Empirical and spectral template based approaches in the analysis of galaxy data

PhD Dissertation

Róbert Beck

Eötvös Loránd University, Faculty of Science
Doctoral School of Physics, Particle Physics and Astronomy Program

Prof. Tamás Tél
Professor, DSc
Head of the Doctoral School

Prof. László Palla
Professor emeritus, DSc
Program Leader

Supervisors:

Dr. István Csabai
Professor, DSc

Dr. László Dobos
Assistant lecturer, PhD

Eötvös Loránd University, Faculty of Science
Department of Physics of Complex Systems

Budapest, 2017
## Contents

1 Introduction 9

2 Machine learning techniques 13
   2.1 Principal component analysis 14
   2.2 $k$-nearest neighbours 14
   2.3 $k$-means clustering 15
   2.4 Support vector machines 15

3 Measurements of galaxies 17
   3.1 Galaxy spectral energy distributions 17
   3.2 Redshift 19
   3.3 Spectroscopy of galaxies 21
   3.4 Photometry of galaxies 22
   3.5 The SDSS database 25

4 Modelling galaxy spectral energy distributions 27
   4.1 Stellar population synthesis models 27
   4.2 Ionization models for emission lines 28
   4.3 The BPT diagram 29
   4.4 Separation of emission lines and stellar continua 30
      4.4.1 Non-negative linear combination of continua 30
      4.4.2 Noise-limited fitting of emission lines 32
      4.4.3 Comparison with other work 32

5 Empirical estimation of emission lines 35
   5.1 Motivation 35
   5.2 Assembling the training set 37
5.2.1 Continuum principal components ................................................ 39
5.2.2 Emission line principal components .............................................. 42
5.3 Local linear regression method .......................................................... 43
  5.3.1 Estimation based on stellar continua ........................................... 44
  5.3.2 Estimation based on photometry .................................................. 48
  5.3.3 Validating the method ............................................................... 48
5.4 Classification of active galaxies ......................................................... 50
  5.4.1 SVM star-forming/AGN classification ......................................... 53
5.5 Stochastic recipe for generating lines ............................................... 55
  5.5.1 \(k\)-means clustering method .................................................. 55
  5.5.2 Modelling the emission line distributions .................................... 56
  5.5.3 Applying the recipe ............................................................... 56
  5.5.4 Shortcomings of the method .................................................... 58
5.6 Discussion of results ........................................................................... 59

6 Photometric redshift estimation ......................................................... 63
  6.1 Motivation ......................................................................................... 63
  6.2 Approaches to photometric redshift estimation ................................... 64
    6.2.1 The empirical approach ........................................................... 64
    6.2.2 The spectral template based approach ....................................... 66
    6.2.3 Difficulties in photometric redshift estimation ............................ 69
  6.3 Photometric redshifts for the SDSS DR12 ......................................... 70
    6.3.1 Estimation method ................................................................. 71
    6.3.2 Spectral template fitting ........................................................... 72
    6.3.3 Assembling the training set ...................................................... 73
    6.3.4 Cross-validation results ........................................................... 79
    6.3.5 Discussion of biases and errors ............................................... 80
    6.3.6 The redshift error map ............................................................ 83
    6.3.7 Best practices in using the database ......................................... 85
    6.3.8 Discussion of results ............................................................... 88
    6.3.9 The photometric redshift tables in SDSS DR12 ............................ 90
  6.4 Photo-z-SQL — redshift estimation in a database ............................... 94
    6.4.1 Estimation method ................................................................. 96
Acknowledgements

First, I would like to sincerely thank my supervisors — István Csabai for always providing deep physical insights, and showing unfettered enthusiasm towards fundamental problems, and László Dobos for his meticulous and detail-oriented work ethic in support of our research, propelling me to strive for excellence.

I would also like to express my gratitude towards Alexander S. Szalay for giving me the chance to study and work at Johns Hopkins University, a focal point of research in astronomy, and towards Tamás Budavári for providing guidance during my time there.

I would like to thank Emille E. O. Ishida, Rafael S. de Souza and Alberto Krone-Martins for organizing the Cosmostatistics Initiative (COIN) workshop, which provides a unique opportunity for astrostatisticians to meet and meaningfully collaborate.

I would like to thank my parents and family for enabling me to become the person that I am.

Finally, I would like to earnestly thank my wife, Adél, for always standing by my side, and always affectionately supporting me through whatever life brings us.

My research was supported by the New National Excellence Program of the Ministry of Human Capacities, the OTKA NN grant 103244, and the NKFI NN grant 114560 of Hungary.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean
Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.
Chapter 1

Introduction

Ever since the ancient Greeks, the study of celestial bodies, and more generally the behaviour and structure of the Universe has been a fascination of scientific-minded people. The early approach of by-eye observations and various related philosophical musings was dramatically transformed with the advent of telescopes in the time of Galilei and Kepler, marking the start of observational astronomy. Later, the work of Newton provided both the mathematical apparatus in the form of calculus, and the theoretical foundations in his laws of gravitation and motion, that enabled a significantly more sophisticated, predictive analysis of the objects in the sky.

Continual technological advancements yielded more and more data about increasingly faint and distant astronomical bodies. In the 20th century, progress became even quicker. Einstein’s general theory of relativity (GR) led to a deep understanding of how gravitation works. In the aftermath of the Shapley–Curtis debate, and thanks to Hubble’s observations, distant nebulae were correctly identified as galaxies far away from our own Galaxy, the Milky Way. Approximate solutions to GR coupled with measurements of the Hubble flow, i.e. the fact that more distant galaxies have larger recession velocities from us, induced the establishment of modern cosmology.

More recent technical innovations include computerization, and semiconductor-based CCD detectors that can collect and digitize incoming electromagnetic waves in an efficient way. These allow ground-based and satellite-mounted telescopes to conduct systematic surveys of the sky, collecting an unprecedented amount of data. Indeed, the expression “astronomical amount” did not gain its meaning by accident.

The deluge of data, both on Galactic and extragalactic sources, brought new challenges in terms of processing it efficiently, but also novel possibilities via statistical
analysis. As a result, astronomers and astrophysicists had to be quick in adopting modern computational and statistical tools, including parallel computation, multivariate methods, and machine learning techniques.

This dissertation is an account of my research on extragalaxies, using public data mainly from the Sloan Digital Sky Survey (SDSS). The work was performed at the Department of Physics of Complex Systems at Eötvös Loránd University, Budapest, Hungary, in part as a MSc student, but mostly as a PhD student, and also as a visiting researcher for a year at the Department of Physics and Astronomy at Johns Hopkins University, Baltimore, USA.

The main topics of the dissertation are a machine learning analysis of emission-line galaxies, and photometric redshift estimation. It is overwhelmingly based on four peer-reviewed publications, Beck et al. (2016a), Beck et al. (2016b), Beck et al. (2017a) and Beck et al. (2017b). The structure of the work is presented below.

In Chap. 2, I provide an introduction to machine learning methods, followed in Chap. 3 by a brief description of the background of galaxy measurements that I will be utilizing. Then, in Chap. 4 I present current approaches to modelling galaxies, and the method of galaxy continuum and line fitting devised and implemented by László Dobos, published in the initial part of Beck et al. (2016b). These could be considered introductory chapters, while subsequent chapters deal with my own research work.

Chap. 5 documents the results of Beck et al. (2016b), and details research on empirically estimating the emission line strengths of active galaxies. In Sect. 5.4, I describe a machine learning classification method that categorizes active galaxies into star-forming and active galactic nucleus host classes. Then, in Sect. 5.5, I introduce a stochastic recipe to generate a realistic distribution of emission lines for stellar population synthesis models. My contributions to the corresponding sections in Beck et al. (2016b) included compiling the sample, selecting and implementing the methods in C++ and R, creating figures, and writing the paper.

Chap. 6 is entirely dedicated to research on photometric redshift estimation.

After an introduction to the field, Sect. 6.3 describes how I created the photometric redshift catalogue of the SDSS Data Release 12 using an empirical machine learning method, published in Beck et al. (2016a). My related work included assembling the training set, significantly modifying the existing C++ code, creating figures, and writing the paper.
Then, in Sect. 6.4 I present a new spectral template fitting photometric redshift code, Photo-z-SQL, that can be integrated into a SQL database, and was documented in Beck et al. (2017a). My contributions to the paper included implementation, validation, plotting figures, and writing.

Finally, in Sect. 6.5 I describe parts of the publication Beck et al. (2017b) pertaining to my research work, including the assembly of a photometric redshift estimation testing catalogue, and results of validation tests related to my photometric redshift approaches. My contributions included the design and compilation of one of two datasets, the implementation and testing of two photometric redshift methods, and writing the paper.

The remaining chapters are comprised of short summaries in English and Hungarian, a list of abbreviations used in the dissertation, and the bibliography.

Throughout the dissertation, when describing collaborative efforts that involved the work or guidance of supervisors and colleagues, I appropriately write in first person plural, i.e. using “we” as opposed to “I”.
Chapter 2

Machine learning techniques

In this chapter, I provide an overview of the statistical and computational methods employed in modern machine learning, with an emphasis on techniques utilized in the dissertation. For a detailed introduction to the field as it relates to astronomy, refer to Ivezić et al. (2014).

Machine learning is an umbrella term, the meaning of which has been continuously evolving in recent years. In essence, it encompasses any statistical or algorithmic approach that allows datasets to be processed in order to arrive at meaningful characterizations or predictions regarding similar data. The datasets are generally comprised of a collection of features, which can be either categorical or continuous variables, belonging to a certain type of objects (e.g. astronomical sources).

There are two main approaches, unsupervised and supervised machine learning. The former attempts to find a useful representation of the input data without using any information external to the data itself, while in case of the latter the desired output of the algorithm is specified, i.e. a fraction of the input data is “labelled” with desired output features (the training set), and the goal is to accurately provide similar labels for unlabelled inputs. Depending on whether the labels in a supervised approach are categorical or continuous, the process can be referred to as a classification or regression task, respectively.

The following sections introduce machine learning methods used in the dissertation. However, there are numerous other sophisticated approaches for both supervised and unsupervised machine learning, such as self-organizing maps, artificial neural networks, random forests, decision trees, etc.
2.1 Principal component analysis

Principal component analysis (PCA) is a widely used multivariate statistical method that solves the eigendecomposition of the covariance matrix of a continuous data matrix, e.g. a collection of continuous features from different measurements. This is also referred to as singular value decomposition (SVD), and can be classified as unsupervised machine learning.

Mathematically, SVD corresponds to a rotation, transforming the initial features into an orthogonal basis that best represents the variance in the data. The so-called singular vectors (basis vectors) that have larger corresponding singular values explain larger proportions of the variance — thus, by only taking into account the first few dimensions in this basis, often a large fraction of the variance in the data can be explained using just a few numbers. PCA is therefore useful in dimensionality reduction.

2.2 $k$-nearest neighbours

The $k$-nearest neighbour method is a supervised machine learning technique. It requires a training set with continuous features (i.e. data vectors) and desired outputs, and attempts to find objects in this training set that are similar to a given unlabelled query point (with only features). Similarity is defined in terms of a distance metric in feature space, usually Euclidean distance. Once the neighbours are found, classification or regression can be performed by e.g. computing a point estimate from the outputs of the neighbours, or taking into account their entire distribution.

The main technical challenge of $k$-nearest neighbour methods is the requirement to efficiently locate the neighbours. Spatial indexing, most often a $kD$-tree index is used for this purpose (Csabai et al., 2007), which is a binary tree where each node represents a hyperplane cut in a given dimension, effectively segmenting the feature space.

The local linear regression method employed in Sect. 5.3 and Sect. 6.2.1 is a more advanced machine learning method which starts with a $k$-nearest neighbour search.
2.3 $k$-means clustering

$k$-means clustering is an unsupervised machine learning classification algorithm. In an astronomical context, it has been applied successfully, for instance, to classify gamma-ray bursts (Chattopadhyay et al., 2007; Veres et al., 2010). The $k$-means algorithm classifies data points based on their distances from cluster centres: points belonging to a cluster must be closer to the centre of mass of that particular cluster than that of any other clusters — essentially, the algorithm constructs a Voronoi tessellation from the data vectors, with seeds being the centres of mass of the clusters. This implicit definition of a cluster makes finding the best exact solution a hard problem, but heuristic, randomized algorithms exist that can find a reasonable clustering relatively fast (Forgy, 1965; MacQueen, 1967; Hartigan & Wong, 1979; Lloyd, 1982). The only inputs of $k$-means clustering are the data vectors and $k$, the number of clusters wanted. The output is the centres of mass of the $k$ clusters. Once the latter are known, new points can be classified simply by measuring their distances from the cluster centres and putting them into the one with the closest centre.

2.4 Support vector machines

A support vector machine (SVM) is a supervised machine learning algorithm that has been used for e.g. star–galaxy separation (Kovács & Szapudi, 2015) and transient detection (Wright et al., 2015). SVMs can be trained to automatically classify multi-dimensional data vectors into two disjoint sets (Vapnik & Vapnik, 1998; Karatzoglou, Meyer & Hornik, 2006). The training phase starts with compiling a training set of data vectors that are tagged as either belonging to class $A$ or class $B$. During learning, the model will find a hyperplane in the space of data vectors which separates the elements of $A$ and $B$ with the largest possible margin. Since this is often not possible in the original space of data vectors because the distributions of the two halves of the training set are non-convex, kernel functions are used to map training set vectors into a higher dimensional space where linear segregation is possible (Schölkopf et al., 2000). Another option to handle non-convex situations is to find a best possible segregation plane which minimizes the overlap. Once the model is trained, it can be used to classify any query point into one of the two classes.
With multiple desired output classes, a set of SVMs can be used, each performing a binary classification, i.e. in the given class, or not in the given class. Additionally, there are extensions to the algorithm that allow SVMs to be used for regression tasks as well.
Chapter 3

Measurements of galaxies

In this chapter I give a brief overview of the physical and observational background of my research on extragalaxies. The focus is on topics that are connected to the applications presented later in the dissertation, i.e. what the data that I subsequently use represent and how they were derived. Accordingly, emphasis is given to the details of the Sloan Digital Sky Survey (SDSS, York et al., 2000), since most of my work relied on data from the SDSS.

The physical basics that I present in this chapter can be found in introductory astronomy textbooks, such as Carroll & Ostlie (2007) or Marik (1989). I refrain from including numerous references to textbooks, so unless an explicit reference is provided to a given paper, the reader can find more information in the foundational books.

3.1 Galaxy spectral energy distributions

The spectral energy distribution (SED), or more concisely, spectrum of a galaxy is the distribution of energy in the form of electromagnetic radiation that is emitted by a galaxy. It can either be a function of the wavelength $\lambda$ or frequency $\nu$, since we can express the energy $E$ of a photon both as $E = h\nu$ and $E = hc/\lambda$, where $h$ is the Planck constant and $c$ is the speed of light in vacuum.

A spectrum is not usually specified directly in energy units, but instead in flux, or more precisely, spectral flux density. The flux is defined as the energy passing through a unit area ($1m^2$ in SI units) of detector, within an infinitesimal wavelength or frequency interval, in unit time (1s in SI units). This way, the area of the detector and the observation time is transformed out, and the integral of the spectrum over the entire
wavelength/frequency range gives the total radiative power per unit area.

The main components of a galaxy regarding radiation are stars, interstellar matter including dust and gas clouds, and potentially an active galactic nucleus (AGN).

The light we can observe from a star comes predominantly from the outer region of the star called the photosphere, which has a given temperature and material composition through which the light must pass, essentially leading to a modified black-body spectrum. The continuous black-body-like component, i.e. the stellar continuum, is superposed with relatively thin line-like features. These are usually absorption lines (but can be emission lines, depending on temperature), located at energies corresponding to certain energy transitions within the gas in the outer layers of the star. The gas particles can capture or emit photons at and near that energy when they jump to a higher excitation state or drop down to a lower one. The ratio of gas particles in given ionization states can be determined via the Saha equation.

The interstellar dust is mainly responsible for extinction, a wavelength-dependent continuous attenuation of electromagnetic energy, and also for re-emission in the near-infrared part of the spectrum. Extinction affects the blue, low-wavelength part of the spectrum more, an effect often referred to as reddening. Here it is important to note that the dust within our own Galaxy also causes so-called Galactic extinction, which is a local phenomenon due to where we are observing, not a feature of the extragalaxy. Dust measurements of the Galaxy, e.g. the dust map of Schlegel, Finkbeiner & Davis (1998), can be used to approximate this effect in a process termed dereddening.

The interstellar gas clouds (or nebulae) within a galaxy can, similarly to gas in the outer regions of a star, be heated and ionized by incident photons. As the gas returns to lower energy states, emission lines are radiated. However, contrary to the dense gas in stars, the low density of nebulae means that metastable states can exist for a long time, therefore emission lines corresponding to forbidden transitions are a major component of nebular emission.

An AGN is a region around the central supermassive black hole (SMBH) of a galaxy where the accretion of matter results in the rapid acceleration of charged particles, leading to a synchrotron-like emission with a continuous power-law spectrum. When this emission outshines that of other components, we refer to the source as a quasar, instead of a galaxy.

It should be noted that the properties of the aforementioned components are not
independent, instead generally they are highly correlated — in fact, the variance in
galaxy spectra is lower than in the spectra of individual stars, therefore the former can
be described by fewer numbers (Yip et al., 2004; McGurk, Kimball & Ivezić, 2010). As
a result, it is often useful to introduce broad categories to characterize galaxies.

The so-called early-type galaxies generally contain old, red stellar populations, little
to no star formation, and only small amounts of interstellar gas. Morphologically, these
are elliptical or lenticular galaxies, having larger mass, and a higher probability of
containing an AGN (Heckman & Best, 2014). On the other hand, the so-called late-
type galaxies usually have younger and bluer stellar populations with ongoing star
formation, a larger amount of interstellar gas, smaller mass and AGN probability, and
spiral or irregular morphology. There is also a less populous, transitional class between
these “red” and “blue” classes, the so-called green valley galaxies (Salim, 2014). Fig. 3.1
illustrates the spectra of different types of galaxies.

The evolution of galaxies is generally assumed to be driven by mergers, broadly
evolving from lower to higher masses, from “blue” to “red”, from late-type to early-type
(Conselice, 2014). Mergers and their effects are usually characterized by the amount
of gas in the two colliding galaxies (dry or wet, Lin et al., 2008), and the scale of the
collision (minor or major).

3.2 Redshift

The incident light from a galaxy can be observed by our ground-based or satellite-
mounted telescopes. However, there is an important distinction between rest-frame
and observed-frame spectra due to the expansion of the Universe. Since the scale factor
\( a(t) \) of spacetime itself increases over time \( t \), so does the wavelength of light traversing
it. Thus, a spectrum emitted in an earlier time \( t_{\text{then}} \) and observed by us in time \( t_{\text{now}} \)
will be shifted towards higher wavelengths, lower frequency, and lower energy. This is
the so-called cosmological redshift \( z \), as the red end of a spectrum refers to higher, and
the blue end refers to lower wavelengths, matching the relationship between the red
and blue optical colours.

The cosmological redshift is defined as \( 1 + z \equiv a(t_{\text{now}})/a(t_{\text{then}}) \), and the corre-
sponding measured wavelength becomes \( \lambda_{\text{now}} = \lambda_{\text{then}} a(t_{\text{now}})/a(t_{\text{then}}) = \lambda_{\text{then}} (1 + z) \).
Since \( a(t_{\text{now}}) > a(t_{\text{then}}) \), the redshift can only be a positive value (or approximately 0
Figure 3.1: The spectra of three different types of galaxies in the rest frame. Fluxes were normalized. A passive red galaxy is plotted in red, a red AGN-powered galaxy with strong [O\textsc{iii}] emission is shown in yellow, while a blue star-forming galaxy with a powerful H\alpha line (extending to 14 arbitrary units) is drawn in blue.
for very close sources).

For small redshifts, the connection between the radial recession velocity $v_\parallel$ of a galaxy and the redshift can be approximated by $z \approx v_\parallel/c$. For larger redshifts and recession velocities, relativistic formulae are required, and the cosmological model has to be taken into account.

Another important effect is the Doppler shift, caused by the so-called peculiar motion of celestial objects, i.e. the physical motion that is detached from the cosmological expansion. This effect is superposed with the cosmological redshift, and can lead to either a redshift or blueshift depending on whether the peculiar motion is directed away from or towards the observer, respectively. In the context of extragalaxies, the Doppler shift can lead to important changes within the spectrum, but is negligible compared to the cosmological redshift, with the exception of the closest, e.g. Local Group galaxies. Therefore, throughout the dissertation I will use the term redshift to refer to the cosmological redshift.

### 3.3 Spectroscopy of galaxies

When the light from a galaxy is directed to a so-called spectrograph, a high-resolution measurement of the spectrum can be made. However, to get spectra representative of the distant source, and comparable to other such measurements, appropriate calibrations need to be done to take into account e.g. the intervening matter in the path of the light.

For observatories in space, this could include gas and dust within our Galaxy, but for ground telescopes the atmosphere of the Earth also plays a major role through the seeing effect and attenuation. All these factors are dependent on the given line of sight, since both the amount of intervening Galactic matter, and the thickness and conditions of the airmass depend on the direction of observation.

Of course, the optical properties of the telescope itself also influence the results of measurements — this is a localized effect, though, and generally not direction-dependent.

The SDSS uses two spectrographs (refer to Smee et al., 2013, for technical details), both of which underwent an update for the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS). Each spectrograph has two CCD cameras, representing a red and blue
channel, with the light split at approximately 6,000 Å. Originally the (observed-frame) wavelength coverage was between 3,800−9,200 Å, which increased to 3,600−10,400 Å after the update. The spectral resolution, i.e. the number of wavelength bins in which the flux is measured, is between 1,560−2,270 for the blue, and 1,850−2,650 for the red channel. They are multi-object spectrographs, with up to 640 (or 1,000 since the update) optical fibers transmitting the light of separate objects, measuring all of them at the same time.

The aforementioned calibration issues are solved by the SDSS in the following way: first, to characterize the optical properties of the telescope itself, three types of lamps with known properties are used for illumination (for flat fields, blue and green emission, and red calibration, respectively); second, utilizing the fact that multiple objects are observed at the same time, each batch of observations includes a number of standard F stars with well-modelled spectra, allowing the calibration of direction- and time-dependent effects.

The result of these procedures is a collection of well-calibrated galaxy spectra with flux measurement points every 1−2 Å within the coverage. This means that emission and absorption lines are resolved, both narrow ones that are mainly widened by the ∼100 km/s collective motion of stars, and wide ones that are caused by the rapid motion of matter around a central SMBH, with velocities that can reach the order of ∼1000 km/s. The SDSS spectral classification pipeline (Bolton et al., 2012) fits and identifies the observable lines — by comparing measured line wavelengths to their laboratory counterparts, a spectroscopic redshift is determined, and also a spectral class is given based on the line properties.

### 3.4 Photometry of galaxies

While a spectroscopic measurement yields a detailed function of flux versus wavelength, considerable time may be required to collect enough photons, especially from faint sources. When minute details are not essential, but an overall view of the properties of an astronomical source is desired, a photometric observation could be a viable alternative.

A photometric observation entails sending the light from a source through a so-called photometric filter, which only transmits the light in a well-defined wavelength
Figure 3.2: The response function — which includes all optical elements — corresponding to the SDSS ugriz filters, from left to right.

range. There are narrow- (∼ 50 Å width), medium- (∼ 500 Å width) and broad-band (∼ 1000 Å width) photometric filters, each category representing a step further away from a spectroscopic-quality resolution. Because light is collected from a much wider wavelength range at the CCD, photometric observations require significantly less telescope time — and thus resources — than spectroscopic ones.

The SDSS photometric system used a set of 5 broad-band photometric filters, with a complete set in all 6 CCD camera columns, to systematically scan the sky (Fukugita et al., 1996; Gunn et al., 1998). The filters constitute the SDSS ugriz filter system, centered on 3,543, 4,770, 6,231, 7,625 and 9,134 Å, respectively, with an observed-frame coverage between 3,000 – 10,500 Å. Fig. 3.2 shows the response function of the SDSS filters as a function of wavelength.

Of course, the calibration issues discussed in Sect. 3.3 are also present in the case of photometric observations. The SDSS underwent several iterations of photometric calibration, with the latest, so-called Ubercalibration method described in Padmanabhan
et al. (2008). It involved the separation of absolute and relative calibration, using a smaller number of standard stars for the former and repeat observations of millions of stars for the latter component.

An additional difficulty regarding galaxies is the fact that they are extended sources, therefore their light is captured on multiple CCD pixels, in a pattern that does not match the point spread function (PSF) of the telescope. For this reason, the SDSS fits multiple brightness profiles (Stoughton et al., 2002), including an exponential profile, a de Vaucouleurs profile, and a best-fitting linear combination of the two. Flux measurements are provided separately for these various profiles, with the labels exp, dev, and cModel, respectively, and also there is a version labelled model that uses the same aperture size (i.e. model radius) in all bands, based on the better-fitting of the exp or dev models in the r-band.

Photometric measurements can be quoted using the measured flux directly, but often they are specified in magnitudes (or mags). There are different magnitude systems, including Vega and AB magnitudes (Oke & Gunn, 1983). Both are given relative to a standard source, with the star Vega defined to be 0 mag in all filter bands for the former, and a source with a 3,631 Jy flat spectral flux density defined to be 0 mag in all bands for the latter. The 3,631 Jy value is also referred to as the flux zeropoint. Thus, AB magnitudes are defined as

$$ m_{AB} = -\frac{5}{2} \log_{10} \left( \frac{F_B}{3631 \text{ Jy}} \right) = -\frac{5}{2} \log_{10} \left( \frac{F_B}{\text{erg s}^{-1} \text{cm}^{-2} \text{Hz}^{-1}} \right) - 48.6, \quad (3.1) $$

where $F_B$ is the spectral flux density in the given filter (or band), and the right-hand version uses cgs units. $F_B$ can be computed using the formula

$$ F_B = \frac{\int_0^\infty F(\nu) R(\nu) (h\nu)^{-1} d\nu}{\int_0^\infty R(\nu) (h\nu)^{-1} d\nu}, \quad (3.2) $$

where $F(\nu)$ is the frequency-dependent flux, $R(\nu)$ is the response function of all optical elements including the CCD and filter, and the $h\nu$ photon energy terms are needed because CCD detectors count photons. Therefore, given a known $F(\nu)$ spectrum and $R(\nu)$ response function, the AB photometric magnitude can be directly computed.

