify them. This problem affects also different models of neural networks, even more specifically focused on feature extraction and dimensionality reduction, like autoencoders [Liou et al., 2008, Gianniotis et al., 2016]. It goes without saying that finding efficient visualization techniques which could be helpful to disentangle the choices and the mechanisms happening into the architecture, in order to unveil the *black box*, could be equally helpful in finding an interpretation for the automatically produced features.

However, the implementation of such models allows to reach performances and to treat amounts of data, which are already well beyond any human capability of inspection and analysis. A striking example of the possibilities open by such technologies, not related to astronomy, is that given by the neural network model able to beat the world champion of the game AlphaGo [Hölldobler et al., 2017]. This model has been subsequently defeated by another neural network [Silver et al., 2017], proving the rapid advancements in the field. Therefore it is worthwhile experimenting with such technologies, also in the view of developing online models. The DCMDN, and all the other models depicted in this thesis, are in fact all based on offline analysis. This means that the data has been already pre-processed and downloaded from the databases where they were stored. This way of processing data is indeed useful for many applications, but in the near future, with the data cascade coming by instruments like SKA or LSST, an online approach in which the data are reduced, analysed and processed in real time will be advisable, if not mandatory.

5.2 Feature based model

The approach depicted in Publication III is meant to combine the advantages given by feature based methods, in particular concerning the physical intepretability, with the attempt to maximize the amount of useful information taken as input by a machine learning model. Despite the attempts reported in Sec. 2.2-2.3, traditionally all the works dedicated to photometric redshift estimation have found magnitudes and colors as best features for this task, even after a feature selection with many other features. The use of all the parameters contained in the SDSS database, and their massive combination, goes in the direction of obtaining a huge number of features by which we are trying to include as much information as possible in the selection process. Many of the features obtained are correlated or redundant; therefore the forward selection allows the model to focus only on the most important features, but with a much wider choice with respect to manual feature selections based on some tens of parameters. Moreover, as it is shown in the paper, the way in which the features are built and combined boosts the performance. It appears that what really makes the difference is not the single feature but the global structure of the sets. Furthermore, as compositions of well known features, whose meaning is clear, one can try to interpret the selected features physically. The proposed method cannot work for all the features and for every physical case, but it can give an indication of the general meaning of the selected features and on the mechanisms going on in the machine, which correspond to specific physical processes characterizing the sources. In this sense, the improvement in performance noticed by using the proposed features was also explained. The new features are able to fully capture the processes going on in the sources, namely quasars, identifying spectral emission lines, which are just partially correlated with the classic features. This is an additional confirmation that the behavior of the model is correct. In fact, as it is stated in the paper, no shape related features have been selected by the model. This is due to the fact that quasars are point-like sources, so radii and ellipticities have much less influence as the kNN explores the feature space with respect to flux-related features. However, it is expected that shape-related features will have a more prominent role when dealing with galaxies, as they are extended objects. Some preliminary results on new experiments I am currently conducting seem to confirm this hypothesis.

Further research into the literature, especially in the computer science and machine learning field, has shown that the important role acquired by ratios and differences of features is understandable in terms of the dependance of the machine learning model from the representation of the feature vector. In this sense, it is not uncommon to adopt this kind of mathematical transformations for feature engineering [Heaton, 2017]. I believe that such evidence makes still stronger the result obtained, as it is possible to establish a connection between a pure data-driven effect and the physics, explained from the corrispondence between feature importance and spectral lines. In other words, the internal structure of the data is reflected in a cleverly engineered feature space, allowing the model to capture relationships and patterns which mirror the physical nature of the sources. It must be noticed that the combinations designed for the chosen parameters, producing the final number of features, was limited (together with the size of the data sample) by the GPU memory amount. In the case of the Nvidia Pascal P40 used for the experiments [1], the memory was of 24 GB. Theoretically there is no other limit, except the memory resources, to the number of combinations that could be investigated, indefinitely increasing the number of features which can be given to the model. A possible extension of the method could reckon, for example, on more complex combinations of the parameters, which could generate still novel and better performing features. This, however, could have the side effect of making their interpretation even more complicated.

The forward selection performed has to be included in the field of the wrapper methods, as already specified in Sec. 3.6. This category of algorithms are based on the use of the predictive model also for the feature selection process. Therefore the feature set found will be particularly adapt to that particular model, giving the best performance in combination with it. On the other hand, this could be considered a weakness, as the feature set could be not general enough to give good results with different models. For this reason the evaluation of the best set has been performed using the RF instead of the kNN. In this case, the performance given by the RF is satisfying enough to assume that the features found are reflecting not just properties related to the specific model adopted, but more general effects connected to physical properties. This is further confirmed by the given physical description. However it has been shown in the paper that the RF is giving a different feature importance with respect to the ordering of the features as selected by the forward selection based on the kNN. This can be explained by the differences occurring between the two methods. It is important to point out that each step of the forward selection is conducted by performing 100 independent experiments, each extracting a random subsample from the data. The random sampling, together with the cross validation, is meant to avoid a selection of features that by chance perform well on a particular dataset. In other words, these tools are useful in preventing overfitting and lack of generalization of the model.

It is evident that the features found with this method are highly related to the specific problem and the sources considered. This is a disadvantage with respect to the DCMDN model, which allows the estimation of photometric redshifts independently from the composition of the catalog with good performances. In Sec. 5.3 the performance of the two models will be compared in more detail, with particular focus on the results obtained with respect to the loss functions used and the PIT.

5.3 Performance comparison

The performances given by both of the methods presented are clearly superior to those obtained with traditional models taken from the literature, like RF or even plain neural networks trained with classic features. The results obtained in the publications for the quasar catalogs can be roughly compared with those, for example, given by Laurino et al. [2011], and they appear to be comparable and in the same order of magnitude and often better. The comparison can be only at a rough level, as the catalogs are not exactly the same, but being mostly composed by the quasars contained in the DR7 of SDSS. However it should be considered that the experiments, in this work, had no pre-processing or data selection based on quality flags, nor post-processing or catastrophic outliers removal has been performed, as was done instead in Laurino et al. [2011]. Removing those objects which are noticeably deviating from the ideal line improves dramatically the performance.

A comparison must also be carried out with the results from Polsterer et al. [2014], Gieseke et al. [2014]. In these works, which are preliminary for what is depicted in Publication III, the fundamentals of the photometric redshift estimation via massive feature selection are established. In particular, in Polsterer et al. [2014] the idea of a massive combination of features is presented, in order to find a possible best combination of features in a set of 55 parameters. A brute force approach is used, testing all the possible 4-features out of $341,055$ combinations, to find the best.

[1] http://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf

In Gieseke et al. [2014] instead, the algorithm for the extreme parallelization in the greedy forward selection is presented, with particular focus on the improvement with respect to the computational time. By adopting this algorithm it was possible to realize the work presented in Publication III, increasing the number of available features from 55 to $4,520$. The forward feature selection was necessary to find the possible best combinations, as described in the paper. The results obtained are not directly comparable with respect to those shown in these two works, due to the differences in the dataset adopted. However, there is a methodological improvement given by the generation of the feature tree and the detailed statistical analysis performed. Furthermore, the given physical interpretation provides an explanation of the behavior of the model in a new fashion, finding a connection with the processes going on in the sources which is beyond a pure data-driven approach.

Another important aspect which I would like to outline is the comparison between the automated model, namely the DCMDN, and the feature based model. By looking at the mere numbers, namely RMSE and CRPS, the forward selection model would appear preferable, as it gives the best performance in terms of pure error. However, if looking at the PIT, it can be noticed that the calibration achieved with the DCMDN is superior to any other model. The reason is that the DCMDN is trained using the CRPS as loss function. Therefore the model is forced to maximize the calibration. In this sense it is important to understand the use case for the photometric redshifts, in order to select the most appropriate model and the form, based on a point estimate or a PDF. In particular this last point should be considered when choosing the preferred model. The DCMDN generates true multimodal PDFs, due to its use of a mixture density network in the fully connected part. The model is meant to predict the parameters (means, variances and weights) defining a Gaussian mixture. The feature based model instead is based on a pipeline which ends by adopting a RF, which is not natively generating PDFs. The only way to obtain a density distribution, in this case, is to fit the Gaussian mixture to the predictions of every single tree in the ensemble. This model is clearly more focused on the study of the features and their behavior with respect to the classic ones. Nothing prevents the use of the selected features with a model meant to deal with density distributions, like a plain mixture density network.

I want to point out once more that in the experiments shown in the three publications I attempt to not use quality flags as they are given in the SDSS documentation[2]. In fact, by adopting these flags, a catalog can be cleaned in order to obtain a perfectly clean photometry. However I believe that such an operation makes an experiment less similar to a realistic use case, which is important in order to develop techniques suitable for future mission implementations and online training tasks. Therefore the catalogs were cleaned only with respect to objects presenting meaningless detections like 'NaNs' or absurd values of magnitudes and errors, which could cause a complete failure of the training phase. Instead problematic objects were kept, as this is the challenge that a well performing and efficient model should deal with. It has been shown in Publication II that a proper cleaning of the data, by adopting post-processing techniques, can dramatically improve the results in terms of the error function. The question is how much such an operation could be legitimate with respect to the development of an affordable model which could be used in a realistic case. Moreover it has to be taken into account the possibility that objects characterized by bad detections or proving to be catastrophic outliers could potentially be the most interesting sources, demanding further investigations. The problem of outliers treatment will be analyzed in more detail in Sec. 5.4.

5.4 Correct estimation of errors

In Chapter 3 I gave an overview of the tools introduced in this thesis to correctly estimate the errors and the quality of the predictions, namely the CRPS and the PIT. This has been motivated by the need to correctly deal with density distributions and to take into account multimodalities. Moreover, the adoption of a proper scoring rule like the CRPS permits to perform the predictions by attempting to maximize the sharpness subject to calibration. All these aspects have been already pointed out in Publication II, but hereby I would like to briefly extend the discussion focusing on the reasons for which the adoption of correct error measures is fundamental in order to obtain affordable predictions.