It should be noted that the $F(\nu)$ incident flux is in the observed frame, hinting at the crucial interplay between redshift and photometric magnitudes. At different
redshifts, different regions of the same source spectrum will fall into the coverage of the
given filters, leading to differing photometric measurements, and forming the basis of
photometric redshift estimation. Additionally, if the spectrum is assumed to be known,
magnitudes measured at a known redshift can be shifted to any other redshift, a process
referred to as $K$-correction.

The SDSS uses the asinh magnitude system (Lupton, Gunn & Szalay, 1999), which
is very similar to AB magnitudes in most cases, but behaves better for very faint objects,
even handling negative fluxes (which can occur after e.g. a background subtraction).

It is important to also mention the term photometric colour, which is defined as the
difference between magnitudes in two different bands. From the formula of magnitudes
it is straightforward to conclude that colours represent the logarithm of flux ratios
between two filters. Thus, colours do not depend on distance, and can be used to
characterize galaxies without absolute calibration. Usually, the ratio of adjacent bands
is chosen, therefore the SDSS measured four independent colours: $u - g$, $g - r$, $r - i$
and $i - z$. The model magnitudes of the SDSS are the most accurate for galaxy colours
due to the matched aperture, while cModel magnitudes are preferred for total flux
measurements of galaxies (Scranton et al., 2005).

### 3.5 The SDSS database

The SDSS database is made up of a number of data tables, incorporating photometric
and spectroscopic measurements, supplementary data, and data products from various
observing programs, of different types of sources. Physically, the database is hosted at
Johns Hopkins University, Baltimore, USA, on a cluster of server computers running
Microsoft SQL Server. Public access to the data is provided through SkyServer\(^1\) via a
SQL interface.

The SDSS photometric catalogue contains broad-band photometric measurements
of over 200 million galaxies, with astrometric data (i.e. position on the celestial sphere),
and $ugriz$ magnitudes corresponding to different brightness profile fits (see Sect. 3.4).
Dereddened magnitudes — i.e. corrected for Galactic extinction — are also provided,
based on the dust map of Schlegel, Finkbeiner & Davis (1998).

Additionally, the SDSS spectroscopic catalogue comprises spectra for over 2 million

\(^1\)http://skyserver.sdss.org/CasJobs/
galaxies, including results from continuum and line fits, spectral classification, and spectroscopic redshifts (see Sect. 3.3).

The SDSS is unique among photometric surveys in the relatively large number of spectroscopic follow-up observations. Upcoming photometric sky surveys such as the LSST (Ivezic et al., 2008) and Pan-STARRS (Tonry et al., 2012) will generate a deluge of photometric data, but currently there is no planned spectroscopic survey that could match their scale. Thus, methods that rely only on photometric data will grow in importance in the future.
Chapter 4

Modelling galaxy spectral energy distributions

In this chapter, I briefly introduce the current approaches in the literature regarding the modelling of galaxy SEDs — stellar population synthesis models for the stellar continuum, and photoionization models for the nebular emission lines. In relation to emission lines, I also mention a widely used galaxy diagnostic diagram, the Baldwin–Phillips–Terlevich (BPT) diagram. Then, in Sect. 4.4, I describe the stellar continuum and emission line fitting algorithm recently developed by László Dobos, which represented the initial part of our research published in Beck et al. (2016b).

4.1 Stellar population synthesis models

Thanks to the large amount of flux-calibrated optical galaxy spectra accumulated by the SDSS, the precision of galaxy spectrum modelling has been improved significantly during the last decade. Stellar population synthesis (SPS) models are the state of the art in galaxy modelling (Bruzual & Charlot, 2003; Maraston & Strömbäck, 2011; Vazdekis et al., 2012). They involve simulating the evolution of an entire population of stars, leading to model spectra that contain the stellar emission of the population (including absorption lines).

The starting point of a SPS model is a star formation history (SFH), which describes the total mass of newly born stars in star formation events at given times. In the simplest case only a single burst of star formation is assumed, i.e. a single stellar population. Then, an initial mass function (IMF) is adopted, which gives the mass
distribution of stars within a single burst of star formation, derived from observations of star-forming regions. The IMFs used in the literature are usually power-law formulae with different cutoffs below one Solar mass (Salpeter, 1955; Miller & Scalo, 1979; Kroupa, 2001; Chabrier, 2003). As the final step, the given population of stars with known masses (and other assumed properties, e.g. metallicity) can be evolved up to the age of the modelled galaxy with the help of stellar evolutionary tracks. These tracks are computed from stellar structure models, and fully describe the time evolution of stellar properties from a known state. The stellar continuum of the galaxy is simply given as the superposition of the spectra of the stars, convolved with a velocity dispersion.

Many software tools and libraries exist to generate realistic stellar continua from a prescribed star formation history and various libraries of single stellar population spectra with a wide range of metallicities and initial mass function choices (Fioc & Rocca-Volmerange, 1997; Bruzual & Charlot, 2003; Maraston & Strömbäck, 2011; Bressan et al., 2012; Vazdekis et al., 2012). Models have also been extended with descriptions of interstellar extinction, the ultraviolet–infrared (UV–IR) balance (Silva et al., 1998; Charlot & Fall, 2000; da Cunha et al., 2010) and the chemical evolution of the gas from which stars can form (Davé, Finlator & Oppenheimer, 2011).

4.2 Ionization models for emission lines

Emission lines of galaxy spectra carry a large amount of information about the abundance and ionization states of elements in the interstellar gas, therefore it is of value to have detailed models of the lines. Population synthesis models generally do not account for nebular emission.

The physical processes that govern the interaction between light and gas particles are well-understood from experiments performed on Earth, in laboratories or particle accelerators. Photoionization models (Stasińska, 1984; Ferland et al., 2013) rely on this knowledge to yield accurate line ratios for any primary radiation spectrum and gas composition by simulating the interactions that occur when photons pass through the medium. To make the problem tractable, significant simplifications are usually adopted, e.g. assuming spherically symmetric gas clouds, isotropic incident radiation, or an identical ionizing spectrum for all gas clouds. Additionally, such models introduce many free parameters to describe the composition of the nebulae, limiting their practical
usage in galaxy modelling.

4.3 The BPT diagram

The two primary radiation sources that ionize nebulae are young, hot, massive stars and active galactic nuclei (AGNs). Their different spectra (thermal and power law, respectively) cause different ionization states and ratios of the most common elements which, in turn, produce well measurable, strong, often broad emission lines: the Balmer series of hydrogen, [O\text{II}], [O\text{III}], [N\text{II}], [S\text{II}], etc.

Based on the ionization ratios of various elements, the source of primary radiation responsible for the excitation of the interstellar medium (ISM) can be characterized. The Baldwin–Phillips–Terlevich (BPT) diagram is defined as having the logarithm of the $\left[\text{N}\text{II}\right]/H\alpha$ line ratio on the $x$-axis, and the logarithm of the $\left[\text{O}\text{III}\right]/H\beta$ line ratio on the $y$-axis (Baldwin, Phillips & Terlevich, 1981). Kewley et al. (2001) found a theoretical maximum starburst line on this diagram, above which starburst galaxy models could not produce line ratios:

$$\log_{10}\left(\frac{[\text{O}\text{III}]}{H\beta}\right) < 0.61 \left[\log_{10}\left(\frac{[\text{N}\text{II}]}{H\alpha}\right) - 0.47\right]^{-1} + 1.19.$$  \hspace{1cm} (4.1)

Kauffmann et al. (2003a) later empirically refined this line, defining star-forming (SF) galaxies as falling below the empirical starburst line of

$$\log_{10}\left(\frac{[\text{O}\text{III}]}{H\beta}\right) < 0.61 \left[\log_{10}\left(\frac{[\text{N}\text{II}]}{H\alpha}\right) - 0.05\right]^{-1} + 1.3.$$  \hspace{1cm} (4.2)

According to the generally accepted classification, galaxies above the line of Eq. 4.1 are AGN hosts, the region between the two lines corresponds to composite galaxies (i.e. having line ratios characteristic of both the AGN and SF classes), and galaxies below the line of Eq. 4.2 are classified as star-forming. Refer to Sect. 5.2 and Fig. 5.1 for a BPT diagram of the emission line galaxy sample used in the dissertation.

Different versions of the BPT diagram exist, incorporating e.g. sulfur lines to subdivide the AGN class, but in the dissertation only the aforementioned, original line ratio diagram is used.
4.4 Separation of emission lines and stellar continua

Principal component analysis (PCA, see Sect. 2.1) is widely used to derive a representative basis from optical spectra of galaxies. When performing PCA on emission line galaxies, the eigenspectra are primarily sensitive to the variations in emission line strengths and only secondarily to continuum features (Connolly et al., 1995a; Yip et al., 2004). Obviously, the slope of the continuum is correlated with emission lines but the variance of the lines is bigger. To run PCA on the pure continua, one has to mask the regions of emission lines, or eliminate the lines completely by subtracting line models from the measured spectra. Line fits have to be precise enough so that the line-subtracted continua contain minimal residuals. We reprocessed the entire set of SDSS DR7 galaxy spectra according to these requirements with our own implementation of the algorithm detailed in this section.

4.4.1 Non-negative linear combination of continua

One frequent method of fitting continua in the optical band is to express the spectrum as a non-negative linear combination of template spectra (Tremonti et al., 2004) while also accounting for the intrinsic attenuation and velocity dispersion. Although more advanced, Bayesian and PCA-based methods exist (Kauffmann et al., 2003b; Chen et al., 2012) to derive physical properties from the continuum, as we were mainly interested in the emission lines, we retained the former technique for continuum subtraction. First, we corrected for galactic extinction, masked emission lines and fitted the continuum using the templates from Bruzual & Charlot (2003) by also fitting the velocity dispersion and intrinsic extinction in parallel. Intrinsic extinction was modelled following Charlot & Fall (2000). Metallicity was taken into account by fitting four sets of templates of differing metallicities and choosing the one with minimal reduced $\chi^2$. Thus, the fitted metallicity can take one of four values: $Z = 0.004, 0.008, 0.02$ or $0.05$. We did not take the nebular continuum emission into account, which, in the case of young starburst galaxies, can contribute a non-negligible flux to the near-infrared part of the spectrum (Leitherer & Heckman, 1995). Since the entire continuum was fitted with stellar templates only, we expect a slight overestimation of absorption lines, and therefore the overestimation of emission lines for starburst galaxies. On the other hand, within the wavelength coverage of SDSS spectroscopy, nebular continuum emission is significant.
Figure 4.1: Illustration of fitting the stellar continuum. In the top left panel the best non-negative least square fit from 10 Bruzual–Charlot templates is plotted, the residual is visible in the bottom left panel. The effect of the low-pass filter on the residual is drawn with a red curve in the bottom left panel; we subtract this curve from the noisy residual prior to fitting emission lines. The top right panel illustrates the best-fitting continuum model, corrected for discrepancies by adding back the low-pass-filtered residual to the stellar population synthesis spectrum. The top right panel shows the high-pass-filtered residual used for fitting the lines.

only in the case of stellar populations younger than 10 Myr or at very low metallicities of $Z \sim 0.0001$ (Mollá, García-Vargas & Bressan, 2009), and only about 0.5 per cent of our sample (see Sect. 5.2) potentially fall into this parameter range.

Due to discrepancies between continuum models and SDSS spectra (Maraston et al., 2009), the continuum-subtracted spectrum consists of three components: the emission lines, the noise and a slowly changing background that originates from the imperfect models. Since the emission lines and noise are high-frequency components, one can easily eliminate the background by a high-pass filter. For this purpose, we used a 50 Å wide rolling median filter. This was wide enough to leave broad AGN lines almost intact, yet remove any residuals of the incorrect background subtraction. Fig. 4.1 illustrates this procedure.
4.4.2 Noise-limited fitting of emission lines

Once the low-frequency background has been removed, lines are fitted using a technique we call noise-limited fitting. To precisely fit all strong emission lines, including those of active galaxies, we use three increasingly complex line models.

- A single Gaussian:
  \[ F(\lambda) = A \cdot e^{-\frac{(\lambda-\lambda_0)^2}{\sigma^2}} \]

- Two Gaussians centred on the same wavelength but with different variance
  \[ F(\lambda) = A \cdot e^{-\frac{(\lambda-\lambda_0)^2}{\sigma_a^2}} + B \cdot e^{-\frac{(\lambda-\lambda_0)^2}{\sigma_b^2}} \]

- Two Gaussians allowing for a small offset \( \Delta \lambda < 5 \, \text{Å} \) between the centres, different variance
  \[ F(\lambda) = A \cdot e^{-\frac{(\lambda-\lambda_a)^2}{\sigma_a^2}} + B \cdot e^{-\frac{(\lambda-\lambda_b)^2}{\sigma_b^2}} \]

While the first model is enough to fit emission lines with typical velocity dispersion, the second model is necessary for lines with broad wings and the third model for asymmetric lines. Our objective is to find the simplest, yet well-fitting model. Overlapping emission lines are — obviously — fitted together, but we do not enforce any correlation on the EWs of lines from the same ion. Also, the velocity dispersions of the lines, even of those from the same ion, are fitted independently. First, we fit the lines with the simplest model, subtract it from the measurement and compare the residual within the region of the emission line with the noise in wavelength ranges without lines. If the rms of the residual inside the region of the line is at least two times larger than elsewhere, we reject the model and attempt to fit the line with a more complex one. Fig 4.2 illustrates how this technique works on asymmetric broad AGN lines, and Tab. 4.1 summarizes the fitted and subtracted emission lines.

4.4.3 Comparison with other work

It is interesting to compare our line fits to those of Brinchmann et al. (2004). In the cited work, the authors used a simpler technique of fitting nebular emission lines of SDSS galaxies with the primary focus on the signal-to-noise ratio of line measurements and not on the minimization of the residuals after line subtraction. As a result, their...
Figure 4.2: Illustration of noise-limited fitting of asymmetric emission lines with increasingly complex models. The top panels show the original continuum-subtracted spectrum in grey and the best-fitting models in black. The bottom panels show the residuals. The left-hand panel corresponds to a single Gaussian fit, the middle panel to two Gaussians centred on the same mean wavelength but with different variance while the right-hand panel shows the results from fitting two Gaussians with slightly different centre wavelengths. In this case, the most complex model is accepted as the line residuals are higher than the average noise for both simpler models, whereas the line residual is comparable to the average noise in the third case.

<table>
<thead>
<tr>
<th>Line</th>
<th>$\lambda_{\text{vac}}$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O\textsc{ii}</td>
<td>3727.09</td>
</tr>
<tr>
<td>O\textsc{iii}</td>
<td>3729.88</td>
</tr>
<tr>
<td>H$\theta$</td>
<td>3798.98</td>
</tr>
<tr>
<td>H$\eta$</td>
<td>3836.47</td>
</tr>
<tr>
<td>H$\zeta$</td>
<td>3890.16</td>
</tr>
<tr>
<td>He\textsc{i}</td>
<td>3971.20</td>
</tr>
<tr>
<td>S\textsc{ii}</td>
<td>4072.30</td>
</tr>
<tr>
<td>H$\delta$</td>
<td>4102.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Line</th>
<th>$\lambda_{\text{vac}}$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H$\gamma$</td>
<td>4341.68</td>
</tr>
<tr>
<td>O\textsc{i}</td>
<td>6302.05</td>
</tr>
<tr>
<td>O\textsc{iii}</td>
<td>4364.44</td>
</tr>
<tr>
<td>N\textsc{i}</td>
<td>6365.54</td>
</tr>
<tr>
<td>O\textsc{iii}</td>
<td>4364.44</td>
</tr>
<tr>
<td>N\textsc{ii}</td>
<td>6529.03</td>
</tr>
<tr>
<td>H$\alpha$</td>
<td>6564.61</td>
</tr>
<tr>
<td>N\textsc{ii}</td>
<td>6549.86</td>
</tr>
<tr>
<td>N\textsc{ii}</td>
<td>6585.27</td>
</tr>
<tr>
<td>O\textsc{iii}</td>
<td>6635.54</td>
</tr>
<tr>
<td>S\textsc{ii}</td>
<td>6718.29</td>
</tr>
<tr>
<td>H$\delta$</td>
<td>6732.67</td>
</tr>
<tr>
<td></td>
<td>6302.05</td>
</tr>
</tbody>
</table>

Table 4.1: List of the fitted nebular emission lines. Wavelengths are quoted in vacuum.
In Fig. 4.3, we compare the EWs of the most prominent emission lines as derived with our technique and with the method of Brinchmann et al. (2004). In the case of strong emission lines, our measurements of line strengths are very similar to the results of Brinchmann et al. (2004), but we estimate weak emission lines significantly higher. This is very likely due to the high-pass filtering applied to the continuum-subtracted spectrum, cf. Sect. 4.4.1. Yet, between 97.9% and 99.1% of our line EWs are within 3σ of Brinchmann et al. (2004), with the exception of Hα and Hβ where only 93.9% and 87.7%, respectively, of the measurements are within 3σ. Also, Brinchmann et al. (2004) measured weak lines by fitting them together with stronger lines of the same ion, imposing a constraint on line ratios, whereas we fitted these lines independently. Weaker lines can easily become undetectable in noisy regions, hence our fitting method introduces some selection bias.
Chapter 5

Empirical estimation of emission lines

In this chapter I describe research on empirically quantifying correlations between properties of the stellar continuum of galaxy spectra, and the strengths of emission lines, published in Beck et al. (2016b). A training set-based method is developed in Sect. 5.3 to empirically estimate emission lines in the optical regime, based on either stellar continua taken from population synthesis models, or photometric magnitudes. Then, in Sect. 5.4, building on the observed correlations, a supervised machine learning method is employed to automatically classify active galaxies. Finally, in Sect. 5.5 I introduce a stochastic recipe for generating a realistic distribution of emission lines for stellar continua.

5.1 Motivation

Stellar population synthesis models are very successful in explaining the spectral energy distribution of galaxies in the optical (Fioc & Rocca-Volmerange, 1997; Bruzual & Charlot, 2003; Maraston & Strömbäck, 2011; Vazdekis et al., 2012) but they do not account for the characteristic emission lines originating from the excited interstellar gas. Starburst galaxies and galaxies with an active nucleus can produce emission lines so strong that can reach 60 per cent of the continuum flux in certain bands, or as much as 1 mag (Atek et al., 2011). As a result, pure population synthesis models are not enough to account for observations made with broad-band photometric filters. Since future large sky surveys will make photometric observations only, accurate modelling of the
emission lines will be essential to estimate physical properties (including photometric redshifts) of galaxies precisely.

To couple stellar population synthesis with models of photoionization, shock-heating of the interstellar gas, emission of the dust etc., the star formation history, several interactions between the stellar populations, the active nucleus, the dust and gas content need to be accounted for. For instance, AGNs are very likely to be responsible for quenching rapid star formation following starburst periods in the galaxy but they also emit ionizing radiation that excites gas, evaporates dust and produces shock waves that heat the ISM. Also, starburst periods are followed by high supernova activity that enriches the ISM with metals, leading to significant chemical evolution which must be reflected in the models of emission lines. Additionally, a recent advancement in stellar population synthesis is the inclusion of stellar rotation and binary evolution effects which have been shown to noticeably influence the strengths of some emission lines (Eldridge & Stanway, 2012; Stanway et al., 2014; Leitherer et al., 2014; Topping & Shull, 2015). Taking everything into account is not possible without detailed hydrodynamic simulation of the galaxies (Jonsson, Groves & Cox, 2010; Kewley et al., 2013) or without making significant simplifications to the models. Various software, notably PÉGASE and BPASS (Fioc & Rocca-Volmerange, 1997; Eldridge & Stanway, 2012), can be used to generate emission lines on top of stellar continua computed from stellar population synthesis. The photoionization part of these software tools, however, introduces a large set of free parameters that describe the distribution and composition of the ISM. A frequently used way of reducing the number of free parameters is to make theoretical or empirical assumptions. Typical theoretical simplifications include the assumption of spherical symmetry or the use of a common ionizing spectrum for all gas clouds (Stasińska, 1984; Fioc & Rocca-Volmerange, 1997; Ferland et al., 2013). If no strict physical considerations can be made, to generate realistic emission lines on top of modelled continua using any photoionization code, one has to estimate the a priori distribution of model parameters by comparing large ensembles of models with observations. For instance, the code Le Phare (Ilbert et al., 2006a) uses the relations of Kennicutt (1998) to parametrize emission lines.

Another route to take to generate realistic emission lines is to work on an entirely empirical basis. Yip et al. (2004) demonstrated that stellar continua of SDSS galaxies form a 1D sequence and thus, can be characterized by a single numerical value, the
**eclass.** The value of eclass for each galaxy spectrum is obtained by expressing the continuum on a basis derived from principal component analysis (PCA, refer to Sect. 2.1). Győry et al. (2011) showed that strong correlations between the eclass (i.e. the stellar continua) of starburst and AGN host galaxies exist. They applied PCA to expand emission line equivalent widths (EWs) of SDSS galaxies on a 3D basis and correlate the principal components with the eclass of the continua. We take their approach a step further: based on the correlations, we give a recipe to automatically generate emission lines with realistic distributions.

### 5.2 Assembling the training set

We started with the entire spectroscopic galaxy sample of the Sloan Digital Sky Survey Data Release 7 (SDSS DR7, Abazajian et al., 2009). As one of our goals was to accurately fit broad AGN lines, we measured line parameters ourselves following the method described in Sect. 4.4.

We selected a smaller sample of $N = 13788$ galaxies from the entire set of continuum and line-fitted spectra that met the following signal-to-noise ratio and line strength criteria:

- observed at a signal-to-noise ratio $S/N > 5$,
- all 11 emission lines listed in Tab. 5.1 are measured and non-zero. These lines are the same as in Győry et al. (2011).

The sample size was further limited to an easily manageable number by choosing a section of the sky (right ascension between $220^\circ$ and $230^\circ$). The Baldwin–Phillips–Terlevich (BPT) diagram of our sample is shown in Fig. 5.1, illustrating that it contains numerous examples of starburst, star-forming, composite and AGN host galaxies (see also Sect. 4.3). The colour coding of Fig. 5.1 is used throughout this chapter.

The requirement that all 11 emission lines should be measurable results in a sample containing galaxies with ongoing star formation or possessing an active nucleus only. Fig. 5.2 shows the selection effects on the distribution of the apparent and absolute $r$-band magnitudes, the redshift and the metallicity. While the cut in signal-to-noise ratio did produce a cutoff around $r = 19$ apparent magnitude and a relative increase
Figure 5.1: The BPT diagram of strong emission line galaxies sampled from SDSS DR7. Galaxies are colour coded based on their loci in the BPT plane: blue galaxies are star-forming, green ones are AGNs and red ones are the intermediate weak SF or weak AGNs in the bottom corner of the distribution. This colour coding based on directly measured emission lines is used in all BPT diagrams and estimation scatterplots throughout this chapter. The dashed curve shows the empirical segregation line between star-forming galaxies and AGNs/composites as defined by Kauffmann et al. (2003a). The solid curve represents the maximum starburst line of Kewley et al. (2001). Refer to Sect. 4.3 for more details regarding the BPT.
of objects towards smaller redshifts, the absolute magnitude distribution shows that our selection method prefers fainter, smaller and younger galaxies. Galaxies with lower metallicities are also selected with higher probability, presumably due to correlations between ongoing strong star formation, metallicity and age. Nevertheless, galaxies with solar and above solar metallicities are still present in the sample.

The SDSS DR7 main galaxy sample, which makes up the majority of our training set, was not selected for morphological type or colour, and thus includes a wide variety of galaxies (Strauss et al., 2002). At larger redshifts, however, different environments, e.g. harder radiation fields and higher ionization parameters (Steidel et al., 2014), can lead to significantly different emission line characteristics. Certainly, the validity of our results is constrained by parameter ranges covered by the sample. By using a data set that goes beyond the types of galaxies observed by the SDSS, one can easily apply our method to a broader range of galaxies.

5.2.1 Continuum principal components

Principal components (see Sect. 2.1) of the stellar continuum were derived from the fitted model spectra instead of the measurements directly. Although the precise line modelling would make it possible to subtract emission lines from the original spectra or run PCA directly on the measurements by masking out emission lines, due to the limited size of the sample which would make eigenspectra noisy, we choose to use the models instead. Fitted continuum models were taken at rest frame, convolved with the best-fitting velocity dispersion kernel and normalized to have equal flux in the following featureless rest-frame wavelength ranges: 4250–4300 Å, 4600–4800 Å, 5400–5500 Å, 5600–5800 Å. PCA was done in the 3722–6761 Å range with 0.6 Å binning. The average continuum was subtracted from the individual spectra prior to calculating the covariance matrix.

Eigenspectra were determined using the Lánczos singular value decomposition (SVD) algorithm from PROPACK (Larsen, 2005). The algorithm calculates only a given number of singular vectors with the largest corresponding singular values. This was very useful in our case as the spectra consisted of 5065 data points, whereas we were interested in the first five principal components only.

The average spectrum and the resulting eigenspectra are plotted in Fig. 5.3. As
Figure 5.2: Normalized histograms of galaxy properties. Our sample is plotted in blue, the grey lines correspond to the entire DR7 spectroscopic galaxy sample, while the black lines show the DR7 sample excluding the deeper LRG sub-sample. The latter provides a better comparison to our data set, since we selected predominantly from the main galaxy sample.
Figure 5.3: The average and the first five eigenvectors of the principal component analysis of galaxy continua, ordered by the corresponding singular values (as displayed in each panel). See the text for the physical interpretation of the eigenspectra.

the average was subtracted, the first eigenspectrum corresponds to the colour of the galaxy. The following two basis vectors are rather similar at first sight but the third one shows more prominent absorption lines. They are together very likely to determine the age and metallicity of the galaxy as the 4000 Å-break is very strong in both of them. The fourth vector probably corresponds to the width of absorption lines thus correlates with velocity dispersion. The magnitude of the fourth and fifth eigenvalues is similar, and they already mark the start of the plateau in the distribution of eigenvalues, therefore taking more eigenspectra into account does not significantly increase the variance explained by them.
Figure 5.4: The first four singular vectors of the correlation matrix of the logarithm of emission line equivalent widths, ordered decreasingly by the corresponding singular values (as displayed in each panel). The fourth vector is very likely to be just noise as [OII] lines should not have different signs.

5.2.2 Emission line principal components

In contrast to what was done by Győry et al. (2011), we calculate principal components (see Sect. 2.1) of the logarithm of emission line EWs. Fig. 5.4 shows the resulting singular vectors. Taking the logarithm is more useful when one is interested in line ratios instead of absolute line strengths and uses linear methods for the analysis. We have to mention, however, that using the logarithm of the EWs also means that the results presented in the rest of the paper will be valid in the logarithmic sense only.
5.3 Local linear regression method

Our goal was to empirically estimate emission line EWs from continuum principal components. If there exists any correlation between the continuum and emission lines of galaxy spectra, it is clearly non-linear. Global linear methods to analyse the correlations are not useful in this case, yet locally linear methods still can be used. Local linear regression using nearest neighbours (Csabai et al., 2007; Kerekes et al., 2013) has been used for physical parameter estimation based on broad-band photometry.

Let us consider an ensemble of measurements where measured values are split into two sets $D = \{d_i\}$ and $R = \{r_i\}$, $i$ indexing the individual measurements. For the sake of simplicity, $r_i$ are taken to be scalars whereas $d_i$ are vectors, thus $D$ forms a metric space of dimension $N$. The Euclidean metric is often used to measure distances among data vectors of $D$ even though it might lack any physical interpretation. Our objective is to characterize known, or predict unknown $r_i$ from the always known $d_i$ vectors. To estimate $r_i$ from $d_i$, first we find the $k$-nearest neighbours of $d_i$ in $D$ (see also Sect. 2.2). Let us denote the set of indices of these nearest neighbours with $NN(d_i, D, k)$, where $i \notin NN$ by definition. Then we express $r_i$ in the following form

$$r_i \approx c_i + a_i d_i. \quad (5.1)$$

Note that both $a_i$ and $d_i$ are vectors and their dot product is taken in the formula above. The $c_i$ constants and the $a_i$ coefficients need to be determined individually for every $(d_i, r_i)$ measurement using standard linear regression by minimizing

$$\chi_i^2 = \sum_{j \in NN} \frac{(r_j - c_i - a_i d_j)^2}{w_j}, \quad (5.2)$$

where $i$ is still the index of the measurement, $j$ runs on the nearest neighbours and $w_j$ is a weight. The expression of $\chi^2$ is similar if $r_i$ are vectors instead of scalars but the $a_i$ coefficients become matrices. Errors in $r_i$ and the components of $d_i$ can be incorporated into the value of $w_i$. Similarly, neighbours in $NN$ can be ordered by distance from $d_i$ and the inverse of (the square of) the distance can be used as a weight in Eq. 5.2.

Local linear regression has many advantages over global non-linear modelling. First of all, global models are usually either too simple to describe the data or prone to overfitting. Local linear models, on the other hand, are simple and can be used to characterize the local estimation errors. For instance, one can measure the goodness of the estimation of $r_i$ by the $\chi^2$ of the local linear fit.
5.3.1 Estimation based on stellar continua

We applied the local linear regression technique to estimate emission line EWs from the stellar continuum principal components. To test whether continuum PCs carry more information regarding the emission lines than broad-band SDSS magnitudes, we will also perform the regression analysis directly on the photometric magnitudes in Sect. 5.3.2. Further tests are done with randomized samples (Sect. 5.3.3) to get a picture of the performance of our method.

By using the notation of Sect. 5.3, \( \mathbf{d}_i \) became the first five continuum principal components and \( r_i \) became the log EWs. Emission line log EWs were fitted individually based on the log EWs of the \( k = 30 \) nearest neighbour galaxies in the continuum PCA space. The \( \chi^2 \) of the fitting was weighted by the inverse-square distance of the neighbours from the query point. The value of \( k = 30 \) was chosen as a rule of thumb: we are fitting \( 5 + 1 \) parameters and the number of data points must be large enough to adequately determine that many parameters but small enough to preserve locality. Modifying this parameter within reasonable limits (e.g. \( 25 - 40 \)) does not significantly impact the results.

Fig. 5.5 shows the reconstructed log EWs of emission lines as functions of the directly measured EWs. EWs reconstructed from the continuum are in reasonably good agreement with directly measured log EWs. The relative flux error \( \sigma_r \) of the line reconstruction is Gaussian but a systematic shift \( \delta \) is visible in the case of [O\text{II}], [O\text{III}] and [N\text{II}] (\( \delta \approx 0.1, 0.15, 0.15 \), respectively). The typical value of the relative error is \( \sigma_r \approx 0.3 \) for hydrogen and sulfur, \( \sigma_r \approx 35\% \) for [O\text{II}] and [N\text{II}], and \( \sigma_r \approx 45\% \) for [O\text{III}]. Col. 3-4 of Tab. 5.1 list the outcome of the correlation analysis for the 11 investigated lines using the local linear regression technique. Pearson’s product-moment correlation coefficient \( \rho \) and the rms error \( \sigma \) were calculated for each line. These numbers also show that fits are most accurate for the hydrogen and sulfur lines (\( \rho > 0.8 \)) whereas oxygen and nitrogen lines are significantly less correlated with direct line measurements.

Fig. 5.6 shows the dependence of the error of the reconstruction on galaxy properties (cf. Fig. 5.2 for histograms of these) for select lines. The brighter and higher metallicity galaxies have a larger fraction of AGNs and are estimated with higher errors, especially in the case of oxygen and sulfur lines. Objects at higher redshifts generally exhibit decreasing accuracy. The error of line reconstruction visibly increases towards the limits of our training set. This is due to the fact that near the edges of the training set
Figure 5.5: Reconstructed log EWs from continuum principal components. Estimated log EWs are plotted as functions of the directly measured log EWs for the 11 emission lines we used. Colour coding of data points is the same as in Fig. 5.1 and reflects the activity class of galaxies.
Figure 5.6: The rms error of emission line log EW reconstruction as the function of various galaxy properties. The colours correspond to the following emission lines: red – [OIII] 5008 Å, black – [NII] 6585 Å, grey – [SII] 6718 Å, and blue – Hα. Wavelengths are quoted in vacuum. See the text for a discussion.

there are fewer galaxies and the nearest neighbours used to estimate the emission lines are generally less similar to each other and to the galaxy whose lines are being fitted.

As we expected, emission lines can be much better reconstructed from the continuum of star-forming galaxies due to the strong connection between the young stellar population and the ISM: young massive stars are responsible for the excitation of interstellar gas clouds. Nevertheless, [OIII] is an important indicator of nuclear activity, its reconstruction from continuum properties in case of AGNs seems rather problematic. It is understandable as AGN activity correlates much less with the properties of the stellar populations than in the star-forming case. Yet, some connections exist as it is visible from [NII] and the hydrogen lines.

One intriguing result is that, while [OIII] is an important indicator of nuclear activity,
Table 5.1: Numerical properties of the various line reconstruction methods, for all 11 emission lines. The four methods are as follows: (1) from continuum principal components, fitting the 30 nearest neighbours, (2) from broad-band magnitudes, fitting the 30 nearest neighbours, (3) from continuum principal components, but instead of using the 30 nearest neighbours we used 30 random galaxies, and (4) from continuum principal components, but with a randomized sample (as a cross-test). For each reconstruction, we calculated the Pearson product-moment correlation coefficient \( \rho \) and rms error \( \sigma \). Emission lines are ordered by reconstructability using the first method. Wavelengths are quoted in vacuum.

<table>
<thead>
<tr>
<th>line</th>
<th>( \lambda [\text{Å}] )</th>
<th>( \rho )</th>
<th>( \sigma )</th>
<th>( \rho )</th>
<th>( \sigma )</th>
<th>( \rho )</th>
<th>( \sigma )</th>
<th>( \rho \times 10^2 )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{H}\alpha )</td>
<td>6565</td>
<td>0.898</td>
<td>0.388</td>
<td>0.842</td>
<td>0.481</td>
<td>0.803</td>
<td>0.561</td>
<td>-1.46</td>
<td>0.961</td>
</tr>
<tr>
<td>( \text{H}\beta )</td>
<td>4863</td>
<td>0.882</td>
<td>0.369</td>
<td>0.840</td>
<td>0.430</td>
<td>0.795</td>
<td>0.535</td>
<td>-1.52</td>
<td>0.854</td>
</tr>
<tr>
<td>( \text{Si}\text{ii} )</td>
<td>6718</td>
<td>0.839</td>
<td>0.416</td>
<td>0.773</td>
<td>0.492</td>
<td>0.798</td>
<td>0.484</td>
<td>-1.23</td>
<td>0.832</td>
</tr>
<tr>
<td>( \text{Si}\text{ii} )</td>
<td>6733</td>
<td>0.827</td>
<td>0.433</td>
<td>0.751</td>
<td>0.516</td>
<td>0.754</td>
<td>0.527</td>
<td>-1.36</td>
<td>0.840</td>
</tr>
<tr>
<td>( \text{H}\gamma )</td>
<td>4342</td>
<td>0.816</td>
<td>0.418</td>
<td>0.773</td>
<td>0.465</td>
<td>0.698</td>
<td>0.772</td>
<td>-2.81</td>
<td>0.790</td>
</tr>
<tr>
<td>( \text{O}\text{iii})</td>
<td>3727</td>
<td>0.749</td>
<td>0.498</td>
<td>0.700</td>
<td>0.547</td>
<td>0.556</td>
<td>0.716</td>
<td>-0.710</td>
<td>0.817</td>
</tr>
<tr>
<td>( \text{O}\text{ii})</td>
<td>5008</td>
<td>0.743</td>
<td>0.784</td>
<td>0.673</td>
<td>0.884</td>
<td>0.673</td>
<td>0.877</td>
<td>-1.10</td>
<td>1.268</td>
</tr>
<tr>
<td>( \text{O}\text{ii})</td>
<td>4960</td>
<td>0.721</td>
<td>0.773</td>
<td>0.659</td>
<td>0.858</td>
<td>0.628</td>
<td>0.890</td>
<td>-1.35</td>
<td>1.208</td>
</tr>
<tr>
<td>( \text{N}\text{ii})</td>
<td>6585</td>
<td>0.680</td>
<td>0.514</td>
<td>0.677</td>
<td>0.527</td>
<td>0.411</td>
<td>0.815</td>
<td>-0.318</td>
<td>0.757</td>
</tr>
<tr>
<td>( \text{N}\text{ii})</td>
<td>6550</td>
<td>0.664</td>
<td>0.570</td>
<td>0.669</td>
<td>0.579</td>
<td>0.367</td>
<td>0.880</td>
<td>0.306</td>
<td>0.820</td>
</tr>
</tbody>
</table>
5.3.2 Estimation based on photometry

To see if continuum PCA is any better than directly estimating emission lines from broad-band magnitudes, we performed the above analysis using the SDSS photometric magnitudes instead of the principal components. For this purpose, we used dereddened model magnitudes without any $K$-correction (see Sect. 3.4 and Sect. 3.5 for more details). The lack of $K$-correction is not supposed to significantly affect the procedure as the redshift distribution of the sampled galaxies is rather sharp.

While broad-band magnitudes are strongly correlated with continuum principal components, it is still interesting to see how lines are reconstructed from them. First of all, magnitudes are highly correlated with each other, whereas PCA eliminates covariance. Also, observed magnitudes are already “contaminated” with emission lines which might result in stronger correlations with EWs. Results of line reconstruction from magnitudes are plotted in Fig. 5.7. Compared with line reconstruction from PCA as plotted in Fig. 5.5, no clear difference can be seen in terms of scatter, perhaps with the exception of more outliers being visible in the photometric case. Thus, log EWs can be reconstructed from magnitudes almost as well as from the principal components. Quantitative results are listed in Col. 5-6 of Tab. 5.1. We have to emphasize here that our sample contained strong emission line galaxies only, thus the strong correlation between magnitudes and log EWs exists only for our sample and cannot be generalized to all galaxies.

5.3.3 Validating the method

To test whether a single global linear model is sufficient to reproduce the lines, we repeated the procedure of estimating log EWs from the continuum principal components as described in Sect. 5.3.1 but instead of using the 30 nearest neighbour galaxies, we randomly selected 30 galaxies from the entire sample. Another difference was that the $\chi^2$ of the fit was not weighted by the inverse-square of the distance from the query point to relax the effect of locality. By looking at Col. 7-8 of Tab. 5.1, it is somewhat surprising that correlation coefficients and rms errors of the individual lines did not get much worse. By looking at panel (c) of Fig. 5.8, one can clearly see, however, that the star-forming branch of the BPT diagram cannot be reconstructed in this way, and the AGN sequence is also greatly distorted. The conclusion is that emission line log EWs
Figure 5.7: Reconstructed log EWs from broad-band SDSS magnitudes. Estimated log EWs are plotted as functions of the directly measured log EWs for the 11 emission lines we used. Colour coding of data points is the same as in Fig. 5.1 and reflects the activity class of galaxies.
cannot be explained by a simple, global linear relationship with continuum principal components. Thus, local fitting from nearest neighbours is necessary to reconstruct the BPT from either continuum principal components or broad-band magnitudes.

As another test, we shuffled the sample and randomly paired continuum principal components with emission line vectors of other galaxies to break the locality in the PCA space. Results are listed in Col. 9-10 of Tab. 5.1. It is not surprising that correlations almost entirely disappear in the shuffled case (note that $\rho$ values are multiplied by $10^2$). This supports that information about the emission lines is indeed encoded in the continuum of a galaxy spectrum.

As the BPT diagram is generally used to classify emission line galaxies into star-forming and AGN, it is very informative to see how well the various methods can recover it solely from the continuum. The difference among the line estimation methods introduced above is obvious once the BPT diagram is plotted from the reconstructed log EWs, as it was done in Fig. 5.8. Local linear fitting of EWs using the nearest neighbours and reconstructing lines from broad-band magnitudes give similarly fair, qualitatively correct BPT diagrams while the randomization of the sample disrupts the diagram entirely.

To further analyse the properties of a reconstructed BPT diagram, we will stick to local linear regression based on the continuum principal components. In Fig. 5.9, we plot the original BPT for reference, the reconstructed diagram for all galaxies, and two diagrams showing AGNs and star-forming galaxies only, as classified by Kauffmann et al. (2003a).

The first thing to see in panels (c) and (d) of Fig. 5.9 is the mixing of weak star-forming galaxies with weak AGNs in the bottom corner of the reconstructed BPT diagram. The mixing is caused by the bad reconstructability of the $\text{[O}^{\text{III}}\text{]}$ line which is most likely due to the lack of a strong correlation between AGN activity level and the stellar continuum.

### 5.4 Classification of active galaxies

In Sect. 5.3, the non-linear correlations between the stellar continua and emission lines of galaxies were utilized to estimate the lines. Existing classifications of active galaxies (Kewley et al., 2001; Kauffmann et al., 2003a) are defined based on the flux ratios of
Figure 5.8: BPT diagrams with log EWs reconstructed from the stellar continuum using different methods and cross-tests. The colour coding of the data points is based on the original BPT as in Fig. 5.1. In each panel, the dashed curve shows the empirical segregation line between star-forming galaxies and AGNs/composites as defined by Kauffmann et al. (2003a). Panel (a) shows reconstructed log EWs from continuum principal components using local linear regression from the 30 nearest neighbours in PCA space. Panel (b) is the reconstruction of lines from broad-band magnitudes by local linear regression from the 30 nearest neighbours. Panel (c) was drawn from lines estimated using the global linear regression technique from 30 randomly selected galaxies. Panel (d) is the cross-test using local linear regression but with shuffled continuum principal components. See the text for a discussion.
Figure 5.9: Original and reconstructed BPT diagrams of our emission line galaxy sample. The colour coding of the data points is based on the original BPT as in Fig. 5.1. In each panel, the dashed curve shows the empirical segregation line between star-forming galaxies and AGNs/composites as defined by Kauffmann et al. (2003a). Panel (a) is plotted from directly measured line EWs. Panel (b) displays the BPT of line log EWs reconstructed from continuum principal components using the local linear regression method with the 30 nearest neighbours in PCA space. Panels (c) and (d) show only galaxies that were originally classified as (c) AGNs, (d) star-forming using directly measured line EWs. While lines of strong AGNs and extreme starburst galaxies can be reconstructed well, there is significant “cross-talk” in the quiescent region.
certain emission lines, but without any dependence on the shape of the continuum. It is therefore an interesting question whether it is possible to refine the star-forming–AGN separation by incorporating information on the stellar continuum into the classification model.

The mixed, low activity–low star formation rate region is located at the bottom corner of the [N\text{II}]/H\alpha–[O\text{III}]/H\beta BPT diagram. Empirically drawn BPT diagrams are noisy enough to smear pure star-forming galaxies and mixed star-forming/AGNs together in this part of the BPT, hence it is worth investigating whether the continuum can help differentiate the classes. By visually inspecting the projections of the 5D continuum PCA space, one can see that while AGNs and star-forming galaxies occupy different loci, they cannot be clearly separated into two disjoint sets by cuts in any principal component dimensions, nor is the distribution of galaxies bimodal. Thus, a more sophisticated classification method is required.

We turned to a machine learning algorithm, a support vector machine (SVM, see Sect. 2.4), to determine an empirical segregation plane between the two classes in the continuum PCA–line PCA space.

### 5.4.1 SVM star-forming/AGN classification

We compiled a training set from our emission line galaxy sample (as described in Sect. 5.2) by selecting galaxies on the BPT that could be classified with high confidence either as pure star-forming or AGN host. To select high-confidence AGNs only, we picked galaxies above the theoretical maximum starburst line of Eq. 4.1 (Kewley et al., 2001). Star-forming galaxies, on the other hand, were selected to fall below the empirical starburst line of Eq. 4.2 (Kauffmann et al., 2003a), and at the same time above the following line, defined by us:

\[
\log_{10} \left( \frac{[O\text{III}]}{H\beta} \right) > 3 \log_{10} \left( \frac{[N\text{II}]}{H\alpha} \right) + 1.55.
\]

The line was drawn empirically to cut out the most reliably identifiable part of the star-forming population. Curves on Fig. 5.10 illustrate these cuts.

We used the first five continuum principal components and the first four log EW principal components of the training set galaxies as input data vectors to the SVM (refer to Sect. 2.4) implementation of R. By combining information from the continuum into
Figure 5.10: BPT diagrams resulting from the SVM-based star-forming/AGN separation. Panels (a) and (b) show star-forming and AGN host galaxies used to train the algorithm. Panels (c) and (d) display the outcome of the automatic classification. The theoretical segregation line of Kewley et al. (2001) is drawn in blue and the empirical one of Kauffmann et al. (2003a) in red. Our star-forming/low-activity line is in black. Colour coding of the data points is based on the original BPT as in Fig. 5.1. See the text for a discussion.

The training, we might hope for a better separation of the two galaxy types in the mixed lower corner of the BPT than simply from the emission lines. As the SVM is a strictly empirical model, we shall not, however, draw far-reaching theoretical conclusions from its outcome. Since our training set was not containing the mixed region, it was directly separable into two disjoint classes by a linear cut. Consequently, data points did not need to be projected into any higher dimensional space by a kernel function — like in most applications of SVMs, we simply ran it on the 5 + 4 dimensional vectors of the continuum + line PCA space.

We plot the results of the SVM-based classification in Fig. 5.10. Panels (a) and (b)
show the two training set classes (star-forming and AGN, respectively) with log EWs of the originally measured emission lines. Panels (c) and (d) show the outcome of the SVM classification. Even though the mixed region was not included in the training set at all, the SVM reproduced the empirical segregation line of Kauffmann et al. (2003a) surprisingly well, with only about 6 per cent of the sample scattered into the opposite region.

### 5.5 Stochastic recipe for generating lines

Based on our findings in Sect. 5.3, we propose a simple stochastic recipe to generate a realistic distribution of emission lines for stellar population synthesis models that provide the continuum only. The algorithm works by expressing the model continuum as a linear combination of the basis vectors derived from PCA of the continua of SDSS galaxies. According to these principal components, the model spectrum is classified into one of the 60 continuum classes. We used $k$-means clustering (see Sect. 2.3) to define the continuum classes, as described in the following section.

#### 5.5.1 $k$-means clustering method

To construct our stochastic model of emission lines, we started from the 5+4-dimensional vector space of continuum and log EW principal components of our high signal-to-noise ratio SDSS galaxy sample (refer to Sect. 5.2 for details about the sample). In this space, we classified galaxies into continuum–log EW classes using $k$-means clustering (introduced in Sect. 2.3), specifically the R implementation and the algorithm of MacQueen (1967).

To choose the right number of clusters, one has to consider the variance of emission line log EWs as functions of the number of the clusters. The variance in each cluster is supposed to be decreasing as the number of clusters is growing since clusters are becoming smaller. The minimum variance is limited by the noise in the data. The $\sigma(k)$ curves for all emission lines are plotted in panel (a) of Fig. 5.11. Hence, to minimize the variance of line strengths within each class, we chose $k = 60$ as all curves get essentially flat above this value. For training sets of different sizes and characteristics, a similar analysis of the variance is advisable to determine the input parameter of
Modelling the emission line distributions

Once k-means clusters in the continuum-PCA–line-PCA space are determined in the way described in Sect. 5.5.1, we have to model the distribution of emission line log EWs within each cluster. If the number of clusters is sufficiently high, clusters will become small enough that the distribution of emission lines within them can be well modelled by a multivariate Gaussian distribution parametrized with \( m_n \) and \( C_n \). We note that a multivariate Gaussian distribution, when its entire covariance matrix is known, does not only account for individual line strengths but also for line ratios, including ratios from the same line series. It is also important to mention that, while we did the k-means classification of galaxies in the 5 + 4-dimensional continuum + log EW space, stellar population synthesis models yield the continuum coefficients only. As a result, when classifying model continua, we measure distances from cluster centres in the 5D continuum-PCA subspace only. This will introduce some mixing among clusters as determined by k-means and cause somewhat larger scatter in the randomly generated log EWs than what it would be based solely on the \( C_n \) covariance matrices. This effect is shown in panel (b) of Fig. 5.11. It is still worth using the entire 5 + 4-dimensional space to run the k-means classification because the resulting variances are still lower than using the continuum principal components only, cf. panel (c) of Fig. 5.11. Also, because the covariance of the lines is treated stochastically, there will be random scatter in line ratios as well.

Applying the recipe

Let us denote the average continuum vector with \( e_{0,\lambda} \) and the PCA basis vectors with \( e_{i,\lambda} \), where \( i \) indexes the five dimensions of the PCA space and \( \lambda \) goes over the wavelength bins. Continuum classes are given by the centre of mass vectors \( c_{n,i} \) where \( n \) indexes the 60 classes. Within each class, model lines are randomly generated from a multivariate Gaussian distribution. The mean line log EWs \( m_n \) and the covariance matrices \( C_n \) of the distributions are pre-calculated from the real galaxy sample and provided for each of the 60 continuum classes.

The detailed recipe for generating realistic emission lines given a stellar continuum
Figure 5.11: Variance of log EW of the 11 strong emission lines, averaged over all clusters, as a function of the number of clusters. Panel (a) refers to the case of considering both the continuum and line principal components for clustering and classification. Variance results from the sum of line measurement errors and uncertainty due to the finite size of the clusters. Panel (b) shows the effect of misclassification when only the continuum principal components are used to classify galaxies, with the clustering done in both the continuum and line log EW PCA space. Misclassification will add extra scatter to the randomly generated log EWs, cf. Sect. 5.5.4 and Fig. 5.12. Panel (c) illustrates the case of performing both the clustering and classification in the continuum principal component space only. Even with the additional variance due to misclassification, using both the continuum and line log EW principal components for the clustering is favourable.
model spectrum is the following.

1. Rebin the rest-frame model spectrum $s_\lambda$ to the grid of the basis vectors and normalize it as described in Sect. 5.2.1 to get $\tilde{s}_\lambda$.

2. Subtract the average continuum $e_{0,\lambda}$ from the normalized spectrum.

3. Express the continuum as a linear combination of the provided basis by calculating the dot products $a_i = \sum_\lambda [e_{i,\lambda} \cdot (\tilde{s}_\lambda - e_{0,\lambda})]$

4. Find the class centre $c_{n,i}$ in the continuum PCA space that is the closest (in Euclidean distance) to the vector $a_i$ of the linear coefficients.

5. Take the covariance matrix $C_n$ and mean line log EW vector $m_n$ of the line distribution within the closest class and generate a random vector of line log EWs from the corresponding multivariate Gaussian distribution.

A direct test of the algorithm is to take the galaxies of our SDSS sample, generate emission lines based on their fitted continua and see how well the BPT diagram can be reconstructed. Results of this procedure are plotted in Fig. 5.12 where we also show the original BPT for reference in panel (a) next to the stochastically generated BPT in panel (b). While the curve of star-forming galaxies and the AGN mixing sequence is reproduced reasonably well, there are also a large number of red data points corresponding to the bottom corner of the original BPT visible in all regions of the plot.

### 5.5.4 Shortcomings of the method

The recipe outlined above yields line EWs only, and is sufficient to reproduce the flux excess caused by emission lines but not line widths. In general, line widths should be taken to be equal to the velocity dispersion, at least in the case of star-forming galaxies. Width distributions of broad lines would need to be investigated to generate AGN lines with realistic breadth distribution.

As we pointed out in Sect. 5.5.2, using only the continuum principal components to generate log EWs introduces additional variance due to the mixing of the classes as defined in the continuum + line space. Additionally, as lines are randomly generated
Figure 5.12: Panel (a) shows the original BPT diagram plotted from the directly measured emission lines for reference. Panel (b) is the BPT diagram plotted from stochastically generated emission lines based on the recipe described in Sect. 5.5. While lines were generated from a multivariate distribution randomly, based on the location of the continuum in the PCA space, the resulting BPT diagram resembles the original one remarkably well, although more scatter and significant mixing of galaxy types is visible. Colour coding of the data points is based on the original BPT as in Fig. 5.1.

based on a multivariate Gaussian distribution, EWs and line ratios are not guaranteed to be correct for individual galaxies, but will be for the entire ensemble of mock galaxies.