[2] `https://www.sdss.org/dr14/tutorials/flags/`

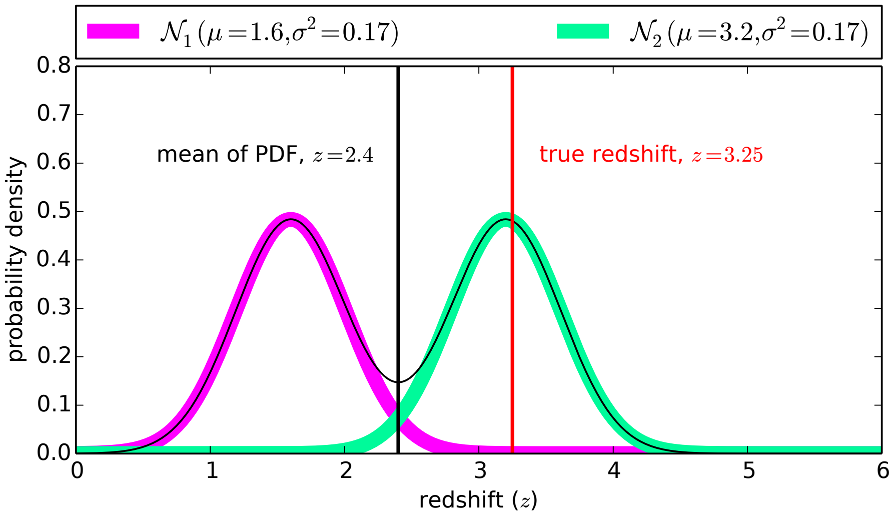
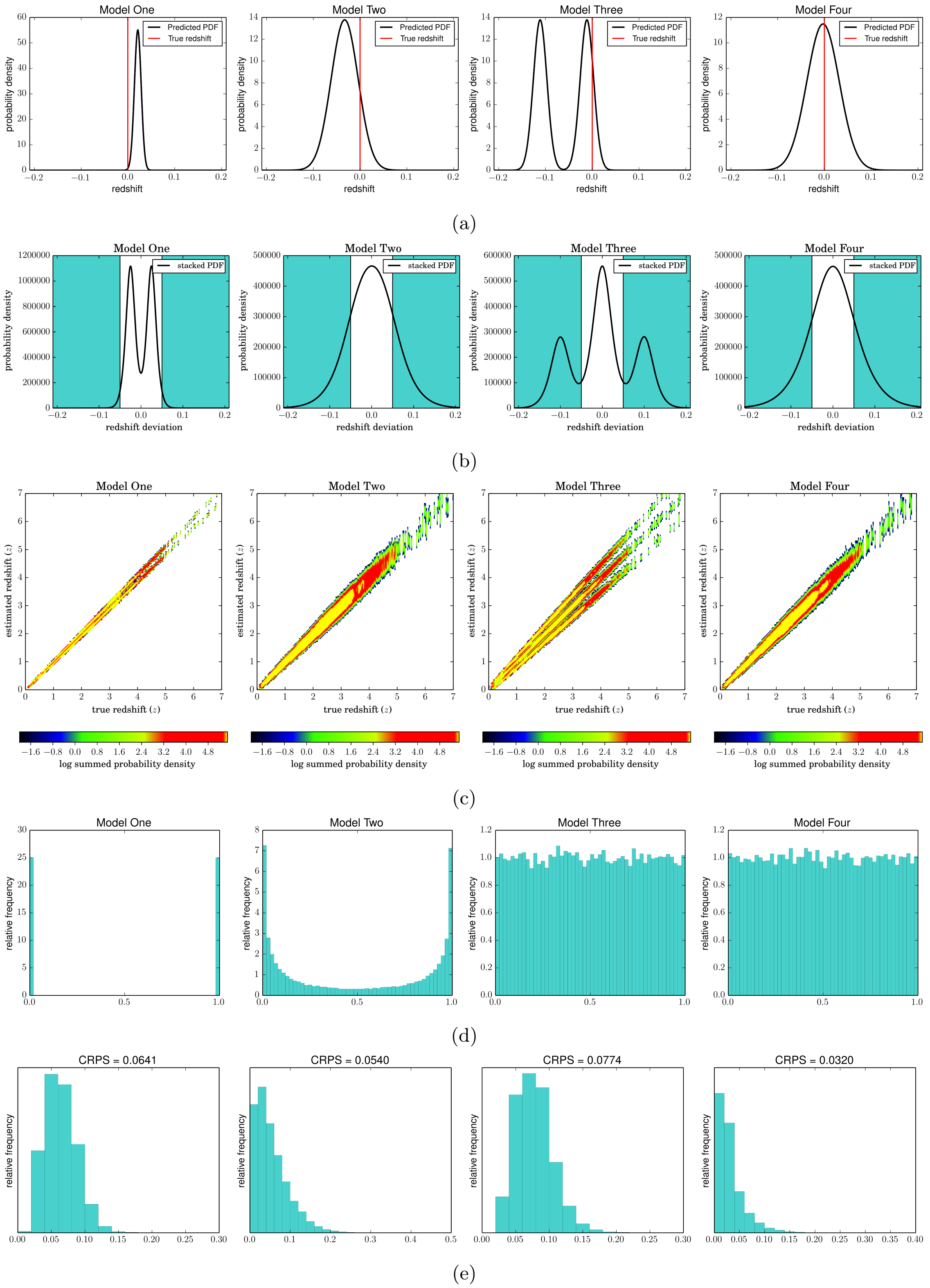


Figure 5.1: Example of misleading result obtained by evaluating the error given a PDF prediction via its mean. The bimodal distribution shown, composed by two normal distributions is here reduced to its mean, and the true redshift value is shown. It is obvious that the value of the true redshift is much closer to the second peak of the distribution than to its mean, which is also falling in a region of low probability density of the distribution. The plot is taken from Polsterer et al. [2016].

Typically the traditional statistics used in the literature [D'Abrusco et al., 2007, Laurino et al., 2011] to estimate errors in the photometric redshift estimation field are the deviation between the true value and the predicted value, or bias, the root mean square error and/or standard deviation and the mean absolute deviation. It is common to apply normalization with respect to redshift in order to express a relative error. These kinds of statistics are well suited to deal with point estimates and their deviations with respect to the true values. However, in case of PDFs predictions, the results given in such a way are not precise enough and can even be completely misleading. The calculation of such statistics in fact requires the reduction of the PDF to a single value in order to perform a comparison with the true value. This is typically done by estimating the mean of the distribution. Unfortunately, this operation, in the presence of multimodal distributions, is highly dangerous, as the results can be completely wrong. A clarifying example is given in Fig. 5.1 where, following Polsterer et al. [2016], a bimodal distribution is shown as a composition of two Gaussian distributions. It is clear that, in such a case, reducing the distribution to its mean to evaluate the performance, by means of the classical scores, is completely misleading. In fact, the true redshift is much closer to one of the peaks than to the mean value. Furthermore the mean of the PDF is in a region of the distribution exhibiting a low probability density, making meaningless to express the predicted redshift by the mean and the error with respect to the true value by the RMSE. This clarifies why the information given by a PDF should be used entirely, withouth oversimplifying the predictions, wasting useful information and producing useless results.

Another important point which should be further discussed is the treatment of outliers. Following the Euclid requirements already reported in Chapter 3 and stated by Laureijs et al. [2011], the number of outliers for the photometric redshift estimation pipeline of the mission should be such that the quantity $\sigma_z/(1+z)$ needs to be less than 5%. This requirement is connecting the analysis of outliers to the error measure and it has to be treated carefully when predicting PDFs. In fact, this operation can generate several problems, as it is shown in Fig. 5.2. It is common in literature [Sadeh et al., 2016, Asorey et al., 2016, Cavuoti et al., 2017] to detect the amount of outliers by adopting a threshold criterion (e.g. 5%). This is done by shifting every PDF of the corresponding true redshift and then co-adding them. The resulting stacked PDF is used to evaluate the amount of predictions deviating too much from the ideal line, or, in other words, which are exceeding the given threshold.

In Fig. 5.2, four different models are depicted. The models are simulations of extreme cases obtained by perturbating true data accordingly. Fig. 5.2a gives an example of a PDF predicted by



(a)

(b)

(c)

(d)

(e)

Figure 5.2: Plots representing different statistical indicators for four different models. From top to bottom, an example of a PDF predicted by the used model, the stacked PDF, the true versus estimated redshift density plot, the PIT histogram and the CRPS are shown. The four models represent, respectively, a sample of predictions characterized by narrow Gaussians (Model One), broad Gaussians (Model Two), bimodal Gaussians mixtures (Model Three) and an *ideal* case with broad Gaussians with the mean very close to the true value (Model Four).

each model, already shifted by the true redshift. For each of them, the stacked PDF (Fig. 5.2b), the density plot of the true redshifts versus the sum of the predicted PDFs (Fig. 5.2c), the PIT (Fig. 5.2d) and the CRPS (Fig. 5.2d) are given. The colored bands in the stacked PDF plots represent the outliers treshold as defined in Laureijs et al. [2011]. In Model One, a sample of narrow Gaussian distributions is used to estimate the redshifts. The PDFs are designed to be characterized by a small σ and shifted to a higher or lower value with respect to the true redshift. Due to the small σ value, the true redshift is not well captured by the predicted distributions. In Model Two, the distributions are broader, as they have been simulated by selecting a larger σ value. In this case, the true redshift is falling in the PDF even if, due to the shift, this is never very well predicted. Model Three instead shows a case in which the PDFs are given by a Gaussian mixture model composed of two Gaussians. The model has been designed to characterize the distributions such that one of the two peaks is representing the true redshift well, while the second peak is further away. Finally, Model Four represents an *almost ideal* case, in which the means of the predicted PDFs are all close to the true redshift, without any shift.

An analysis based merely on the stacked PDF and the true versus estimated redshift plot gives results that can be highly misleading, as proven by calculating the PIT and the CRPS. In Model One the stacked PDF has a relative minimum correspondending to the true redshift and because of the narrow distributions never capture it, the PIT is extremely underdispersed. Even if there are almost zero outliers, this cannot be considered a good prediction. Model Two improves the situation, as the PDFs are broader and the true redshift is typically falling into the distribution. However, due to the shift, none of the predictions are predicting the true value, and this still generates a underdispersed PIT. Moreover the amount of outliers in this case, as indicated by the stacked PDF, is not negligible. Model Three, due to the bimodal configuration, generates a extremely high number of outliers and this is shown also from the very poor CRPS value. Despite this, the PIT is perfectly uniform. In fact, the PIT is just expressing the calibration of the predicted distributions. As in this model the true redshift is well captured by one of the two peaks, the histogram gives a uniform distribution. Model Four, as already said, is a almost ideal case, giving good performance in terms of the CRPS and a uniform PIT plot. This model is barely distinguishable from Model Two if only the stacked PDF and the density plot are taken into account.

From the comparison of these models, it becomes clear that stacking PDFs is not a good solution for detecting and defining outliers and estimating the performance. In fact, following the definition given in Eq. 3.1, the stacked PDF is not a proper score. The treshold at the tails of the distribution, which is fixed to remove outliers, can potentially cut the minimum of the natural distribution that makes the statement valid. In other words, reporting the natural probability distribution does not give the minimum expected penalty. Outliers, intended as problematic sources, should be instead treated by adopting a probabilistic description and not defining them by the relative difference between the true redshift and parts of the PDF, nor based on the area of the stacked PDF [Polsterer et al., 2016]. The use of a probabilistic indicator (e.g. the likelihood) as outlier criterion is recommended instead. In this sense, a comparison between the use of the likelihood and/or the CRPS as scoring and loss functions has been already done in Publication II. However, a good analysis on this subject is given by Gebetsberger et al. [2017], where likelihood and CRPS are adopted as score functions in the context of non-homogenous regression models for post-processing ensemble weather forecasts. As it has already been stated, the likelihood is more focused on the spatiality. This means that it is more sensitive to outliers with respect to the CRPS [Selten, 1998]. As shown in Gneiting et al. [2005] this sensitivity leads to generating overdispersed predictions.