If the goal is to generate realistic emission lines for individual model continua, the local linear regression method as described in Sect. 5.3.1 is suggested. While that technique yields more accurate emission line estimates, it also requires a much larger input data set and more heavyweight algorithms.

Data necessary to generate random lines are published online\(^1\).

5.6 Discussion of results

In Sect. 5.3, I have shown that optical emission line log EWs can be reasonably well reconstructed from both the optical stellar continuum and broad-band magnitudes of galaxies.

The main practical use case of our method is to generate emission lines for stellar continua from stellar population synthesis models, provided that the models fall into

\(^1\)http://vo.elte.hu/papers/2015/emissionlines/
the wavelength and physical parameter coverage of our training set — in our case the
physical parameters are redshift, metallicity, luminosity, continuum and line properties.
Extrapolation capabilities of empirical techniques to parameter ranges outside the cov-
erage is usually poor compared to theoretical models. While this certainly constrains
the applicability of our results to strong emission line galaxies of the Sloan Digital Sky
Survey (with all 11 prominent lines measured), the methods itself can be easily extended
to galaxies outside the investigated sample by augmenting the training set. However,
further research is necessary to use our line reconstruction method for galaxies with
fewer and weaker lines: correlations between the stellar continuum and the probability
of the very presence of weak emission lines need to be taken into account.

Another application of our method is to estimate emission lines of photometric
galaxies. The technique readily works for the SDSS ugriz filter set, but the existing
training set can be adapted to other filter systems as well. While one simple way to
do this is to compute synthetic photometry from the spectra, building a new training
set by cross-matching the photometric measurements made with the other filter set to
our spectroscopic sample is a better option (provided that the survey overlaps with
the SDSS), since it would automatically account for the unknown systematics in spec-
trophotometry. With the outlined modifications, our technique will be of great value for
analysing data from large photometric surveys like Pan-STARRS (Tonry et al., 2012)
and LSST (Ivezic et al., 2008).

Additionally, by correcting for the contributions of strong emission lines to broad-
band magnitudes, our method can be useful in improving template-based photometric
redshift estimation algorithms to narrow the performance gap between the theoretical
and the empirical approach.

In Sect. 5.4, I described how we used a supervised machine learning algorithm, a
SVM, to verify the empirical demarcation line between star-forming galaxies and AGNs
defined by Kauffmann et al. (2003a). Even though we used only extreme starburst
galaxies and strong AGNs to train the algorithm, the SVM yielded a result very similar
to the analytical segregation curve, only about 6 per cent of galaxies in the bottom
corner of the BPT diagram were classified differently. A future application to SVMs
would be to revisit the Seyfert/LINER separation as it was done in Kewley et al. (2006).

Finally, in Sect. 5.5, I presented a very simple recipe to generate random emis-
sion lines with realistic EWs on top of stellar continua generated by stellar population
synthesis models. I have demonstrated that, despite its simplicity, the method can qualitatively reconstruct the BPT. Our model has its application when the objective is not the accurate modelling of the emission lines of individual galaxies, but rather generating stochastic mock catalogues with more realistic broad-band magnitudes.
Chapter 6

Photometric redshift estimation

In this chapter, I present my research on photometric redshift (photo-z) estimation, based on three peer-reviewed publications. Between Sect. 6.1 and Sect. 6.2.3, I give a general introduction to photo-z approaches that will be of import in later sections. Sect. 6.3 deals with how the official empirical photo-z catalogue for the SDSS Data Release 12 (DR12, Alam et al., 2015) was assembled, as detailed in Beck et al. (2016a). In Sect. 6.4, I describe Photo-z-SQL, a recent spectral template based photo-z implementation documented in Beck et al. (2017a) that can be integrated into a database server. Finally, in Sect. 6.5 I present results from Beck et al. (2017b) that relate to how my photo-z implementations handle more challenging situations, specifically extrapolation and high photometric measurement errors.

6.1 Motivation

To map the large scale structure of the Universe for cosmological research, or to perform weak gravitational lensing studies, three spatial coordinates are required for extragalactic objects. While celestial coordinates are relatively straightforward to obtain, the measurement of distance is far from trivial, and it usually involves measuring the spectral energy distribution, identifying emission or absorption lines to arrive at a redshift, then calculating the distance using a given cosmological model (refer to Sect. 3.2 and Sect. 3.3 for more details). However, spectroscopic observations are very time-consuming compared to broad-band photometric measurements (see also Sect. 3.4).

In recent years, photometric redshift estimation has become a vital and widespread tool in the astronomical repertoire. Large amounts of photometric data are produced
by sky surveys such as the SDSS (York et al., 2000; Eisenstein et al., 2011; Alam et al.,
2015), but spectroscopic measurements are more scarce due to the longer observing
times required. For upcoming and future photometric sky surveys, e.g. Pan-STARRS
(Tonry et al., 2012) and LSST (Ivezic et al., 2008), this issue will only be more acute
as there is no corresponding large-scale spectroscopic survey. Therefore it is important
to accurately estimate the redshift — and thus the distance — of objects based just on
their photometry.

6.2 Approaches to photometric redshift estimation

In the literature, there are two distinct approaches to estimating redshifts from broad-
band photometry — the empirical approach starts from a training set with known
redshifts and uses a machine learning method to perform the estimation, while the
template-based approach fits a spectral template using synthetic photometry, i.e. by
computing the broad-band magnitudes corresponding to known filter transmission curves
and a known spectrum at a given redshift (see Sect. 3.4 and Eq. 3.2). Various public
photo-z implementations are available for both families of methods\(^1\).

The template-based approach has notable advantages over the empirical one: an
extensive training set with known redshifts is not required, and additional physical
properties are implicitly estimated, since the entire template spectral energy distribu-
tion (SED) is known. However, unknown systematics in the photometric measurements
are not accounted for, as opposed to empirical methods, where these are contained
within the training set. Additionally, empirical techniques generally perform consider-
ably better than template-based ones within the object type and redshift coverage of
the training set (Csabai et al., 2003). Still, the extrapolating capabilities of empirical
methods are typically poor, and they are significantly less flexible in adapting to e.g.
observations in different filters, or from different instruments.

6.2.1 The empirical approach

Empirical photo-z methods generally utilize a supervised machine learning algorithm to
find patterns in a training set — with both broad-band magnitude and redshift values

— that allow prediction for cases when the redshift is not known. The 'similarity' of galaxies is usually defined in a metric space, the dimensions of which are some combination of the broad-band magnitudes and colours, perhaps with some scaling applied — I will refer to this as the colour and magnitude space, or, even more concisely, the colour space. The metric is generally chosen to be Euclidean distance within the colour space. Galaxies with a small distance between them — i.e. local galaxies — are considered to be similar, therefore their redshifts are also assumed to be similar. This assumption is then used in an algorithm for estimating galaxies with unknown redshifts.

Examples of machine learning tools used for this purpose include simple linear regression (Wang, Bahcall & Turner, 1998) or quadratic fitting (Connolly et al., 1995b), support vector machines (Wadadekar, 2005), artificial neural networks (Collister et al., 2007; Reis et al., 2012; Brescia et al., 2014), local polynomial fits (Csabai et al., 2007), random forests (Carliles et al., 2010), and boosted decision trees (Gerdes et al., 2010).

Local linear regression method

Following Csabai et al. (2007) and earlier SDSS releases, for DR12 we adopted a local (or piecewise) linear model to describe how the redshifts of galaxies depend on broad-band colours and magnitudes (Beck et al., 2016a). The locality allows the model to follow the complex relationship between these properties, while using a polynomial of just the first order means that a relatively small number of galaxies is enough to fit the parameters. Taking just a few neighbouring galaxies into account helps preserve the local aspect of the model, and, as opposed to a simple average of the neighbours, the linear fit can follow subtle colour-dependent trends in the redshift (cf. Sect. 2.2).

The method described here is mathematically (if not in implementation) the same as the one used for emission line estimation in Sect. 5.3, with the exception that here the estimation error will also be computed. Still, the equations are introduced again to provide the photo-z-specific notations.

Let $i$ be the index of a galaxy in the set $Q$ of galaxies to be estimated (query set), and let us denote the redshift of the $i$-th galaxy with $z_i$ and its coordinates in the $D$-dimensional colour and magnitude space with the vector $d_i$. Let us use $j$ to index galaxies in the training set $T$, which is a collection of galaxies with both coordinate and redshift measurements — $d_j$ and $z_j$, respectively. Thus, our local linear model can be
formulated in the following way:

\[ z_i \approx c_i + a_i d_i = z_{\text{phot},i} \]  

(6.1)

\( z_{\text{phot},i} \) denotes the photometric redshift estimate. The parameter \( c_i \) is a constant offset, while components of the vector \( a_i \) are linear coefficients. These parameters describe our model in the local neighbourhood of galaxy \( i \) — to determine them, we need to extract the local empirical relationship present in the training set, \( T \). We do this by first finding the \( k \)-nearest neighbours of galaxy \( i \) within \( T \), i.e. the \( k \) galaxies whose \( d_j \) coordinates are the closest to \( d_i \) in terms of Euclidean distance. Let us denote the set of nearest neighbours by \( \text{NN} \). The parameters can then be determined using standard linear regression, by minimising the expression

\[ \chi_i^2 = \sum_{j \in \text{NN}} \left( \frac{z_j - c_i - a_i d_j}{w_j} \right)^2 \]  

(6.2)

where \( w_j \) is a weight that could e.g. represent uncertainties in \( z_j \) and \( d_j \), or it could be a function of the distance between \( d_i \) and \( d_j \). The summation runs over the nearest neighbours, and the \( \chi_i^2 \)-minimisation has to be done for every galaxy \( i \) within \( Q \). The error of the photometric redshift \( z_{\text{phot},i} \) can be estimated by how well the thus fitted hyperplane reproduces the \( z_j \) redshifts of the nearest neighbours — we compute the RMS of the deviations from the fit:

\[ \delta z_{\text{phot},i} \approx \sqrt{\frac{\sum_{j \in \text{NN}} (z_j - c_i - a_i d_j)^2}{k}} \]  

(6.3)

6.2.2 The spectral template based approach

The template-based approach generally starts with a set of spectral templates and filter transmission curves, computes synthetic photometric magnitudes from them at various redshifts, and records the redshifts of templates that best reproduce the observed photometry. The choice of spectral templates is a crucial element of these methods. Galaxy templates can be computed theoretically using stellar population synthesis models (Fioc & Rocca-Volmerange, 1997; Bruzual & Charlot, 2003; Maraston & Strömbäck, 2011; Vazdekis et al., 2012). The modelling of emission lines in such models — which can contribute significantly to broad-band magnitudes (Atek et al., 2011) — is a difficulty
because of the number of extra parameters needed to model the interstellar medium, but additional theoretical assumptions (Stasińska, 1984; Fioc & Rocca-Volmerange, 1997; Ferland et al., 2013) or empirical line estimation (Győry et al., 2011; Beck et al., 2016b) can still be used. Alternatively, sets of measured galaxy spectra can be used to compile a library of empirical spectral templates, where the inclusion of measured lines is relatively straightforward (Yip et al., 2004; Dobos et al., 2012b; Marchetti et al., 2013).

Within the family of template-based photometric redshift estimation methods, there are two main branches: those that search for the maximum likelihood, best-fitting SED template and record the redshift of this single SED (Bolzonella, Miralles & Pelló, 2000; Csabai et al., 2000; Arnouts et al., 2002; Ilbert et al., 2006b), and Bayesian methods that aim to reproduce the full posterior redshift probability distribution based on the observations (Benítez, 2000; Coe et al., 2006; Brammer, van Dokkum & Coppi, 2008; Budavári, 2009). The latter method provides a more detailed answer, but it is computationally more expensive, and it does need more input information in the form of a prior (or, with a flat prior, the peak of the probability distribution gives the same result as the maximum likelihood method). Additional refinements include template corrections based on objects with known redshifts (Budavári et al., 2000; Csabai et al., 2000; Feldmann et al., 2006), and using a linear combination of templates along with wavelength-dependent weighting of template errors (Brammer, van Dokkum & Coppi, 2008).

I follow up with describing the two branches of template-based algorithms.

**Maximum likelihood method**

The maximum likelihood method can be summarized with the formulae

\[
\chi^2(z', t', m'_0) = \frac{1}{2} \sum_{ij} (m_i - (s_i (z', t') - m'_0)) C_{ij}^{-1} (m_j - (s_j (z', t') - m'_0)) \quad (6.4)
\]

and

\[
(z_{\text{phot}}, t, m_0) = \arg \min_{(z', t', m'_0)} \chi^2(z', t', m'_0), \quad (6.5)
\]

where \(z'\) and \(t'\) are the redshift and the template SED, \(m'_0\) is a factor that scales the total flux of the template spectrum, \(m\) is the measured broad-band magnitude vector.
of an object, \( s(z', t') \) is the synthetic magnitude vector of template \( t' \) at redshift \( z' \), and finally, \( C \) is the covariance matrix of magnitude errors between filters.

The fitting process involves iterating over a set of redshifts and templates, but not over \( m'_0 \), since for a given \( z' \) and \( t' \) the \( m'_0 \) value that yields the lowest \( \chi^2 \) can be determined algebraically. \( (z_{\text{phot}}, t, m_0) \) denote the set of best-fitting parameters, corresponding to the lowest \( \chi^2 \) value found during the iteration. \( z_{\text{phot}} \) is the photometric redshift estimate.

Most applications assume that magnitude errors are uncorrelated, in addition to being Gaussian with a standard deviation matching the estimated measurement error for the given object. Under the uncorrelated assumption, the \( C \) covariance matrix takes the form

\[
C_{ij} = \sigma^2_i \delta_{ij},
\]

where \( \sigma \) is the estimated magnitude error vector of the object, and \( \delta \) is the Kronecker delta. The expression for \( \chi^2 \) can be simplified under this assumption to yield

\[
\chi^2 (z', t', m'_0) = \frac{1}{2} \sum_i \left( m_i - (s_i (z', t') - m'_0) \right)^2 / \sigma^2_i.
\]

**Bayesian method**

Using the notation introduced in Sect. 6.2.2, the Bayesian approach starts from the expression

\[
P (z, t, m_0|m, C) = \frac{P (z, t, m_0) P (m, C|z, t, m_0)}{\int dz' \int dt \int dm_0 P (z', t, m_0) P (m, C|z', t, m_0)},
\]

which specifies the probability \( P (z, t, m_0|m, C) \) of a redshift and a template SED – scaled to a flux – given the data, i.e. the measured magnitudes and magnitude errors (optionally taking the covariance between filters into account). \( P (z, t, m_0) \) is the prior probability of the given template SED at the given redshift. \( P (m, C|z, t, m_0) \) is the likelihood that the given template and redshift produce the data. The denominator contains a factor that normalizes the total probability to 1, but in practice this is not computed, and instead the final probability distribution is normalized. We note that we use the \( \int dt \) term as shorthand for integrating over the model parameters that describe a template SED, \( t \) does not represent a real number.

The likelihood term can be formulated as
\[ P(z|m, C|z, t, m_0) = \exp\left(-\chi^2(z, t, m_0)\right), \quad (6.9) \]

where \( \chi^2 \) is the expression defined in Eq. 6.4. Again, assuming uncorrelated magnitude errors, the simplified version in Eq. 6.7 can be applied.

The method aims to extract the posterior redshift probability distribution, \( P(z|m, C) \), which can be obtained by integrating out the nuisance parameters in Eq. 6.8 such that

\[ P(z|m, C) = \int dt \int dm_0 P(z, t, m|z, t, m_0). \quad (6.10) \]

### 6.2.3 Difficulties in photometric redshift estimation

There are two main factors that are detrimental to the accuracy of photometric redshift estimation, regardless of the specific approach taken: the overlap in photometric colour space between different galaxy types, and the measurement errors in the photometry. While these are of different origin, their effect is very much intertwined.

The first factor, overlap in broad-band colour space, is a purely physical phenomenon. When the available colours cannot differentiate between morphological types, i.e. when different galaxy types have the same colours at different redshifts, there simply is not enough data to give an unequivocal answer to the question of what the redshift is. In such cases, the assumption that the broad-band magnitudes and colours uniquely determine the redshift does not hold, there are degeneracies in the colour–redshift relation (Benítez, 2000).

The second factor, photometric measurement errors, is a major issue. The measurement errors can greatly exacerbate the effects of overlap, blurring the divisions in colour space between different galaxy types, and also between galaxies of the same type but with differing redshifts (Benítez, 2000). Additionally, when the measurement errors are not estimated accurately, or when photometric errors in different bands are correlated (Scranton et al., 2005), the assumption of uncorrelated Gaussian errors, used in many methodologies, simply does not hold (Budavári, 2009).

These issues can be mitigated by improvements in the instrumentation. A better camera and telescope can reduce photometric errors (Ivezic et al., 2008; Tonry et al., 2012), while a large selection of filters (Wolf et al., 2003), or filters designed specifically for photometric redshift estimation (Budavári et al., 2001) can remove degeneracies.
6.3 Photometric redshifts for the SDSS DR12

The Sloan Digital Sky Survey is one of the largest public collections of both photometric and spectroscopic measurements, with 208,478,448 galaxies in its photometric catalog (York et al., 2000; Gunn et al., 1998; Doi et al., 2010), and, as of Data Release 12 (Alam et al., 2015), 2,274,081 galaxy spectra in the continually expanded spectroscopic catalog (Eisenstein et al., 2011; Smee et al., 2013).

The purpose of this section is to give a detailed description of the methods and data we used in creating the official photometric redshift database of SDSS DR12, released to the public in January 2015, and documented in Beck et al. (2016a). We chose an empirical technique, local linear regression, to estimate the redshift and its error, utilizing a training set of 1,976,978 elements, assembled from DR12 spectroscopy and data from other spectroscopic surveys (listed in Sect. 6.3.3). Additionally, we computed the maximum likelihood spectral template fit to the photometry, using the composite spectrum atlas of Dobos et al. (2012b), to obtain additional information such as K-corrections, spectral type, and rest-frame absolute magnitudes.

The main goal of our photometric redshift catalog is to complement the estimated redshift with a reasonable assessment of the estimation error for the wide variety of galaxies in the SDSS photometric survey. The inclusion of spectroscopic data from other surveys means that we have more reference points for distant and faint bluer objects, up to $z \approx 0.8$ and $r \approx 21.5$ mag, which would be less well-represented in SDSS spectroscopy due to target selection (Eisenstein et al., 2001; Dawson et al., 2013). We also published an error map in support of this goal, as it highlights problematic regions in the space of galaxy colours where there are overlapping galaxies at different redshifts, leading to reduced accuracy.

Throughout the section, broad-band magnitudes are quoted in the SDSS asinh magnitude system (Lupton, Gunn & Szalay, 1999), and are dereddened according to Schlegel, Finkbeiner & Davis (1998). Following the recommendations of Scranton et al. (2005), for galaxy magnitudes we use the SDSS $cModelMag$ magnitudes, and scale magnitude errors according to Eq. 15 in Scranton et al. (2005), while for galaxy colours we use SDSS $modelMag$ magnitudes. Similarly to other SDSS applications, we adopt WMAP 5-year + SNe + BAO best-fitting cosmological parameters: $\Omega_{\Lambda} = 0.726$, $\Omega_m = 0.2739$, $\Omega_r = 0.0001$ and $H_0 = 70.5 km/s/Mpc$ (Hinshaw et al., 2009). The
photometric database can be accessed via SkyServer\(^2\).

### 6.3.1 Estimation method

As in previous SDSS data releases, to utilize the extensive spectroscopic sample of the SDSS, we elected to use an empirical method (as opposed to template-based) for estimating the redshift and its error, local linear regression. Specifics of the method are detailed in Sect. 6.2.1. To get the best of both worlds, we combined this with a template fitting step that uses the photometric redshift, yielding additional physical information (see Sect. 6.3.2).

In our current implementation, we have \( D = 5 \) dimensions, and the components of the vectors \( \mathbf{d}_i \) and \( \mathbf{d}_j \) are the \( r \)-band magnitude, and the \( u - g, g - r, r - i, i - z \) colours (the notation in this section follows that of Sect. 6.2.1). All five dimensions are scaled to have zero mean and unit standard deviation. The nearest neighbours are weighted equally, \( w_j = 1 \) for every \( j \). These choices were made to optimise the accuracy of the photo-z estimation. We use \( k = 100 \) to have enough data points to determine the parameters and the error, but still preserve the locality of the model. The exact choice of \( k \) does not significantly impact the results, however.

We assume that the error of the spectroscopic redshift is negligible, i.e. \( z_j = z_{\text{spec},j} \). Generally, this is a reasonable approximation because spectroscopic redshifts are much more accurate than photometric redshift estimates. However, it is important to note that there is a non-negligible percentage of spectroscopic redshift failures corresponding to a given quality cut in a survey (see Sect. 6.3.3 for a discussion of failure rates). If the failures are correlated with spectral type and colour, this systematic error in \( z_{\text{spec}} \) will be included in our training set, and thus propagate through to our \( z_{\text{phot}} \) estimates. Still, our best reference points for estimation are the redshifts published by spectroscopic surveys.

As an additional refinement of our method, when there are neighbours with outlying redshifts, we perform the computations twice to eliminate them. Neighbours that satisfy \( 3\delta z_{\text{phot},i} < |z_j - c_i - \mathbf{a}_i \mathbf{d}_j| \) are discarded from the set \( NN \) (cf. Eq. 6.1 and Eq. 6.3), and the fit is redone for the limited set of \( l < k \) nearest neighbours, as needed. Also, we flag galaxies that lie outside the bounding box of the nearest neighbours in the \( D\)-
dimensional colour and magnitude space. In such cases, we perform an extrapolation using the fitted hyperplane as opposed to an interpolation, therefore we can expect less reliable results (see Sect. 6.3.5 for more details).

Once the photometric redshift of the query point has been determined using this empirical method, we follow up with a spectral template fitting step, as described in the following section.

### 6.3.2 Spectral template fitting

Traditional maximum likelihood template-based photometric redshift estimation methods (Bolzonella, Miralles & Pelló, 2000; Csabai et al., 2000; Arnouts et al., 2002; Ilbert et al., 2006b) solve the problem of Eq. 6.7 and Eq. 6.5, as shown in Sect. 6.2.2.

In our hybrid approach, instead of iterating over every redshift \( z' \) in consideration, we use the empirically determined photometric redshift, as described in Sect. 6.2.1. This way, we enjoy the benefit of higher redshift accuracy due to the extensive training set, while also fitting a galaxy template with a known SED. Thus, instead of Eq. 6.7 and Eq. 6.5, the expression we solve for every galaxy \( i \) becomes:

\[
(t_i, m_{0,i}) = \arg \min_{(t', m'_0)} \left( \frac{1}{2} \sum_{p=1}^{D} \left( \frac{m_{p,i} - (s_p(z' = z_i, t') - m'_0)}{\sigma_{p,i}} \right)^2 \right), \tag{6.11}
\]

where \( z_i \) is computed using Eq. 6.1, and we followed the notation of Sect. 6.2.2 but changed the running index to \( p \). As for the list of templates, we use the composite spectrum atlas of Dobos et al. (2012b), which has been assembled from SDSS spectra, takes emission lines into account, and contains extreme red and blue galaxy types in addition to the more frequently occurring ones. Dobos et al. (2012b) also published synthetic photometric magnitudes in the SDSS filter set for a grid of redshift values. Fig. 6.1 shows the coverage of the templates in \( g - r, r - i \) colours — the dense galaxy regions are well-covered by the composite spectrum atlas. For the set of all photometric galaxies, the fitted synthetic magnitudes are within \( 3\sigma \) of the measured \( m \) for 82.0%, 89.7%, 96.2%, 97.2% and 97.3% of cases, respectively, for the \( u, g, r, i \) and \( z \) broadband magnitudes. The normalized error distributions are roughly Gaussian, with the exception of the \( u \)-band, where it is asymmetric. Considering the redshift estimation errors and outlier rate in the unfiltered galaxy set (see Sect. 6.3.4 and Tab. 6.3 for...
Figure 6.1: The colour space coverage of the spectral templates from Dobos et al. (2012b) in $g - r$, $r - i$ dimensions. Blue dots show templates with $z \leq 0.35$, while red dots correspond to redshifts $0.35 < z < 0.7$. The template colours are superimposed on a grayscale density map of SDSS photometric measurements.

more details), the within-3σ ratios are relatively high, which shows that the templates adequately describe the fitted galaxies.

Once we have found the best-fitting spectral template $t_i$, we determine other values of physical interest. The $DM$ distance modulus and $D_L$ luminosity distance are computed using the redshift and our assumed cosmology. Knowing the SED of the template, we also calculate observed-frame synthetic magnitudes, K-corrections to redshifts 0 and 0.1, and rest-frame absolute magnitudes (see Sect. 6.3.9 for exact definitions of these).

6.3.3 Assembling the training set

Our training set initially consisted of the entire spectroscopic galaxy catalog of the Sloan Digital Sky Survey Data Release 12. This includes the earlier main galaxy and
Table 6.1: Information about the external spectroscopic surveys we used to expand our training set.

<table>
<thead>
<tr>
<th>Survey Name</th>
<th>References</th>
<th>Quality flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>2dF</td>
<td>Colless et al. (2001, 2003)</td>
<td>$Quality = 4, 5$</td>
</tr>
<tr>
<td>6dF</td>
<td>Jones et al. (2004, 2009)</td>
<td>$Q = 3, 4$</td>
</tr>
<tr>
<td>DEEP2</td>
<td>Davis et al. (2003); Newman et al. (2013)</td>
<td>$ZQUALITY = 3, 4$</td>
</tr>
<tr>
<td>GAMA</td>
<td>Driver et al. (2011); Baldry et al. (2014)</td>
<td>$NQ = 4$</td>
</tr>
<tr>
<td>PRIMUS</td>
<td>Coil et al. (2011); Cool et al. (2013)</td>
<td>$Q = 4$</td>
</tr>
<tr>
<td>VIPERS</td>
<td>Garilli et al. (2014); Guzzo et al. (2014)</td>
<td>$[zflg] \mod 10 = 3, 4$</td>
</tr>
<tr>
<td>VVDS</td>
<td>Le Fèvre et al. (2004); Garilli et al. (2008)</td>
<td>$ZFLAGS \mod 10 = 3, 4$</td>
</tr>
<tr>
<td>WiggleZ</td>
<td>Drinkwater et al. (2010); Parkinson et al. (2012)</td>
<td>$Qop = 4, 5$</td>
</tr>
<tr>
<td>zCOSMOS</td>
<td>Lilly et al. (2007, 2009)</td>
<td>$[Class] \mod 10 = 3, 4$</td>
</tr>
</tbody>
</table>

LRG samples, and also the more recent BOSS sample. The main galaxy sample consists of a wide variety of galaxies, with no cuts on colour, although it is rather limited in terms of redshift (Strauss et al., 2002). The LRG sample provides an expanded redshift coverage, however, it has specifically targeted luminous red galaxies (Eisenstein et al., 2001). The BOSS sample extends much deeper than the former two, and has somewhat relaxed the sharp colour cuts of the LRG sample, but it is still targeted towards massive galaxies (Dawson et al., 2013), likely resulting in non-negligible selection effects.

On the other hand, the photometric galaxy catalog of the SDSS has no such selection effects, and our goal is to provide photometric redshifts for the entire catalog, not just a subset of morphological types or colour. Since it would be advantageous to have a wider selection of galaxy types and colours even at higher redshifts, we decided to extend our training set by cross-matching galaxies in the Sloan photometric catalog with spectroscopic measurements of other, publicly available surveys.

**Data from other surveys**

The spectroscopic surveys with which we extended our training set are listed in Tab. 6.1, with references. For each survey, we used the published redshift quality flag to select only the reasonably confident redshift measurements, with confidences $\geq 95\%$ (with the exception of PRIMUS, where $\geq 92\%$).

We cross-matched the galaxies from other surveys with SDSS primary photometric
galaxy measurements by using J2000 right ascension and declination coordinates and published astrometric errors. We followed the probabilistic methodology of Budavári & Szalay (2008), assumed Gaussian errors, and calculated the Bayes factor of Eq. 16 in Budavári & Szalay (2008), which is the ratio of the likelihood that the two measurements are of the same source, and the likelihood that they are of separate sources:

\[
B = \frac{L(\text{same source})}{L(\text{separate sources})} = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp \left\{ -\frac{\psi^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}
\] (6.12)

Here \(\sigma_1\) and \(\sigma_2\) are the astrometric errors of two given galaxies, and \(\psi\) is the angular separation between them. We accepted matches with \(B > 10,000\), thus ensuring that we only used rather certain matches.

Galaxies with existing SDSS spectrometry were excluded from the cross-match, and where we found multiple matches for the same Sloan galaxy, we selected the one with the smallest redshift error.

In total, we found 168,834 matches with reasonable redshift confidence. However, later filtering steps greatly limited the number we could utilize, to 76,193.

Multiple matches provide an opportunity to test whether the published spectroscopic redshift failure rates are correct and whether the cross-match itself works reliably. Of the total of 1,012 multiple matches in the filtered sample, 171 were between two PRIMUS measurements, 769 had one PRIMUS object, and 72 did not include PRIMUS. (We are handling PRIMUS separately because of its lower confidence level and higher redshift error compared to other surveys.)

The only-PRIMUS set had a standard deviation of \(\sigma(\Delta z_{\text{spec}}) = 0.00473\) and a 3\(\sigma\) outlier rate of \(P_o = 9.36\%\) (outliers were removed iteratively). Individual PRIMUS measurements typically have an accuracy of \(\sim 0.005\), therefore the deviation is even below what one would expect, and the outlier rate is also well below the theoretically expected \(1 - 0.92 \times 0.92 = 15.4\%\).

The PRIMUS–other survey set is described by the numbers \(\sigma(\Delta z_{\text{spec}}) = 0.00472\) and \(P_o = 8.97\%\). The deviation is roughly the accuracy of PRIMUS, just as expected, while the outlier rate is again below the expected \(1 - 0.95 \times 0.92 = 12.6\%\).

The non-PRIMUS set had a standard deviation of \(\sigma(\Delta z_{\text{spec}}) = 0.00071\) and an outlier rate of \(P_o = 29.2\%\). The deviation is negligible for our purposes, but the outlier rate is significantly larger than the expected \(1 - 0.95 \times 0.95 = 9.75\%\). The observed discrepancies might be due to a number of reasons, below we list a few.
• The spectroscopic redshift failures could be correlated, which would reduce the combined outlier rate. Especially the only-PRIMUS set could be affected, where the same survey measured the same object twice.

• Galaxies could be erroneously cross-matched due to an underestimation of astrometric accuracy — DEEP2, the survey responsible for 71.4% of outliers in the non-PRIMUS set, uses the Canada France Hawaii Telescope, which quotes the USNOA 2.0 astrometric error of 0.5″ (Coil et al., 2004), the highest value of all the external surveys.

• Overlapping galaxies in the field of view could compromise spectroscopic redshifts, and could also lead to incorrect cross-matches.

Furthermore, it is important to note that the non-PRIMUS set only had 72 matches, of which 21 were outliers. This is a rather small sample size, and might not be representative. On the whole, the confidence levels derived from multiple matches are in line with — or better than — the expectations based on the published numbers of the surveys.

Filtering the training set

While our goal was to assemble a training set with as wide a coverage in redshift and colour space as possible, the inclusion of objects with too large photometric errors would diminish our ability to find the most similar reference galaxies. The ‘true’ nearest neighbours may be scattered away due to errors, with less similar galaxies taking their place. To alleviate this problem, we introduced photometric error cuts to the training set. Additionally, we filtered out galaxies with outlying colours, which both eliminates erroneous measurements, and also more clearly defines the boundaries of our training set in colour space.

The exact parameters of the cuts were determined empirically, with the following criteria in mind:

• optimise the photometric redshift estimation results,

• leave no region empty in the space of broad-band colours, if otherwise within the coverage of the training set,
• keep fainter and higher-redshift measurements of sufficient accuracy.

The final values of the photometric error and colour cuts are as follows:

\[
\begin{align*}
\Delta r &< 0.15 \\
\Delta(g - r) &< 0.225 \\
\Delta(r - i) &< 0.15 \\
\Delta(i - z) &< 0.25 \\
-0.911 &< (u - g) < 5.597 \\
0.167 &< (g - r) < 2.483 \\
0.029 &< (r - i) < 1.369 \\
-0.452 &< (i - z) < 0.790
\end{align*}
\]  

(6.13)

Magnitudes are in the SDSS \texttt{ugriz} filter system, with errors scaled following Scran-
ton et al. (2005) (see also Sect. 6.3). The colour cuts correspond to filtering out the
highest and lowest 0.5\% of data for the \((u - g)\) colour, and 1\% for the other three
colours. The reason for having no limit on \(\Delta u\), and for having relaxed criteria on
\((u - g)\) compared to other colours is that the errors of the SDSS \(u\)-band are generally
much larger than that of other bands, and even galaxies with fairly secure photometric
redshifts can have very large \(u\)-band errors.

Only those galaxies were included in the training set that fulfilled all of Eq. 6.13. Additionally, SDSS galaxies with unsecure spectroscopic redshifts were also cut: the
spectroscopic error flag \texttt{SpeczWarning} had to either take the value \texttt{OK} or \texttt{MANY\_OUTLIERS}
(the latter rarely signifying a real error according to the documentation). This spectro-
scopic error flag cut filters out a higher and higher fraction of galaxies as the redshift
increases — with more distant galaxies typically having lower signal-to-noise spectra
— but there is no indication of a specific redshift being preferentially eliminated, which
otherwise could have pointed to a systematic incompleteness in our training set. The
redshift distribution of the finalised training set is shown in Fig. 6.2.
Figure 6.2: The redshift distribution of our entire DR12 photo-z training set, and subsets of it: the BOSS spectroscopic sample, the pre-BOSS spectroscopic sample that includes the main galaxy sample and the LRG sample (MGS+LRG), and the additional galaxies cross-matched from other surveys. In the top right corner of each panel, we indicate the corresponding subset, and the total number of galaxies within that subset. Note the different scale of the cross-match subset.
Figure 6.3: The photometric redshift ($z_{\text{phot}}$) as a function of spectroscopic redshift ($z_{\text{spec}}$), $z_{\text{phot}} - z_{\text{spec}}$ as a function of $z_{\text{spec}}$, and $(z_{\text{phot}} - z_{\text{spec}})/\delta z_{\text{phot}}$ as a function of $z_{\text{spec}}$, where $\delta z_{\text{phot}}$ is the photometric redshift error estimate. The galaxy density of our training set is shown in grayscale — we took the logarithm of galaxy counts so that even individual galaxies can be seen. The red solid, dashed and dotted lines represent the median, 68% and 95% confidence regions of the training set, respectively. The green line shows $z_{\text{spec}} = z_{\text{phot}}$, i.e. what would be the perfect estimation. See the text for a discussion.

6.3.4 Cross-validation results

To evaluate the performance of our method, we randomly divided the training set into two equal-sized subsets, and performed cross-validation, estimating the photometric redshifts of one half using the other half as the training set (and vice versa). The resulting photometric redshifts could then be contrasted with the spectroscopic redshifts. Fig. 6.3 shows the photometric redshift $z_{\text{phot}}$, the estimation error $z_{\text{phot}} - z_{\text{spec}}$, and the estimation error divided by the reported photometric redshift error $\delta z_{\text{phot}}$, as functions of the spectroscopic redshift.

Using the normalized redshift estimation error $\Delta z_{\text{norm}} = \frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}}$, we achieve an average bias of $\overline{\Delta z_{\text{norm}}} = 5.84 \times 10^{-5}$, a standard deviation of $\sigma(\Delta z_{\text{norm}}) = 0.0205$, and an outlier rate of $P_o = 4.11\%$. Outliers are defined as $|\Delta z_{\text{norm}}| > 3\sigma(\Delta z_{\text{norm}})$, and are removed iteratively. While most galaxies in the training set have fairly small estimation errors, on Fig. 6.3 it is apparent that there are redshift ranges where there is a non-negligible bias, up to $\Delta z = 0.01$ or $0.5 \delta z_{\text{phot}}$.

Still, in biased regions between 58% and 76% of galaxies are within $\pm 1 \delta z_{\text{phot}}$, and
between 86% and 98% of galaxies are within $\pm 2\delta z_{\text{phot}}$. Thus, the confidence interval $z_{\text{phot}} \pm \delta z_{\text{phot}}$ can be reasonably used in applications, as it will contain a fairly high fraction of galaxies even when there is bias in the estimation, and the distribution is not centered on $z_{\text{phot}}$.

Additionally, on Fig. 6.4, we plotted the probability density function of $(z_{\text{phot}} - z_{\text{spec}})/\delta z_{\text{phot}}$ alongside a standard normal distribution. There is a small overall bias, but otherwise the two distributions match rather well, highlighting that our method for estimating the error of the photometric redshift gives a fair assessment of the estimation accuracy.

Another issue visible on Fig. 6.3 is that the estimation accuracy declines dramatically from around $z = 0.6$, where the number count of the training set falls off. These high-redshift galaxies occupy sparsely sampled regions in colour space, as evidenced by the fact that 94% of them are above the 50th, and 68% are above the 75th percentile of nearest neighbour bounding box volume. Sparse regions are more likely to include a non-negligible amount of galaxies scattered there due to high photometric errors. In the case of high-redshift galaxy regions, the scattered galaxies are also more likely to have lower redshifts, hence the negative estimation bias.

In the following section, we will go into more detail concerning the biases and errors.

6.3.5 Discussion of biases and errors

Here we discuss how the issues outlined in Sect. 6.2.3, namely overlap and errors in the photometry, affect our photo-z estimation results on the SDSS DR12.

When galaxies of different redshifts overlap in colour space, the nearest neighbours that we find with our algorithm will have a bimodal, or even multimodal redshift distribution. In this case, the estimated redshift will lie between the different peaks in the distribution, and the estimated error will increase accordingly. Additionally, galaxies belonging to a peak at a smaller redshift will have a positive estimation bias, while galaxies that correspond to a peak at a higher $z$ will be estimated with a negative bias. When the overlap is ‘perfect’, it is impossible to decide which is the appropriate galaxy group, but when the loci of groups in colour space have a slight offset between them, the closer galaxy group will more strongly constrain the fitted hyperplane locally.

Assuming a mixture of two Gaussian distributions of equal weight but with different
Figure 6.4: The normalized histogram of the scaled redshift error, \((z_{\text{phot}} - z_{\text{spec}}) / \delta z_{\text{phot}}\), is plotted in black for our training set. The blue line shows the standard normal distribution. See the text for a discussion.
means, our method will estimate the redshift closer to the correct peak, as opposed to a simple $k$-nearest neighbour average, which would give the centerpoint, the average of the means. This effect becomes less noticeable when one of the groups is underrepresented in the training set, or when photometric errors are large enough to sufficiently mix the groups in colour space. To remove some of the degeneracy, in addition to the colours, we also used the $r$-band magnitude in our local linear regression. In Sect. 6.3.6, we describe an error map that helps quantify the effects of overlap.

Photometric measurement errors strongly limit the accuracy of photometric redshifts in the SDSS catalog — both the training set and the query set are affected, especially the fainter galaxies. We introduced several photometric error classes for the galaxies to quantify the dependence of redshift estimation errors on errors in the photometry: class 1 matches the error limits of the training set, and the subsequent classes $(2-7)$ contain galaxies with progressively higher errors. The exact limits were determined empirically based on the photometric error and redshift error distributions, with the aim of giving a sequence from useful photometric redshifts to highly inaccurate ones. The class identifiers have a negative sign if the local linear regression was an extrapolation, i.e. the estimated galaxy lay outside the bounding box of the nearest neighbours. For example, class $-1$ denotes galaxies that match the error limits of the training set, but were estimated with an extrapolation. This way, class $-1$ also includes galaxies that did not satisfy the colour cuts of Eq. 6.13, and therefore are not within the training set. Classes $2-7$ and $(-2)-(-7)$ do contain galaxies with spectroscopic redshifts, specifically those that did not fulfill the error limits of Eq. 6.13 — these galaxies can be used to test the redshift estimation accuracy in a given class.

Tab. 6.2 gives the photometric error limits used for each class, while Tab. 6.3 lists the redshift estimation bias, standard deviation, outlier rate, and the spectroscopic and photometric galaxy count in the classes. It is clear that higher photometric errors correspond to sharply increasing biases and deviations.

We emphasise here that, since the training set only contains galaxies of class 1 or -1, the redshift error estimate ($\delta z_{phot}$) is expected to be an accurate representation of the estimation error only when the query galaxy also belongs to class 1 or -1 (and satisfies Eq. 6.13, if class -1). As we show in Sect. 6.3.6, the redshift estimation errors are dependent on the position in colour space. A higher photometric error class leads to additional variance in $z_{phot}$, which therefore should ideally be characterized as a
function of the position in colour space. However, as shown in Tab. 6.3, there are relatively few spectroscopic galaxies in the higher error classes, and we do not have a good enough coverage to allow a detailed treatment of this phenomenon. As a crude first approximation for other classes, the class-wide extra variance with respect to class 1 can be added according to Tab. 6.3.

### 6.3.6 The redshift error map

As described in the previous section, the presence of biases and higher errors in the redshift estimation is strongly dependent on the position of a given galaxy in the space of broad-band colours. To provide a tool for filtering out regions in the colour space where these issues are the most prominent, we compiled and published an error map.

The error map gives the redshift estimation results — as computed on the training set — for a 3D grid in $r$-band magnitude, and $g - r$, $r - i$ colours. For each bin in the grid, we report the galaxy count, the average $z_{\text{spec}}$, the average $z_{\text{phot}}$, the rms of $z_{\text{phot}} - z_{\text{spec}}$, the average $\Delta z_{\text{phot}}$, and the average standard deviation of the redshifts of the neighbours, $\sigma (z_{NN})$. With the help of this map, it is possible to flag galaxies in sparsely populated regions, or in regions with high estimation errors (which also indicate possible biases).

To illustrate, we computed these measures for a 2D projection of the 3D map, where the $r$-band magnitude has been summed over, and the grid remains in $g - r$, $r - i$ colours.

<table>
<thead>
<tr>
<th>Class</th>
<th>$\Delta r_{\text{max}}$</th>
<th>$\Delta (g - r)_{\text{max}}$</th>
<th>$\Delta (r - i)_{\text{max}}$</th>
<th>$\Delta (i - z)_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.15</td>
<td>0.225</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0.25</td>
<td>0.18</td>
<td>0.28</td>
</tr>
<tr>
<td>3</td>
<td>0.21</td>
<td>0.30</td>
<td>0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>4</td>
<td>0.24</td>
<td>0.35</td>
<td>0.24</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>0.27</td>
<td>0.40</td>
<td>0.27</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.30</td>
<td>0.45</td>
<td>0.30</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 6.2: The maximum photometric error values that a galaxy belonging to a photometric error class was allowed to have. The errors were scaled following Scranton et al. (2005). Each galaxy is placed in the lowest possible class. Class 7 contains galaxies that could not be placed in any other class.
<table>
<thead>
<tr>
<th>Class</th>
<th>$\Delta z_{\text{norm}}$</th>
<th>$\sigma (\Delta z_{\text{norm}})$</th>
<th>$P_o$</th>
<th>$N_{\text{spec}}$</th>
<th>$N_{\text{phot}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$6.11 \times 10^{-5}$</td>
<td>0.0204</td>
<td>4.07%</td>
<td>1,957,234</td>
<td>42,410,836</td>
</tr>
<tr>
<td>2</td>
<td>-0.0033</td>
<td>0.0333</td>
<td>4.03%</td>
<td>77,281</td>
<td>5,657,368</td>
</tr>
<tr>
<td>3</td>
<td>-0.0057</td>
<td>0.0331</td>
<td>3.93%</td>
<td>68,610</td>
<td>5,240,766</td>
</tr>
<tr>
<td>4</td>
<td>-0.0082</td>
<td>0.0369</td>
<td>4.45%</td>
<td>36,218</td>
<td>3,955,814</td>
</tr>
<tr>
<td>5</td>
<td>-0.0107</td>
<td>0.0412</td>
<td>5.00%</td>
<td>19,110</td>
<td>2,970,417</td>
</tr>
<tr>
<td>6</td>
<td>-0.0127</td>
<td>0.0486</td>
<td>5.33%</td>
<td>10,674</td>
<td>2,232,881</td>
</tr>
<tr>
<td>7</td>
<td>-0.0222</td>
<td>0.0823</td>
<td>3.80%</td>
<td>16,563</td>
<td>6,950,249</td>
</tr>
<tr>
<td>-1</td>
<td>$4.22 \times 10^{-4}$</td>
<td>0.0289</td>
<td>5.71%</td>
<td>19,744</td>
<td>2,001,544</td>
</tr>
<tr>
<td>-2</td>
<td>-0.0051</td>
<td>0.0549</td>
<td>11.2%</td>
<td>5,940</td>
<td>1,421,618</td>
</tr>
<tr>
<td>-3</td>
<td>-0.0081</td>
<td>0.0514</td>
<td>8.97%</td>
<td>10,262</td>
<td>2,848,424</td>
</tr>
<tr>
<td>-4</td>
<td>-0.0104</td>
<td>0.0567</td>
<td>7.65%</td>
<td>11,200</td>
<td>4,098,896</td>
</tr>
<tr>
<td>-5</td>
<td>-0.0150</td>
<td>0.0643</td>
<td>6.68%</td>
<td>10,917</td>
<td>5,118,595</td>
</tr>
<tr>
<td>-6</td>
<td>-0.0165</td>
<td>0.0728</td>
<td>6.06%</td>
<td>10,350</td>
<td>5,862,776</td>
</tr>
<tr>
<td>-7</td>
<td>-0.0488</td>
<td>0.1410</td>
<td>2.30%</td>
<td>86,574</td>
<td>117,703,892</td>
</tr>
</tbody>
</table>

Table 6.3: The average redshift estimation bias $\Delta z_{\text{norm}}$, standard deviation $\sigma (\Delta z_{\text{norm}})$, outlier rate $P_o$, number of spectroscopic galaxies $N_{\text{spec}}$ and number of photometric galaxies $N_{\text{phot}}$ in each photometric error class, with $\Delta z_{\text{norm}} = \frac{z_{\text{phot}} - z_{\text{spec}}}{1+z_{\text{spec}}}$. Outliers are defined to have $|\Delta z_{\text{norm}}| > 3\sigma (\Delta z_{\text{norm}})$, and are removed iteratively. We indicate extrapolation in the local linear regression with a negative sign in front of the class identifier.
In Fig. 6.5, the galaxy count distribution is shown as a function of the two colours. The pronounced discontinuity — a diagonal line — is a target selection effect produced by the colour cut of the CMASS subsample in the BOSS survey (Dawson et al., 2013), leading to a sparsely populated region below the cut. In Fig. 6.6, we plotted three measures of the redshift error, all of which show similar behaviour. The estimation error is highest where the redshifts of the neighbours have a larger deviation, i.e. where there is overlap of galaxies with differing redshifts. Since the photometric error limit of even error class 1 is as high as $\Delta(g-r)_{\text{max}} = 0.225$ and $\Delta(r-i)_{\text{max}} = 0.15$ in these two dimensions, such mixing between different redshifts is to be expected. The reported error follows the actual error closely, drawing the same overall picture of the dependence of estimation errors on the location in colour space, which also supports our result in Sect. 6.3.4 that our redshift error estimates are accurate. Additionally, sparsely populated regions in Fig. 6.5 correspond to higher errors in Fig. 6.6. Since sparse regions could be occupied either by exotic galaxy types or galaxies that were scattered there due to high photometric errors, it is not surprising that their redshift estimation is inaccurate.

6.3.7 Best practices in using the database

The photometric redshift database has been made public along with the SDSS DR12. It can be accessed via SkysServer\(^3\), it is the Photoz table within the DR12 context. The redshift error map is contained in the table PhotozErrorMap, also in the DR12 context. Refer to Sect. 6.3.9 for a description of each column in these tables.

Here we intend to give a few suggestions on how our database can be best utilised.

As a first step, it is recommended to only use galaxies with a photometric error class of 1, because that is when the redshift error estimate is expected to be accurate. If more galaxies are desired, the instructions at the end of Sect. 6.3.5 and Tab. 6.3 detail how additional error classes can be included.

When nearest neighbours in the local linear regression are outliers, they are excluded from the fit. However, having too many outliers may indicate that the given galaxy is difficult to estimate, therefore a limit should be put on the minimum number of nearest neighbours used out of the total of $k = 100$, e.g. a minimum of $l = 97$. Additionally,

\(^3\)http://skyserver.sdss.org/CasJobs/
Figure 6.5: The galaxy count distribution of the training set, on a 2D grid of $g - r$, $r - i$ broad-band colours. The published 3D map also includes the $r$-band magnitude, this 2D version is used here for illustrative purposes. The discontinuity is caused by the colour cut of the CMASS sample (Dawson et al., 2013). See the text for a discussion.
Figure 6.6: Photometric redshift estimation results for the training set, on a 2D grid of $g - r$, $r - i$ broad-band colours. The published 3D map also includes the $r$-band magnitude, this 2D version is used here for illustrative purposes. Panel (a) shows the average standard deviation of the redshifts of the nearest neighbours ($\sigma(z_{NN})$), panel (b) displays the rms of $z_{\text{phot}} - z_{\text{spec}}$, the actual estimation error, while panel (c) shows the average reported estimation error ($\delta z_{\text{phot}}$). We note that the outliers have not been removed from the rms computation, therefore panel (b) is noisier. See the text for a discussion.
the desired accuracy may be achieved with a cut based on the redshift error estimate, e.g. only using galaxies with $\delta z_{\text{phot}} < 0.03$.

The linear fitting algorithm can fail when it encounters a singular or near-singular matrix — such cases are indicated by $z_{\text{phot}} = -9999$, therefore those galaxies should be excluded. If required, the redshift of the first nearest neighbour, and the average redshift of the 100 nearest neighbours are still available, however, in this case, there is no redshift error estimate (also flagged with $\delta z_{\text{phot}} = -9999$).

When small biases are a critical issue, the redshift error map of Sect. 6.3.6 should be used for leaving out galaxies that are located in a high-error region in colour space, but that otherwise have a low reported $\delta z_{\text{phot}}$.

Additionally, the volume of the bounding box of the nearest neighbours in the colour space could also be used for filtering — a volume that is very large means that galaxies of very different colours are used in the local linear regression, therefore the estimated redshift could be compromised. To give an idea of potential limits, a bounding box volume cut of $\text{nnVol} < 2$ filters out $\approx 2.5\%$ of training set galaxies that are in the sparsest colour space regions, $\text{nnVol} < 1$ eliminates $\approx 5\%$, while $\text{nnVol} < 0.45$ cuts $\approx 10\%$. The galaxies thus filtered out have estimation accuracies of $\sigma (\Delta z_{\text{norm}}) = 0.0464$, 0.0400 and 0.0342, respectively, with the $3\sigma$ outlier rate around 7% for all three cases.

In Fig. 6.7, we illustrate the hazards of using non-filtered photometric redshift data. Without implementing any cuts, the errors and the number of catastrophic failures are visibly much larger. Also, it is important to remember that we are only analysing spectroscopic measurements, which on average have much more accurate photometry than the rest of the SDSS photometric catalog — the fraction of catastrophic outliers is expected to be much larger for the unfiltered photometric sample. On the other hand, using too stringent cuts might unnecessarily limit the redshift coverage, or the colour space coverage of the sample. For this reason, it is recommended to experiment with different filtering choices to find the one most appropriate for the task at hand.

### 6.3.8 Discussion of results

In Sect. 6.3, I described in detail how we created the photometric redshift database of SDSS DR12.