As shown in Fig. 5.2, it is evident that, in order to estimate the performance and to deal with outliers, the combination of different tools, like the CRPS and the PIT, are necessary. None of these tools alone, in fact, are able to correctly express the quality of the predictions. The previously described ambiguities are mainly created by the use of a measurement like the RMSE, which is not ideal in the case of distributions, and by the stacked PDF, which is not a proper score. Instead, the requirement of a probabilistic description is related to the importance assumed in such a framework by calibration and sharpness, which are connected to the adoption of proper scores and to the quality of the predicted distributions.

A description based on proper scores, like the likelihood or the CRPS, and a correct errors estimation is also fundamental to correctly take into account multimodalities. In particular, the effect

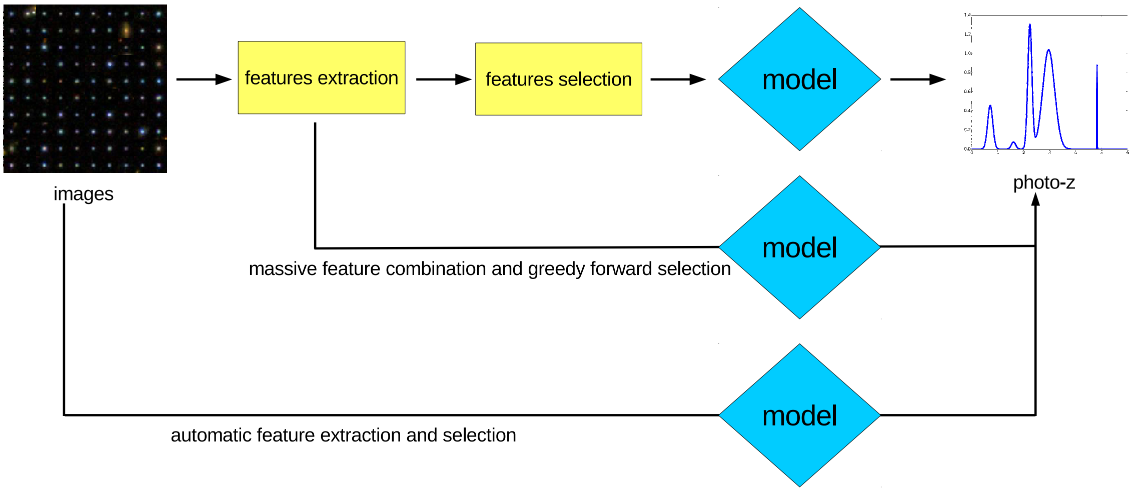

images
features extraction
features selection
model
photo-z
model
massive feature combination and greedy forward selection
model
automatic feature extraction and selection

Figure 5.3: Workflow depicting the traditional process of photometric redshift estimation, in the different steps which bring from raw images to the redshift. The two methods presented in this thesis are indicated on distinct paths, as different alternatives to improve the original workflow.

of multimodalities becomes important to distinguish between sources characterized by degeneracy, and so exhibiting multiple density peaks, and true outliers, which deviate from the ideal line for other reasons. A description based on a Gaussian mixture model, like the one presented in this thesis, allows the association of every source to multiple redshift values, corresponding to the different peaks of the PDF, each with a specific probability.

Taking all these considerations into account, it would be appropriate that the tools presented could become a standard, as already happened in Pasquet et al. [2018] and in Tanaka et al. [2018], which adopted the CRPS and PIT to estimate the quality of their predictions in the field of photometric redshift estimation.

5.5 A comprehensive overview

The two models presented in this thesis follow a common scheme, which starts from raw images and ends with photometric redshift estimation. Such workflow is depicted in Fig. 5.3. From a very general point of view the process always begins from photometric images, which are used to extract fluxes and magnitudes, with their errors. Traditionally such parameters constitute the basis for features, which can be the same magnitudes and/or colors. In the literature there are several works in which other bands besides the optical ones have been used, like the infrared. The feature extraction phase is followed by the feature selection, which can be manual or automated, and then a model is applied to predict photometric redshifts. It must be noted that when using template based models, instead of machine learning techniques, the workflow stays unchanged. In fact, template models are based on the estimation of photometric features, which are required to fit the spectral energy distribution. The models proposed are inserted in these schemes, allowing automatizing, or to improve the efficiency and performance of certain steps in the workflow.

The DCMDN is a fully automated model. Therefore, starting directly from images, a single model takes care of the feature extraction and selection phase, by means of the convolutional part, and of the photometric redshift estimation through the fully-connected part of the network. The model is fully probabilistic, pre-classification less and takes into account multimodalities. Moreover, it is trained using a proper score like the CRPS, maximizing the sharpness of the predicted PDFs.

The feature based model instead requires a feature extraction step, which is done by the massive combination of the 90 parameters taken from the SDSS. The feature selection is then performed applying the forward selection in order to generate a feature tree. In this sense, this method employs two different machine learning models: the kNN, which is used during the feature selection phase, and the RF, which is adopted to test the different branches of the tree and find the best feature combination. The feature selection step in this case is not automated and still requires human knowledge and intervention, but it is developed in a novel and more efficient way. This permits us to make better use of the information contained in the data, originally extracted from the images.

In Fig. 5.3, both alternative paths to the general workflow cross a *model box*, which delivers the final photometric redshift estimation. In the case of the DCMDN, this model is represented by the fully connected part of the network, while for the feature based model, it is just the RF used to predict the redshifts with the selected features. However, the workflow is not model dependent and in both cases different models and/or different architectures could be adopted. One can even increase the complexity of the workflow by using ensembles in order to find the best configuration or the best model with respect to the selected task.

In conclusion, the presented scheme can have many different interpretation and variations, but the general idea and the steps to do will be always more or less the same, even if embedded into the models adopted. The variations on the workflow are meant to improve the performance, the affordability and the statistical consistency of the evaluation process, to speed up and facilitate the calculation and, in the end, to develop a model which could give photometric redshift estimates for as many sources as possible.

The work presented in this thesis has mainly a methodological purpose. It is meant to give to the community affordable methods and techniques to deal with the problem of photometric redshift estimation, to correctly treat the errors and to manage with features extraction, selection and interpretation. It goes without saying that the concept, and even the two methods presented in the thesis, are quite general, and can be applied to many other astronomical problems, whether they are regression or classification tasks.

Chapter 6

Conclusion

The work presented in this thesis is dedicated to the development of novel methodologies for photometric redshift estimation with machine learning methods. Photometric redshifts are important as a fundamental measurement to retrieve distances in the cosmological field and as a major step of the cosmic distance ladder. Traditionally obtained from spectroscopy, photometric redshift estimation became a necessity in the recent years, in order to obtain distances of a larger sample of objects. This constitutes a major subject due to many forthcoming projects and missions, which will investigate several open questions in astronomy. Studies concerning galaxy formation and evolution, dark matter, dark energy, gravitational lensing, and many more, will be highly dependent on the availability of redshift estimates for a huge number of sources. Spectroscopy alone could never satisfy such a need. Photometric redshifts constitute a reasonable compromise in this sense, allowing us to obtain the required estimates at the cost of some precision.

On the other hand, the *Big Data* explosion experienced by astronomy, with the availability of many synoptic all-sky surveys, triggered an increasing interest in the community on this topic. The huge amount of photometric information available nowadays increased as well as the interest in an approach based on machine learning techniques. Machine learning, and in particular deep learning, became increasingly popular in the recent years in many fields of research, not just astronomy related. This is due to the big improvement in computational capabilities and the availability of hardware resources able to deal with the demands of such technologies. Nowadays the astronomical literature is full of machine learning applications, both for regression and classification problems, adopting supervised or unsupervised models. The benefit from the adoption of such methodologies is evident: the pipeline of the photometric redshift estimation process, depicted in Fig. 5.3, can be partially or totally automated. This potentially allows, once the chosen model is properly trained, to quickly obtain photometric redshifts for millions of sources.

The discussion in the community is nowadays focused on several topics concerning photometric redshift estimation. The most obvious subject is devolved to the improvement of the global quality of the predictions, reducing as much as possible the error between the estimated values and the spectroscopic counterparts. This aspect is based on the implementation of new methods and models, but also on different usage of the available data. However other themes are directly connected and involved in the discussion. It has been extensively discussed that photometric redshifts are often degenerate and can highly benefit from a description based on probability density functions and taking into account multimodalities. For this reason, the error treatment in presence of PDF estimates consitutes a major problem in the field. Another important topic, more related to the machine learning aspect, is the interpretability of the features used for the redshfit estimation, when different from plain magnitudes and colors and in particular when obtained by means of automated models.

The two models presented in the thesis, coupled with the statistical tools and techniques introduced, are meant to deal with these issues. The DCMDN model, based on the combination of a convolutional neural network with a mixture density network, is able to predict photometric redshifts directly from images, in the form of PDFs, fully probabilistic, multimodal and source independent. The model is fully automated, so features are extracted and selected automatically by the neural network. The introduction of the CRPS as a loss function consents to predict well calibrated and sharp PDFs. The advantage of such an approach has been already recognized by the community, as it was recently applied both in astronomy, for photometric redshift estimation, and in the weather forecast field too, by using the CRPS as a loss function for a neural network. However the model has two major disadvantages: it tends to be a black box, as the user can lose control of the internal behavior of the network, and it is hard to give an interpretation for the features automatically extracted by the convolutional part.

The problem of interpretability is well known to those who deal with deep learning applications. At the current stage, there are no substantial improvements in understanding the behavior of convolutional neural network based models and the features produced by them, despite several attempts involving visualization and statistical techniques. On the other hand, features based on known physical parameters and selected by means of a feature selection algorithm can be understood and interpreted, at least from a theoretical point of view. The second model presented, based on a massive feature generation and a selection performed adopting a forward selection model, is meant to improve the performance in the prediction of the photometric redshift estimates and, at the same time, to guarantee a certain degree of interpretability in the selected features. The improvement is given by the combination of the advantages obtained from a human-based knowledge together with the high amount of information brought in by the massive combination of parameters and the forward selection done on such a huge number of features. Some of the features obtained have been interpreted by finding a correspondence between the feature importance calculated with the RF and quasar emission lines. It has been proven that this correspondence is only partially captured by classic features, namely magnitudes and colors.

Therefore the two proposed methods allow to predict photometric redshifts improving the performances with respect to other methods and solving several issues which traditionally affect the models in the literature of the field. The comparison between the two models has shown that the feature-based one obtains superior performances with respect to the error function, while the analysis of the PIT proves that the DCMDN is able to predict better calibrated PDFs. An important part of the work is dedicated to correctly estimating the quality of the predictions, in particular when dealing with PDFs. In this sense, the CRPS and the PIT are fundamental tools in guaranteeing the prediction of well calibrated and sharp density functions. The importance of a probabilistic description for photometric redshifts and the necessity to take into account multimodalities have been also deeply discussed. This discussion is particular important due to upcoming missions like Euclid. In particular, concerning the Euclid mission, it has been demonstrated that the requirements established could potentially lead to misleading results.

This entire work has been accomplished by making intensive use of GPU computing, in order to parallelize the models adopted. Such technologies in the recent years allowed for a satisfying and efficient implementation of deep learning models. In fact, as the theory behind them was known already for several decades, only the increment in computational power of the last ten years allowed for the implementation of them for practical purposes. Nowadays, deep learning technologies, thanks to hardware improvement and data explosion, constitute the milestone for many applications which have great impact on our everyday life. Self-driving cars, real-time automated translation, and facial recognition are just few examples of the important role acquired by deep learning in our society. In this sense, astronomy is surely benefitting from the implementation of machine learning in solving a manifold of different problems. On the other hand, astronomy itself is a good playground to test such applications in a *safe* environment. Surely, in the field of the *X-informatics*, astronomy is between those scientific fields which are experiencing the most radical transformation in terms of the application of the *fourth paradigm*. The work presented in this thesis, apart the improvements and advantages that could bring in the field, constitutes also a good use case to show the power of machine learning application for astronomical problems. This is why the methods and the models presented have been developed in the most general way, in order to allow the community to apply them for different problems and research fields.

Let the data speak for themselves has been said. To do this we need instruments able to hear their voice. I would like to think that this could constitute a small step toward the revolution of this new astronomy.

Appendix

In the following I report the reprint permission granted from ESO for Publication II and Publication III.

Astronomy and Astrophysics
Editor in Chief: T. Forveille

T. Forveille
Astronomy & Astrophysics
Observatoire de Paris
61, avenue de l'Observatoire
75014 Paris, France

merging
Annales d'Astrophysique
Arkiv for Astronomi
Bulletin of the Astronomical Institutes of the Netherlands
Bulletin Astronomique
Journal des Observateurs
Zeitschrift fur Astrophysik
Bulletin of the Astronomical Institutes of Czechoslovakia

Tel.: 33 0(1) 43 29 05 41
Fax: 33 0(1) 43 29 05 57
e-mail: aanda.paris@obspm.fr
Web: http://www.aanda.org

Paris, July 5, 2018

Reprint Permission

Material:
Article by D'Isanto & Polsterer, 2018, A&A, 609, A111
Article by D'Isanto et al. 2018, A&A, in press, DOI 10.1051/0004-6361/201833103

To be used in:
PhD thesis, Heidelberg Institute for Theoretical Studies

Permission granted to:
Antonio D'Isanto
antonio.disanto@h-its.org

I hold copyright on the material referred to above, and hereby grant permission for its use as requested herewith.

The article should be reproduced in the same format as that published in A&A (for example, in an appendix). In particular, the present permission rules do not allow copy-and-pasting parts of the article into the main text of the thesis.

Credit should be given as follows:
Credit: Author, A&A, vol, page, year, reproduced with permission © ESO.

Thierry Forveille
A&A Editor-in-Chief

Sponsored by Argentina, Armenia, Austria, Belgium, Bulgaria, Chile, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, and Switzerland.
Produced and distributed by EDP Sciences for ESO.

Lists

List of Figures

List of Tables

Bibliography

K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, and et al. The Seventh Data Release of the Sloan Digital Sky Survey. , 182:543–558, June 2009. doi: 10.1088/0067-0049/182/2/543.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al. Adhikari. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, 116:061102, Feb 2016. doi: 10.1103/PhysRevLett.116.061102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.116.061102`.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and V. B. et al. Adya. A gravitational-wave standard siren measurement of the Hubble constant. , 551:85–88, November 2017. doi: 10.1038/nature24471.

T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, J. Annis, J. Asorey, S. Avila, O. Ballester, M. Banerji, and et al. Barkhouse. The Dark Energy Survey Data Release 1. *ArXiv e-prints*, January 2018.

F. B. Abdalla, A. Amara, P. Capak, E. S. Cypriano, O. Lahav, and J. Rhodes. Photometric redshifts for weak lensing tomography from space: the role of optical and near infrared photometry. , 387:969–986, July 2008. doi: 10.1111/j.1365-2966.2008.13151.x.

B. Abolfathi, D. S. Aguado, G. Aguilar, C. Allende Prieto, A. Almeida, T. Tasnim Ananna, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, and et al. The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. , 235:42, April 2018. doi: 10.3847/1538-4365/aa9e8a.

Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), pages 483–485, New York, NY, USA, 1967. ACM. doi: 10.1145/1465482.1465560. URL `http://doi.acm.org/10.1145/1465482.1465560`.

S. Andreon, G. Gargiulo, G. Longo, R. Tagliaferri, and N. Capuano. Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification. , 319:700–716, December 2000. doi: 10.1046/j.1365-8711.2000.03700.x.

S. Arnouts and O. Ilbert. LePHARE: Photometric Analysis for Redshift Estimate. Astrophysics Source Code Library, August 2011.

J. Asorey, M. Carrasco Kind, I. Sevilla-Noarbe, R. J. Brunner, and J. Thaler. Galaxy clustering with photometric surveys using PDF redshift information. , 459:1293–1309, June 2016. doi: 10.1093/mnras/stw721.

A. Avati, T. Duan, K. Jung, N. H. Shah, and A. Ng. Countdown Regression: Sharp and Calibrated Survival Predictions. *ArXiv e-prints*, June 2018.

N. M. Ball and R. J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19:1049–1106, 2010. doi: 10.1142/S0218271810017160.

N. M. Ball, R. J. Brunner, A. D. Myers, and D. Tcheng. Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. , 650:497–509, October 2006. doi: 10.1086/507440.

W. A. Baum. Photoelectric Magnitudes and Red-Shifts. In G. C. McVittie, editor, *Problems of Extra-Galactic Research*, volume 15 of *IAU Symposium*, page 390, 1962.

Richard E Bellman. *Adaptive control processes: a guided tour.* Princeton university press, 1961.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *ArXiv e-prints*, June 2012.

N. Benítez. Bayesian Photometric Redshift Estimation. , 536:571–583, June 2000. doi: 10.1086/308947.

N. Benítez, M. Moles, J. A. L. Aguerri, E. Alfaro, T. Broadhurst, J. Cabrera-Caño, F. J. Castander, J. Cepa, M. Cerviño, D. Cristóbal-Hornillos, A. Fernández-Soto, R. M. González Delgado, L. Infante, I. Márquez, V. J. Martínez, J. Masegosa, A. Del Olmo, J. Perea, F. Prada, J. M. Quintana, and S. F. Sánchez. Optimal Filter Systems for Photometric Redshift Estimation. , 692:L5–L8, February 2009. doi: 10.1088/0004-637X/692/1/L5.

Longo Tino Biehl, Bunte. Machine learning and data analysis in astroinformatics. In *ESANN*, 2018.

Christopher Bishop. Mixture density networks. 01 1994.

Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA, 1995. ISBN 0198538642.

M. Bolzonella, J.-M. Miralles, and R. Pelló. Photometric redshifts based on standard SED fitting procedures. , 363:476–492, November 2000.

R. Bordoloi, S. J. Lilly, and A. Amara. Photo-z performance for precision cosmology. , 406: 881–895, August 2010. doi: 10.1111/j.1365-2966.2010.16765.x.

W. J. Borucki, D. Koch, G. Basri, N. Batalha, T. Brown, D. Caldwell, J. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, A. K. Dupree, T. N. Gautier, J. C. Geary, R. Gilliland, A. Gould, S. B. Howell, J. M. Jenkins, Y. Kondo, D. W. Latham, G. W. Marcy, S. Meibom, H. Kjeldsen, J. J. Lissauer, D. G. Monet, D. Morrison, D. Sasselov, J. Tarter, A. Boss, D. Brownlee, T. Owen, D. Buzasi, D. Charbonneau, L. Doyle, J. Fortney, E. B. Ford, M. J. Holman, S. Seager, J. H. Steffen, W. F. Welsh, J. Rowe, H. Anderson, L. Buchhave, D. Ciardi, L. Walkowicz, W. Sherry, E. Horch, H. Isaacson, M. E. Everett, D. Fischer, G. Torres, J. A. Johnson, M. Endl, P. MacQueen, S. T. Bryson, J. Dotson, M. Haas, J. Kolodziejczak, J. Van Cleve, H. Chandrasekaran, J. D. Twicken, E. V. Quintana, B. D. Clarke, C. Allen, J. Li, H. Wu, P. Tenenbaum, E. Verner, F. Bruhweiler, J. Barnes, and A. Prsa. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327:977, February 2010. doi: 10.1126/science.1185402.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA, 1984.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

M. Brescia, S. Cavuoti, G. Longo, A. Nocella, M. Garofalo, F. Manna, F. Esposito, G. Albano, M. Guglielmo, G. D'Angelo, A. Di Guido, S. G. Djorgovski, C. Donalek, A. A. Mahabal, M. J. Graham, M. Fiore, and R. D'Abrusco. DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining. , 126:783, August 2014. doi: 10.1086/677725.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL `https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2`.

R. J. Brunner, A. J. Connolly, A. S. Szalay, and M. A. Bershady. Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. , 482:L21–L24, June 1997. doi: 10.1086/310674.

Robert J. Brunner, S. George Djorgovski, Thomas A. Prince, and Alex S. Szalay. *Massive Datasets in Astronomy*, pages 931–979. Springer US, Boston, MA, 2002. ISBN 978-1-4615-0005-6. doi: 10.1007/978-1-4615-0005-6_27. URL `https://doi.org/10.1007/978-1-4615-0005-6_27`.

G. Bruzual A. Spectral evolution of galaxies. I - Early-type systems. , 273:105–127, October 1983. doi: 10.1086/161352.

G. Bruzual A. and S. Charlot. Spectral evolution of stellar populations using isochrone synthesis. , 405:538–553, March 1993. doi: 10.1086/172385.

A. Buzzoni. Evolutionary population synthesis in stellar systems. I - A global approach. , 71: 817–869, December 1989. doi: 10.1086/191399.