I specified the local linear regression method that we use for the redshift and redshift
Figure 6.7: The photometric redshift ($z_{\text{phot}}$) as a function of spectroscopic redshift ($z_{\text{spec}}$), for three different subsets of all available spectroscopic measurements. The galaxy density is shown in grayscale — we took the logarithm of galaxy counts so that even individual galaxies can be seen. The red solid, dashed and dotted lines represent the median, 68% and 95% confidence regions of the data, respectively. On panel (a), there is no selection, all galaxies are shown. On panel (b), we only included galaxies of photometric error class 1 or -1, and with a reported redshift error of $\delta z_{\text{phot}} < 0.03$. On panel (c), galaxies of photometric error class 1 and with $\delta z_{\text{phot}} < 0.02$ are shown. We note that on panel (c), the more biased region around $z_{\text{spec}} = 0.4$ has been almost completely filtered out. See the text for a discussion.
error estimation, and described the spectral template fitting step that followed it. I gave an account of the data and methods that went into assembling the training set. I detailed how we evaluated the accuracy of our estimation via cross-validation on the training set, then discussed the errors and biases that we encountered. I presented the photometric error classes that were introduced, and the 3D redshift error map that helps quantify the errors and filter out inaccurately estimated galaxies. I also provided recommendations for using the database, and choosing appropriate filtering criteria.

Our photometric redshift estimates are relatively accurate, with a standard deviation of \( \sigma(\Delta z_{\text{norm}}) = 0.0205 \), and an acceptable 3\( \sigma \) outlier rate of \( P_o = 4.11\% \). The reported redshift error is a realistic estimate of the actual redshift estimation error (see Fig. 6.4). While we observed redshift-dependent biases of up to \( \Delta z = 0.01 \), the \( z_{\text{phot}} \pm \delta z_{\text{phot}} \) confidence intervals provide a reasonably good approximation of the spectroscopic redshift (see Sect. 6.3.4). However, from \( z \approx 0.6 \), the coverage of our training set drops sharply, therefore so does the accuracy of our photometric redshifts.

In addition to the redshift error estimate, we provide further tools that allow users to select measurements of the desired accuracy. These include the photometric error class, the 3D redshift error map, and the bounding box volume of the nearest neighbours (see Sect. 6.3.5, Sect. 6.3.6 and Sect. 6.3.7, respectively).

As opposed to purely empirical methods, our hybrid method fits a spectral template, which allows us to provide K-corrections and absolute magnitudes (see Sect. 6.3.2 and Sect. 6.3.9).

In later SDSS data releases, we intend to expand our training set as new data comes available, and also review our training set and methods with the intention of reducing biases and extending the useful coverage to higher redshifts. Having a less sparse sampling of high-redshift galaxies in photometric colour space would help reduce the pronounced negative bias in their redshift estimation.

The photometric redshift database, corresponding documentation and tools are available online on the appropriate SDSS DR12 webpages.

### 6.3.9 The photometric redshift tables in SDSS DR12

Here I give a description of each column in the published tables, either referencing a concept used in this section, or detailing it here. With \{ugriz\} I denote that there
is a column for each of the five SDSS ugriz broad-band magnitudes, with the single corresponding (capitalized) letter present in the column name.

The Photoz table

- **objID** — the SDSS objID of the query galaxy.
- **z** — $z_{\text{phot},i}$ in Eq.6.1, i.e. the photometric redshift. It takes the value $-9999$ when there was an error in the fitting algorithm.
- **zErr** — $\delta z_{\text{phot},i}$ in Eq.6.3, i.e. the photometric redshift error estimate. It takes the value $-9999$ when there was an error in the fitting algorithm.
- **nnCount** — $l$, the number of nearest neighbours used in the local linear regression, with outliers excluded from the total of $k = 100$, as described in Sect. 6.3.1. It takes the value $-9999$ when there was an error in the fitting algorithm.
- **nnVol** — the volume of the bounding box of the $k = 100$ nearest neighbours.
- **photoErrorClass** — the photometric error class described in Sect. 6.3.5, Tab. 6.2 and Tab. 6.3.
- **nnObjID** — the SDSS objID of the first nearest neighbour.
- **nnSpecz** — the spectroscopic redshift ($z_{\text{spec}}$) of the first nearest neighbour.
- **nnFarObjID** — the SDSS objID of the farthest, 100th nearest neighbour.
- **nnAvgZ** — the average redshift of the $k = 100$ nearest neighbours.
- **distMod** — the distance modulus ($DM_i$) corresponding to $z$, if available, or $nnAvgZ$. See the end of Sect. 6.3 for the adopted cosmology.
- **lumDist** — the luminosity distance in $Mpc$ corresponding to $z$, if available, or $nnAvgZ$. See the end of Sect. 6.3 for the adopted cosmology.
- **chisq** — the $\chi^2$ value of the spectral template fit, i.e. $\chi^2 = \sum_{p=1}^{D} \left( \frac{m_{p,i} - (s_p(z=z_i,t=t_i) - m_{0,i})}{\sigma_{p,i}} \right)^2$, using the notation of Eq. 6.11.
• **rnorm** — the residual Euclidean norm of the spectral template fit, i.e.
\[
\left(\sum_{p=1}^{D} (m_{p,i} - (s_p(z = z_i, t = t_i) - m_{0,i}))^2\right)^{0.5},
\] using the notation of Eq. 6.11.

• **bestFitTemplateID** — \(t_i\), the identifier of the best-fitting spectral template. See Tab. 6.4 for the corresponding names in Dobos et al. (2012b).

• **synth\{ugriz\}** — the synthetic magnitude of the best-fitting spectral template, i.e. \(s_p(z = z_i, t = t_i) - m_{0,i}\), using the notation of Eq. 6.11.

• **kcorr\{ugriz\}** — the \(K\)-correction to \(z = 0\), i.e. \(K_{p,i}(z = 0) = s_p(z = z_i, t = t_i) - s_p(z = 0, t = t_i)\), using the notation of Eq. 6.11.

• **kcorr\{ugriz\}01** — the \(K\)-correction to \(z = 0.1\), i.e. \(K_{p,i}(z = 0.1) = s_p(z = z_i, t = t_i) - s_p(z = 0.1, t = t_i)\), using the notation of Eq. 6.11.

• **absMag\{ugriz\}** — the rest-frame absolute magnitude of the galaxy, i.e. \(m_{p,i} - K_{p,i}(z = 0) - DM_i\), using the notation of Eq. 6.11.

The **PhotozErrorMap** table

• **CellID** — The unique identifier of the cell in the grid. The grid spans the \(r\)-band magnitude, and the \(g - r\), \(r - i\) colours.

• **rMag** — The centerpoint of the cell in \(r\)-band magnitude. Linear size of a cell: 0.5.

• **gMag_Minus_rMag** — The centerpoint of the cell in \(g - r\) colour. Linear size of a cell: 0.01.

• **rMag_Minus_iMag** — The centerpoint of the cell in \(r - i\) colour. Linear size of a cell: 0.01.

• **countInCell** — The number of training set galaxies within the cell (denoted below with \(N\)).

• **avgPhotoZ** — The average photometric redshift of training set galaxies in the cell, i.e. \(\frac{\sum_{i=1}^{N} \zeta_{\text{phot},i}}{N}\), using the notation of Sect. 6.3.1.
<table>
<thead>
<tr>
<th>$t_i$</th>
<th>Name</th>
<th>$t_i$</th>
<th>Name</th>
<th>$t_i$</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Red P</td>
<td>15</td>
<td>Blue P</td>
<td>29</td>
<td>RED 1</td>
</tr>
<tr>
<td>2</td>
<td>Red H$\alpha$</td>
<td>16</td>
<td>Blue H$\alpha$</td>
<td>30</td>
<td>RED 2</td>
</tr>
<tr>
<td>3</td>
<td>Red SF</td>
<td>17</td>
<td>Blue SF</td>
<td>31</td>
<td>RED 3</td>
</tr>
<tr>
<td>4</td>
<td>Red A+HII</td>
<td>18</td>
<td>Blue A+HII</td>
<td>32</td>
<td>RED 4</td>
</tr>
<tr>
<td>5</td>
<td>Red L</td>
<td>19</td>
<td>Blue L</td>
<td>33</td>
<td>RED 5</td>
</tr>
<tr>
<td>6</td>
<td>Red S</td>
<td>20</td>
<td>Blue S</td>
<td>34</td>
<td>SF 1</td>
</tr>
<tr>
<td>7</td>
<td>Red all</td>
<td>21</td>
<td>Blue all</td>
<td>35</td>
<td>SF 2</td>
</tr>
<tr>
<td>8</td>
<td>Green P</td>
<td>22</td>
<td>All P</td>
<td>36</td>
<td>SF 3</td>
</tr>
<tr>
<td>9</td>
<td>Green H$\alpha$</td>
<td>23</td>
<td>All H$\alpha$</td>
<td>37</td>
<td>SF 4</td>
</tr>
<tr>
<td>10</td>
<td>Green SF</td>
<td>24</td>
<td>All SF</td>
<td>38</td>
<td>SF 5</td>
</tr>
<tr>
<td>11</td>
<td>Green A+HII</td>
<td>25</td>
<td>All A+HII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Green L</td>
<td>26</td>
<td>All L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Green S</td>
<td>27</td>
<td>All S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Green all</td>
<td>28</td>
<td>All all</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: The name in Dobos et al. (2012b) that corresponds to the $t_i$ (or bestFitTemplateID) template identifier used in the SDSS DR12 Photoz table.
• **avgSpectroZ** — The average spectroscopic redshift of training set galaxies in the cell, i.e. \( \frac{\sum_{i=1}^{N} z_{\text{spec},i}}{N} \), using the notation of Sect. 6.3.1.

• **avgRMS** — The rms of the redshift estimation error for training set galaxies in the cell, i.e. \( \left( \frac{\sum_{i=1}^{N} (z_{\text{phot},i} - z_{\text{spec},i})^2}{N} \right)^{0.5} \), using the notation of Sect. 6.3.1.

• **avgEstimatedError** — The average redshift error estimate for training set galaxies in the cell, i.e. \( \frac{\sum_{i=1}^{N} \delta z_{\text{phot},i}}{N} \), using the notation of Sect. 6.3.1.

• **avgNeighborZStDev** — The average standard deviation of the redshifts of the \( k = 100 \) nearest neighbours, for every training set galaxy in the cell. Denoting the standard deviation of the \( z_{\text{spec}} \) of the neighbours with \( \sigma_i (z_{\text{NN}}) \), it is \( \frac{\sum_{i=1}^{N} \sigma_i (z_{\text{NN}})}{N} \).

### 6.4 Photo-z-SQL — redshift estimation in a database

Large photometric catalogs are most often stored in relational databases (e.g. SkyServer for SDSS), and cross-matching tools that connect such databases also use the relational model (Dobos et al., 2012a; Budavari, Dobos & Szalay, 2013; Han et al., 2016). It would be advantageous to have a photo-z implementation that can work directly on data in the servers, and on results from cross-matches, by integrating into the well-established SQL language. This section presents research work performed with this goal in mind, published in Beck et al. (2017a).

There are various software packages available that utilize either the template-based or the empirical approach to perform photometric redshift estimation (see Sect. 6.2 for an introduction). However, to the best of our knowledge, none of the public codes have the ability to perform the computations within the database (DB) itself. Thus, they require moving the photometric catalogs outside of the database, and the results also have to be loaded into the DB separately. This is a cumbersome process, especially when the amount of data is in the PB range, as will be the case with upcoming surveys such as the LSST (Ivezic et al., 2008) and Pan-STARRS (Tonry et al., 2012). The problem is also acute when on-demand photo-z would be needed to quickly process cross-match results.

Additionally, existing template-based photometric redshift estimation codes use a predefined set of broad-band photometric filters, and compute synthetic magnitudes in all of these bands before the actual photo-z calculations. This approach is not ideal for heterogeneous observations, such as the Hubble Source Catalog (Whitmore et al., 2016), where each object can have a different corresponding filter set, and the total number of filters is much larger than the number of available filters for an object. Also, this issue can become especially prominent if we consider that filter transmission curves can change over time — if this time-dependence is to be properly taken into account, precomputing synthetic magnitudes is simply not an option.

Even with all these difficulties, the template-based approach is at least feasible in such situations, while building a training set for empirical estimation is almost impossible when the input filter set is not static, and not defined beforehand. Moreover, cross-match results are generally not extensive enough to provide complete training set coverage in redshift and colour-magnitude space, and empirical methods usually do not perform well when substantial extrapolation is necessary (Beck et al., 2017b).

Thus, we decided to adopt the template-based photo-z approach (refer to Sect. 6.2.2 for more details), and sought to address the issues of existing codes with our own implementation, named Photo-z-SQL (Beck et al., 2017a). We apply a computational innovation to deal with variable filter sets — we use lazy evaluation for synthetic magnitudes, computing them only when they are needed, and caching them to be reused.

The code has been developed in the C# language, which can be run stand-alone, but can also be integrated into Microsoft SQL Server as a set of user-defined functions. Microsoft SQL Server is a commercial DB server application, which has been chosen because it enables using a general-purpose computer language via CLR integration (see Dobos et al., 2011, for another astronomy-related use case), and also because it is used by e.g. the SDSS and Pan-STARRS collaborations.

The database integration has the added benefit of utilizing the computational capabilities of DB servers — i.e. bringing the computations directly to the data —, which often have multiple processors specifically to handle computationally intensive tasks. Parallel execution is fully supported by our code, taking advantage of modern multicore systems. The objective-oriented power of C# also leads to flexibility in defining and providing priors, templates and filters.
6.4.1 Estimation method

The Photo-z-SQL code implements both template-based photo-z estimation approaches, i.e. the maximum likelihood and Bayesian methods, as described in Sect. 6.2.2 and Sect. 6.2.2, respectively.

The usually adopted assumption of uncorrelated magnitude errors may often simply not hold (Scranton et al., 2005), therefore our application does support using the full covariance matrix C, and thus the two expressions for $\chi^2$, Eq. 6.4 and Eq. 6.7. To our knowledge, this is a missing feature in other public codes.

In the actual Photo-z-SQL implementation for the Bayesian analysis, the integrals in Eq. 6.10 are discretized, replaced by a summation over a predetermined set of template parameters, a range in flux, and the points of a predetermined redshift grid, to finally yield the expression

$$P(z|m, C) = \frac{\sum_t \sum_{m_0} P(z, t, m_0) \exp(-\chi^2(z, t, m_0))}{\sum_{z'} \sum_t \sum_{m_0} P(z', t, m_0) \exp(-\chi^2(z', t, m_0))}. \quad (6.14)$$

The exact form of prior to use with Photo-z-SQL has to be chosen from the list currently available (see Sect. 6.4.2 and Sect. 6.4.4), but further priors could be easily added to the code. The output result is the $P(z|m, C)$ posterior redshift probability distribution, normalized to an integral probability of 1 on the given redshift grid.

6.4.2 Database integration

Microsoft SQL Server provides a unique opportunity among relational database management systems in its Common Language Runtime (CLR) integration feature. Compiled assemblies produced by any Common Language Infrastructure (CLI) programming language, which include C++, C# and F#, can be loaded into a DB on a server, thus granting access to code with arbitrarily complex functionality.

The main difficulty in SQL-CLR integration pertains to the SQL interface of the code, which has to translate between storage types of the programming language and the types available in SQL, and additionally it has to conform to the limitations of user-defined functions and user-defined stored procedures in SQL. Conversions between basic types such as integers, floating-point numbers and strings are fairly trivial, but more complex containers do need to be implemented on a case-by-case basis, relying on the binary storage types of SQL, or potentially on user-defined types. The publicly
available SqlArray library (Dobos et al., 2011; Dobos et al., 2012c) is a good example of this process, providing variable-length array functionality in SQL.

Most algorithms need to store data, maintain a state between different execution steps. In a SQL interface, functions called in subsequent queries would normally only be able to share data by storing it on a disk, in DB tables using SQL types. However, in our implementation we bypass this limitation, and maintain a state in-memory without needing to convert to SQL types (see Sect. 6.4.2).

The Photo-z-SQL library has been coded in C#. The interface consists of a set of user-defined SQL functions which can be used to first configure the photo-z calculation engine, and then perform the computations themselves. Parallel execution is fully supported to utilize the multi-core processors of typical DB servers. We provide install and uninstall scripts to make the installation process straightforward.

The remainder of this section gives more details about our implementation and interface in relation to the database integration.

**Implementation**

The data needs of the algorithm are the following:

- observed fluxes/magnitudes and errors for every object (optionally with a Schlegel, Finkbeiner & Davis, 1998, extinction map value),
- the transmission curves corresponding to the broad-band filters,
- the spectral templates that are considered in the fit,
- information about the prior,
- additional configuration details such as the resolution in redshift and luminosity space.

In addition to the fluxes/magnitudes, the filter set may also change from object to object, but the rest of the data are static in a single estimation run, and thus can be pre-loaded and stored in the DB server. Additionally, the synthetic magnitude cache corresponding to the given configuration also has to be stored.

There is a way to perform this storage in-memory, without moving data into temporary tables: a static C# class that uses the singleton pattern will retain its state in
Microsoft SQL Server between function calls. Therefore we created a singleton wrapper class that stores the photo-z configuration and synthetic magnitudes. Currently only a single, global configuration can exist in a DB, so multiple users would have to share the set of spectral templates, redshift grid and prior when estimating (but not the set of photometric filters, that can vary). In the future this can easily be extended to allow separate configurations, e.g. linked to a globally unique identifier (GUID) that corresponds to a user.

The DB server needs access to filter and template curves. While this could be solved by locally loading the required data into tables, it would put an extra burden on users to create them, and to ensure that the photo-z code accesses the data correctly. Instead, we decided to use the Spectrum Services and Filter Profile Services of the Virtual Observatory (VO)\(^5\). Users can choose from the existing templates and filters, or they can upload their own. This solution has the added benefit that the identifiers used in the Virtual Observatory can uniquely link the fluxes/magnitudes of objects to corresponding filters. The filter curves are also cached, thus they have to be downloaded from the VO only once.

The per-object data is expected to come from tables in the local DB server, supplied to the SQL functions of the Photo-z-SQL library. In order to allow variable-length array inputs of fluxes/magnitudes (and their errors), we used the \texttt{SqlArray} library (Dobos et al., 2011; Dobos et al., 2012c), which can also be installed into a DB server via a simple script.

The overall setup of Photo-z-SQL is illustrated in Fig. 6.8.

**SQL interface**

Interaction with the Photo-z-SQL code is achieved with a collection of user-defined SQL functions that are called from the DB. There are three types of functions, Config functions that set up the photo-z configuration, Compute functions that perform the photo-z estimation itself, and Util functions which e.g. help interfacing with \texttt{SqlArray}. When functions require a filter or template, they can be specified either with their VO integer identifier, or a complete URL address. These two options correspond to two versions of the given function, with a \_ID or \_URL suffix in the function name,

\(^5\text{http://voservices.net/}\)
Figure 6.8: An illustration of the architecture of Photo-z-SQL. Computing refers to the dynamic memory and computing cores of the server, while storage refers to the disk-based data storage. Mag is shorthand for magnitude, but fluxes could also be used.

respectively. Here we briefly list the main functions and their role, and an example query of a complete setup is provided in Sect. 6.4.8.

- **Config.SetupTemplateList** Specifies the list of SED templates to use in the photo-z estimation, along with the resolution in redshift and luminosity. Either the number of steps around the best-fitting luminosity can be given, or alternatively a range of physical luminosities (LuminositySpecified suffix).

- **Config.SetupExtinctionLaw** Specifies the reference spectrum, $R_V$ dust parameter and extinction law to use when correcting for Galactic extinction (see Sect. 6.4.3 for more details).

- **Config.RemoveExtinctionLaw** Removes the data associated with the extinction law – Galactic extinction correction is no longer applied.

- **Config.SetupFlatPrior** Specifies that a flat prior should be used (this is the default prior).

- **Config.SetupBenitezHDFPrior** Specifies that the HDF-N prior of Benítez (2000) should be used (see Sect. 6.4.4 for more details).

- **Config.SetupAbsoluteMagnitudeLimitPrior** Specifies a maximum limit in absolute magnitude as the prior, in the provided reference filter. Cosmological parameters are also needed to define the distance–redshift relation.
• **Config.SetupTemplateTypePrior** Specifies a prior that assigns a given probability to each of the used SED templates.

• **Config.AddMissingValueSpecifier** For convenience, specifies a value that denotes missing or otherwise invalid inputs. All such inputs are ignored.

• **Config.RemoveInitialization** Removes all data connected to the photo-z configuration, including cached filters, templates and synthetic magnitudes.

• **Compute.PhotoZMinChiSqr** Performs maximum likelihood photo-z estimation, returning a single scalar $z_{\text{phot}}$ value (see Sect. 6.2.2 and Sect. 6.4.1). The prior is not used here, and luminosity is only evaluated at the best-fitting value. Magnitudes (or fluxes) and their errors, and corresponding filter identifiers are needed. An extinction map value can be specified, and the fit itself can be done in either magnitude or flux space. Also, an extra uncorrelated variance term can be added to the observational errors (see Sect. 6.4.4).

• **Compute.PhotoZBayesian** Performs Bayesian photo-z estimation, returning the entire posterior redshift probability distribution within the specified coverage as a SQL table (see Sect. 6.2.2 and Sect. 6.4.1 for more details). Input parameters are the same as in Compute.PhotoZMinChiSqr.

### 6.4.3 Implementation highlights

In this section, we detail some of the features of the implementation that could be noteworthy to users.

**Dynamic filter handling**

The computation of synthetic magnitudes is relatively expensive, as it involves integrating a filter curve with an SED curve. This has to be done – ideally only once – for every filter and SED pairing, at every considered redshift. Available photo-z software generally solve this issue by precomputing the entire synthetic magnitude table before the actual photo-z computations. As we noted in Sect. 6.4, when the selection of filters changes between objects, or there is a large number of filters from which only a few
are available at a time, or especially when filter time-dependence is to be taken into account, this choice is far from optimal.

In our implementation, we instead chose to set up a synthetic magnitude cache. The memory addresses of the filters serve as keys in a hash table, while the corresponding values are arrays of synthetic magnitudes. The arrays are indexed by the given template parameters and redshift. Whenever a synthetic magnitude is needed, it is retrieved from the cache if available, or computed and then stored in the cache if not. Therefore values are computed only when necessary, and only once.

With our approach, there is no requirement to know beforehand which filters will be needed. While it does include an additional hash table lookup whenever a synthetic magnitude is accessed, that is an inexpensive operation. Additionally, all functions working on the cache have been programmed to support parallel access. Currently, the cache has to be cleared by the user via a function call, but it would be possible to implement e.g. the deletion of old synthetic magnitudes after a time or cache size limit has been passed.

**Magnitudes and fluxes**

All of our photo-z algorithms have been programmed to be able to perform computations using either magnitudes or fluxes. When fluxes are used, the additive $m_0$ flux scaling parameter is replaced by a multiplicative $f_0$ parameter, but the best-fitting value of $f_0$ can still be determined algebraically for a given template SED and redshift (see Eq. 6.4 and Sect. 6.2.2).

Currently, our code supports using the AB magnitude system (Oke & Gunn, 1983), and the SDSS asinh magnitude system (Lupton, Gunn & Szalay, 1999), but additional systems could be readily included. Functions are available for converting magnitudes and magnitude errors into fluxes and flux errors, and vice versa – conversions between magnitude systems also use this mechanism. The conversions assume a Gaussian distribution for errors, therefore users should take care not to perform such conversions when this assumption does not hold, e.g. in the case of faint objects where the flux errors are Gaussian, but the magnitude errors are not.
Zeropoint calibration and error scaling

Ideally, all photometric magnitude measurements should conform to the theoretical flux zeropoints of the given magnitude system, which, for example, is 3631 Jy for the AB magnitude system (Oke & Gunn, 1983). However, there are cases when this assumption is incorrect, and the actual zeropoint of a broad-band filter is slightly different (Doi et al., 2010).

To mitigate this issue, we implemented an optional zeropoint calibration algorithm which uses objects with known redshifts. The uncorrelated assumption is adopted here (see Eq. 6.6), and the best-fitting template is determined at the given redshift using Eq. 6.7 and Eq. 6.5 for every object $j$ in the $T$ training set (refer to Sect. 6.2.2). Then, for each of the broad-band filters, indexed by $i$, a similar algebraic minimization is performed to find the filter-dependent zeropoint shift in magnitude, expressed by

$$m_i = \arg \min_{m'\in T} \sum_{j\in T} \frac{(m_{i,j} - m'_i - (s_i(z_j, t_j) - m_{0,j}))^2}{\sigma_{i,j}^2}.$$  \hspace{1cm} (6.15)

Since the zeropoint shift can change the best-fitting template $t_j$ and the corresponding $m_{0,j}$, the two minimization steps are repeated iteratively – with the latest $m_i$ shift applied to the $m_{i,j}$ measured magnitudes – until a chosen zeropoint precision is achieved.

There is an additional, optional refinement to the calibration that can be applied after the iteration. If our assumptions are correct, and the data are well-described by the template spectra, the residuals scaled by the estimated photometric errors should follow the standard normal distribution. This will not be the case if the photometric errors are underestimated in a band, but the errors could be scaled by a factor

$$\alpha_i = \sqrt{\frac{1}{N_T-1} \sum_{j\in T} \frac{(m_{i,j} - m_i - (s_i(z_j, t_j) - m_{0,j}))^2}{\sigma_{i,j}^2}}$$  \hspace{1cm} (6.16)

to ensure that the scaled residual distribution

$$\sigma_{i,j}^{\text{scaled}} = \alpha_i \sigma_{i,j}$$  \hspace{1cm} (6.17)

has a standard deviation of 1. Above, $N_T$ is the number of objects in the training set, and the error scaling factor is linked to a given filter $i$. Thus, problematic passbands with ill-estimated errors or a higher proportion of badly matching templates can be downweighted during the actual photo-z estimation.
We note that the same calibration algorithm can also be performed using fluxes, in that case the \( f_i \) zeropoint flux multiplier is estimated for each filter.

**Galactic extinction**

We provide a built-in implementation for correcting broad-band magnitudes for Galactic extinction using the IR dust map of Schlegel, Finkbeiner & Davis (1998).


The reference galaxy spectrum used in the computation can be specified freely, but it should have coverage in the entire wavelength range of the broad-band filters for which the correction is applied.