N. Cappelluti, P. Predehl, H. Böhringer, H. Brunner, M. Brusa, V. Burwitz, E. Churazov, K. Dennerl, A. Finoguenov, M. Freyberg, P. Friedrich, G. Hasinger, E. Kenziorra, I. Kreykenbohm, G. Lamer, N. Meidinger, M. Mühlegger, M. Pavlinsky, J. Robrade, A. Santangelo, J. Schmitt, A. Schwope, M. Steinmitz, L. Strüder, R. Sunyaev, and C. Tenzer. eROSITA on SRG. A X-ray all-sky survey mission. *Memorie della Societa Astronomica Italiana Supplementi*, 17:159, 2011.

S. Carliles, T. Budavári, S. Heinis, C. Priebe, and A. S. Szalay. Random Forests for Photometric Redshifts. , 712:511–515, March 2010. doi: 10.1088/0004-637X/712/1/511.

M. Carrasco Kind and R. J. Brunner. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. , 432:1483–1501, June 2013. doi: 10.1093/mnras/stt574.

M. Carrasco Kind and R. J. Brunner. SOMz: photometric redshift PDFs with self-organizing maps and random atlas. , 438:3409–3421, March 2014. doi: 10.1093/mnras/stt2456.

S. Cavuoti, M. Brescia, G. Longo, and A. Mercurio. Photometric redshifts with the quasi Newton algorithm (MLPQNA) Results in the PHAT1 contest. , 546:A13, October 2012. doi: 10.1051/0004-6361/201219755.

S. Cavuoti, M. Brescia, V. De Stefano, and G. Longo. Photometric redshift estimation based on data mining with photoraptor. *Experimental Astronomy*, 39(1):45–71, 2015. ISSN 09226435. doi: 10.1007/s10686-015-9443-4. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-84945301895&doi=10.1007%2fs10686-015-9443-4&partnerID=40&md5=17c5a8e9a11300778e66692667f27faf`.

S. Cavuoti, M. Brescia, V. Amaro, C. Vellucci, G. Longo, and C. Tortora. Probability density estimation of photometric redshifts based on machine learning. *ArXiv e-prints*, June 2017.

S. Charlot and A. G. Bruzual. Stellar population synthesis revisited. , 367:126–140, January 1991. doi: 10.1086/169608.

J. Charlton and C. Churchill. *Quasistellar Objects: Intervening Absorption Lines*, page 2366. November 2000. doi: 10.1888/0333750888/2366.

J. Chevallard and S. Charlot. Modelling and interpreting spectral energy distributions of galaxies with BEAGLE. , 462:1415–1443, October 2016. doi: 10.1093/mnras/stw1756.

G. D. Coleman, C.-C. Wu, and D. W. Weedman. Colors and magnitudes predicted for high redshift galaxies. , 43:393–416, July 1980. doi: 10.1086/190674.

A. A. Collister and O. Lahav. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. , 116:345–351, April 2004. doi: 10.1086/383254.

A. J. Connolly, I. Csabai, A. S. Szalay, D. C. Koo, R. G. Kron, and J. A. Munn. Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. , 110:2655, December 1995. doi: 10.1086/117720.

J. Coupon, S. Arnouts, L. van Waerbeke, T. Moutard, O. Ilbert, E. van Uitert, T. Erben, B. Garilli, L. Guzzo, C. Heymans, H. Hildebrandt, H. Hoekstra, M. Kilbinger, T. Kitching, Y. Mellier, L. Miller, M. Scodeggio, C. Bonnett, E. Branchini, I. Davidzon, G. De Lucia, A. Fritz, L. Fu, P. Hudelot, M. J. Hudson, K. Kuijken, A. Leauthaud, O. Le Fèvre, H. J. McCracken, L. Moscardini, B. T. P. Rowe, T. Schrabback, E. Semboloni, and M. Velander. The galaxy-halo connection from a joint lensing, clustering and abundance analysis in the CFHTLenS/VIPERS field. , 449: 1352–1379, May 2015. doi: 10.1093/mnras/stv276.

R. D'Abrusco, A. Staiano, G. Longo, M. Brescia, M. Paolillo, E. De Filippis, and R. Tagliaferri. Mining the SDSS Archive. I. Photometric Redshifts in the Nearby Universe. , 663:752–764, July 2007. doi: 10.1086/518020.

K. S. Dawson, D. J. Schlegel, C. P. Ahn, S. F. Anderson, É. Aubourg, S. Bailey, R. H. Barkhouser, J. E. Bautista, A. Beifiori, A. A. Berlind, V. Bhardwaj, D. Bizyaev, C. H. Blake, M. R. Blanton, M. Blomqvist, A. S. Bolton, A. Borde, J. Bovy, W. N. Brandt, H. Brewington, J. Brinkmann, P. J. Brown, J. R. Brownstein, K. Bundy, N. G. Busca, W. Carithers, A. R. Carnero, M. A. Carr, Y. Chen, J. Comparat, N. Connolly, F. Cope, R. A. C. Croft, A. J. Cuesta, L. N. da Costa, J. R. A. Davenport, T. Delubac, R. de Putter, S. Dhital, A. Ealet, G. L. Ebelke, D. J. Eisenstein, S. Escoffier, X. Fan, N. Filiz Ak, H. Finley, A. Font-Ribera, R. Génova-Santos, J. E. Gunn, H. Guo, D. Haggard, P. B. Hall, J.-C. Hamilton, B. Harris, D. W. Harris, S. Ho, D. W. Hogg, D. Holder, K. Honscheid, J. Huehnerhoff, B. Jordan, W. P. Jordan, G. Kauffmann, E. A. Kazin, D. Kirkby, M. A. Klaene, J.-P. Kneib, J.-M. Le Goff, K.-G. Lee, D. C. Long, C. P. Loomis, B. Lundgren, R. H. Lupton, M. A. G. Maia, M. Makler, E. Malanushenko, V. Malanushenko, R. Mandelbaum, M. Manera, C. Maraston, D. Margala, K. L. Masters, C. K. McBride, P. McDonald, I. D. McGreer, R. G. McMahon, O. Mena, J. Miralda-Escudé, A. D. Montero-Dorta, F. Montesano, D. Muna, A. D. Myers, T. Naugle, R. C. Nichol, P. Noterdaeme, S. E. Nuza, M. D. Olmstead, A. Oravetz, D. J. Oravetz, R. Owen, N. Padmanabhan, N. Palanque-Delabrouille, K. Pan, J. K. Parejko, I. Pâris, W. J. Percival, I. Pérez-Fournon, I. Pérez-Ràfols, P. Petitjean, R. Pfaffenberger, J. Pforr, M. M. Pieri, F. Prada, A. M. Price-Whelan, M. J. Raddick, R. Rebolo, J. Rich, G. T. Richards, C. M. Rockosi, N. A. Roe, A. J. Ross, N. P. Ross, G. Rossi, J. A. Rubiño-Martin, L. Samushia, A. G. Sánchez, C. Sayres, S. J. Schmidt, D. P. Schneider, C. G. Scóccola, H.-J. Seo, A. Shelden, E. Sheldon, Y. Shen, Y. Shu, A. Slosar, S. A. Smee, S. A. Snedden, F. Stauffer, O. Steele, M. A. Strauss, A. Streblyanska, N. Suzuki, M. E. C. Swanson, T. Tal, M. Tanaka, D. Thomas, J. L. Tinker, R. Tojeiro, C. A. Tremonti, M. Vargas Magaña, L. Verde, M. Viel, D. A. Wake, M. Watson, B. A. Weaver, D. H. Weinberg, B. J. Weiner, A. A. West, M. White, W. M. Wood-Vasey, C. Yeche, I. Zehavi, G.-B. Zhao, and Z. Zheng. The Baryon Oscillation Spectroscopic Survey of SDSS-III. , 145:10, January 2013. doi: 10.1088/0004-6256/145/1/10.

Rina Dechter. Learning while searching in constraint-satisfaction-problems. In *AAAI*, pages 178–185, 01 1986.

M. Dickinson, M. Giavalisco, and GOODS Team. The Great Observatories Origins Deep Survey. In R. Bender and A. Renzini, editors, *The Mass of Galaxies at Low and High Redshift*, page 324, 2003. doi: 10.1007/10899892_78.

S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. , 450:1441–1459, June 2015. doi: 10.1093/mnras/stv632.

A. D'Isanto, S. Cavuoti, M. Brescia, C. Donalek, G. Longo, G. Riccio, and S. G. Djorgovski. An analysis of feature relevance in the classification of astronomical transients with machine learning methods. , 457:3119–3132, April 2016. doi: 10.1093/mnras/stw157.

S. G. Djorgovski, A. J. Drake, A. A. Mahabal, M. J. Graham, C. Donalek, R. Williams, E. C. Beshore, S. M. Larson, J. Prieto, M. Catelan, E. Christensen, and R. H. McNaught. The Catalina Real-Time Transient Survey (CRTS). *ArXiv e-prints*, February 2011.

C. Donalek, A. Arun Kumar, S. G. Djorgovski, A. A. Mahabal, M. J. Graham, T. J. Fuchs, M. J. Turmon, N. Sajeeth Philip, M. T.-C. Yang, and G. Longo. Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets. *ArXiv e-prints*, October 2013.

D. J. Eisenstein, J. Annis, J. E. Gunn, A. S. Szalay, A. J. Connolly, R. C. Nichol, N. A. Bahcall, M. Bernardi, S. Burles, F. J. Castander, M. Fukugita, D. W. Hogg, Ž. Ivezić, G. R. Knapp, R. H. Lupton, V. Narayanan, M. Postman, D. E. Reichart, M. Richmond, D. P. Schneider, D. J. Schlegel, M. A. Strauss, M. SubbaRao, D. L. Tucker, D. Vanden Berk, M. S. Vogeley, D. H. Weinberg, and B. Yanny. Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample. , 122:2267–2280, November 2001. doi: 10.1086/323717.

Pablo A. Estévez. Big data era challenges and opportunities in astronomy—how som/lvq and related learning methods can contribute? In Erzsébet Merényi, Michael J. Mendenhall, and

Patrick O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 267–267, Cham, 2016. Springer International Publishing. ISBN 978-3-319-28518-4.

A. Fernández-Soto, K. M. Lanzetta, and A. Yahil. A New Catalog of Photometric Redshifts in the Hubble Deep Field. , 513:34–50, March 1999. doi: 10.1086/306847.

A. Finoguenov, L. Guzzo, G. Hasinger, N. Z. Scoville, H. Aussel, H. Böhringer, M. Brusa, P. Capak, N. Cappelluti, A. Comastri, S. Giodini, R. E. Griffiths, C. Impey, A. M. Koekemoer, J.-P. Kneib, A. Leauthaud, O. Le Fèvre, S. Lilly, V. Mainieri, R. Massey, H. J. McCracken, B. Mobasher, T. Murayama, J. A. Peacock, I. Sakelliou, E. Schinnerer, J. D. Silverman, V. Smolčić, Y. Taniguchi, L. Tasca, J. E. Taylor, J. R. Trump, and G. Zamorani. The XMM-Newton Wide-Field Survey in the COSMOS Field: Statistical Properties of Clusters of Galaxies. , 172:182–195, September 2007. doi: 10.1086/516577.

M. Fioc and B. Rocca-Volmerange. PEGASE.2, a metallicity-consistent spectral evolution model of galaxies: the documentation and the code. *ArXiv Astrophysics e-prints*, December 1999.

A. E. Firth, O. Lahav, and R. S. Somerville. Estimating photometric redshifts with artificial neural networks. , 339:1195–1202, March 2003. doi: 10.1046/j.1365-8711.2003.06271.x.

Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California Univ Berkeley, 1951.

A. Fontana, S. D'Odorico, F. Poli, E. Giallongo, S. Arnouts, S. Cristiani, A. Moorwood, and P. Saracco. Photometric Redshifts and Selection of High-Redshift Galaxies in the NTT and Hubble Deep Fields. , 120:2206–2219, November 2000. doi: 10.1086/316803.

M. Gebetsberger, J. W. Messner, G. J. Mayr, and A. Zeileis. Estimation Methods for Non-Homogeneous Regression - Minimum CRPS vs Maximum Likelihood. In *EGU General Assembly Conference Abstracts*, volume 19 of *EGU General Assembly Conference Abstracts*, page 5573, April 2017.

D. W. Gerdes, A. J. Sypniewski, T. A. McKay, J. Hao, M. R. Weis, R. H. Wechsler, and M. T. Busha. ArborZ: Photometric Redshifts Using Boosted Decision Trees. , 715:823–832, June 2010. doi: 10.1088/0004-637X/715/2/823.

N. Gianniotis, S. D. Kügler, P. Tiňo, and K. L. Polsterer. Model-Coupled Autoencoder for Time Series Visualisation. *ArXiv e-prints*, January 2016.

Fabian Gieseke, Kai Lars Polsterer, Cosmin E. Oancea, and Christian Igel. Speedy greedy feature selection: Better redshift estimation via massive parallelism. In *ESANN*, 2014.

C. Gini. *Variabilità e mutabilità*. 1912.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. Technical report, Journal of the American Statistical Association, 2004.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld III, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: 10.1175/MWR2904.1. URL `https://doi.org/10.1175/MWR2904.1`.

Tilmann Gneiting, Balabdaoui Fadoua, and Raftery Adrian E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x`.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. de Haas, Ž. Ivezić, G. Knapp, R. Lupton, G. Pauls, R. Simcoe, R. Hirsch, D. Sanford, S. Wang, D. York, F. Harris, J. Annis, L. Bartozek, W. Boroski, J. Bakken, M. Haldeman, S. Kent, S. Holm, D. Holmgren, D. Petravick, A. Prosapio, R. Rechenmacher, M. Doi, M. Fukugita, K. Shimasaku, N. Okada, C. Hull,

W. Siegmund, E. Mannery, M. Blouke, D. Heidtman, D. Schneider, R. Lucinio, and J. Brinkman. The Sloan Digital Sky Survey Photometric Camera. , 116:3040–3081, December 1998. doi: 10.1086/300645.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944968.

J. Heaton. An Empirical Analysis of Feature Engineering for Predictive Modeling. *ArXiv e-prints*, January 2017.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. 15:559–570, 10 2000a.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000b. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, October 2009. ISBN 978-0-9825442-0-4. URL https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

H. Hildebrandt, S. Arnouts, P. Capak, L. A. Moustakas, C. Wolf, F. B. Abdalla, R. J. Assef, M. Banerji, N. Benítez, G. B. Brammer, T. Budavári, S. Carliles, D. Coe, T. Dahlen, R. Feldmann, D. Gerdes, B. Gillis, O. Ilbert, R. Kotulla, O. Lahav, I. H. Li, J.-M. Miralles, N. Purger, S. Schmidt, and J. Singal. PHAT: PHoto-z Accuracy Testing. , 523:A31, November 2010. doi: 10.1051/0004-6361/201014885.

S. Ho, A. Cuesta, H.-J. Seo, R. de Putter, A. J. Ross, M. White, N. Padmanabhan, S. Saito, D. J. Schlegel, E. Schlafly, U. Seljak, C. Hernández-Monteagudo, A. G. Sánchez, W. J. Percival, M. Blanton, R. Skibba, D. Schneider, B. Reid, O. Mena, M. Viel, D. J. Eisenstein, F. Prada, B. A. Weaver, N. Bahcall, D. Bizyaev, H. Brewinton, J. Brinkman, L. Nicolaci da Costa, J. R. Gott, E. Malanushenko, V. Malanushenko, B. Nichol, D. Oravetz, K. Pan, N. Palanque-Delabrouille, N. P. Ross, A. Simmons, F. de Simoni, S. Snedden, and C. Yeche. Clustering of Sloan Digital Sky Survey III Photometric Luminous Galaxies: The Measurement, Systematics, and Cosmological Implications. , 761:14, December 2012. doi: 10.1088/0004-637X/761/1/14.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9: 1735–80, 12 1997.

F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *ArXiv e-prints*, January 2018.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. ISSN 0027-8424. doi: 10.1073/pnas.79.8.2554. URL http://www.pnas.org/content/79/8/2554.

B. Hoyle. Measuring photometric redshifts using galaxy images and Deep Neural Networks. *Astronomy and Computing*, 16:34–40, July 2016. doi: 10.1016/j.ascom.2016.03.006.

B. Hoyle, M. M. Rau, K. Paech, C. Bonnett, S. Seitz, and J. Weller. Anomaly detection for machine learning redshifts applied to SDSS galaxies. , 452:4183–4194, October 2015a. doi: 10.1093/mnras/stv1551.

B. Hoyle, M. M. Rau, R. Zitlau, S. Seitz, and J. Weller. Feature importance for machine learning redshifts applied to SDSS galaxies. , 449:1275–1283, May 2015b. doi: 10.1093/mnras/stv373.

W. Hu. Power Spectrum Tomography with Weak Lensing. , 522:L21–L24, September 1999. doi: 10.1086/312210.

E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15:168–173, March 1929. doi: 10.1073/pnas.15.3.168.

Steffen Hölldobler, Sibylle Möhle, and Anna Tigunova. Lessons learned from alphago. In *Proceedings of the Second Young Scientist's International*, 06 2017.

O. Ilbert, S. Arnouts, H. J. McCracken, M. Bolzonella, E. Bertin, O. Le Fèvre, Y. Mellier, G. Zamorani, R. Pellò, A. Iovino, L. Tresse, V. Le Brun, D. Bottini, B. Garilli, D. Maccagni, J. P. Picat, R. Scaramella, M. Scodeggio, G. Vettolani, A. Zanichelli, C. Adami, S. Bardelli, A. Cappi, S. Charlot, P. Ciliegi, T. Contini, O. Cucciati, S. Foucaud, P. Franzetti, I. Gavignaud, L. Guzzo, B. Marano, C. Marinoni, A. Mazure, B. Meneux, R. Merighi, S. Paltani, A. Pollo, L. Pozzetti, M. Radovich, E. Zucca, M. Bondi, A. Bongiorno, G. Busarello, S. de La Torre, L. Gregorini, F. Lamareille, G. Mathez, P. Merluzzi, V. Ripepi, D. Rizzo, and D. Vergani. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. , 457:841–856, October 2006. doi: 10.1051/0004-6361:20065138.

K. Janocha and W. M. Czarnecki. On Loss Functions for Deep Neural Networks in Classification. *ArXiv e-prints*, February 2017.

M. J. Jee, J. A. Tyson, M. D. Schneider, D. Wittman, S. Schmidt, and S. Hilbert. Cosmic Shear Results from the Deep Lens Survey. I. Joint Constraints on Ω_M and σ_8 with a Two-dimensional Analysis. , 765:74, March 2013. doi: 10.1088/0004-637X/765/1/74.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, 09 1952. doi: 10.1214/aoms/1177729392. URL `https://doi.org/10.1214/aoms/1177729392`.

Will Knight. The dark secret at the heart of ai, 2018. URL `https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/`.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, Jan 1982. ISSN 1432-0770. doi: 10.1007/BF00337288. URL `https://doi.org/10.1007/BF00337288`.

D. C. Koo. Optical multicolors - A poor person's Z machine for galaxies. , 90:418–440, March 1985. doi: 10.1086/113748.

O. Lahav, A. Naim, L. Sodré, Jr., and M. C. Storrie-Lombardi. Neural computation as a tool for galaxy classification: methods and examples. , 283:207, November 1996. doi: 10.1093/mnras/283.1.207.

C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, I. Davidzon, P. Capak, G. Hasinger, J. D. Silverman, C. Pichon, J. Coupon, H. Aussel, D. Le Borgne, K. Caputi, P. Cassata, Y.-Y. Chang, F. Civano, J. Dunlop, J. Fynbo, J. S. Kartaltepe, A. Koekemoer, O. Le Fèvre, E. Le Floc'h, A. Leauthaud, S. Lilly, L. Lin, S. Marchesi, B. Milvang-Jensen, M. Salvato, D. B. Sanders, N. Scoville, V. Smolcic, M. Stockmann, Y. Taniguchi, L. Tasca, S. Toft, M. Vaccari, and J. Zabl. The COSMOS2015 Catalog: Exploring the 1 z 6 Universe with Half a Million Galaxies. , 224:24, June 2016. doi: 10.3847/0067-0049/224/2/24.

K. M. Lanzetta, A. Yahil, and A. Fernández-Soto. Star-forming galaxies at very high redshifts. , 381:759–763, June 1996. doi: 10.1038/381759a0.

R. Laureijs, J. Amiaux, S. Arduini, J. . Auguères, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, and et al. Euclid Definition Study Report. *ArXiv e-prints*, October 2011.

O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio. Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation. , 418:2165–2195, December 2011. doi: 10.1111/j.1365-2966.2011.19416.x.

O. Le Fèvre, G. Vettolani, B. Garilli, L. Tresse, D. Bottini, V. Le Brun, D. Maccagni, J. P. Picat, R. Scaramella, M. Scodeggio, A. Zanichelli, C. Adami, M. Arnaboldi, S. Arnouts, S. Bardelli, M. Bolzonella, A. Cappi, S. Charlot, P. Ciliegi, T. Contini, S. Foucaud, P. Franzetti, I. Gavignaud, L. Guzzo, O. Ilbert, A. Iovino, H. J. McCracken, B. Marano, C. Marinoni, G. Mathez, A. Mazure, B. Meneux, R. Merighi, S. Paltani, R. Pellò, A. Pollo, L. Pozzetti, M. Radovich, G. Zamorani, E. Zucca, M. Bondi, A. Bongiorno, G. Busarello, F. Lamareille, Y. Mellier, P. Merluzzi, V. Ripepi, and D. Rizzo. The VIMOS VLT deep survey. First epoch VVDS-deep survey: 11 564 spectra with $17.5 \leq IAB{\leq}24, andtheredshiftdistributionover0{\leq}z{\leq}5., 439: 845--862, September2005.doi: .$

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In Michael A. Arbib, editor, *Handbook of Brain Theory and Neural Networks*, page 3361. MIT Press, 1995.