### 6.4.4 Configuration details

For the tests described in Sect. 6.4.5 and Sect. 6.4.6, we adopted some of the more successful approaches in the literature (Hildebrandt et al., 2010; Dahlen et al., 2013). In our setups, we use two different sets of galaxy template SEDs, the Hubble UDF set of Coe et al. (2006), denoted by \( BPZ \), and the COSMOS set of Ilbert et al. (2009), indicated by \( LP \). In the case of the \( BPZ \) set, following Coe et al. (2006) we linearly interpolated 9 galaxies between each of the 8 neighboring templates to generate a total of 71 SEDs. The wavelength coverage of these templates ends at 25600 Å, after which they were linearly extrapolated. For the \( LP \) set, adopting the choices of Ilbert et al. (2009) we used the Le PHARE code (Arnouts et al., 2002; Ilbert et al., 2006b) to add emission lines of different fluxes, and to apply a selection of extinction laws with different parameters to the templates, yielding a total of 641 SEDs.

The Bayesian estimation method was selected. The resolution of the redshift grid was taken to be 0.01. We use two different priors, a simple flat prior (indicated by \( Flat \)), and the apparent \( I \)-band magnitude, galaxy type and redshift prior of Benítez (2000), empirically calibrated on HDF-N data (denoted by \( HDF \)). The latter prior has been adapted to the \( LP \) template set based on galaxy type, distributing the total probability
of a given type evenly among its instances. Since a measured $I$-band magnitude may not be available, the synthetic $I$-band magnitude corresponding to the given parameters is used as a proxy to allow applying this prior in all situations.

We found that adding an independent magnitude variance term to the measured magnitude variance can improve the estimation results. This extra error term can represent uncertainties in Galactic extinction, or in the template SEDs themselves, and it prevents single filters with very low errors from placing too stringent limits on the fitted templates. Thus, as an additional refinement, we may add an extra 0.01 mag of error when using the $LP$ set, or 0.02 mag in the case of the $BPZ$ set (shown by the tag $Err$). The exact values were chosen based on the redshift estimation results of the $PHAT$ dataset (see Sect. 6.4.5), but have not been fine-tuned.

The combination of the two template sets, two priors, and whether or not extra error is added yields a total of 8 different configurations that we present.

### 6.4.5 Application results

There have been two recent large publications that allowed the comparison of photometric redshift estimation methods by publishing blind datasets for testing purposes. These two are “Photo-z Accuracy Testing” or $PHAT$ by Hildebrandt et al. (2010), and “A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation” by Dahlen et al. (2013), which we will refer to as $CAPR$. We use the public datasets of these articles to demonstrate the performance of our code.

**PHAT**

The PHAT1 dataset (Hildebrandt et al., 2010) contains 515 galaxies that have a spectroscopic redshift, with magnitude measurements in 14 different broad-band filters that span the wavelength range between 3000Å and 25000Å (here we do not use the extra 4 IRAC filters). The measures that we report are the $\Delta z_{\text{norm}}$ bias and $\sigma (\Delta z_{\text{norm}})$ standard deviation — excluding outliers — of the normalized redshift error, $\Delta z_{\text{norm}} = \frac{z_{\text{spec}} - z_{\text{phot}}}{1+z_{\text{spec}}}$, and also the $P_o$ percentage of outliers. Outliers are defined as having $|\Delta z_{\text{norm}}| > 0.15$. $z_{\text{phot}}$ is chosen to be the highest-probability redshift in the posterior redshift distribution defined in Eq. 6.14.

The results are presented in Tab. 6.5, while the spectroscopic–photometric redshift
Figure 6.9: The $z_{\text{phot}}$ photometric redshift as a function of the $z_{\text{spec}}$ spectroscopic redshift, for all the Photo-z-SQL configurations we ran when estimating the PHAT dataset. The text in the top left corner of each panel indicates the given setup. Outlying galaxies with $|\Delta z_{\text{norm}}| > 0.15$ are shown in light red, non-outlying galaxies in blue.

scatterplots are shown in Fig. 6.9. Applying the HDF prior is slightly beneficial for the smaller BPZ template set in that it reduces scatter while only marginally changing other measures, and it is somewhat detrimental in the case of the detailed LP SED set, mainly because of an increased bias. Adding the extra error term reduces the estimation bias considerably while also slightly reducing the scatter, however, the outlier rate is marginally increased. All in all, our results are comparable to those of the better-performing methods in Table 5 of Hildebrandt et al. (2010), whose typical values were: $\overline{\Delta z_{\text{norm}}} \approx 0.004 - 0.0011$, $\sigma (\Delta z_{\text{norm}}) \approx 0.038 - 0.048$ and $P_o \approx 9.2\% - 13.5\%$.

We note that the PHAT dataset is not large enough to warrant performing the calibration detailed in Sect. 6.4.3 – with such a small sample, the calibration generally converges to a local minimum in filter zeropoints (as opposed to the global one), improving performance on the calibration set, but actually decreasing it on the validation set. Additionally, the Schlegel, Finkbeiner & Davis (1998) dust map value corresponding to the galaxies was not published, and neither were the coordinates on the sky, therefore Galactic extinction has not been taken into account.
<table>
<thead>
<tr>
<th>Configuration</th>
<th>$\Delta z_{\text{norm}}$</th>
<th>$\sigma (\Delta z_{\text{norm}})$</th>
<th>$P_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPZ Flat</td>
<td>0.0114</td>
<td>0.0494</td>
<td>9.91%</td>
</tr>
<tr>
<td>BPZ HDF</td>
<td>0.0119</td>
<td>0.0484</td>
<td>9.94%</td>
</tr>
<tr>
<td>BPZ Flat Err</td>
<td>0.0029</td>
<td>0.0486</td>
<td>10.34%</td>
</tr>
<tr>
<td>BPZ HDF Err</td>
<td>0.0026</td>
<td>0.0462</td>
<td>10.77%</td>
</tr>
<tr>
<td>LP Flat</td>
<td>0.0049</td>
<td>0.0466</td>
<td>9.36%</td>
</tr>
<tr>
<td>LP HDF</td>
<td>0.0055</td>
<td>0.0467</td>
<td>9.38%</td>
</tr>
<tr>
<td>LP Flat Err</td>
<td>0.0013</td>
<td>0.0445</td>
<td>9.88%</td>
</tr>
<tr>
<td>LP HDF Err</td>
<td>0.0025</td>
<td>0.0448</td>
<td>10.08%</td>
</tr>
</tbody>
</table>

Table 6.5: Photo-z-SQL numerical results for the PHAT dataset. The table lists the $\Delta z_{\text{norm}}$ average bias and $\sigma (\Delta z_{\text{norm}})$ standard deviation of the $\Delta z_{\text{norm}} = \frac{z_{\text{spec}} - z_{\text{phot}}}{1 + z_{\text{spec}}}$ normalized redshift estimation error, and the $P_o$ outlier rate with $|\Delta z_{\text{norm}}| > 0.15$ outliers, for each configuration we ran.

**CAPR**

The CAPR dataset (Dahlen et al., 2013) contains 589 galaxies intended for redshift estimation, along with a training set of 580 galaxies intended for calibration. There are flux measurements in 14 different passbands, of which we do not use the the IRAC 5.8$\mu$m and 8.0$\mu$m channels, similarly to code H (Le PHARE) in Dahlen et al. (2013). This way, the wavelength coverage is between 3000Å and 50000Å. The IRAC passbands are problematic because they probe wavelength ranges where the spectral templates at low redshifts are not as reliable, and where Galactic extinction is a more significant factor (Dahlen et al., 2013). As with the PHAT dataset, there were no published IR dust map values, therefore we did not take Galactic extinction into account.

First, our code was executed without performing any calibration. Again, we report the same numeric measures as in Sect. 6.4.5, collated in Tab. 6.6. Refer to Fig. 6.10 for the redshift estimation scatterplots. Whether or not the HDF prior was applied does not significantly impact the results, and the additional error term leads to a small improvement in the case of the LP template set, but slightly worse bias and scatter for the BPZ template set. When compared with the results in Table 2 of Dahlen et al. (2013) (bias, $\sigma_O$ and OLF columns), where typical values were: $\Delta z_{\text{norm}} \approx 0.005 - 0.023$, $\sigma (\Delta z_{\text{norm}}) \approx 0.034 - 0.064$ and $P_o \approx 3.9\% - 9.3\%$, the BPZ setups are among the
Table 6.6: Photo-z-SQL numerical results for the CAPR dataset, without calibration. The table lists the $\Delta z_{\text{norm}}$ average bias and $\sigma (\Delta z_{\text{norm}})$ standard deviation of the $\Delta z_{\text{norm}} = \frac{z_{\text{spec}} - z_{\text{phot}}}{1 + z_{\text{spec}}}$ normalized redshift estimation error, and the $P_o$ outlier rate with $|\Delta z_{\text{norm}}| > 0.15$. 15 outliers, for each configuration we ran.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$\Delta z_{\text{norm}}$</th>
<th>$\sigma (\Delta z_{\text{norm}})$</th>
<th>$P_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPZ Flat</td>
<td>-0.0359</td>
<td>0.0732</td>
<td>19.84%</td>
</tr>
<tr>
<td>BPZ HDF</td>
<td>-0.0352</td>
<td>0.0727</td>
<td>20.49%</td>
</tr>
<tr>
<td>BPZ Flat Err</td>
<td>-0.0428</td>
<td>0.0747</td>
<td>17.76%</td>
</tr>
<tr>
<td>BPZ HDF Err</td>
<td>-0.0421</td>
<td>0.0742</td>
<td>18.45%</td>
</tr>
<tr>
<td>LP Flat</td>
<td>-0.0083</td>
<td>0.0514</td>
<td>8.54%</td>
</tr>
<tr>
<td>LP HDF</td>
<td>-0.0083</td>
<td>0.0514</td>
<td>8.54%</td>
</tr>
<tr>
<td>LP Flat Err</td>
<td>-0.0083</td>
<td>0.0492</td>
<td>7.65%</td>
</tr>
<tr>
<td>LP HDF Err</td>
<td>-0.0081</td>
<td>0.0490</td>
<td>7.65%</td>
</tr>
</tbody>
</table>

worse performers, while the LP configurations are in the middle of the pack. However, it should be noted that the better performers all included some sort of training using the spectroscopic sample.

For the sake of a better comparison, we tested our calibration algorithm, utilizing the published training set. We selected the BPZ HDF Err and LP Flat Err configurations to illustrate the effects of calibration – the prior does not have a significant influence on results, and the added error was necessary because a few galaxies with very low estimated errors can adversely impact the calibration of a filter. As discussed in Sect. 6.4.3, we can either only perform zeropoint calibration (ZP), or both zeropoint calibration and photometric error scaling (ZP+E). The results are presented in Tab. 6.7, and the estimation scatterplots are shown in Fig. 6.11. Even with the calibration, the BPZ template set appears inadequate in describing the data, probably because its “proper” wavelength coverage ends at 25600Å. However, after calibration the results of the LP set are similar to what the better performers in Dahlen et al. (2013) can produce — the only exception is the higher outlier fraction, which is dominated by high-redshift galaxies: 49% of outliers have $z > 3$, as opposed to only 8% of the whole sample. In fact, the fraction of $z > 3$ outliers jumps from 39% to 52% because of the zeropoint calibration, and to 76% with the error scaling applied. This is because the training
Figure 6.10: The $z_{\text{phot}}$ photometric redshift as a function of the $z_{\text{spec}}$ spectroscopic redshift, for all the Photo-z-SQL configurations we ran when estimating the CAPR dataset, without calibration. The text in the top left corner of each panel indicates the given setup. Outlying galaxies with $|\Delta z_{\text{norm}}| > 0.15$ are shown in light red, non-outlying galaxies in blue.
Table 6.7: Photo-z-SQL numerical results for the CAPR dataset, including calibration. The table lists the $\Delta z_{\text{norm}}$ average bias and $\sigma (\Delta z_{\text{norm}})$ standard deviation of the $\Delta z_{\text{norm}} = \frac{z_{\text{spec}} - z_{\text{phot}}}{1+z_{\text{spec}}}$ normalized redshift estimation error, and the $P_o$ outlier rate with $|\Delta z_{\text{norm}}| > 0.15$ outliers. ZP denotes only zeropoint calibration, while ZP+E indicates zeropoint calibration and error scaling.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$\Delta z_{\text{norm}}$</th>
<th>$\sigma (\Delta z_{\text{norm}})$</th>
<th>$P_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPZ HDF Err</td>
<td>-0.0421</td>
<td>0.0742</td>
<td>18.45%</td>
</tr>
<tr>
<td>BPZ HDF Err ZP</td>
<td>0.0117</td>
<td>0.0615</td>
<td>15.45%</td>
</tr>
<tr>
<td>BPZ HDF Err ZP+E</td>
<td>0.0200</td>
<td>0.0631</td>
<td>15.62%</td>
</tr>
<tr>
<td>LP Flat Err</td>
<td>-0.0083</td>
<td>0.0492</td>
<td>7.65%</td>
</tr>
<tr>
<td>LP Flat Err ZP</td>
<td>0.0076</td>
<td>0.0445</td>
<td>9.85%</td>
</tr>
<tr>
<td>LP Flat Err ZP+E</td>
<td>0.0123</td>
<td>0.0402</td>
<td>12.05%</td>
</tr>
</tbody>
</table>

set is also mainly made up of low-redshift galaxies, and at different redshifts different passbands become important. Additionally, with low-redshift galaxies dominating the calibration, the less well-modeled higher wavelength ranges in the templates will lead to a larger than needed assigned uncertainty in the case of higher wavelength filters, downweighting filters that should constrain the photo-z estimation at high $z$. This issue could be solved by upweighting less well populated redshift bins in the calibration, and by applying a wavelength-dependent template error correction e.g. following Brammer, van Dokkum & Coppi (2008).

6.4.6 Performance analysis

In this section, we provide an analysis of the computational performance and memory usage of the Photo-z-SQL code. We test how these factors scale when varying the details of the adopted configuration.

The test system was a slightly outdated server computer with a 16-core, 2.2 GHz Intel Xeon processor running Microsoft SQL Server 2012. With hyper-threading, this system can support 32 parallel threads.

The server could and was used by multiple users at the time, but no other queries were running that were as computationally intensive or lengthy as the Photo-z-SQL testing queries, leading to relatively stable benchmark results.
Figure 6.11: The $z_{\text{phot}}$ photometric redshift as a function of the $z_{\text{spec}}$ spectroscopic redshift, for the Photo-z-SQL calibration tests we ran on the CAPR dataset. The text in the top left corner of each panel indicates the given setup. Outlying galaxies with $|\Delta z_{\text{norm}}| > 0.15$ are shown in light red, non-outlying galaxies in blue.
When first specifying the photo-z configuration, the synthetic magnitude cache is empty, and it will be filled during the starting phase of the first estimation query. Any subsequent queries that use the same filters will then work with the existing magnitude cache. For this reason, for each configuration we execute the same query twice, recording two times, $T_{\text{first}}$ and $T_{\text{second}}$. $T_{\text{first}}$ includes the synthetic magnitude computation for the first few galaxies, the estimation time for the rest of the galaxies, and any overhead from the initial setup and the transmission of results. $T_{\text{second}}$ includes the estimation time for all the galaxies, and again any overhead. Therefore $T_{\text{cache}} \equiv T_{\text{first}} - T_{\text{second}}$ is a good measure of the time required to fill the magnitude cache, and using the number of galaxies $N_{\text{gal}}$, $T_{\text{gal}} \equiv T_{\text{second}}/N_{\text{gal}}$ is a reasonable estimate of the time required to perform photo-z on a single galaxy.

One of our goals is to quantify memory usage, however, C# is a garbage-collected language with automatically managed memory, and it would be very difficult to track the memory used by different objects in a detailed way. Still, we can measure the total CLR memory usage $\text{MEM}_{\text{total}}$, which is the final metric in our analysis.

We use the PHAT dataset in our tests, with $N_{\text{gal}} = 515$ (see Sect. 6.4.5). Each test run was performed five times, and we provide the average and standard deviation of the resulting values. In Tab. 6.8, we report the metrics defined above for the different configurations introduced in Sect. 6.4.4. We only mention the four Err configurations, since the increased error values do not impact performance. As expected, having $\approx 9$ times as many template SEDs in the LP configurations scales up runtime and memory usage by a correspondingly large factor of $\approx 8$. Additionally, applying the HDF prior, which has to be evaluated in the innermost loop, more than doubles the per-galaxy estimation times. The synthetic magnitude computation time is not that much affected by the prior, since in that case the filter–spectrum integration is the most expensive operation.

We follow up our tests of the predetermined configurations with an analysis of how the metrics scale when different parameters are varied. Starting from the LP Flat Err configuration, again on the PHAT dataset, we modify the number of templates used between 30 and 630, the number of broad-band filters considered between 4 and the original 14, the number of redshift grid points between 100 and the original 600, and finally the number of allowed parallel threads between 1 and the original 32.

Results from these scaling tests are presented in Fig. 6.12. The most intriguing
Figure 6.12: The scaling of the execution time $T_{\text{second}}$ (in blue, left axes), the cache filling time $T_{\text{cache}}$ (black, left axes), and the total memory usage $MEM_{\text{total}}$ (light red, right axes) with respect to Photo-z-SQL configuration parameters. To allow the information in $T_{\text{cache}}$ and $T_{\text{gal}}$ to share the same y axis, we show $T_{\text{second}}$ in place of $T_{\text{gal}}$, since $T_{\text{second}}$ can be divided by 515 to yield $T_{\text{gal}}$. See the text for a discussion.
<table>
<thead>
<tr>
<th>Configuration</th>
<th>$T_{\text{first}}$ (s)</th>
<th>$T_{\text{second}}$ (s)</th>
<th>$T_{\text{cache}}$ (s)</th>
<th>$T_{\text{gal}}$ (s)</th>
<th>MEM$_{\text{total}}$ (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPZ Flat Err</td>
<td>49.7 ± 3.5</td>
<td>18.1 ± 1.1</td>
<td>31.6 ± 3.7</td>
<td>0.035 ± 0.002</td>
<td>101 ± 3</td>
</tr>
<tr>
<td>BPZ HDF Err</td>
<td>77.4 ± 1.7</td>
<td>41.4 ± 2.8</td>
<td>36.0 ± 3.3</td>
<td>0.080 ± 0.005</td>
<td>107 ± 3</td>
</tr>
<tr>
<td>LP Flat Err</td>
<td>296 ± 29</td>
<td>147 ± 10</td>
<td>148 ± 31</td>
<td>0.286 ± 0.020</td>
<td>760 ± 19</td>
</tr>
<tr>
<td>LP HDF Err</td>
<td>502 ± 20</td>
<td>348 ± 5</td>
<td>154 ± 21</td>
<td>0.676 ± 0.010</td>
<td>823 ± 18</td>
</tr>
</tbody>
</table>

Table 6.8: Photo-z-SQL performance results for four configurations on the PHAT dataset. The table lists the execution time $T_{\text{first}}$ of the Bayesian query for the first time (which includes synthetic magnitude computation), the execution time $T_{\text{second}}$ of the same query for the second time (with synthetic magnitudes already cached), the cache filling time $T_{\text{cache}}$, the photo-z estimation time $T_{\text{gal}}$ for a single galaxy, and the total CLR memory usage MEM$_{\text{total}}$. See the text for a discussion.

Behavior can be observed in case of the thread number: the per-galaxy execution time $T_{\text{gal}}$ drops sharply up to 16 threads as expected, but remains roughly the same for 32 threads. Since the number of cores in the server is 16, this shows that hyper-threading is not an advantage for this algorithm. Also, the cache filling time $T_{\text{cache}}$ remains roughly the same with an increasing thread number (in fact, it gets slightly worse due to more collisions), which is caused by the fact that all galaxies in our sample have the same filter list, filling the cache in the same order. However, this is a one-time cost, and could be parallelized if it ever becomes an issue. The allocated memory also sharply increases with more threads, which is expected since every thread requires resources to work on.

The execution time and memory usage functions with respect to redshift grid points, template number and filter number all follow the same, linearly growing trend. This also matches our expectations, since the storage unit sizes and number of required function evaluations scale linearly with these parameters. While the $T_{\text{cache}}$ curve apparently breaks from the trend at filter numbers above 10, this is due to the fact that those are infrared filters with a smaller resolution, hence the synthetic magnitudes are computed more quickly.
6.4.7 Discussion of Photo-z-SQL

In Sect. 6.4, I presented a new C# photometric redshift estimation code, Photo-z-SQL. I detailed the photo-z approaches implemented by the code, and listed the most important available features. I demonstrated the performance of our code on two public datasets that had been used for photo-z method comparison, PHAT and CAPR (Hildebrandt et al., 2010; Dahlen et al., 2013). I also provided an analysis of how execution time scales with various configuration parameter choices.

Our photo-z estimation results are on par with those of the better performers in the literature, as expected from our adopted configurations, e.g. choice of template SEDs matching Ilbert et al. (2009). While we do not introduce anything inherently new from a photo-z estimation point of view, merely adopt some of the more successful methods in the literature, our implementation does possess the following important technical advantages:

- Capability to be directly integrated into Microsoft SQL Server, eliminating the need to move photometric data outside the DB, and giving users and administrators an integrated platform for photo-z computations.

- Dynamic, on-demand handling of the set of broad-band photometric filters, which is a requirement for heterogeneous databases such as the Hubble Source Catalog (Whitmore et al., 2016).

- Ability to utilize the full covariance matrix between filters.

We should note that, while from a modeling standpoint it is better not to neglect the covariance between filters, in practice it may be difficult to appropriately determine the covariance matrix when no direct measurements are available. An empirical approach similar to the one in Sect. 6.4.3 might be adopted, but that solution will depend on the adopted template set and calibration sample.

Our stand-alone C# code can be compiled and executed on Windows, Unix and Mac OS X systems, but the SQL-CLR integration is strictly tied to Microsoft SQL Server – in this sense, the implementation is not portable. However, SDSS, Pan-STARRS and the Hubble Source Catalog all use Microsoft SQL Server, therefore there are plenty of useful applications even with this limitation.
The implementation is actively in development, especially regarding priors, calibration techniques and the SQL interface. The planned applications include the assembly of a photo-z table for the Hubble Source Catalog, and integration into the SDSS SkyServer, or the database cross-matching platform SkyQuery\textsuperscript{6} (Dobos et al., 2012a; Budavari, Dobos & Szalay, 2013).

The main unsolved challenge in running photo-z on heterogeneous, cross-matched data is the handling of different aperture sizes used in different catalogs. The differing apertures will result in a per-object magnitude shift between photometric bands, dependent on the apparent size of the object on the sky, and on what portion of it is covered by the apertures. Without going to the source images, these magnitude shifts cannot be transformed out – unless there are multiple available aperture sizes for the same band, and some form of profile fitting can be utilized. More research will be needed on how strongly this issue affects photo-z results, and on how to mitigate the effects.

Another avenue for continuing this research could be the implementation of an empirical photo-z technique in C#, to be integrated into Microsoft SQL Server. While the interface for specifying the training set and the learning step in SQL would presumably be more difficult to implement and to use than the interface of the template-based method, in principle there are no technical obstacles since a pre-loaded state can be stored in a static C# class. Furthermore, the 2016 version of Microsoft SQL Server will provide support for R, potentially making it straightforward to perform empirical photo-z calculations directly in a DB.

The Photo-z-SQL code is available for download at https://github.com/beckrob/Photo-z-SQL.

\subsection*{6.4.8 Example SQL queries}

In this section, I provide a few example SQL queries to illustrate how the Photo-z-SQL code can be called from a DB.

The first example sets up a photo-z configuration. First it clears the potentially existing photo-z setup, specifies that −9999 in the input data denotes a missing value, and chooses a flat prior. Then, it creates a string with VO template identifiers corre-

\textsuperscript{6}http://www.sciserver.org/tools/skyquery/
sponding to the LP template set (see Sect. 6.4.4), and parses it into a SqlArray object. Finally, this template list is given to the photo-z code, with a redshift coverage between 0.001 – 1.001 and a linear step size of 0.01. 11 steps will be taken in luminosity space around the best-fitting luminosity.

SELECT PhotoZSQL.Config.RemoveInitialization()

SELECT PhotoZSQL.Config.AddMissingValueSpecifier(-9999)

SELECT PhotoZSQL.Config.SetupFlatPrior()

--To get a SqlArray int array of template IDs, a string
--has to be set up with this format: '[511,512, (...), 1151]'
DECLARE @IDString varchar(max) = '['
DECLARE @id INT = 511
WHILE @id < 1151
BEGIN
    SET @IDString = @IDString + CAST(@id AS varchar(max)) + ','
    SET @id = @id + 1
END
SET @IDString = @IDString + CAST(@id AS varchar(max)) + ']

--Parsing the string into an integer array
DECLARE @TemplateIDArray varbinary(max) = 
    SqlArray.IntArrayMax.ParseInvariant(@IDString)

SELECT PhotoZSQL.Config.SetupTemplateList_ID(@TemplateIDArray,
    1.0, 0.001, 1.001, 0.01, 0, 11)

The second example performs maximum likelihood photo-z estimation on a sample of SDSS data, using the 5 SDSS filters. The fit is performed in magnitude space, with no additional extinction correction, and with a 0.01 mag independent error term added.

DECLARE @FilterIDArray varbinary(max) = 
    SqlArray.IntArrayMax.Vector_5(14, 15, 16, 17, 18)

SELECT TOP 100 gal.objID,
    PhotoZSQL.[Compute].PhotoZMinChiSqr_ID(
        SqlArray.FloatArrayMax.Vector_5(gal.dered_u,
            gal.dered_g,
            gal.dered_r,
            gal.dered_i,
            gal.dered_z)
The third example performs Bayesian photo-z estimation, otherwise with the same particulars as the previous query.

```
DECLARE @FilterIDArray varbinary(max) = 
  SqlArray.IntArrayMax.Vector_5(14,15,16,17,18)

SELECT TOP 100 gal.objID,pz.*
FROM DR12.Galaxy AS gal
CROSS APPLY PhotoZSQL.[Compute].PhotoZBayesian_ID(
  SqlArray.FloatArrayMax.Vector_5(gal.dered_u,gal.dered_g,gal.dered_r,gal.dered_i,gal.dered_z),
  SqlArray.FloatArrayMax.Vector_5(gal.modelMagErr_u,gal.modelMagErr_g,gal.modelMagErr_r,gal.modelMagErr_i,gal.modelMagErr_z),
  1,@FilterIDArray,0.0,0,0.01) AS pz
```

### 6.5 Realistic photo-z validation

This section describes parts of the publication Beck et al. (2017b) that relate to research work that I performed, and the performance of photo-z methods that I implemented.