Yann Lecun, Leon Bottou, Y Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. 86:2278 – 2324, 12 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Cheng-Yuan Liou, Jau-Chi Huang, and Wen-Chie Yang. Modeling word perception using the elman network. *Neurocomputing*, 71(16):3150 – 3157, 2008. ISSN 0925-2312. Advances in Neural Information Processing (ICONIP 2006) / Brazilian Symposium on Neural Networks (SBRN 2006).

E. D. Loh and E. J. Spillar. Photometric redshifts of galaxies. , 303:154–161, April 1986. 10.1086/164062.

LSST Science Collaborations, P. Marshall, T. Anguita, F. B. Bianco, E. C. Bellm, N. Brandt, W. Clarkson, A. Connolly, E. Gawiser, Z. Ivezic, L. Jones, M. Lochner, M. B. Lund, A. Mahabal, D. Nidever, K. Olsen, S. Ridgway, J. Rhodes, O. Shemmer, D. Trilling, K. Vivas, L. Walkowicz, B. Willman, P. Yoachim, S. Anderson, P. Antilogus, R. Angus, I. Arcavi, H. Awan, R. Biswas, K. J. Bell, D. Bennett, C. Britt, D. Buzasi, D. I. Casetti-Dinescu, L. Chomiuk, C. Claver, K. Cook, J. Davenport, V. Debattista, S. Digel, Z. Doctor, R. E. Firth, R. Foley, W.-f. Fong, L. Galbany, M. Giampapa, J. E. Gizis, M. L. Graham, C. Grillmair, P. Gris, Z. Haiman, P. Hartigan, S. Hawley, R. Hlozek, S. W. Jha, C. Johns-Krull, S. Kanbur, V. Kalogera, V. Kashyap, V. Kasliwal, R. Kessler, A. Kim, P. Kurczynski, O. Lahav, M. C. Liu, A. Malz, R. Margutti, T. Matheson, J. D. McEwen, P. McGehee, S. Meibom, J. Meyers, D. Monet, E. Neilsen, J. Newman, M. O'Dowd, H. V. Peiris, M. T. Penny, C. Peters, R. Poleski, K. Ponder, G. Richards, J. Rho, D. Rubin, S. Schmidt, R. L. Schuhmann, A. Shporer, C. Slater, N. Smith, M. Soares-Santos, K. Stassun, J. Strader, M. Strauss, R. Street, C. Stubbs, M. Sullivan, P. Szkody, V. Trimble, T. Tyson, M. de Val-Borro, S. Valenti, R. Wagoner, W. M. Wood-Vasey, and B. A. Zauderer. Science-Driven Optimization of the LSST Observing Strategy. *ArXiv e-prints*, August 2017. 10.5281/zenodo.842712. URL `https://doi.org/10.5281/zenodo.842712`.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL `https://projecteuclid.org/euclid.bsmsp/1200512992`.

P. Madau. Radiative transfer in a clumpy universe: The colors of high-redshift galaxies. , 441: 18–27, March 1995. 10.1086/175332.

A. Mahabal, K. Sheth, F. Gieseke, A. Pai, S. G. Djorgovski, A. Drake, M. Graham, and the CSS/CRTS/PTF Collaboration. Deep-Learnt Classification of Light Curves. *ArXiv e-prints*, September 2017.

R. Mandelbaum, U. Seljak, C. M. Hirata, S. Bardelli, M. Bolzonella, A. Bongiorno, M. Carollo, T. Contini, C. E. Cunha, B. Garilli, A. Iovino, P. Kampczyk, J.-P. Kneib, C. Knobel, D. C. Koo, F. Lamareille, O. Le Fèvre, J.-F. Le Borgne, S. J. Lilly, C. Maier, V. Mainieri, M. Mignoli, J. A. Newman, P. A. Oesch, E. Perez-Montero, E. Ricciardelli, M. Scodeggio, J. Silverman, and L. Tasca. Precision photometric redshift calibration for galaxy-galaxy weak lensing. , 386:781–806, May 2008. 10.1111/j.1365-2966.2008.12947.x.

C. Maraston. Evolutionary population synthesis: models, analysis of the ingredients and application to high-z galaxies. , 362:799–825, September 2005. 10.1111/j.1365-2966.2005.09270.x.

C. Maraston, G. Strömbäck, D. Thomas, D. A. Wake, and R. C. Nichol. Modelling the colour evolution of luminous red galaxies - improvements with empirical stellar spectra. , 394:L107–L111, March 2009. 10.1111/j.1745-3933.2009.00621.x.

P. Martí, R. Miquel, A. Bauer, and E. Gaztañaga. Photo-z quality cuts and their effect on the measured galaxy clustering. , 437:3490–3505, February 2014. 10.1093/mnras/stt2152.

Warren S. McCulloch and Walter Pitts. Neurocomputing: Foundations of research. chapter A Logical Calculus of the Ideas Immanent in Nervous Activity, pages 15–27. MIT Press, Cambridge, MA, USA, 1943. ISBN 0-262-01097-6. URL http://dl.acm.org/citation.cfm?id=65669.104377.

G Mclachlan and K Basford. *Mixture Models: Inference and Applications to Clustering*, volume 38. 01 1988. 10.2307/2348072.

T. Miyaji, G. Hasinger, M. Salvato, M. Brusa, N. Cappelluti, F. Civano, S. Puccetti, M. Elvis, H. Brunner, S. Fotopoulou, Y. Ueda, R. E. Griffiths, A. M. Koekemoer, M. Akiyama, A. Comastri, R. Gilli, G. Lanzuisi, A. Merloni, and C. Vignali. Detailed Shape and Evolutionary Behavior of the X-Ray Luminosity Function of Active Galactic Nuclei. , 804:104, May 2015. 10.1088/0004-637X/804/2/104.

A. D. Myers, M. White, and N. M. Ball. Incorporating photometric redshift probability density information into real-space clustering measurements. , 399:2279–2287, November 2009. 10.1111/j.1365-2966.2009.15432.x.

J. A. Newman, A. Abate, F. B. Abdalla, S. Allam, S. W. Allen, R. Ansari, S. Bailey, W. A. Barkhouse, T. C. Beers, M. R. Blanton, M. Brodwin, J. R. Brownstein, R. J. Brunner, M. Carrasco Kind, J. L. Cervantes-Cota, E. Cheu, N. E. Chisari, M. Colless, J. Comparat, J. Coupon, C. E. Cunha, A. de la Macorra, I. P. Dell'Antonio, B. L. Frye, E. J. Gawiser, N. Gehrels, K. Grady, A. Hagen, P. B. Hall, A. P. Hearin, H. Hildebrandt, C. M. Hirata, S. Ho, K. Honscheid, D. Huterer, Ž. Ivezić, J.-P. Kneib, J. W. Kruk, O. Lahav, R. Mandelbaum, J. L. Marshall, D. J. Matthews, B. Ménard, R. Miquel, M. Moniez, H. W. Moos, J. Moustakas, A. D. Myers, C. Papovich, J. A. Peacock, C. Park, M. Rahman, J. Rhodes, J.-S. Ricol, I. Sadeh, A. Slozar, S. J. Schmidt, D. K. Stern, J. Anthony Tyson, A. von der Linden, R. H. Wechsler, W. M. Wood-Vasey, and A. R. Zentner. Spectroscopic needs for imaging dark energy experiments. *Astroparticle Physics*, 63: 81–100, March 2015. 10.1016/j.astropartphys.2014.06.007.

E P. Grimit, T Gneiting, Veronica Berrocal, and N A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. 132: 2925 – 2942, 10 2006.

Tapio Pahikkala, Antti Airola, and Tapio Salakoski. Speeding up greedy forward selection for regularized least-squares. *2010 Ninth International Conference on Machine Learning and Applications*, pages 325–330, 2010.

J. Pasquet, E. Bertin, M. Treyer, S. Arnouts, and D. Fouchez. Photometric redshifts from SDSS images using a Convolutional Neural Network. *ArXiv e-prints*, June 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pixar. The renderman interface specification, Sep 1989. URL http://www.redrabbit-studios.com/coursework/renderman/prman/RISpec/index.html.

K. L. Polsterer. Dealing with Uncertain Multimodal Photometric Redshift Estimations. In M. Brescia, S. G. Djorgovski, E. D. Feigelson, G. Longo, and S. Cavuoti, editors, *Astroinformatics*, volume 325 of *IAU Symposium*, pages 156–165, June 2017. 10.1017/S1743921316013089.

K. L. Polsterer, P.-C. Zinn, and F. Gieseke. Finding new high-redshift quasars by asking the neighbours. , 428:226–235, January 2013. 10.1093/mnras/sts017.

K. L. Polsterer, F. Gieseke, C. Igel, and T. Goto. Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection. In N. Manset and P. Forshay, editors, *Astronomical Data Analysis Software and Systems XXIII*, volume 485 of *Astronomical Society of the Pacific Conference Series*, page 425, May 2014.

K. L. Polsterer, A. D'Isanto, and F. Gieseke. Uncertain Photometric Redshifts. *ArXiv e-prints*, August 2016.

Kai Polsterer, Fabian Gieseke, Christian Igel, Bernd Doser, and Nikolaos Gianniotis. Parallelized rotation and flipping invariant kohonen maps (pink) on gpus, 04 2016.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986. ISSN 1573-0565. 10.1007/BF00116251. URL `https://doi.org/10.1007/BF00116251`.

S. Rasp and S. Lerch. Neural networks for post-processing ensemble weather forecasts. *ArXiv e-prints*, May 2018.

B. A. Reid, W. J. Percival, D. J. Eisenstein, L. Verde, D. N. Spergel, R. A. Skibba, N. A. Bahcall, T. Budavari, J. A. Frieman, M. Fukugita, J. R. Gott, J. E. Gunn, Ž. Ivezić, G. R. Knapp, R. G. Kron, R. H. Lupton, T. A. McKay, A. Meiksin, R. C. Nichol, A. C. Pope, D. J. Schlegel, D. P. Schneider, C. Stoughton, M. A. Strauss, A. S. Szalay, M. Tegmark, M. S. Vogeley, D. H. Weinberg, D. G. York, and I. Zehavi. Cosmological constraints from the clustering of the Sloan Digital Sky Survey DR7 luminous red galaxies. , 404:60–85, May 2010. 10.1111/j.1365-2966.2010.16276.x.