The main issue that limits the accurate validation of photo-z methods is the fact that objects on which we can test photo-z accuracy are the ones with spectroscopic redshifts,
but in the actual use case photo-z values are required, somewhat self-evidently, for sources without spectroscopic measurements. The characteristics of spectroscopic and photometric samples are generally noticeably different, with the latter being on average fainter, having larger photometric errors, and covering a wider range of observable parameters.

The main goal of Beck et al. (2017b) was to provide a public test bench that allows researchers to evaluate their photo-z approaches in situations mimicking the realistic use case. For this purpose, two galaxy catalogues were created, which we named TEDDY and HAPPY, aiming at testing extrapolation capability, and the handling of different photometric error distributions, respectively.

The TEDDY catalogue was assembled by Chieh-An Lin, starting from SDSS DR12 spectroscopic data that I compiled by filtering out redshift failures and other issues, e.g. missing values. It has four subsets, from A to D, with A functioning as the (spectroscopic) training set, and the rest being validation sets. TEDDY B has the same properties as A, C covers the same range of parameters as A (colours/magnitudes, and redshift) but with a different underlying distribution, and D represents extrapolation, having a wider coverage than A in both colour space and redshift. Refer to Beck et al. (2017b) for specific details.

The HAPPY catalogue was designed to allow a clear assessment of the impact of photometric errors on photo-z estimation, while at the same time closely resembling the colour space differences between the original SDSS DR12 spectroscopic and photometric data sets. A description of how I assembled it is provided in the following section.

### 6.5.1 The Happy catalogue

The first goal was to reproduce the colour-magnitude space distribution of the SDSS DR12 photometric sample, only with objects with measured spectroscopic redshifts, so that the photo-z methods could be properly evaluated. However, as the DR12 spectroscopic set does not contain objects with more extreme colours and fainter magnitudes, it is not an adequate source of example galaxies. Thus, we chose to extend the DR12 spectroscopic set (S1) by cross-matching SDSS photometric measurements with galaxies from other spectroscopic surveys. This approach provides a deeper sample of spectroscopic galaxies, while also keeping the use of actual SDSS photometry and its inherent
systematics.

We followed the cross-matching methodology described in Sect. 6.3.3 — the same methods and data were used here, with the important distinction that photometric colour and error cuts were not applied. The procedure enabled us to find 168,834 matches, which extended the total number of galaxies with spectroscopic redshifts to 2,209,299, and also extended the colour-magnitude space coverage of the sample such that the parameter range of the SDSS photometric set was covered.

In order to build, from the new extended spectroscopic sample (E1), a subset which follows the same colour-magnitude distribution as the original SDSS DR12 photometric set, we randomly selected 75,000 objects from the SDSS DR12 photometric sample (S2)\(^7\) and performed a nearest neighbour (NN) search in E1 (in colour/r-magnitude space).

For each object in S2 we search for its 1\(^{st}\) NN to include into our new set, HAPPY D. To avoid duplicate entries, if the given NN was already included, we select the next closest NN (2\(^{nd}\), 3\(^{rd}\), etc.) which was not already in HAPPY D.

Then, we similarly constructed two new subsets to represent the DR12 spectroscopic sample, HAPPY A and HAPPY B. The former will act as a training set/spectroscopic sample, while the latter will be a test set that has the same distribution of photometric properties as the training set. Thus, we randomly selected 2 \(\times\) 75,000 objects from S1, and searched for their nearest neighbour in E1 using the method outline above, again avoiding any duplicates within HAPPY sets.

Finally, to create an intermediate sample that is between the photometric error properties of S1 and S2, we decided to perform a photometric error cut. Our goal was to reproduce the same range of photometric errors as in S1, but with a distribution that resembles S2, being more weighted towards higher errors. Thus, the cut was chosen to be at the 98\(^{th}\) percentile (to discard outliers) of the photometric error distribution of S1 for each observed feature. We randomly selected 150,000 objects from S2, searched for their nearest neighbours in E1 following the same procedure, and applied the error cut. This yielded the set HAPPY C, composed of \(\approx\) 60,000 galaxies. We note that contrary to all other HAPPY set pairings, HAPPY C and HAPPY D were allowed to overlap to avoid excessively selecting from the less populated faint end of E1.

Cutting the photometric error range of HAPPY C to match that of HAPPY A and

\(^7\)To be precise, the objects were selected from a 2 million element random sub-sample of S2.
B had an important side effect: the magnitude and colour ranges (note: the range, not the shape of the distribution) were also essentially cut to the limits of HAPPY A and B. This shows that the effects of photometric error and magnitude/colour coverage are in fact very much intertwined, one cannot modify one without affecting the other. There are two feasible explanations for why the colour ranges shrink because of the error cut. First, this observation could indicate that the wider colour distributions in the photometric sample are mainly caused by the higher photometric errors smearing the distribution, not by containing physically different extreme galaxies that are missing from the spectroscopic sample. Second, it could be a consequence of galaxy types with extreme colours being significantly more likely to have high measurement errors, therefore these would be preferentially eliminated by the cut, and also would not be present in the spectroscopic sample. Fig. 6.13 shows the $r$-magnitude and colour PDFs for all samples in HAPPY.

Similarly to the subsets of the TEDDY catalogue (see Sect. 6.5 and Beck et al., 2017b), we built HAPPY A to act as a spectroscopic sample and, as such, should be used for training. HAPPY B is completely representative of HAPPY A and so mirrors the equivalent scenario of traditional photo-z validation exercises. HAPPY C illustrates a photometric sample that has been cut to conform to the training sample, with a larger proportion of objects having high photometric errors. HAPPY D serves as a complete photometric sample, with both a wider parameter coverage and higher measurements errors. Thus, HAPPY C and D must only be used for testing, representing increasing degrees of complexity and similarity with the real photometric situation.

### 6.5.2 Validation results

In Beck et al. (2017b), a selection of photometric redshift estimation methods was validated on the TEDDY and HAPPY catalogues. The respective A subsets acted as training sets, while the B, C and D subsets were the validation sets representing increasingly difficult situations.

Following earlier works that compare photo-z methods (Hildebrandt et al., 2010; Dahlen et al., 2013), we selected four summary statistics to quantify the photo-z estimation quality of the various methods tested here. We consider the normalized redshift error, $\Delta z_{\text{norm}} = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$, and from the distribution of $\Delta z_{\text{norm}}$, we
Figure 6.13: Distributions for magnitude in $r$-band and 4 colours for the Happy catalogue compared with the full spectroscopic (red) and photometric (purple) distributions from SDSS DR12.

compute its mean (which is also the average bias), standard deviation (std), median absolute deviation (MAD), and outlier rate. Outliers are defined by $|\Delta z_{\text{norm}}| > 0.15$. The median absolute deviation $MAD = \text{median}(|\Delta z_{\text{norm}}|)$ is computed with outliers included; and the mean $\overline{\Delta z_{\text{norm}}}$ and standard deviation $\sigma(\Delta z_{\text{norm}})$ are computed with the outliers removed from the samples.

**Photo-z estimation methods**

Four empirical, and four template-based photo-z approaches were chosen (refer to Sect. 6.2 for an overview). The empirical methods included the following:

- ANNz, an artificial neural network implementation (Collister & Lahav, 2004),
- a generalized additive model (GAM), a new approach that fits a rather general “global” functional formula,
- a random forest implementation,
- and the local linear regression (LLR) method (as described in Sect. 6.2.1 and Sect. 6.3.1).

Template-based approaches were represented by the following implementations:
Table 6.9: Photo-z results for the TEDDY catalogue.

- Bayesian Photometric Redshifts (BPZ, Benítez, 2000; Coe et al., 2006),
- Easy and Accurate Redshifts from Yale (EAZY, Brammer, van Dokkum & Coppi, 2008),
- Photo-z-SQL using the configuration BPZ HDF Err ZP (SQL BPZ, see Sect. 6.4 and Sect. 6.4.4),
- and Photo-z-SQL using the configuration LP Flat Err (SQL LP, see Sect. 6.4 and Sect. 6.4.4).

Refer to Beck et al. (2017b) for a detailed technical description of each adopted method.

Results on Teddy

Fig. 6.14 shows the estimation results of empirical photo-z methods on the TEDDY catalogue, and Tab. 6.9 lists the numerical results for all photo-z methods.

All the methods that we tested perform relatively well when the training set matches the validation set (TEDDY B), or when the underlying distribution of parameters in the validation set differs from that of the training set, as long as there is sufficient training set coverage (TEDDY C). However, when explicit extrapolation is needed (TEDDY D), purely empirical machine learning methods such as the neural network or the random forest fail spectacularly. Empirical methods that fit a functional formula, i.e. GAM
Figure 6.14: Results on three testing sets of the TEDDY catalogue (columns) obtained from four empirical photo-z methods (lines). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (center) and the limits for being considered outliers. Numerical results are shown in Tab. 6.9 - left panel.
### Table 6.10: Photo-z results for the HAPPY catalogue.

<table>
<thead>
<tr>
<th>Method</th>
<th>Set</th>
<th>Diagnostics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(×10⁻²)</td>
</tr>
<tr>
<td>ANNz</td>
<td>B</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.52</td>
</tr>
<tr>
<td>GAM</td>
<td>B</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.51</td>
</tr>
<tr>
<td>LLR</td>
<td>B</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.79</td>
</tr>
<tr>
<td>Random Forest</td>
<td>B</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Set</th>
<th>Diagnostics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(×10⁻²)</td>
</tr>
<tr>
<td>BPZ</td>
<td>B</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-1.56</td>
</tr>
<tr>
<td>EAZY</td>
<td>B</td>
<td>-3.66</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-4.11</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-4.23</td>
</tr>
<tr>
<td>SQL BPZ</td>
<td>B</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-1.82</td>
</tr>
<tr>
<td>SQL LP</td>
<td>B</td>
<td>-0.47</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-0.51</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-1.33</td>
</tr>
</tbody>
</table>

The results of the template-based methods on the TEDDY sample are as expected, which is why I included no figure of them in the dissertation — overall, their scatter and especially bias is larger than that of the machine learning methods when no extrapolation is needed, but on TEDDY D, they perform as well as the better-extrapolating empirical methods. This illustrates that template-based methods do not depend heavily on a training set. The SQL LP approach provided the best results, especially in terms of bias, with SQL BPZ being second-best.

**Results on Happy**

For the HAPPY catalogue, photo-z results are shown in Fig. 6.15 for the empirical, and in Fig. 6.16 for the spectral template based methods. Numerical results are provided in Tab. 6.10.

All empirical methods provide reasonably accurate redshifts on HAPPY B, where the estimated galaxies have the same distribution of magnitude, colour and photometric error as the spectroscopic training set (HAPPY A). Interestingly, even with the same range of photometric errors, and within the coverage of other properties, on HAPPY C the scatter and proportion of outliers are significantly larger across the board due to the different error distribution (weighted towards higher errors). As expected, photo-z
Figure 6.15: Results from applying empirical photo-z algorithms (lines) to the 3 testing samples of the HAPPY catalogue (columns). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (middle) and the limits for being considered as outliers. Numerical results for all 4 diagnostics are shown in Tab. 6.10 - left panel.
Figure 6.16: Template fitting photo-z results obtained from the 4 adopted methods (lines) for the three testing subsamples of the HAPPY catalogue (columns). The colour gradient shows the logarithmic density. The dashed lines define the perfect prediction (middle) and the limits for being considered as outliers. Numerical results for all 4 diagnostics are shown in Tab. 6.10 - right panel.
accuracy drops even more for all methods on HAPPY D, where many objects are outside the coverage of the training set in all respects. In this scenario, all empirical methods produced outlier rates > 10%.

For the GAM method, an unwanted feature shows up on HAPPY C, a linear broadening of the well-populated region between $z_{\text{spec}} \approx 0.2 - 0.6$ that is estimated to be at $z_{\text{phot}} \approx 0.3$. The feature broadens even further on the sample HAPPY D. This would suggest that the global fitting is more and more disrupted as the photometric errors increase, whereas for the other local methods such an effect is not observed. The numerical results also show that indeed GAM performs worse than the other empirical methods when photometric errors get higher: the MAD value goes from being $\approx 0.005$ worse than the other methods to being $\approx 0.012$ worse, while the outlier rate goes from 0.3% worse to 2% worse. LLR provides results that are similar, if slightly worse than those of the ANNz and random forest methods.

Regarding template-based approaches, on the HAPPY B sample, all methods perform reasonably well, but with a notable overall positive bias for the two BPZ template cases, and a negative bias for EAZY, all on the order of $\approx 0.02$. The SQL LP case had the least bias, and the best results of the four. On HAPPY C, with the photometric errors increasing, the scatter also jumps for all methods, but the degraded performance is best illustrated by the outlier rate skyrocketing to 8 – 18%. On HAPPY D this trend only continues, with outlier rates becoming unmanageable, between 20 – 34%.

We note that the many extreme outliers of the SQL LP case are a result of overfitting the errors – the Le PHARE template set it uses is rather varied (641 templates in this configuration), containing young, dusty starburst galaxies with different dust models, thus more extreme colour combinations can still be fitted. Using a prior, the number of extreme cases could have been greatly mitigated, but that would have increased bias on HAPPY B, which is the sample we chose to optimise for.

While the spectral template fitting methods clearly lag behind empirical ones in this test, we note that even the results of the best-performing method leave much to be desired.
6.5.3 Discussion of results

In Sect. 6.5, I introduced two public photo-z testing galaxy catalogues, **TEDDY** and **HAPPY**, which enable photo-z methods to be tested regarding extrapolation capability, and the handling of different photometric error distributions, respectively (Beck et al., 2017b). Then, I described results from evaluating various empirical and template-based photo-z methods on these catalogues.

Based on results from the **TEDDY** catalogue, we confirmed that most methods can adequately handle a difference in feature distribution shape as long as there is sufficient coverage in the training sample. In this setup, template fitting methods perform worse, especially in terms of bias. However, when training set coverage is not available and thus extrapolation must be performed, local machine learning methods fail, while empirical methods that fit a functional formula achieve reasonable results. Also, in the extrapolation case template fitting methods with the proper configuration can perform comparably to the better machine learning methods. It should be noted that our extrapolation test sample (**TEDDY D**) still has a relatively large number of training set points, and although their colour coverage is rather limited, we do not have to extrapolate very far in terms of redshift in this test. More extensive extrapolation might prove to be difficult for all kinds of machine learning methods.

Additionally, there are two main takeaways from the results on the **HAPPY** catalogue. First, even if an error or colour cut is performed on the photometric sample to make it cover the same parameter range as the spectroscopic training set (this was done in **HAPPY C**), the results on a spectroscopic validation set (**HAPPY B**) will not be representative of results on such a cut photometric sample, due to the differing shape of the error distribution. Ultimately, to accurately determine the photo-z estimation accuracy of an object, its individual photometric error has to be taken into account along with the typical photo-z error of its nearest neighbour galaxies (those with similar colour and magnitude). It follows that any attempt at dealing with the mismatch between spectroscopic and photometric samples must also include their photometric error distribution differences in the calculation of population diagnostics, independently of their feature space coverage. Otherwise, even domain adaptation approaches like calculating appropriate weights for the training sample will output too optimistic diagnostics (refer to Beck et al., 2017b, for more details). The **HAPPY** catalogue provides for the first time an environment where such new approaches can be directly tested.
Second, based on the methods tested here, it appears that “global” model fitting methods perform worse in the presence of large photometric errors than local empirical methods. Neural networks, while in essence fitting an arbitrary functional formula, behave similarly to nearest neighbour methods in this regard.

A more detailed quantitative analysis of the results is presented in Beck et al. (2017b).
Chapter 7

Summary

Recent sky surveys such as the Sloan Digital Sky Survey (SDSS) provide astronomers with extensive datasets of various astronomical sources. To efficiently extract physical information from the sheer amount of data, it becomes necessary to use novel statistical and computational approaches, including modern machine learning methods.

The dissertation presents my research on extragalaxies, specifically the machine learning analysis of emission-line galaxies, and photometric redshift estimation. The work is based on four peer-reviewed publications.

The beginning chapters of the dissertation provide a brief introduction to the machine learning techniques, background of galaxy measurements, and approaches to galaxy modelling that pertain to the applications presented in later sections.

I follow up with an analysis of emission line galaxies — I describe a method to empirically estimate emission line strengths based on a training set, from either galaxy stellar continua, or broad-band photometry. I also propose a more practical stochastic recipe to fulfil the same purpose. These methods could be useful for supplementing stellar population synthesis models with emission lines. Additionally, I introduce a machine learning method to automatically classify star-forming and AGN host galaxies.

The later sections of the dissertation deal with research on photometric redshift estimation. I detail how I assembled the official machine learning photo-z catalogue of the SDSS DR12, with an emphasis on providing unbiased measures of estimation accuracy. I also present and evaluate the performance of a new template-based code, Photo-z-SQL, which can be integrated into a database. Finally, I describe two photo-z testing catalogues that aim to provide difficult tests for methods, and facilitate research on accurate photo-z validation. I present results from various photo-z methods on these catalogues, including two of my own.
Chapter 8

Összefoglalás – Hungarian summary

A közelült égterképezései, mint a Sloan Digitális Égterképezés (SDSS) hatalmas mennyiségű adatot bocsának a csillagászok rendelkezésére. Hogy hatékonyan tudjunk fizikai információt kinyerni ezen adathalmazból, szükséges válík újszerű statisztikai és számítástechnikai eljárások felhasználása, beleértve a modern gépi tanulási módszereket.

Az értekezés bemutatja az extragalaxisokról végzett kutatómunkámat, pontosabban az emissziós vonalakat tartalmazó galaxisok vizsgálatát gépi tanulási módszerekkel, illetve a fotometrikus vöröseltolódás-becslés kutatását. A dolgozat négy, bírált tudományos folyóiratban megjelent, vagy közlésre elfogadott cikken alapszik.

A disszertáció első fejezetei bevezetést nyújtanak azon gépi tanulási módszerekbe, galaxis-megfigyelések hátterébe, és galaxis-modellezési módszerekbe, melyek a későbbi fejezetekben felsorolt alkalmazások szempontjából fontosak lehetnek.

Ezután az emissziós vonalas galaxisok elemzése kerül bemutatásra — bemutatok egy módszert, mellyel tanítóálmaz alapján becsühető az emissziós vonalak fluxusa, akár csillagok, akár szélesszámú fotometriai adatok alapján. Emellett összefekvők egy könnyebben felhasználható, sztochasztikus módszert, mely ugyanazt a szerepet tölti be. Ezekkel az eljárásokkal csillagpopuláció-szintézis modellekhez lehet emissziós vonalakat hozzárendelni. Továbbá létre egy gépi tanulási módszert, mely felhasználásával galaxisokat osztályozó és aktív magvú kategóriákba.

A disszertáció későbbi fejezetei a fotometrikus vöröseltolódás-becsléssel kapcsolatos kutatómunkámat dokumentálják. Részletezem, hogyan állítottam össze az SDSS DR12 hivatalos, gépi tanulási fotometrikus vöröseltolódás-katalógusát, külön figyelmet fordítva a torzítatlan hibabecslés végzésére. Emellett bemutatom a Photo-z-SQL-t, mely egy új, spektrummintázó-illesztésen alapuló, adatbázisba integrálható photo-z implementáció, és elemzem a teljesítményét is. Végül ismertetek két, photo-z tesztelésre szánt adathalmazt, melyek célja a módszerek alapos kihívás elé állítása, és a pontos photo-z hibabecslést célzó kutatások elősegítése. Bemutatom többféle photo-z módszer eredményeit az adathalmazokon, beleértve két saját implementációt is.
## Chapter 9

### List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGN</td>
<td>active galactic nucleus</td>
</tr>
<tr>
<td>BOSS</td>
<td>Baryon Oscillation Spectroscopic Survey</td>
</tr>
<tr>
<td>BPT</td>
<td>Baldwin–Phillips–Terlevich (diagram)</td>
</tr>
<tr>
<td>CAPR</td>
<td>A Critical Assessment of Photometric Redshift Methods</td>
</tr>
<tr>
<td>CCD</td>
<td>charge-coupled device</td>
</tr>
<tr>
<td>DB</td>
<td>database</td>
</tr>
<tr>
<td>DR7/12</td>
<td>Data Release 7/12</td>
</tr>
<tr>
<td>EW</td>
<td>equivalent width</td>
</tr>
<tr>
<td>GAM</td>
<td>generalized additive model</td>
</tr>
<tr>
<td>GR</td>
<td>general relativity</td>
</tr>
<tr>
<td>HDF-N</td>
<td>Hubble Deep Field North (survey)</td>
</tr>
<tr>
<td>IMF</td>
<td>initial mass function</td>
</tr>
<tr>
<td>IR</td>
<td>infrared</td>
</tr>
<tr>
<td>ISM</td>
<td>interstellar medium</td>
</tr>
<tr>
<td>mag</td>
<td>magnitude</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PHAT</td>
<td>Photo-z Accuracy Testing</td>
</tr>
<tr>
<td>photo-z</td>
<td>photometric redshift</td>
</tr>
<tr>
<td>PSF</td>
<td>point spread function</td>
</tr>
<tr>
<td>SDSS</td>
<td>Sloan Digital Sky Survey</td>
</tr>
<tr>
<td>SED</td>
<td>spectral energy distribution</td>
</tr>
<tr>
<td>SF</td>
<td>star-forming (galaxy)</td>
</tr>
<tr>
<td>SFH</td>
<td>star formation history</td>
</tr>
<tr>
<td>SMBH</td>
<td>supermassive black hole</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>SPS</td>
<td>stellar population synthesis</td>
</tr>
<tr>
<td>SQL</td>
<td>structured query language</td>
</tr>
<tr>
<td>SVD</td>
<td>singular value decomposition</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
</tbody>
</table>
Chapter 10

Bibliography


Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2017a, Astronomy and Com-
puting, DOI: 10.1016/j.ascom.2017.03.002


Bolton A. S. et al., 2012, AJ, 144, 144


Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A.,


Budavári T., Dobos L., Szalay A. S., 2013, Computing in Science and Engineering, 15, 12


Csabai I. et al., 2003, AJ, 125, 580


Doi M. et al., 2010, AJ, 139, 1628


Eisenstein D. J. et al., 2011, AJ, 142, 72


Ferland G. J. et al., 2013, Revista Mexicana de Astronomía y Astrofísica, 49, 137


Forgy E. W., 1965, Biometrics, 21, 768


Han B., Zhang Y., Zhong S., Zhao Y., 2016, ArXiv e-prints

Hartigan J. A., Wong M. A., 1979, Applied statistics, 100


Hildebrandt H. et al., 2010, A&A, 523, A31


Ilbert O. et al., 2006a, A&A, 457, 841

Ilbert O. et al., 2006b, A&A, 457, 841


Ivezić Z. et al., 2008, E-print, arXiv:0805.2366


Kerekes G., Csabai I., Dobos L., Trenčsényi M., 2013, Astronomische Nachrichten, 334, 1012


Lilly S. J. et al., 2009, ApJS, 184, 218

Lilly S. J. et al., 2007, ApJS, 172, 70


MacQueen J., 1967, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1, 281


Marik M., 1989, Csillagászat. Akadémiai Kiadó


Salim S., 2014, Serbian Astronomical Journal, 189, 1


Smee S. A. et al., 2013, AJ, 146, 32


Stasińska G., 1984, A&AS, 55, 15

143


Vazdekis A., Ricciardelli E., Cenarro A. J., Rivero-González J. G., Díaz-García L. A.,


Yip C. W. et al., 2004, AJ, 128, 585

ADATLAP

a doktori értekezés nyilvánosságra hozatalához*

I. A doktori értekezés adatai
A szerző neve: Beck Róbert
MTMT-azonosító: 10054702
A doktori értekezés címe és alcíme: (az értekezés nyelve: angol)
Empirical and spectral template based approaches in the analysis of galaxy data
Empirikus és spektrumminta-alapú módszerek a galaxisadatok elemzésében
DOI-azonosító: 10.15476/ELTE.2017.052
A doktori iskola neve: ELTE TTK, Fizika Doktori Iskola
A doktori iskolán belüli doktori program neve: Részecskefizika és csillagászat program
A témavezető neve és tudományos fokozata: Dr. Csabai István (DSc) és Dr. Dobos László (Phd)
A témavezető munkahelye: ELTE TTK, Komplex Rendszerek Fizikája Tanszék

II. Nyilatkozatok

1. A doktori értekezés szerzőjéként
   a) hozzájárulok, hogy a doktori fokozat megtervezését követően a doktori értekezésem és a tézisek
   nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a
   Természettudományi kar Dékáni Hivatal Doktori, Habilitációs és Nemzetközi Ügyek
   Csoportjának ügyintézőjét, hogy az értekezést és a téziseket feltöltse az ELTE Digitális
   Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.
   b) kérem, hogy a mellékelt kérelmekben részletezett szabadalmi, illetőleg oltalmi bejelentés
      közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az
      ELTE Digitális Intézményi Tudástárban;
   c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a
      minősítés (datum)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban
      és az ELTE Digitális Intézményi Tudástárban;
   d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori
      értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az
      ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyük közzé. Ha a
      könyv a fokozatszerzést követően egy évig nem jelenik meg, hozzájárulok, hogy a doktori
      értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE
      Digitális Intézményi Tudástárban.

2. A doktori értekezés szerzőjéként kijelentem, hogy
   a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját
      eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői
      jogait;
   b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón
      benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjéként hozzájárulok a doktori értekezés és a tézisek szövegének
   plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: 2017.04.05.

*a doktori értekezés szerzőjének aláírása

*ELTE SZMSZ SZMR 12. sz. melléklet