G. T. Richards, A. D. Myers, A. G. Gray, R. N. Riegel, R. C. Nichol, R. J. Brunner, A. S. Szalay, D. P. Schneider, and S. F. Anderson. Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ˜1,000,000 Quasars from Data Release 6. , 180:67–83, January 2009. 10.1088/0067-0049/180/1/67.

J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. , 733:10, May 2011. 10.1088/0004-637X/733/1/10.

David P. Rodgers. Improvements in multiprocessor system design. *SIGARCH Comput. Archit. News*, 13(3):225–231, June 1985. ISSN 0163-5964. 10.1145/327070.327215. URL `http://doi.acm.org/10.1145/327070.327215`.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6. URL `http://dl.acm.org/citation.cfm?id=65669.104451`.

I. Sadeh, F. B. Abdalla, and O. Lahav. ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. , 128(10):104502, October 2016. 10.1088/1538-3873/128/968/104502.

M. Salvato, G. Hasinger, O. Ilbert, G. Zamorani, M. Brusa, N. Z. Scoville, A. Rau, P. Capak, S. Arnouts, H. Aussel, M. Bolzonella, A. Buongiorno, N. Cappelluti, K. Caputi, F. Civano, R. Cook, M. Elvis, R. Gilli, K. Jahnke, J. S. Kartaltepe, C. D. Impey, F. Lamareille, E. Le Floc'h, S. Lilly, V. Mainieri, P. McCarthy, H. McCracken, M. Mignoli, B. Mobasher, T. Murayama, S. Sasaki, D. B. Sanders, D. Schiminovich, Y. Shioya, P. Shopbell, J. Silverman, V. Smolčić, J. Surace, Y. Taniguchi, D. Thompson, J. R. Trump, M. Urry, and M. Zamojski. Photometric Redshift and Classification for the XMM-COSMOS Sources. , 690:1250–1263, January 2009. 10.1088/0004-637X/690/2/1250.

M. Salvato, O. Ilbert, G. Hasinger, A. Rau, F. Civano, G. Zamorani, M. Brusa, M. Elvis, C. Vignali, H. Aussel, A. Comastri, F. Fiore, E. Le Floc'h, V. Mainieri, S. Bardelli, M. Bolzonella, A. Bongiorno, P. Capak, K. Caputi, N. Cappelluti, C. M. Carollo, T. Contini, B. Garilli, A. Iovino, S. Fotopoulou, A. Fruscione, R. Gilli, C. Halliday, J.-P. Kneib, Y. Kakazu, J. S. Kartaltepe, A. M. Koekemoer, K. Kovac, Y. Ideue, H. Ikeda, C. D. Impey, O. Le Fevre, F. Lamareille, G. Lanzuisi, J.-F. Le Borgne, V. Le Brun, S. Lilly, C. Maier, S. Manohar, D. Masters, H. McCracken, H. Messias, M. Mignoli, B. Mobasher, T. Nagao, R. Pello, S. Puccetti, E. Perez-Montero, A. Renzini, M. Sargent, D. B. Sanders, M. Scodeggio, N. Scoville, P. Shopbell, J. Silvermann, Y. Taniguchi, L. Tasca, L. Tresse, J. R. Trump, and E. Zucca. Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples. , 742:61, December 2011. 10.1088/0004-637X/742/2/61.

M. Salvato, O. Ilbert, and B. Hoyle. The many flavours of photometric redshifts. *Nature Astronomy*, June 2018. 10.1038/s41550-018-0478-0.

M. J. Sawicki, H. Lin, and H. K. C. Yee. Evolution of the Galaxy Population Based on Photometric Redshifts in the Hubble Deep Field. , 113:1–12, January 1997. 10.1086/118231.

Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–61, Jun 1998. ISSN 1573-6938. 10.1023/A:1009957816843. URL `https://doi.org/10.1023/A:1009957816843`.

R. K. Sheth. On estimating redshift and luminosity distributions in photometric redshift surveys. , 378:709–715, June 2007. 10.1111/j.1365-2966.2007.11812.x.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 10 2017.

C. C. Steidel, M. Giavalisco, M. Pettini, M. Dickinson, and K. L. Adelberger. Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at Redshifts Z 3. , 462:L17, May 1996. 10.1086/310029.

C. C. Steidel, K. L. Adelberger, M. Dickinson, M. Giavalisco, and M. Pettini. Lyman Break Galaxies at z˜3 and Beyond. *ArXiv Astrophysics e-prints*, December 1998.

A. Szalay. Amdahl's Laws and Extreme Data-Intensive Scientific Computing. In I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots, editors, *Astronomical Data Analysis Software and Systems XX*, volume 442 of *Astronomical Society of the Pacific Conference Series*, page 405, July 2011.

Alexander Szalay and Jim Gray. Science in an exponential world. *Nature*, 440:413–414, 01 2006.

M. Tanaka. Photometric Redshift with Bayesian Priors on Physical Properties of Galaxies. , 801: 20, March 2015. 10.1088/0004-637X/801/1/20.

M. Tanaka, J. Coupon, B.-C. Hsieh, S. Mineo, A. J. Nishizawa, J. Speagle, H. Furusawa, S. Miyazaki, and H. Murayama. Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. , 70:S9, January 2018. 10.1093/pasj/psx077.

The Theano Development Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. Bleecher Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, A. de Brébisson, O. Breuleux, P.-L. Carrier, K. Cho, J. Chorowski, P. Christiano, T. Cooijmans, M.-A. Côté, M. Côté, A. Courville, Y. N. Dauphin, O. Delalleau, J. Demouth, G. Desjardins, S. Dieleman, L. Dinh, M. Ducoffe, V. Dumoulin, S. Ebrahimi Kahou, D. Erhan, Z. Fan, O. Firat, M. Germain, X. Glorot, I. Goodfellow, M. Graham, C. Gulcehre, P. Hamel, I. Harlouchet, J.-P. Heng, B. Hidasi, S. Honari, A. Jain, S. Jean, K. Jia, M. Korobov, V. Kulkarni, A. Lamb, P. Lamblin, E. Larsen, C. Laurent, S. Lee, S. Lefrancois, S. Lemieux, N. Léonard, Z. Lin, J. A. Livezey, C. Lorenz, J. Lowin, Q. Ma, P.-A. Manzagol, O. Mastropietro, R. T. McGibbon, R. Memisevic, B. van Merriënboer, V. Michalski, M. Mirza, A. Orlandi, C. Pal, R. Pascanu, M. Pezeshki, C. Raffel, D. Renshaw, M. Rocklin, A. Romero, M. Roth, P. Sadowski, J. Salvatier, F. Savard, J. Schlüter, J. Schulman, G. Schwartz, I. Vlad Serban, D. Serdyuk, S. Shabanian, É. Simon, S. Spieckermann, S. Ramana Subramanyam, J. Sygnowski, J. Tanguay, G. van Tulder, J. Turian, S. Urban, P. Vincent, F. Visin, H. de Vries, D. Warde-Farley, D. J. Webb, M. Willson, K. Xu, L. Xue, L. Yao, S. Zhang, and Y. Zhang. Theano: A Python framework for fast computation of mathematical expressions. *ArXiv e-prints*, May 2016.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL `http://arxiv.org/abs/1605.02688`.

Steve Upstill. *RenderMan Companion: A Programmer's Guide to Realistic Computer Graphics*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0201508680.

C. van Breukelen and L. Clewley. A reliable cluster detection technique using photometric redshifts: introducing the 2TecX algorithm. , 395:1845–1856, June 2009. 10.1111/j.1365-2966.2009.14692.x.

J.T. Vanderplas, A.J. Connolly, Ž. Ivezić, and A. Gray. Introduction to astroml: Machine learning for astrophysics. In *Conference on Intelligent Data Understanding (CIDU)*, pages 47 –54, oct. 2012. 10.1109/CIDU.2012.6382200.

E. Vanzella, S. Cristiani, A. Fontana, M. Nonino, S. Arnouts, E. Giallongo, A. Grazian, G. Fasano, P. Popesso, P. Saracco, and S. Zaggia. Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. , 423:761–776, August 2004. 10.1051/0004-6361:20040176.

A. Vazdekis, E. Casuso, R. F. Peletier, and J. E. Beckman. A New Chemo-evolutionary Population Synthesis Model for Early-Type Galaxies. I. Theoretical Basis. , 106:307, October 1996. 10.1086/192340.

Y. Wadadekar. Estimating Photometric Redshifts Using Support Vector Machines. , 117:79–85, January 2005. 10.1086/427710.

Y. Wang, N. Bahcall, and E. L. Turner. A Catalog of Color-based Redshift Estimates for Z ˜ 4 Galaxies in the Hubble Deep Field. , 116:2081–2085, November 1998. 10.1086/300592.

R. E. Williams, B. Blacker, M. Dickinson, W. V. D. Dixon, H. C. Ferguson, A. S. Fruchter, M. Giavalisco, R. L. Gilliland, I. Heyer, R. Katsanis, Z. Levay, R. A. Lucas, D. B. McElroy, L. Petro, M. Postman, H.-M. Adorf, and R. Hook. The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry. , 112:1335, October 1996. 10.1086/118105.

C. W. Yip, A. J. Connolly, A. S. Szalay, T. Budavári, M. SubbaRao, J. A. Frieman, R. C. Nichol, A. M. Hopkins, D. G. York, S. Okamura, J. Brinkmann, I. Csabai, A. R. Thakar, M. Fukugita, and Ž. Ivezić. Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey. , 128: 585–609, August 2004. 10.1086/422429.

Y. Zhang, Y. Zhao, and C. Cui. Data mining and knowledge discovery in database of astronomy. *Progress in Astronomy*, 20:312–323, December 2002.

Acknowledgements

I gratefully thank the Klaus Tschira Stiftung for the support.

This thesis was typeset using the LaTeX typesetting system originally developed by Leslie Lamport, based on TeX created by Donald Knuth.

Many thanks to Kai Polsterer for his constant support, for all the things he teached me and for his effort in trying to make me a true scientist, for having been, in these four long years a guide, a teacher, a friend.

Thanks also to Joachim Wambsganß and Coryn Bailer-Jones for all the useful suggestions and discussions about the thesis and my work.

I surely have to thank all my collegues for the constant support, for the everyday debates and for their contribution in creating such a pleasant working environment. In particular, many thanks to Nikos Gianniotis and Erica Hopkins for proofreading the thesis and to Isabel Lacurie for helping me with the translation of the German abstract.

I also want to gratefully thank the collegues who contributed to the publications, the already mentioned Kai Polsterer, Stefano Cavuoti and Fabian Gieseke.

Obviously I cannot forget my family. My father, for his dedication to me and for everything he has done and continues to do for me. My mother, for her never-ending love. I would not be half the man I am without them.

Finally, last but not least, I want to thank my wife Simona. It is not easy to share your life with a scientist, but she had the courage to follow me in this adventure. I would have never been able to arrive where I am without her.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 14.11.2018 ....